2015 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2015, BOSTON, USA

A BAYESIAN NON PARAMETRIC APPROACH TO LEARN DICTIONARIES WITH ADAPTED NUMBERS OF ATOMS

Hong Phuong Dang, Pierre Chainais

Ecole Centrale Lille, CRIStAL CNRS UMR 9189
INRIA Lille-Nord Europe, SequeL
CS 20048, 59651 Villeneuve d'Ascq, France
hong_phuong.dang, pierre.chainais@ec-lille.fr

ABSTRACT

Learning redundant dictionaries for sparse representation from sets of patches has proven its efficiency in solving inverse problems. In many methods, the size of the dictionary is fixed in advance. Moreover the optimization process often calls for the prior knowledge of the noise level to tune parameters. We propose a Bayesian non parametric approach which is able to learn a dictionary of adapted size : the adequate number of atoms is inferred thanks to an Indian Buffet Process prior. The noise level is also accurately estimated so that nearly no parameter tuning is needed. Numerical experiments illustrate the relevance of the resulting dictionaries.

Index Terms— sparse representations, dictionary learning, inverse problems, Indian Buffet Process

1. INTRODUCTION

Inverse problems in image processing (denoising, inpainting, deconvolution, super resolution...) are most often ill-posed problems so that the set of solutions is not a singleton. Some prior information or regularization is necessary. This can be based on the use of sparse representations. Such representations can be either some family of mathematical functions (DCT, wavelets...) or a dictionary learnt from the data, typically over a set of patches (e.g., 8×8 blocks). Several works have proposed to learn redundant dictionaries where the number K of atoms may be larger than the dimension P of the space, e.g, $K > P = 64$ for 8×8 patches. The richer the dictionary, the sparser the representation. An oversized dictionary leads to overfitting while too small it becomes useless. The choice of the size of a dictionary is crucial. A few works have elaborated on the seminal K-SVD approach [1] to propose dictionary learning (DL) methods that infer the size of the dictionary. They automatically determine the 'efficient' number of atoms to represent image patches like enhanced K-SVD [2], subclustering K-SVD [3] or stagewise K-SVD [4]. These strategies essentially alternate between two steps to either increase or decrease the size of the dictionary thanks to some modification of the K-SVD approach. A fast online approach is Clustering based Online Learning of Dictionaries (COLD) [5] which elaborates on the work in [6] by adding a mean-shift clustering step in the dictionary update step. Another strategy was proposed in [7] that starts from 2 atoms only. Then atoms are recursively bifurcated aiming at a compromise between the reconstruction error and the sparsity of the representation. In these optimization methods sparsity is typically promoted by L0 or L1 penalty terms on the set of encoding coefficients.

In [8], a Bayesian DL method is proposed thanks to a Beta-Bernoulli model where sparsity is promoted through an adapted Beta-Bernoulli prior to enforce many encoding coefficients to zero. Note that this corresponds to a parametric approximation of the Indian Buffet Process since this approach works with a (large) fixed number of atoms. The present contribution belongs to the same family but the size of the dictionary is no more fixed in advance. This is made possible thanks to the use of a Bayesian non parametric prior, namely an Indian Buffet Process (IBP) [9, 10] to both promote sparsity and deal with an adaptive number of atoms. The proposed method learns atoms starting from an empty dictionary, except the constant atom to treat the DC component apart as usual. Gibbs sampling is used for inference. The proposed method does not need to tune parameters since the level of noise, which determines the regularization level for sparse encoding, is also estimated during the dictionary learning. This makes the method truely *non parametric*, only some crude intialization is needed. We illustrate the relevance of this approach on a set of denoising experiments.

Section 2 briefly recalls on the problem of dictionary learning. Section 3 first presents the Indian Buffet Process (IBP) prior, then the proposed model and the Gibbs sampling algorithm for inference. Section 4 illustrates the relevance of our DL approach on classical image denoising experiments in comparison with other DL methods. Section 5 concludes.

Thanks to the BNPSI ANR project no ANR-13-BS-03-0006-01 and to the Fondation Ecole Centrale Lille for funding.

978-1-4673-7454-5/15/$31.00

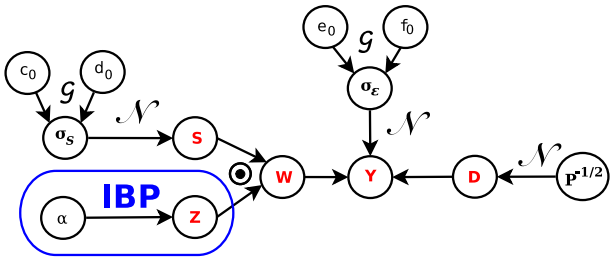

Fig. 1: Graphical model for IBP-DL.

2. DICTIONARY LEARNING (DL)

Here is a brief introduction to the DL optimization problem, see [11] for a review. Let matrix $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N] \in \mathbb{R}^{P \times N}$ a set of N image patches of size $\sqrt{P} \times \sqrt{P}$, ordered lexicographically as column vectors $\mathbf{y} \in \mathbb{R}^P$. Let dictionary of K atoms $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_K] \in \mathbb{R}^{P \times K}$. In presence of some additive noise $\boldsymbol{\varepsilon} \in \mathbb{R}^{P \times N}$, the data is modeled by

$$\mathbf{Y} = \mathbf{DW} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_N] \in \mathbb{R}^{K \times N}$ are the encoding coefficients. Each observation $\mathbf{y}_i$ should be described by a sparse set of coefficients $\mathbf{w}_i$. Usually, when working on image patches of size 8×8 (in dimension $P = 64$), a set of $K = 256$ or 512 atoms is learnt [1, 8, 11]. The noise is generally assumed to be Gaussian i.i.d. (reconstruction error = quadratic error). Sparsity is typically imposed through a L0 or L1-penalty in the mixed optimization problem (other formulations are possible):

$$(\mathbf{D}, \mathbf{W}) = \text{argmin}_{(\mathbf{D},\mathbf{W})} \quad \frac{1}{2}\|\mathbf{Y} - \mathbf{DW}\|_2^2 + \lambda\|\mathbf{W}\|_1 \tag{2}$$

Various approaches have been proposed to solve this problem by an alternate optimization on $\mathbf{D}$ and $\mathbf{W}$, including K-SVD (batch DL) [1, 11] and ODL (online DL) [6]. Note that the choice of the regularization parameter λ is of importance and should be decreasing with the noise level $\sigma_{\boldsymbol{\varepsilon}}$. We consider a Bayesian formulation of this problem and Gibbs sampling for inference; moreover the noise level is estimated simultaneously so that no parameter tuning is necessary.

3. PROPOSED APPROACH : IBP-DL

The present approach uses the Indian Buffet Process (IBP) [9, 10] as a Bayesian non parametric prior on sparse binary matrices. The IBP prior can be understood as a prior on sparse binary matrices with a potentially infinite number of rows, which is key to the learning of a dictionary for sparse representation with adaptive (potentially infinite) size. We only briefly recall about the IBP (see [10] for details) before describing the model and Gibbs sampling inference.

$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: Gaussian distribution of expectation $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

3.1. Indian Buffet Process (IBP)

The IBP was introduced in [9, 10] to deal with latent features models in a Bayesian non parametric framework. It can be built as the limit of a finite Beta-Bernoulli model with an infinite number of features. IBP provides a prior on infinite binary feature-assignement matrices $\mathbf{Z}$: $\mathbf{Z}(k, i) = 1$ if observation i owns feature k (0 otherwise). It combines two interesting properties for dictionary learning. IBP generates binary matrices that are *sparse* and *potentially infinite*. Therefore such a prior on the support of coefficients of a sparse representation with an adaptive number of atoms may be relevant. The properties of IBP are usually introduced thanks to the following 'history'. A sequence of customers (observations) taste dishes (features) in an infinite buffet. Customer i tastes dish k with probability m_k/i where m_k is the number of previous customers who have tasted dish k : this behaviour induces some clustering of customers'choices who exploits previous customers decisions. The customer i also tastes Poisson(α/i) new dishes, which allows for exploration. Taking into account the exchangeability of customers and the invariance to the ordering of features, IBP is characterized by a distribution on equivalence classes of binary matrices [9] :

$$P[\mathbf{Z}] = \frac{1}{\prod_{h=1}^{2^N-1} K_h!} \exp(-\alpha H_N) \prod_{k=1}^{\mathrm{K}_+} \frac{(N - m_k)!(m_k - 1)!}{N!} \tag{3}$$

where $H_N = \sum_{i=1}^{N} \frac{1}{i}$, m_k is the number of observations using feature k, K_+ is the number of features for which $m_k > 0$, K_h is the number of features with the same 'history' $\mathbf{Z}(k, \cdot) = h$. Parameter $\alpha > 0$ controls the expected total number of features that $K_+ \sim$ Poisson(αH_N) hence $\mathbf{E}[K_+] = \alpha H_N \simeq \alpha \log N$. The IBP permits to both deal with a variable sized dictionary (potentially infite but penalized) and promote sparsity (like a Bernoulli-Gaussian model).

3.2. The Bayesian non parametric model: IBP-DL

Fig. 1 shows the graphical model which may be expressed as

$$\begin{aligned}
\mathbf{y}_i &= \mathbf{D}\mathbf{w}_i + \boldsymbol{\varepsilon}_i, \forall 1 \leq i \leq N && (4)\\
\mathbf{w}_i &= \mathbf{z}_i \odot \mathbf{s}_i, \forall 1 \leq i \leq N && (5)\\
\mathbf{d}_k &\sim \mathcal{N}(0, P^{-1}\mathbb{I}_P), \forall k \in \mathbb{N} && (6)\\
\mathbf{Z} &\sim IBP(\alpha) && (7)\\
\mathbf{s}_i &\sim \mathcal{N}(0, \sigma_s^2\mathbb{I}_K), \forall 1 \leq i \leq N && (8)\\
\boldsymbol{\varepsilon}_i &\sim \mathcal{N}(0, \sigma_{\boldsymbol{\varepsilon}}^2\mathbb{I}_P), \forall 1 \leq i \leq N && (9)
\end{aligned}$$

where y_i is a column vector of dimension P, $\odot$ represents the Hadamard product. We place priors on $\mathbf{D}$, $\mathbf{W}$ and $\boldsymbol{\varepsilon}$. The vector $\mathbf{z}_i \in \{0, 1\}^K$ denotes wich of the K columns of $\mathbf{D}$ are used for representation of $\mathbf{y}_i$; $\mathbf{s}_i \in \mathbb{R}^K$ represents the coefficients used for this representation. The representation

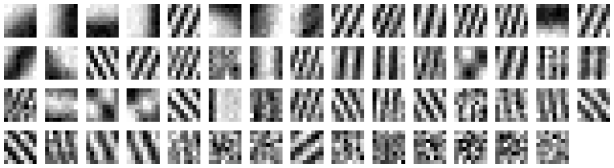

Fig. 2: Barbara, $\sigma_{\boldsymbol{\varepsilon}} = 40$: IBP-DL dictionary of 59 atoms.

$\mathbf{w}_i$ is defined : z_{ki}=0 $\Rightarrow$ w_{ki}=0 and z_{ki}=1 $\Rightarrow$ w_{ki}=s_{ki}, as in a parametric Bernoulli-Gaussian model. Hence, the sparsity properties of $\mathbf{W}$ are induced by the sparsity of $\mathbf{Z}$ thanks to the IBP prior. The present model also deals with a potentially infinite number of atoms $\mathbf{d}_k$ so that the size of the dictionary is not limited a priori. However, the IBP prior plays the role of a regularization term that tends to penalize the number K of active (non zero) rows in $\mathbf{Z}$; we have seen that $\mathbf{E}[K] \simeq \alpha \log N$ in the IBP. Except for σ_D^2 that is fixed to $1/P$, conjugate priors are used for parameters $\boldsymbol{\theta} = (\sigma_S^2, \sigma_{\boldsymbol{\varepsilon}}^2, \alpha)$: inverse Gamma distributions for variances with very small hyperparameters ($c_0 = d_0 = e_0 = f_0 = 10^{-6}$) to make hyperpriors vague; a $\mathcal{G}(1,1)$ for α associated to a Poisson law in the IBP. Detailed expressions of posterior distributions are given below. We emphasize that the noise variance $\sigma_{\boldsymbol{\varepsilon}}^2$ is estimated as well during inference, making the approach very close to truely non parametric. Fig 2 shows an example of a result of IBP-DL.

3.3. Algorithm for Gibbs sampling

Now we briefly describe the Gibbs sampling strategy to sample the posterior distribution $P(\mathbf{D}, \mathbf{S}, \mathbf{Z}, \boldsymbol{\theta}|\mathbf{Y})$.

Sampling $\mathbf{Z} \sim IBP(\alpha)$. $\mathbf{Z}$ is a matrix with an infinite number of rows, but only non-zero rows are kept in memory. Let $m_{k,-i}$ the number of observations other than i using atom k. One possible Gibbs sampling of the IBP goes in 2 steps [10] : 1) update the $z_{ki} = \mathbf{Z}(k,i)$ for 'active' atoms k such that $m_{k,-i} > 0$ (at least 1 patch other than i uses $\mathbf{d}_k$); 2) add new rows to $\mathbf{Z}$ which corresponds to activating new atoms in dictionary $\mathbf{D}$. In practice, one deals with finite matrices $\mathbf{Z}$ and $\mathbf{S}$ despite their theoretically potentially infinite size.

Update active atoms : The prior term is $p(z_{ki} = 1|\mathbf{Z}_{k,-i}) = m_{k,-i}/N$. The likelihood $p(\mathbf{Y}|\mathbf{D}, \mathbf{Z}, \mathbf{S}, \boldsymbol{\theta})$ is easily computed from the Gaussian noise model. Thanks to conjugacy of the prior on dictionary $\mathbf{D}$, we can marginalize $\mathbf{D}$ out. Then

$$p(\mathbf{Y}|\mathbf{Z}, \mathbf{S}, \sigma_{\boldsymbol{\varepsilon}}^2, \sigma_D^2) = p(\mathbf{Y}|\mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}^2, \sigma_D^2) = \frac{\exp\left\{-\frac{1}{2\sigma_{\boldsymbol{\varepsilon}}^2}\mathrm{tr}[\mathbf{Y}(\mathbb{I} - \mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \frac{\sigma_{\boldsymbol{\varepsilon}}^2}{\sigma_D^2}\mathbb{I})^{-1}\mathbf{W})\mathbf{Y}^T]\right\}}{(2\pi)^{NP/2}\sigma_{\boldsymbol{\varepsilon}}^{(N-K)P}\sigma_D^{KP}|\mathbf{W}\mathbf{W}^T + \frac{\sigma_{\boldsymbol{\varepsilon}}^2}{\sigma_D^2}\mathbb{I}|^{P/2}} \tag{10}$$

In [10], from Bayes' rule:

$$p(z_{ki}|\mathbf{Y}, \mathbf{Z}_{k,-i}, \mathbf{S}, \sigma_{\boldsymbol{\varepsilon}}^2, \sigma_D^2) \propto p(\mathbf{Y}|\mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}^2, \sigma_D^2) p(z_{ki}|\mathbf{Z}_{k,-i}) \tag{11}$$

$\mathcal{G}(x; a, b) = x^{a-1}b^a \exp(-bx)/\Gamma(a)$ pour $x > 0$

Init. : K=0, $\mathbf{Z}$=$\emptyset$, $\mathbf{D}$=$\emptyset$, α=1, σ_D^2=P^{-1}, σ_S^2=1, $\sigma_{\boldsymbol{\varepsilon}}$
Result: $\mathbf{D} \in \mathbb{R}^{P\times K}$, $\mathbf{Z} \in \{0;1\}^{K\times P}$, $\mathbf{S} \in \mathbb{R}^{K\times P}$, $\sigma_{\boldsymbol{\varepsilon}}$
for *iteration t=1:T* **do**
  Sample $\mathbf{Z} \sim IBP(\alpha)$
    for *data i=1:N* **do**
      for *atom k=1:K* **do**
        Sample $\mathbf{Z}(k,i)$ according to (11)
      end
      Sample k_{new} (# of new atoms) acc. to (13)
      Complete $\mathbf{Z}$ with k_{new} rows
      Complete $\mathbf{S}$ with k_{new} rows $\sim \mathcal{N}(0, \sigma_S^2)$
      Update K $\leftarrow$ size($\mathbf{Z}$,1)
    end
  for *intern loop f=1:F (e.g.* $1 \le F \le 10$*)* **do**
    for *atoms k=1:K* **do**
      Sample $\mathbf{d}_k \sim \mathcal{N}(\boldsymbol{\mu}_{dk}, \boldsymbol{\Sigma}_{dk})$ (14)
      Sample $\mathbf{S}(k, \mathbf{z}_k \neq 0) \sim \mathcal{N}(\mu_{sk}, \Sigma_{sk})$ (15)
    end
    Sample σ_S according to (16)
    Sample $\sigma_{\boldsymbol{\varepsilon}}$ according to (17)
  end
  Sample α according to (18)
end

Algorithm 1: Pseudo-algorithm of the IBP-DL method.

If row $\mathbf{Z}(k,\cdot)$=**0**, we suppress this row and the atom $\mathbf{d}_k$ in $\mathbf{D}$.

Activate new atoms : Following [12], we use a Metropolis-Hastings method to sample the number k_{new} of new atoms. This is equivalent in fact to deal with rows of $\mathbf{Z}$ such that $m_{k,-i} = 0$: this happens either when an atom is not used (inactive, not stored) or when it is used by 1 patch only. Rows with *singletons* have a unique coefficient 1 and zeros elsewhere: $z_{ki} = 1$ and $m_{k,-i} = 0$. To sample the number of new atoms amounts to sample the number of singletons since when a new atom is activated, it creates a new singleton. Let k_{sing} the number of such singletons in matrix $\mathbf{Z}$. Let $k_{prop} \in \mathbb{N}$ a proposal for the new number of singletons according to the same distribution as the prior on k_{sing} that is a Poisson law with parameter $\frac{\alpha}{N}$ in the IBP model:

$$J(k_{prop}) = \mathcal{P}(k_{prop}; \alpha/N) \tag{12}$$

Then the acceptance threshold is simply governed by the likelihood ratio after integrating new atoms $\mathbf{d}_k$ out. The proposal is accepted so that $k_{new} = k_{prop}$ if a uniform random variable $u \in (0,1)$ verifies

$$u \le \min\left(1, \frac{p(Y|k_{prop}, \text{rest})}{p(Y|k_{sing}, \text{rest})}\right) \tag{13}$$

Let $\gamma_{\boldsymbol{\varepsilon}}$=1/$\sigma_{\boldsymbol{\varepsilon}}^2$, γ_D=1/σ_D^2, γ_S=1/σ_S^2. Sampling $\mathbf{D}$, $\mathbf{S}$ and $\boldsymbol{\theta} = (\sigma_S^2, \sigma_{\boldsymbol{\varepsilon}}^2, \alpha)$ are done according to

$$\mathbf{D}\begin{cases} p(\mathbf{d}_k|\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{D}_{-k},\boldsymbol{\theta}) \propto \mathcal{N}(\boldsymbol{\mu}_{\mathbf{d}_k},\boldsymbol{\Sigma}_{\mathbf{d}_k}) \\ \boldsymbol{\Sigma}_{\mathbf{d}_k} = \left(\gamma_D \mathbb{I}_p + \gamma_{\boldsymbol{\varepsilon}} \sum_{i=1}^{N} w_{ki}^2\right)^{-1} \\ \boldsymbol{\mu}_{\mathbf{d}_k} = \gamma_{\boldsymbol{\varepsilon}} \boldsymbol{\Sigma}_{\mathbf{d}_k} \sum_{i=1}^{N} w_{ki}(\mathbf{y}_i - \mathbf{D}\mathbf{w}_i + \mathbf{d}_k w_{ki}) \end{cases} \quad (14)$$

$$\mathbf{S}\begin{cases} p(s_{ki}|\mathbf{Y},\mathbf{D},\mathbf{Z},\mathbf{S}_{k,-i},\boldsymbol{\theta}) \propto \mathcal{N}(\mu_{s_{ki}},\Sigma_{s_{ki}}) \\ z_{ki}=1 \Rightarrow \begin{cases} \Sigma_{s_{ki}} &= \left(\gamma_{\boldsymbol{\varepsilon}} \mathbf{d}_k^T \mathbf{d}_k + \gamma_S\right)^{-1} \\ \mu_{s_{ki}} &= \gamma_{\boldsymbol{\varepsilon}} \Sigma_{s_{ki}} \mathbf{d}_k^T (\mathbf{y}_i - \mathbf{D}\mathbf{w}_i + \mathbf{d}_k s_{ki}) \end{cases} \\ z_{ki}=0 \Rightarrow \begin{cases} \Sigma_{s_{ki}} &= \sigma_S^2 \\ \mu_{s_{ki}} &= 0 \end{cases} \end{cases} \quad (15)$$

$$\frac{1}{\sigma_S^2} \sim \mathcal{G}\left(c_0 + \frac{KN}{2}, d_0 + \frac{1}{2}\sum_{i=1}^{N} \mathbf{s}_i^T \mathbf{s}_i\right) \quad (16)$$

$$\frac{1}{\sigma_{\boldsymbol{\varepsilon}}^2} \sim \mathcal{G}\left(e_0 + \frac{NP}{2}, f_0 + \frac{1}{2}\sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{D}\mathbf{w}_i]\|_2^2\right) \quad (17)$$

$$\alpha \sim \mathcal{G}\left(1+K, 1+\sum_{j=1}^{N} 1/j\right) \quad (18)$$

An optional intern loop of F iterations can be used to sample $(\mathbf{D},\mathbf{S},\sigma_S,\sigma_{\boldsymbol{\varepsilon}}|\mathbf{Z})$ at fixed support $\mathbf{Z}$ (see Algo. 1).

4. NUMERICAL EXPERIMENTS

Dictionary learning (DL) provides an adapted representation to solve inverse problems. Even though there exist better state of the art methods for denoising, e.g. BM3D [13], one simple and usual way to compare the relevance of different dictionary learning methods is to compare their denoising performances. Results from BM3D are recalled for information only since we do not expect to obtain better results here. Present experiments aim at checking the relevance of the dictionaries obtained from the proposed IBP-DL in the simple setting for comparisons. In our experiments, 9 images of size 512×512 (8 bits) are processed for 2 noise levels $\sigma_{\boldsymbol{\varepsilon}}$=25 or 40. The initial value of $\widehat{\sigma}_{\boldsymbol{\varepsilon}}$ is set to a crude estimate of twice the true one in Algo. 1. There are $(512-7)^2 = 255025$ overlapping patches in each image. Here IBP-DL works with 16129 50%-overlapping patches only. The DC component (the mean value) is kept apart; it is associated to the constant atom $\mathbf{d}_0 = (1,...,1)$. The denoising method [14][1] use Orthogonal Matching Pursuit (OMP) and K-SVD dictionary. It averages pixel estimates from overlapping patches reconstructed by OMP with maximum number of coefficients set to ten. The images are denoised by using this method with the same default initialization: the maximum tolerance of representation error is set to $1.15\sigma_{\boldsymbol{\varepsilon}}$ and Lagrangian multiplier $\lambda = 30/\sigma_{\boldsymbol{\varepsilon}}$. But the K-SVD dictionary is here replaced by IBP-DL and the noise level σ_ε is its IBP-DL estimate $\widehat{\sigma}_{\boldsymbol{\varepsilon}}$.

[1]Matlab code by R. Rubinstein is available at `http://www.cs.technion.ac.il/~ronrubin/software.html`

Image	Case σ_ϵ=25	Case σ_ϵ=40
Initial value	$\sigma_{init}=51$	$\sigma_{init}=76.5$
Barbara	25.86	40.76
Boat	25.82	40.64
Cameraman	26.10	41.16
Fingerprint	25.79	40.90
GoldHill	25.89	40.70
House	25.53	40.36
Lena	25.45	40.24
Mandrill	27.57	42.50
Peppers	25.76	40.58

Table 1: Estimation results of noise level estimates for 2 noise levels in 9 images.

We illustrate the relevance of IBP-DL by comparing denoising results with BM3D (state of the art as a top reference) and several DL based methods [1]:

1. BM3D as a state of the art reference
2. K-SVD with K=256 learnt from all available patches,
3. K-SVD with K=256 learnt from the same reduced dataset as IBP-DL,
4. DLENE [7], an adaptive approach to learn overcomplete dictionaries with efficient numbers of elements.

Fig.3 gathers numerical results. It gives denoising performances and the dictionary size of IBP-DL, and results of other methods. The main observation is that IBP-DL performances are comparable to K-SVD, 0.3dB below at worst. Since our purpose is not to achieve the best denoising but to validate our dictionary learning approach, this is a good indication that IBP-DL dictionaries are as relevant as K-SVD ones. Note that the results using K-SVD are presented in the best conditions, that is when the parameters are set to their optimal values. This is possible in particular when an accurate estimate of the level of noise is available. However we emphasize that in IBP-DL the noise level is part of the set of estimated parameters so that the method does not call for any parameter tuning. Another important observation with respect to K-SVD is its sensitivity to the training set. It appears that denoising performances drop dramatically when a reduced training set is used which indicates a worse learning efficiency than IBP-DL of which performances are similar using a full or reduced training set. Here, to reduce computational time, IBP-DL works with a reduced set of 50% overlapping patches (16129 patches). Fig.3 shows that K-SVD performs much worse when using this same dataset in place of the full set of 255025 patches.

It is noticeable that IBP-DL dictionaries sometimes feature $K < 64$ atoms : the adaptive sized dictionary is not

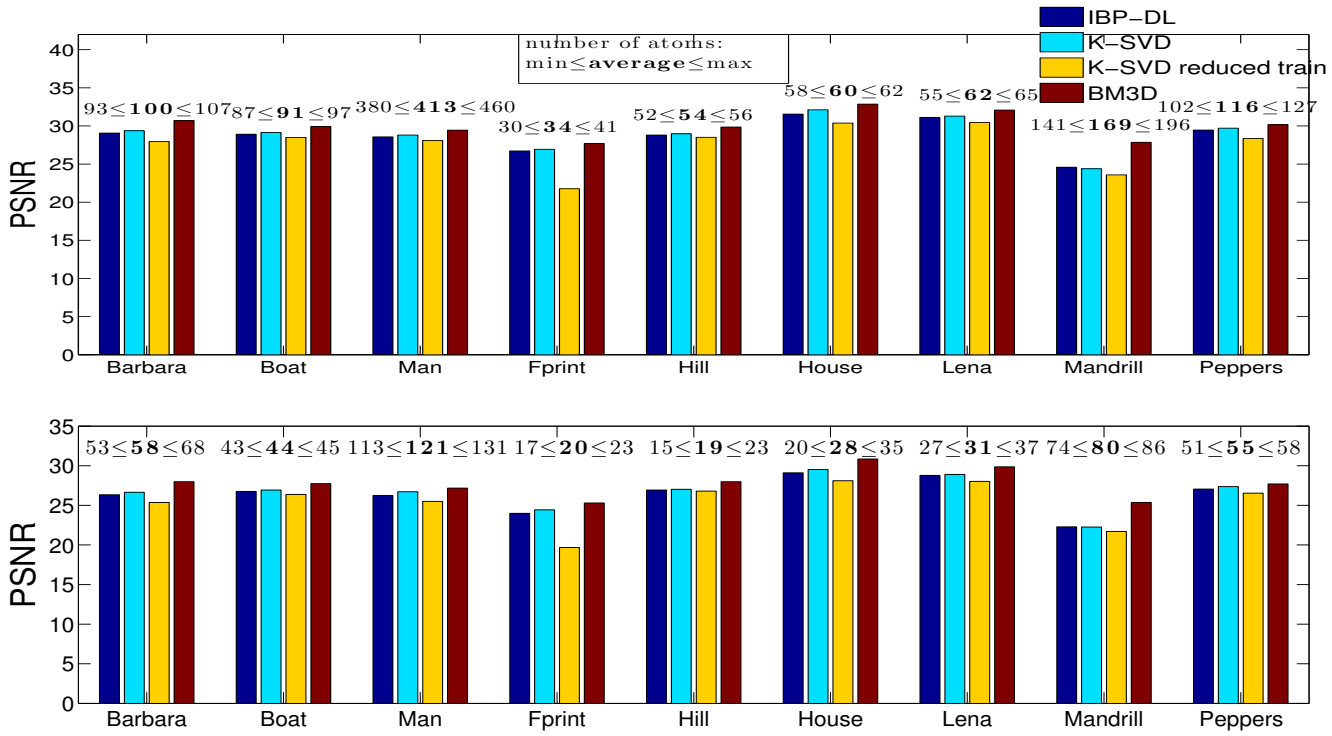

Fig. 3: Denoising results and sizes of IBP-DL dictionaries for noise level (top) $\sigma_\varepsilon = 25$, (bottom) $\sigma_\varepsilon = 40$. The text above each group of bars is the IBP-DL dictionary size. From left to right are the PSNR using IBP-DL, K-SVD with 256 atoms learnt from the full set of available patches, K-SVD with 256 atoms learnt from the reduced training set (as IBP-DL), BM3D.

always redundant. However, the denoising performance remains comparable with K-SVD that learns a larger redundant dictionary of 256 atoms : the IBP-DL dictionary well captures a reduced and efficient representation of the image content. One can observe that the size of the dictionary seems to increase for a smaller noise level. This is understandable since in the limit of no noise, the dictionary should ideally comprise all the original patches of the image (up to 255025) in a 1-sparse representation while in the limit of large noise, more and more patches must be combined to reduce noise in less atoms by averaging.

Fig. 4 shows the evolution of the size K of the dictionary over iterations. The final size (around 100 in this example) is attained typically after about 15 iterations only; implicitly, it also means that α converges to about $K/\log N \simeq 10$. Using a really non-parametric approach like IBP-DL, it appears that the size of the dictionary can considerably vary from one image to the other, for instance from dozens to hundreds at the same level of noise, see Table 3. The approach in [8] works with a finite but 'large' dictionary in which only a subset of atoms are used in the end. In practice, the size of the 'large' dictionary is most often fixed to 256 and the method generally yields dictionaries of size slightly smaller than 256, typically about 200 atoms. To this respect, IBP-DL improves on the previous method [8] and our observations support the interest of a non parametric approach that is more adaptive to the actual content of the image.

We briefly compare our results with DLENE [7], a recent work which also adapts the size of the dictionary. DLENE also uses the reduced training set. It targets a compromise between reconstruction error and sparsity by adapting the number of atoms. For Barbara with σ_ε=25, DLENE yields $\text{PSNR}_{\text{DLENE}}$= 28.82 dB while we get $\text{PSNR}_{\text{IBP-DL}}$=29.06 dB; for Peppers with σ_ε=40, $\text{PSNR}_{\text{DLENE}}$=27.27dB and $\text{PSNR}_{\text{IBP-DL}}$=27.07dB; see results in [7] for other comparisons. In general, IBP-DL performs as well as DLENE for denoising. Again this supports the relevance of the dictionaries by IBP-DL. One limitation of our algorithm is its computational cost because of Gibbs sampling. Indeed, the complexity per-iteration of the IBP sampler is $O(N^3(K^2 + KP))$. Sampling IBP is expensive even though the accelerated sampling [15] is implemented, reducing the complexity to $O(N(K^2 + KP))$, and a reduced dataset is used. For example, the Barbara image costs 1 hour for 30 iterations. There is room for a significant improvement on this aspect maybe by using a different kind of inference.

Finally, note that the noise level σ_ε is inferred with good accuracy. Fig. 5 shows the evolution of the sampled σ_ε over iterations on an example. After 15 iterations, the sampled value is very close to the ground truth and converges to it. Table 1 presents noise level estimates for a set of images. The estimation error is at most of a few percents only, from 2% to 10% when $\sigma_\varepsilon = 25$ and from 1% to 6% when $\sigma_\varepsilon = 40$. This accurate estimate is an essential profit of this approach.

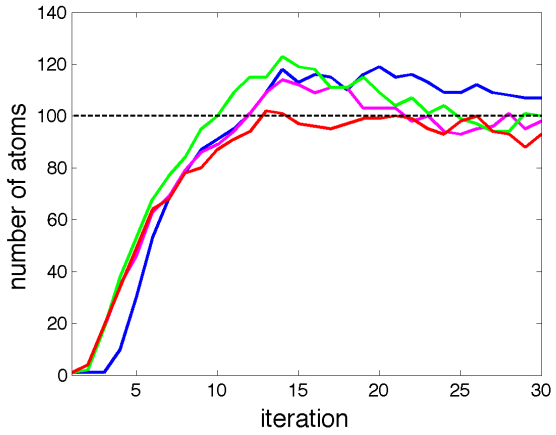

Fig. 4: Evolution of the number of atoms in the dictionary across iterations of IBP-DL on Barbara for 4 different noise realizations with $\sigma_\varepsilon = 25$.

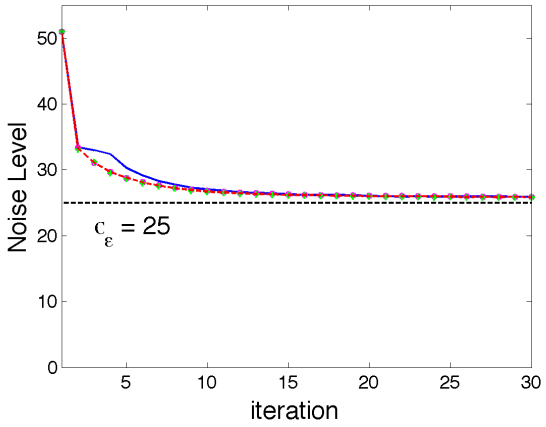

Fig. 5: Evolution of the noise level sampled over iterations of IBP-DL on Barbara when $\sigma_\varepsilon = 25$ and $\sigma_{init} = 51$.

5. CONCLUSION

The present Bayesian non parametric (BNP) approach learns a dictionary of adaptive size from noisy images. To illustrate and compare the relevance of the proposed IBP-DL with respect to other DL methods, numerical experiments study the denoising performances of the proposed IBP-DL: they are similar to those of other DL approaches such as K-SVD (in its optimal setting) [1] for fixed size or DLENE [7] for an adaptive size of the dictionary. The proposed approach simultaneously infers the size of the dictionary starting from an empty one, as well as other parameters of the model such as the noise level that is a crucial input to later solution of any inverse problem. We emphasize that IBP-DL appears as a *non parametric* method with an adaptive number of degrees of freedom and no parameter tuning. Future work will explore other inference methods for scalability.

6. 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] R. Mazhar and P.D. Gader, "Ek-svd: Optimized dictionary design for sparse representations," in *19th International Conference on Pattern Recognition*, Dec 2008, pp. 1–4.

[3] J. Feng, L. Song, X. Yang, and W. Zhang, "Sub clustering k-svd: Size variable dictionary learning for sparse representations," in *6th IEEE International Conference on Image Processing (ICIP)*, Nov 2009, pp. 2149–2152.

[4] C. Rusu and B. Dumitrescu, "Stagewise k-svd to design efficient dictionaries for sparse representations," *IEEE Signal Processing Letters*, vol. 19, no. 10, pp. 631–634, Oct 2012.

[5] N. Rao and F. Porikli, "A clustering approach to optimize online dictionary learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 1293–1296.

[6] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010.

[7] M. Marsousi, K. Abhari, P. Babyn, and J. Alirezaie, "An adaptive approach to learn overcomplete dictionaries with efficient numbers of elements," *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3272–3283, June 2014.

[8] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin, "Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images," *IEEE Transactions on Image Processing*, vol. 21, pp. 130–144, 2012.

[9] T. Griffiths and Z. Ghahramani, "Infinite latent feature models and the indian buffet process," in *Advances in NIPS 18*, pp. 475–482. MIT Press, 2006.

[10] T.L. Griffiths and Z. Ghahramani, "The indian buffet process: An introduction and review," *Journal of Machine Learning Research*, vol. 12, pp. 1185–1224, 2011.

[11] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27 –38, march 2011.

[12] D A Knowles and Z Ghahramani, "Nonparametric Bayesian sparse factor models with application to gene expression modeling," *The Annals of Applied Statistics*, vol. 5, no. 2B, pp. 1534–1552, 2011.

[13] K.. Dabov, A.. Foi, V.. Katkovnik, and K.. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080 –2095, aug. 2007.

[14] 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.

[15] F. Doshi-Velez and Z. Ghahramani, "Accelerated sampling for the indian buffet process," in *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, ICML '09, pp. 273–280, ACM.