GREEDY BAYESIAN DOUBLE SPARSITY DICTIONARY LEARNING

Juan G. Serra[*1], *Salvador Villena*[*2], *Rafael Molina*[*1], *Aggelos K. Katsaggelos*[†]

[*] University of Granada, [1] Dept. of Computer Science and AI, [2] Dept. of Languages and Inf. Systems.
[†] Northwestern University, Dept. of Electrical Engineering and Computer Science.

ABSTRACT

This work presents a greedy Bayesian dictionary learning (DL) algorithm where not only the signals but also the dictionary representation matrix accept a sparse representation. This double-sparsity (DS) model has been shown to be superior to the standard sparse approach in some image processing tasks, where sparsity is only imposed on the signal coefficients. We present a new Bayesian approach which addresses typical shortcomings of regularization-based DS algorithms: the prior knowledge of the true noise level and the need of parameter tuning. Our model estimates the noise and sparsity levels as well as the model parameters from the observations and frequently outperforms state-of-the-art dictionary based techniques by taking into account the uncertainty of the estimates. Additionally, we introduce a versatile notation which generalizes denoising, inpainting and compressive sensing problem formulations. Finally, theoretical results are validated with denoising experiments on a set of images.

Index Terms— Sparse Representation, Dictionary Learning, Bayesian Inference.

1. INTRODUCTION

Natural signals, such as images, have an economy of representation over DCT, Wavelets or Curvelets, among others, that noise and artificial signals just do not share. This concept was first applied in signal compression with remarkable results and later evolved thanks to the notion that, for a given set of natural signals, a better basis can be learned yielding even compacter representations. This process laid the foundations of sparse dictionary learning, where the dictionary is an overcomplete matrix of basis signals learned from a particular dataset.

In the standard dictionary learning problem, we are interested in learning both the overcomplete dictionary $\mathbf{D} = [\mathbf{d}_1 \dots \mathbf{d}_K] \in \mathbb{R}^{P\times K}$ and Q signal representations $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_Q] \in \mathbb{R}^{K\times Q}$ from a set of Q observed signals, concatenated columnwise in $\mathbf{Y} = [\mathbf{y}_1 \dots \mathbf{y}_Q] \in \mathbb{R}^{P\times Q}$. Both the dictionary $\mathbf{D}$ and the sparse representation matrix $\mathbf{X}$ can be recovered by solving an optimization problem where we seek the best reconstruction of signals $\mathbf{y}_q$ given a maximum number of non-zero entries allowed for each $\mathbf{x}_q$. We can mathematically express this as

$$\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \text{ , s.t. } \|\mathbf{x}_q\|_0 \le T, \forall q. \tag{1}$$

Since the objective function is not convex in $\mathbf{D}$ and $\mathbf{X}$ jointly, but bi-convex in $\mathbf{D}$ and $\mathbf{X}$ individually, (1) can be addressed by alternating minimization over $\mathbf{D}$ and $\mathbf{X}$ separately. However, exact minimization over $\mathbf{X}$ is NP-hard, so approximate methods, K-SVD [1] being the most popular, are used to overcome this problem. The sparsity constraint can be relaxed substituting the ℓ_0 pseudo-norm in (1) by the ℓ_1 norm, which can be solved by convex optimization and Bayesian techniques. See, for instance, [1–6].

The DL problem in (1) and its relaxation have been extensively applied to image processing tasks: denoising and inpainting [1, 2, 4, 7, 8], superresolution [9, 10], deblurring [11], face recognition [12]; and machine learning: classification and clustering [13].

Learned dictionaries are highly structured [14], which suggests that dictionary atoms themselves may have a sparse representation over a fixed primary dictionary $\mathbf{\Psi} \in \mathbb{R}^{P\times M}$ (such as an overcomplete DCT), this is, $\mathbf{D} = \mathbf{\Psi}\mathbf{A}$, where $\mathbf{A} \in \mathbb{R}^{M\times K}$ is the sparse representation of the dictionary. We can now define the complete constrained DS dictionary learning problem as

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{X}} \|\mathbf{Y} - \mathbf{\Psi}\mathbf{A}\mathbf{X}\|_F^2 \\ &\text{s.t. } \|\mathbf{x}_q\|_0 \le T, \forall q, \text{ and } \|\mathbf{a}_k\|_0 \le S, \forall k. \end{aligned} \tag{2}$$

The strength of this model over the traditional one is that the dictionary representation constraint acts as a regularizer, allowing for feature selection from $\mathbf{\Psi}$ and helps reduce overfitting [15]. Applications of this model include image denoising [16], superresolution [17], remote sensing image compression [18], detection of activated voxels in fMRI [19], and face recognition [20].

K-SVDS [14] is a greedy technique based on the standard K-SVD algorithm to solve the ℓ_0 DS learning problem in (2). Here, minimization over $\mathbf{A}$ and $\mathbf{X}$ is alternated. First, the Orthogonal Matching Pursuit (OMP) algorithm determines the support of each signal in $\mathbf{X}$, whose updated value admits a close-form expression, next, the minimization over $\mathbf{A}$ can be shown to be equivalent to a simple sparse coding problem.

However, one potential drawback of this approach is that it requires the true noise level, which is hardly ever available in real experiments.

We propose a Bayesian approximate algorithm that solves the whole ℓ_1 DS learning problem, that is, it automatically estimates the noise variance and sparsity levels of both dictionary atoms and signals. The method takes into account the uncertainty of the estimates which leads to improved performance. We apply our method to image denoising and compare to the state-of-the-art technique K-SVDS.

The paper is organized as follows. Section 2 presents the Bayesian modeling of the DS problem with hierachical priors on $\mathbf{A}$ and $\mathbf{X}$. In section 3 variational inference is used to elaborate an optimal algorithm. Based on the inference procedure from section 3, we develop a computationally efficient implementation based on Empirical Bayes in section 4. Finally, we present some experimental results in section 5 and summarize the most important conclusions of the presented work in section 6.

Research funded by the Spanish Ministry of Economy and Competitiveness (MINECO) projects TIN2013-43880-R and DPI2016-77869-C2-2-R, and the US Department of Energy grant DE-NA0002520.

2. BAYESIAN MODELING

The ℓ_1 relaxation of the DS learning problem in (2) naturally admits a hierarchical probabilistic modeling analogous to the one in [21]. We first introduce the observation model

$$\mathrm{p}(\mathbf{Y}|\beta, \mathbf{A}, \mathbf{X}) \propto \beta^{\frac{\sum_q N_q}{2}} e^{-\frac{\beta}{2}\sum_{q=1}^{Q}||\mathbf{M}_q(\mathbf{y}_q - \mathbf{\Psi}_q \mathbf{A}\mathbf{x}_q)||^2}, \qquad (3)$$

where $\mathbf{M}_q$ is a diagonal matrix whose (n,n)-th entry is 1 if the corresponding y_{nq} is observed, 0 otherwise; $N_q = ||diag(\mathbf{M}_q)||_0$ and $\mathbf{\Psi}_q = \mathbf{\Phi}_q\mathbf{\Psi}$, where $\mathbf{\Psi}$ denotes an overcomplete basis. Notice that this general notation covers denoising, inpainting and compressive sensing problems depending on the choice of matrices $\mathbf{M}_q$ and $\mathbf{\Phi}_q$, see table 1.

$\mathbf{M}_q$	$\mathbf{\Phi}_q$	**Application**
$\mathbf{I}_P$	$\mathbf{I}_P$	denoising
$\mathbf{M}_q$	$\mathbf{I}_P$	inpainting
$\mathbf{I}_P$	$\mathbf{\Phi}_q$	compressive sensing

Table 1. Matrices $\mathbf{M}_q$ and $\mathbf{\Phi}_q$ for different image processing problems.

To promote sparsity on the columns of $\mathbf{A}$ and $\mathbf{X}$ we use Laplace priors. Unfortunately, the non-conjugacy of the Laplace prior to the Gaussian observation model in (3) renders exact inference intractable. To solve this problem, we utilize a hierarchical representation of the Laplace distribution.

First, we assume an independent zero-mean Gaussian prior with unknown diagonal covariance matrix on the columns of $\mathbf{A}$

$$\mathrm{p}(\mathbf{a}_k|\boldsymbol{\omega}_k) = \prod_{m=1}^{M}\mathcal{N}(a_{mk}|0,\omega_{mk}) = \mathcal{N}(\mathbf{a}_k|\mathbf{0}_M, \mathbf{\Omega}_k), \qquad (4)$$

where $\mathbf{\Omega}_k = diag(\boldsymbol{\omega}_k)$, being ω_{mk} the variance associated to the mth entry of $\mathbf{a}_k$. These variances share a common Gamma hyperprior given by

$$\mathrm{p}(\boldsymbol{\omega}_k|\rho_k) = \prod_{m=1}^{M}\Gamma(\omega_{mk}|1,\rho_k/2). \qquad (5)$$

Notice that the distribution of $\mathbf{a}_k$ given ρ_k can be obtained by marginalization as

$$\mathrm{p}(\mathbf{a}_k|\rho_k) = \int \mathrm{p}(\mathbf{a}_k|\boldsymbol{\omega}_k)\mathrm{p}(\boldsymbol{\omega}_k|\rho_k)\mathrm{d}\boldsymbol{\omega}_k = \frac{\rho_k^{M/2}}{2^M}\exp\bigl(-\sqrt{\rho_k}\|\mathbf{a}_k\|_1\bigr). \qquad (6)$$

Finally, we place the following distribution on ρ_k

$$\mathrm{p}(\rho_k|\eta_k) = \Gamma(\rho_k|\eta_k/2,\eta_k/2). \qquad (7)$$

An identical prior hierarchy (4, 5, 7) is imposed on the columns of $\mathbf{X}$

$$\mathrm{p}(\mathbf{x}_q|\boldsymbol{\gamma}_q) = \prod_{k=1}^{K}\mathcal{N}(x_{kq}|0,\gamma_{kq}) = \mathcal{N}(\mathbf{x}_q|\mathbf{0}_K,\mathbf{\Gamma}_q) \qquad (8)$$

$$\mathrm{p}(\boldsymbol{\gamma}_q|\lambda_q) = \prod_{k=1}^{K}\Gamma(\gamma_{kq}|1,\lambda_q/2) \qquad (9)$$

$$\mathrm{p}(\lambda_q|\nu_q) = \Gamma(\lambda_q|\nu_q/2,\nu_q/2), \qquad (10)$$

with $\mathbf{\Gamma}_q = diag(\boldsymbol{\gamma}_q)$.

To complete the model, we assume the noise precision β to be Gamma distributed

$$\mathrm{p}(\beta) = \Gamma(\beta|a_\beta, b_\beta) \propto \beta^{a_\beta - 1}\exp(-b_\beta\beta), \qquad (11)$$

with positive scalars a_β and b_β being the shape and inverse scale parameters respectively.

The joint distribution condenses all the knowledge on the problem

$$\begin{aligned}\mathrm{p}(\mathbf{Y},\mathbf{\Theta}) =& \mathrm{p}(\mathbf{Y}|\beta,\mathbf{A},\mathbf{X})\mathrm{p}(\beta)\Big[\mathrm{p}(\mathbf{A}|\mathbf{\Omega})\mathrm{p}(\mathbf{\Omega}|\boldsymbol{\rho})\mathrm{p}(\boldsymbol{\rho}|\boldsymbol{\eta})\mathrm{p}(\boldsymbol{\eta})\Big] \\ &\times \Big[\mathrm{p}(\mathbf{X}|\mathbf{\Gamma})\mathrm{p}(\mathbf{\Gamma}|\boldsymbol{\lambda})\mathrm{p}(\boldsymbol{\lambda}|\boldsymbol{\nu})\mathrm{p}(\boldsymbol{\nu})\Big], \qquad (12)\end{aligned}$$

where we assume $\mathrm{p}(\boldsymbol{\eta})$ and $\mathrm{p}(\boldsymbol{\nu})$ to be flat improper priors, $\mathbf{\Theta} = \{\mathbf{A},\mathbf{\Omega},\boldsymbol{\rho},\boldsymbol{\eta},\mathbf{X},\mathbf{\Gamma},\boldsymbol{\lambda},\boldsymbol{\nu},\beta\}$ denotes the entire set of unknowns, $\mathbf{\Omega} = [\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_K]$ and $\mathbf{\Gamma} = [\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_Q]$.

3. VARIATIONAL BAYESIAN INFERENCE

We aim at estimating the posterior distribution $\mathrm{p}(\mathbf{\Theta}|\mathbf{Y})$, which is not analytically feasible due to the intractability of the marginal of $\mathbf{Y}$. Consequently, we use an approximate procedure based on the following mean-field factorization

$$\begin{aligned}\mathrm{q}(\mathbf{\Theta}) =& \mathrm{q}(\beta)\Big[\prod_k \mathrm{q}(\mathbf{a}_k)\Big]\mathrm{q}(\mathbf{\Omega})\mathrm{q}(\boldsymbol{\rho})\mathrm{q}(\boldsymbol{\eta}) \\ &\times \Big[\prod_q \mathrm{q}(\mathbf{x}_q)\Big]\mathrm{q}(\mathbf{\Gamma})\mathrm{q}(\boldsymbol{\lambda})\mathrm{q}(\boldsymbol{\nu}), \qquad (13)\end{aligned}$$

where the posteriors of $\mathbf{\Omega}$, $\boldsymbol{\rho}$, $\boldsymbol{\eta}$, $\mathbf{\Gamma}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ are assumed to be degenerate. Notice also that columnwise independence was assumed for the posterior distributions of $\mathbf{A}$ and $\mathbf{X}$.

Applying calculus of variations, for each $\theta_i \in \mathbf{\Theta}$ where $\mathrm{q}(\theta_i)$ is non-degenerate we have

$$\log\mathrm{q}(\theta_i) = \langle\log\mathrm{p}(\mathbf{Y},\mathbf{\Theta})\rangle_{\mathbf{\Theta}\backslash\theta_i} + C, \qquad (14)$$

where $\langle\bullet\rangle_{\mathbf{\Theta}\backslash\theta_i}$ denotes the expectation taken with respect to all approximating factors in $\mathrm{q}(\mathbf{\Theta})$ except for $\mathrm{q}(\theta_i)$.

In the case of degenerate distributions, the concrete value taken by the distribution $\mathrm{q}(\theta_i)$ is

$$\hat{\theta}_i = \arg\max_{\theta_i}\langle\log\mathrm{p}(\mathbf{Y},\mathbf{\Theta})\rangle_{\mathbf{\Theta}\backslash\theta_i}. \qquad (15)$$

The posterior of $\mathbf{A}$ can be estimated using eq. (14). Focusing on a single column of $\mathbf{A}$, $\mathbf{a}_k$, denoting by A the set of indexes q such that the qth column of $\mathbf{X}$ satisfies $\langle x_{kq}^2\rangle > 0$, we have

$$\begin{aligned}\log\mathrm{q}(\mathbf{a}_k) =& -\frac{\beta}{2}\sum_q\langle\|\mathbf{M}_q(\mathbf{y}_q - \mathbf{\Psi}_q\mathbf{A}\mathbf{x}_q)\|^2\rangle_{\mathbf{\Theta}\backslash\mathbf{a}_k} \\ &- \frac{1}{2}\mathbf{a}_k^{\mathrm{T}}\mathbf{\Omega}_k^{-1}\mathbf{a}_k + C \\ =& -\frac{\hat{\beta}}{2}\|\mathbf{u}_k - \mathbf{U}_k\mathbf{a}_k\|^2 - \frac{1}{2}\mathbf{a}_k^{\mathrm{T}}\hat{\mathbf{\Omega}}_k^{-1}\mathbf{a}_k + C, \qquad (16)\end{aligned}$$

where $\mathbf{u}_k$ and $\mathbf{U}_k$ are block-row matrices whose blocks have the form $1/\langle x_{kq}^2\rangle^{1/2}\mathbf{M}_q(\mathbf{y}_q\langle x_{kq}\rangle - \mathbf{\Psi}_q\sum_{j\neq k}\hat{\mathbf{a}}_j\langle x_{jq}x_{kq}\rangle)$ and $\langle x_{kq}^2\rangle^{1/2}\mathbf{M}_q\mathbf{\Psi}_q$, $q \in A$ respectively.

which leads to a Gaussian distribution $\mathrm{q}(\mathbf{a}_k)$, with mean and covariance matrix

$$\hat{\mathbf{a}}_k = \hat{\beta}\boldsymbol{\Sigma}_{\mathbf{a}_k}\mathbf{U}_k^{\mathrm{T}}\mathbf{u}_k \tag{17}$$

$$\boldsymbol{\Sigma}_{\mathbf{a}_k} = (\hat{\beta}\mathbf{U}_k^{\mathrm{T}}\mathbf{U}_k + \boldsymbol{\Omega}_k^{-1})^{-1}, \tag{18}$$

Next, we can find the updates of the hyperparameters related to $\mathbf{A}$ solving eq. (15). Specifically, for $\boldsymbol{\omega}_{mk}$, we need to maximize

$$-\frac{1}{2}\log\omega_{mk} - \frac{1}{2}\frac{\langle a_{mk}^2\rangle}{w_{mk}} - \frac{1}{2}\langle\rho_k\rangle\omega_{mk} + C. \tag{19}$$

Setting the derivative w.r.t. ω_{mk} equal to zero, we obtain

$$\hat{\omega}_{mk} = \frac{-1+\sqrt{1+4\hat{\rho}_k(\hat{a}_{mk}^2 + \boldsymbol{\Sigma}_{\mathbf{a}_k}(m,m))}}{2\hat{\rho}_k}. \tag{20}$$

Following a similar procedure, the estimated value of ρ_k yields

$$\hat{\rho}_k = \frac{\hat{\eta}_k + 2M - 2}{\hat{\eta}_k + \sum_{m=1}^{M}\hat{\omega}_{mk}}, \tag{21}$$

and lastly, $\hat{\eta}_k$ is obtained by maximizing

$$\frac{\eta_k}{2}\log\frac{\eta_k}{2} + (\log\hat{\rho}_k - \hat{\rho}_k)\frac{\eta_k}{2} - \log\Gamma(\frac{\eta_k}{2}), \tag{22}$$

which must be solved numerically.

Inference on each column of $\mathbf{X}$ leads, yet again, to a Gaussian distribution

$$\begin{aligned}\log\mathrm{q}(\mathbf{x}_q) &= -\frac{\hat{\beta}}{2}\sum_q\langle\|\mathbf{M}_q(\mathbf{y}_q - \boldsymbol{\Psi}_q\mathbf{A}\mathbf{x}_q)\|^2\rangle_{\boldsymbol{\Theta}\backslash\mathbf{x}_q}\\ &- \frac{1}{2}\mathbf{x}_q^{\mathrm{T}}\hat{\boldsymbol{\Gamma}}_q^{-1}\mathbf{x}_q + C\\ &= -\frac{\hat{\beta}}{2}\|\mathbf{y}_q - \mathbf{V}_q\mathbf{x}_q\|^2 - \frac{1}{2}\mathbf{x}_q^{\mathrm{T}}(\mathbf{Z}_q + \hat{\boldsymbol{\Gamma}}_k^{-1})\mathbf{x}_q + C,\end{aligned} \tag{23}$$

whose mean $\hat{\mathbf{x}}_q$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x}_q}$ are given by

$$\hat{\mathbf{x}}_q = \hat{\beta}\boldsymbol{\Sigma}_{\mathbf{x}_q}\mathbf{V}_q\mathbf{y}_q \tag{24}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}_q} = (\hat{\beta}\mathbf{V}_q^{\mathrm{T}}\mathbf{V}_q + \mathbf{Z}_q + \boldsymbol{\Gamma}_q^{-1})^{-1}, \tag{25}$$

where $\mathbf{V}_q = \mathbf{M}_q\boldsymbol{\Psi}_q\hat{\mathbf{A}}$ and $\mathbf{Z}$ is a diagonal matrix with entries $\mathbf{Z}_q(k,k) = \hat{\beta}\,\mathrm{Tr}(\boldsymbol{\Psi}_q^{\mathrm{T}}\mathbf{M}_q\boldsymbol{\Psi}_q\boldsymbol{\Sigma}_{\mathbf{a}_k})$.

The updates on the hyperparameters associated with $\mathbf{x}_q$ have equivalent expressions to (20), (21), and (22).

Finally, to estimate the noise precision we simply apply (14) to eq. (12) considering only the terms that depend on β. We have

$$\begin{aligned}\log\mathrm{q}(\beta) &= \frac{\sum_q N_q}{2}\log\beta + (a_\beta - 1)\log\beta - b_\beta\beta\\ &- \frac{\beta}{2}\sum\nolimits_q\langle\|\mathbf{M}_q(\mathbf{y}_q - \boldsymbol{\Psi}_q\mathbf{A}\mathbf{x}_q)\|^2\rangle_{\boldsymbol{\Theta}\backslash\beta} + C,\end{aligned} \tag{26}$$

which corresponds to a Gamma distribution on β with mean

$$\hat{\beta} = \langle\beta\rangle = \frac{\sum_q N_q + 2a_\beta}{\sum_q\langle\|\mathbf{M}_q(\mathbf{y}_q - \boldsymbol{\Psi}_q\mathbf{A}\mathbf{x}_q)\|^2\rangle_{\boldsymbol{\Theta}\backslash\beta} + 2b_\beta}. \tag{27}$$

4. FAST INFERENCE

The inference procedure described above, albeit mathematically sound, has two practical drawbacks. First, the computation of $\boldsymbol{\Sigma}_{\mathbf{a}_k}$ and $\boldsymbol{\Sigma}_{\mathbf{x}_q}$ requires K $M \times M$ and Q $K \times K$ matrix inversions at each iteration, which can be computationally expensive and memory intensive. Secondly, the Laplace prior does not provide exact sparse solutions, but simply close to zero.

To reduce the computational complexity and alleviate memory usage, we propose a fast inference procedure based on the approach proposed in [21] for reconstruction based on compressed sensing observations, see also [22].

Concretely, the support of $\mathbf{a}_k$ and $\mathbf{x}_q$ is calculated sequentially, starting from empty matrices and iteratively adding components to the model. The correspondent hyperparameters at these non-zero locations are obtained via MAP estimation. In this way, memory usage drops drastically due to sparsity. The procedure provides an efficient scheme for the update of the covariance matrices $\boldsymbol{\Sigma}_{\mathbf{a}_k}$ and $\boldsymbol{\Sigma}_{\mathbf{x}_q}$ and uses the analytical formulas for $\hat{\mathbf{a}}_k$ and $\hat{\mathbf{x}}_q$ obtained above at the estimated support.

4.1. Fast Bayesian Inference for $\boldsymbol{\omega}_k$

Notice that the first term on the right-hand side of eq. (16) can be interpreted as the likelihood $\mathrm{p}(\mathbf{u}_k|\mathbf{a}_k)$ of having $\mathbf{u}_k$ as the observed value given $\mathbf{a}_k$, which is Gaussian distributed, and the second term represents the prior $\mathrm{p}(\mathbf{a}_k|\boldsymbol{\omega}_k)$. We can express the posterior of $\boldsymbol{\omega}_k$ given $\mathbf{u}_k$, holding ρ_k fixed at its most recent estimated value, via marginalization

$$\mathrm{p}(\boldsymbol{\omega}_k|\mathbf{u}_k) \propto \mathrm{p}(\boldsymbol{\omega}_k|\hat{\rho}_k)\int\mathrm{p}(\mathbf{u}_k|\mathbf{a}_k)\mathrm{p}(\mathbf{a}_k|\boldsymbol{\omega}_k)\mathrm{d}\mathbf{a}_k, \tag{28}$$

which yields

$$\mathrm{p}(\boldsymbol{\omega}_k|\mathbf{u}_k) = \mathcal{N}(\mathbf{u}_k|\mathbf{0},\mathbf{C}_k)\mathrm{p}(\boldsymbol{\omega}_k|\hat{\rho}_k), \tag{29}$$

with $\mathbf{C}_k = \hat{\beta}^{-1}\mathbf{I} + \mathbf{U}_k\boldsymbol{\Omega}_k\mathbf{U}_k^{\mathrm{T}}$. We can obtain the optimal values of $\boldsymbol{\omega}_k$ via MAP estimation. Taking the logarithm of (29) and considering only the terms that depend on $\boldsymbol{\omega}_k$ we have

$$\mathcal{L}(\boldsymbol{\omega}_k) = \frac{1}{2}\log|\mathbf{C}_k| - \frac{1}{2}\mathbf{u}_k^{\mathrm{T}}\mathbf{C}_k^{-1}\mathbf{u}_k - \frac{\hat{\rho}_k}{2}\sum_{m=1}^{M}\omega_{mk}. \tag{30}$$

It is now interesting to see that $\mathbf{C}_k$ can be decomposed to isolate the contribution of the mth entry as $\mathbf{C}_k = {}^{-m}\mathbf{C}_k + \omega_{mk}\mathbf{u}_{m,k}\mathbf{u}_{m,k}^{\mathrm{T}}$. Next, calculating $|\mathbf{C}_k|$, see [21], and applying the matrix inversion lemma to $\mathbf{C}_k^{-1}$, we can easily decompose $\mathcal{L}(\boldsymbol{\omega}_k)$ in $\mathcal{L}({}^{-m}\boldsymbol{\omega}_k) + \ell(\boldsymbol{\omega}_{mk})$, which leads to a constructive process in which the support in $\mathbf{a}_k$ and $\boldsymbol{\Sigma}_{\mathbf{a}_k}$ is not only added incrementally, but it is also possible to re-estimate or delete previously added support depending on which produces the highest increase of $\mathcal{L}(\boldsymbol{\omega}_k)$. See [21] for more details.

4.2. Fast Bayesian Inference for $\boldsymbol{\gamma}_q$

Notice that the first term on the right-hand side of (23) can be recognized as the energy of a Gaussian distribution that we denote $\mathrm{n}(\mathbf{y}_q|\mathbf{x}_q)$, whereas the second term may be regarded as a modified prior on $\mathbf{x}_q$, $\mathrm{n}(\mathbf{x}_q)$. With this notation, we can express the posterior $\boldsymbol{\gamma}_q$ given $\mathbf{y}_q$ as

$$\mathrm{p}(\boldsymbol{\gamma}_q|\mathbf{y}_q) \propto \mathrm{p}(\boldsymbol{\gamma}_q|\hat{\lambda}_q)\int\mathrm{n}(\mathbf{y}_q|\mathbf{x}_q)\mathrm{n}(\mathbf{x}_q)\mathrm{d}\mathbf{x}_q, \tag{31}$$

which produces

$$\mathrm{p}(\boldsymbol{\gamma}_q|\mathbf{y}_q) = \mathcal{N}(\mathbf{y}_q|\mathbf{0}, \mathbf{D}_q)\mathrm{p}(\boldsymbol{\gamma}_q|\hat{\lambda}_q), \qquad (32)$$

with $\mathbf{D}_q = \hat{\beta}^{-1} + \mathbf{V}_q(\mathbf{Z}_q + \boldsymbol{\Gamma}_q^{-1})^{-1}\mathbf{V}^{\mathrm{T}}$, where we can again isolate the contribution of a single $\boldsymbol{\gamma}_{kq}$ and build a fast inference procedure analogous to the one with $\mathbf{a}_k$ and $\boldsymbol{\omega}_k$.

4.3. Fast Inference based on addition only

Computational speed can be further increased by only allowing the addition of new components to the model. That is, in subsections 4.1 and 4.2 we only consider the addition of terms ω_{mk} and γ_{kq} respectively, and do not neither update nor remove atoms already used in the representation. We have experimentally observed that this modification has little influence on the algorithm performance, but results in a convenient speed-up.

5. EXPERIMENTS

This section is dedicated to assess the performance of the proposed method. Due to lack of space, we will focus on the denoising problem. Three standard 512×512 images, namely 'Barbara', 'Boat' and 'Peppers' were used for testing. Each of these images was corrupted with zero-mean Gaussian noise with standard deviation ranging from 10 to 50 in 10-unit intervals. For training, we randomly extract $Q = 2000$ square image blocks of size 8×8 rearranged in column vectors of length $P = 64$ and we use an overcomplete DCT base dictionary $\boldsymbol{\Psi} \in \mathbb{R}^{64\times100}$. We test two different implementations of our algorithm, the full version described in secs. 4.1 and 4.2 (BDS), and the only-addition technique in sec. 4.3 (BDSA). These techniques are compared with the state-of-the-art methods K-SVDS and K-SVD. These methods require explicit knowledge of the true noise and sparsity levels. To make a fair comparison, they are provided with the estimation of the noise and sparsity produced by our algorithm. We conducted 5 independent realizations of each experiment and present the corresponding average values.

Fig. 1 (left) shows that the estimated σ produced by the proposed technique is very accurate for small values. For larger values, the estimation precision decreases due to excessive degradation on the image. Enhanced precision can be achieved by the use of a higher number of training signals and/or a larger patch size.

Regarding the reconstruction error, as we can see in fig. 1 (right), all techniques perform similarly for smaller values of σ, although the two proposed methods slightly outperform K-SVDS and K-SVD for larger noise deviation.

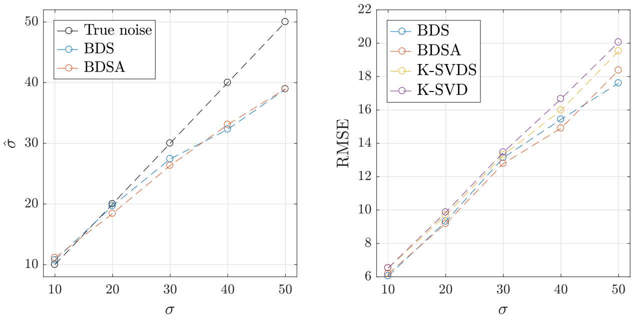


Fig. 1. Estimated σ (left) at different noise levels and RMSE reconstruction errors (right).

Table 2. Average PSNR and SSIM values.

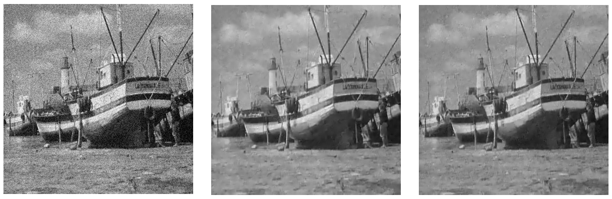
	Barbara		Boat		Peppers		
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	σ
BDS	29.09	0.91	**32.19**	**0.86**	23.14	0.85	10
BDSA	29.05	0.91	31.53	0.85	**23.69**	**0.86**	
K-SVDS	**31.12**	**0.92**	31.66	0.85	23.57	**0.86**	
K-SVD	29.07	0.88	29.27	0.82	24.53	0.84	
BDS	27.23	**0.86**	**29.13**	**0.79**	**23.93**	**0.81**	20
BDSA	26.81	**0.86**	28.17	**0.79**	22.31	0.80	
K-SVDS	**28.08**	**0.86**	28.37	**0.79**	21.44	0.80	
K-SVD	26.18	0.80	26.12	0.72	21.91	0.77	
BDS	**26.61**	**0.76**	**26.51**	**0.73**	**24.26**	**0.70**	30
BDSA	25.97	**0.76**	25.80	0.72	21.77	0.66	
K-SVDS	25.17	**0.76**	26.02	0.72	21.86	0.66	
K-SVD	23.33	0.69	23.17	0.64	20.64	0.63	
BDS	23.66	0.63	**23.95**	0.61	**22.27**	**0.64**	40
BDSA	22.92	0.68	23.24	**0.63**	21.60	0.55	
K-SVDS	**23.74**	**0.69**	23.31	0.62	21.60	0.56	
K-SVD	22.09	0.63	21.91	0.56	20.54	0.53	
BDS	**21.56**	**0.55**	21.81	0.51	**19.99**	**0.48**	50
BDSA	20.43	0.51	21.88	**0.52**	19.81	0.41	
K-SVDS	20.41	0.50	**21.91**	0.51	19.35	0.41	
K-SVD	19.77	0.46	20.96	0.46	19.25	0.39	

Table 2 shows the superiority of the proposed method in PSNR and SSIM. BDS presents the best results, being better than K-SVDS in 10 out of 15 test cases in terms of PSNR. The faster BDSA achieves competitive results, while accelerating the original algorithm. Notice also the better general behaviour of the double sparsity algorithms, compared to the standard single-sparse technique.

We conclude this section with a graphical sample of the algorithm's performance. Fig. 2 shows a corrupted image (a) along with the denoised output of the proposed method (b) and K-SVDS (c).

6. CONCLUSIONS

We have presented a new approximate Bayesian algorithm for ℓ_1 DS dictionary learning for denoising, inpainting and compressive sensing problems. In contrast to deterministic approaches, our method takes into account the uncertainty of the model which, we believe, gives it significant practical value in applications. We have shown that, unlike deterministic approaches which require explicit knowledge of the true noise level and the number of atoms used for the model in the sparse representations, our approach automatically estimates all the parameters involved. Two different strategies for the fast update of the representations have been proposed, both of them achieve very competitive PSNR and SSIM values. Due to space limitations, the model has been applied to image denoising only.



(a) Noisy image (b) BDSA (c) K-SVDS

Fig. 2. 'Boat' image denoising, $\sigma = 30$.

7. REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[2] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin, "Non-parametric Bayesian dictionary learning for sparse image representations," *Advances in Neural Information Processing Systems*, 2009.

[3] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on MachineLearning*, 2009, pp. 689–696.

[4] M. Lázaro-Gredilla and M. K. Titsias, "Spike and slab variational inference for multi-task and multiple kernel learning," in *Advances in neural information processing systems*, 2011, pp. 2339–2347.

[5] A. Szlam, K. Gregor, and Y. LeCun, *Fast Approximations to Structured Sparse Coding and Applications to Object Classification*, pp. 200–213, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[6] C. Bao, H. Ji, Y. Quan, and Z. Shen, "Dictionary learning for sparse coding: Algorithms and convergence analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, pp. 1356–1369, 2016.

[7] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, Dec 2006.

[8] B. Dumitrescu and P. Irofti, "Regularized K-SVD," *IEEE Signal Processing Letters*, vol. PP, no. 99, pp. 1–1, 2017.

[9] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, Nov 2010.

[10] X. Zhang, W. Zhou, and Z. Duan, "Image super-resolution reconstruction based on fusion of k-svd and semi-coupled dictionary learning," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Dec 2016, pp. 1–5.

[11] M. Elad, M. A. T. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 972–982, June 2010.

[12] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 2691–2698.

[13] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 3501–3508.

[14] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1553–1564, March 2010.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer Series in Statistics. Springer New York, 2009.

[16] R. Liang, Z. Zhao, and S. Li, "Image denoising using learned dictionary based on double sparsity model," in *2011 4th International Congress on Image and Signal Processing*, Oct 2011, vol. 2, pp. 691–695.

[17] F. Li and S. Zhang, "Double sparse dictionary learning for image super resolution," in *2016 Chinese Control and Decision Conference (CCDC)*, May 2016, pp. 4344–4348.

[18] X. Zhan, R. Zhang, D. Yin, A. Hu, and W. Hu, "Remote sensing image compression based on double-sparsity dictionary learning and universal trellis coded quantization," in *2013 IEEE International Conference on Image Processing*, Sept 2013, pp. 1665–1669.

[19] S. Li and H. Qi, "Compressed dictionary learning for detecting activations in fmri using double sparsity," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2014, pp. 434–437.

[20] M. Abavisani, M. Joneidi, S. Rezaeifar, and S. B. Shokouhi, "A robust sparse representation based face recognition system for smartphones," in *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, Dec 2015, pp. 1–6.

[21] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Bayesian compressive sensing using Laplace priors," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 53–63, Jan 2010.

[22] Z. Chen, R. Molina, and A. K. Katsaggelos, "Automated recovery of compressedly observed sparse signals from smooth background," *IEEE Signal Processing Letters*, vol. 21, no. 8, pp. 1012–1016, Aug 2014.