Towards Dictionaries of Optimal Size: A Bayesian Non Parametric Approach

Hong-Phuong Dang, Pierre Chainais

▶ **To cite this version:**

Hong-Phuong Dang, Pierre Chainais. Towards Dictionaries of Optimal Size: A Bayesian Non Parametric Approach. Journal of Signal Processing Systems, 2018, 90 (2), pp.221-232. 10.1007/s11265-016-1154-1 . hal-01433621v1

HAL Id: hal-01433621

https://hal.science/hal-01433621v1

Submitted on 12 Jan 2017 (v1), last revised 15 Oct 2021 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Noname manuscript No.
(will be inserted by the editor)

Towards dictionaries of optimal size: a Bayesian non parametric approach

Hong Phuong Dang · Pierre Chainais

Received: date / Accepted: date

Abstract Solving inverse problems usually calls for adapted priors such as the definition of a well chosen representation of possible solutions. One family of approaches relies on learning redundant dictionaries for sparse representation. In image processing, dictionary learning is applied to sets of patches. Many methods work with a dictionary with a number of atoms that is fixed in advance. Moreover optimization methods often call for the prior knowledge of the noise level to tune regularization parameters. We propose a Bayesian non parametric approach that is able to learn a dictionary of adapted size. The use of an Indian Buffet Process prior permits to learn an adequate number of atoms. The noise level is also accurately estimated so that nearly no parameter tuning is needed. We illustrate the relevance of the resulting dictionaries on numerical experiments.

Keywords sparse representations · dictionary learning · inverse problems · Indian Buffet Process

1 Introduction

Ill-posed inverse problems such as denoising, inpainting, deconvolution or super resolution in image processing do not have a unique solution. The choice of some adequate representation space thanks to some prior information or regularization is necessary so that one can identify a unique and relevant solution. In recent years, many works have proposed to use sparse representations [1]. A signal admits a sparse representation in some dictionary if it can be reconstructed by using a small subset of a redundant set of atoms, a redundant

Hong Phuong Dang, Pierre Chainais
Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France
E-mail: hong_phuong.dang, pierre.chainais@ec-lille.fr

dictionary. In image processing, this principle has led to two main classes of approaches.

One possibility is to use some family of mathematical functions that share generic geometrical properties of images like discrete-cosine-transform (DCT) [2] , wavelets [3], ... Such family of functions can be either orthogonal or redundant. The choice of the dictionary remains crucial and greatly influences the quality of the results. This paper adopts the second option which is dictionary learning: a redundant dictionary of atoms is learnt from a set of reference signals. The most simple technique is Principal Component Analysis (PCA) that learns an orthonormal basis through matrix factorization. Redundant dictionaries gather a number of atoms K that is greater than the dimension P of the physical, inpired by the seminal work by Olshausen and Field 1996 [4].

Many dictionary learning methods solve an optimization problem. The approaches in [5, 6], [7, 8] propose an optimal dictionary by setting in advance a large size (256 or 512) of the dictionary. A fast online approach is Clustering based Online Learning of Dictionaries (COLD) [9] which elaborates on the work in [10] by adding a mean-shift clustering step in the dictionary update step. The choice of the size of a dictionary is crucial. A few works have therefore elaborated on the seminal K-SVD approach [5] 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 [11], subclustering K-SVD [12] or stagewise K-SVD [13]. 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. Another strategy called DLENE [14] 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. However, they suffer from some limitations. They often fix in advance the noise level, the size of the dictionary or the sparsity level.

Bayesian approaches have been much less studied. In [15], a Bayesian DL method called BPFA is proposed thanks to a Beta-Bernoulli model. The BPFA method promotes sparsity 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 proposes a Bayesian non parametric approach where the size of the dictionary is no more fixed in advance thanks to the use of an Indian Buffet Process (IBP) prior [16, 17] to both promote sparsity and deal with an adaptive number of atoms. The proposed method starts from an empty dictionary, except the constant atom to treat the DC component apart as usual. Gibbs sampling is used for inference. Itdoes 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* since only some crude intialization

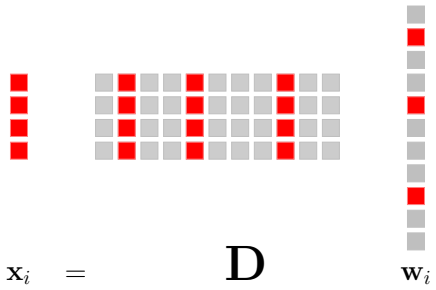


Fig. 1 $\mathbf{x}_i$ is sparse on $\mathbf{D}$: only few coefficients are active in $\mathbf{w}_i$.

is needed. We illustrate the relevance of this approach on a set of denoising experiments.

The paper is organized as follows. 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 a synthetic dataset as well as on the reconstruction of a clean image (without noise) and classical image denoising experiments in comparison with other methods. Section 5 concludes and evokes some directions for future work.

2 Dictionary learning (DL)

Let matrix $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N] \in \mathbb{R}^{P \times N}$ a set of N observations $\mathbf{y}_i$. In image processing, each vector $\mathbf{y}_i \in \mathbb{R}^P$ represents a patch of size $\sqrt{P} \times \sqrt{P}$, in lexicographic order as column vectors. Let matrix $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{P \times N}$ represent the patches of the initial image. In presence of some additive noise $\boldsymbol{\varepsilon} \in \mathbb{R}^{P \times N}$, the data is modeled by

$$\begin{cases} \mathbf{Y} = \mathbf{X} + \boldsymbol{\varepsilon} \\ \mathbf{X} = \mathbf{D}\mathbf{W} \end{cases} \tag{1}$$

where $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_N] \in \mathbb{R}^{K \times N}$ are the encoding coefficients and $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_K] \in \mathbb{R}^{P \times K}$ is the dictionary of K atoms. Each $\mathbf{x}_i$ is described by a sparse set of coefficients $\mathbf{w}_i$, see Fig. 1. When working on image patches of size 8×8 (in dimension $P = 64$), a set of $K = 256$ or 512 atoms is typically learnt [1, 5, 15]. 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):

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

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, 5] and ODL (online DL) [10]. Note that the choice of the regularization parameter λ is of importance and should decrease as the noise level $\sigma_{\boldsymbol{\varepsilon}}$ increases.

Fig. 2 Dictionary of 59 atoms learnt from Barbara using IBP-DL with a noise level of $\sigma_{\boldsymbol{\varepsilon}} = 40$.

In the Bayesian framework, the problem is typically written in the form of a likelihood built according to the model (1):

$$p(\mathbf{Y} \mid \mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}) = \frac{1}{(2\pi\sigma_{\boldsymbol{\varepsilon}}^2)^{NP/2}} \exp(-\frac{1}{2\sigma_{\boldsymbol{\varepsilon}}^2} \mathrm{tr}[(\mathbf{Y} - \mathbf{DW})^{\mathrm{T}}(\mathbf{Y} - \mathbf{DW})]) \quad (3)$$

The prior $p(\mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}})$ plays the role of regularization and the joint posterior writes:

$$p(\mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}) p(\mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}) \quad (4)$$

Using Gibbs sampling for inference, for example, the problem can be solved by sampling alternately:

$$p(\mathbf{W} \mid \mathbf{Y}, \mathbf{D}, \sigma_{\boldsymbol{\varepsilon}}) \propto p(\mathbf{Y} \mid \mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}) p(\mathbf{W}) \quad (5)$$

$$p(\mathbf{D} \mid \mathbf{Y}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}) \propto p(\mathbf{Y} \mid \mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}) p(\mathbf{D}) \quad (6)$$

$$p(\sigma_{\boldsymbol{\varepsilon}} \mid \mathbf{Y}, \mathbf{D}, \mathbf{W}) \propto p(\mathbf{Y} \mid \mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}) p(\sigma_{\boldsymbol{\varepsilon}}) \quad (7)$$

In the parametric framework, the size of the dictionary must be set in advance. Taking benefit from the Bayesian non parametric framework, we propose a learning method without setting the size of dictionary in advance thanks to an Indian Buffet Process prior [17]. The noise level is estimated simultaneously so that no parameter tuning is necessary. The method is called IBP-DL for Indian Buffet Process in Dictionary Learning. Fig. 2 shows an example of a result of IBP-DL on Barbara image.

3 Proposed approach : IBP-DL

The present approach uses the Indian Buffet Process (IBP) [16,17] as a Bayesian non parametric prior on sparse binary matrices with a potentially infinite number of rows. This model is key to the learning of a dictionary for sparse representation with adaptive (potentially infinite) size. We only briefly recall about the IBP, see [17] for details, before describing the model and Gibbs sampling inference.

3.1 Indian Buffet Process (IBP)

The IBP was introduced in [16, 17] to deal with latent feature models in a Bayesian non parametric framework. As a reminder, in the case of latent class

Fig. 3 One realization of the IBP with α=10.

models each observation belongs to a single class only. In the latent feature models each observation can be associated to a vector of latent features. For a finite number K of binary of latent features, each element of the vector of latent features may be considered as a Bernoulli random variable. The weights of features drawn from a conjugate Beta prior.

The IBP can be built as the limit of a finite Beta-Bernoulli model with K features when K$\rightarrow \infty$. It provides a prior on infinite binary feature-assignement matrices $\mathbf{Z}$ such that $\mathbf{Z}(k,i) = 1$ 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'corresponding to the Polya's urn description. A sequence of customers (observations) tastes 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. This customer then also tastes Poisson(α/i) new dishes, which allows for exploration and innovation. Fig. 3 illustrates a realization of the IBP with 20 customers and α=10.

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 [16] :

$$P[\mathbf{Z}] = \frac{\alpha^{\mathrm{K}_+}}{\prod_{h=1}^{2^N-1} K_h!} \exp(-\alpha H_N) \prod_{k=1}^{\mathrm{K}_+} \frac{(N-m_k)!(m_k-1)!}{N!} \tag{8}$$

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' $h = \mathbf{Z}(k,\cdot)$. The parameter $\alpha > 0$ controls the expected total number of features. It appears that $K_+ \sim$ Poisson(αH_N), hence

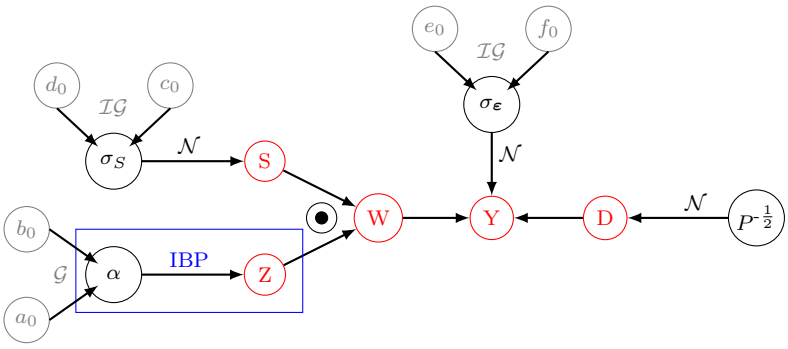


Fig. 4 Graphical model for IBP-DL.

$\mathbb{E}[K_+]=\alpha H_N \simeq \alpha \log N$. Fig. 3 illustrates the regularization effect if the IBP through the logarithmic growth of the number of dishes K_+ with the number of customers N. Some dishes are often used ; for example third dish is used by all customers, and other is rarely used, only one customer chooses dish 35. This shows a sparse effect. 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

The model is described by[1] :

$$\mathbf{y}_i = \mathbf{D}\mathbf{w}_i + \boldsymbol{\varepsilon}_i, \forall 1 \leq i \leq N \tag{9}$$
$$\mathbf{w}_i = \mathbf{z}_i \odot \mathbf{s}_i, \forall 1 \leq i \leq N \tag{10}$$
$$\mathbf{d}_k \sim \mathcal{N}(0, P^{-1}\mathbb{I}_P), \forall k \in \mathbb{N} \tag{11}$$
$$\mathbf{Z} \sim IBP(\alpha) \tag{12}$$
$$\mathbf{s}_i \sim \mathcal{N}(0, \sigma_s^2\mathbb{I}_K), \forall 1 \leq i \leq N \tag{13}$$
$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma_{\boldsymbol{\varepsilon}}^2\mathbb{I}_P), \forall 1 \leq i \leq N \tag{14}$$

where $\mathbf{y}_i$ is a column vector of dimension P, $\odot$ represents the Hadamard product. The vector $\mathbf{z}_i \in \{0,1\}^K$ encodes which of the K columns of $\mathbf{D}$ are used to represent $\mathbf{y}_i$; $\mathbf{s}_i \in \mathbb{R}^K$ represents the coefficients used for this representation. The representation coefficients are defined as $w_{ki} = z_{ki}s_{ki}$, in the spirit of a parametric Bernoulli-Gaussian model. 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. The IBP prior plays the role of a regularization term that penalizes the number K of active (non zero) rows in $\mathbf{Z}$ since

[1] $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: Gaussian distribution with expectation $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

$\mathbb{E}[K] \simeq \alpha \log N$ in the IBP. Except for σ_D^2 that is fixed to $1/P$, conjugate priors are used for parameters $\boldsymbol{\theta} = (\sigma_S^2, \sigma_{\boldsymbol{\varepsilon}}^2, \alpha)$: vague inverse Gamma distributions[2] for variances with very small hyperparameters ($c_0 = d_0 = e_0 = f_0 = 10^{-6}$) are used for $\sigma_{\boldsymbol{\varepsilon}}^2$, σ_S^2, and a $\mathcal{G}(1,1)$ for α associated to a Poisson law in the IBP. Posterior distributions are detailed on the next section. We emphasize that the noise variance $\sigma_{\boldsymbol{\varepsilon}}^2$ is estimated during inference, making the approach very close to truely non parametric. Fig. 4 shows the graphical model.

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 [17] :

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. We now describe these steps in more detail.

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. Hence, with $\mathbf{W} = \mathbf{Z} \odot \mathbf{S}$, we obtain the collapsed likelihood

$$p(\mathbf{Y}|\mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}^2, \sigma_D^2) = \frac{\exp\left\{-\frac{1}{2\sigma_{\boldsymbol{\varepsilon}}^2}\mathrm{tr}[\mathbf{Y}(\mathbb{I} - \mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \frac{\sigma_{\boldsymbol{\varepsilon}}^2}{\sigma_D^2}\mathbb{I})^{-1}\mathbf{W})\mathbf{Y}^T]\right\}}{(2\pi)^{NP/2}\sigma_{\boldsymbol{\varepsilon}}^{(N-K)P}\sigma_D^{KP}|\mathbf{W}\mathbf{W}^T + \frac{\sigma_{\boldsymbol{\varepsilon}}^2}{\sigma_D^2}\mathbb{I}|^{P/2}} \quad (15)$$

From Bayes' rule:

$$p(z_{ki}|\mathbf{Y}, \mathbf{Z}_{-(k,i)}, \mathbf{S}, \sigma_{\boldsymbol{\varepsilon}}^2, \sigma_D^2) \propto p(\mathbf{Y}|\mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}^2, \sigma_D^2)p(z_{ki}|\mathbf{Z}_{-(k,i)}) \quad (16)$$

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

Activate new atoms : Following [18], 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. We choose to integrate out $\mathbf{D}$ then we don't need to propose the new atoms in $\mathbf{D_{new}}$.

[2] $\mathcal{G}(x; a, b) = x^{a-1}b^a \exp(-bx)/\Gamma(a)$ for $x > 0$.

Note that we can choose to integrate out $\mathbf{D}$ or $\mathbf{S}$, but not both. Let k_{sing} the number of such singletons in matrix $\mathbf{Z}$, S_{sing} the coefficients corresponding to k_{sing}. Let $k_{prop} \in \mathbb{N}$ a proposal for the new number of singletons and S_{prop} the new proposed coefficients corresponding to k_{prop}. Thus the proposal is $\zeta_{prop} = \{k_{prop}, S_{prop}\}$ and we propose a move $\zeta_{sing} \to \zeta_{prop}$ with a probability having the form :

$$J(\zeta_{prop}) = J_K(k_{prop}) J_S(S_{prop}) \tag{17}$$

The proposal is accepted with probability min (1, $a_{\zeta_{sing} \to \zeta_{prop}}$) where

$$a_{\zeta_{sing} \to \zeta_{prop}} = \frac{P(\zeta_{prop} \mid \mathbf{Y}, rest) J(\zeta_{sing})}{P(\zeta_{sing} \mid \mathbf{Y}, rest) J(\zeta_{prop})} = \frac{p(\mathbf{Y}|\zeta_{prop}, rest)}{p(\mathbf{Y}|\zeta_{sing}, rest)} a_K a_S \tag{18}$$

where

$$a_K = \frac{\mathcal{P}oisson(k_{prop}; \alpha/N) J_K(k_{sing})}{\mathcal{P}oisson(k_{sing}; \alpha/N) J_K(k_{prop})} \ , \ a_S = \frac{\mathcal{N}(S_{prop}; 0, \sigma_S^2) J_S(S_{sing})}{\mathcal{N}(S_{sing}; 0, \sigma_S^2) J_S(S_{prop})} \tag{19}$$

The simplest proposal would be to use the prior on ζ_{prop}, i.e.

$$J_K(k_{prop}) = \mathcal{P}oisson(k_{prop}; \alpha/N) \text{ then } a_K = 1 \tag{20}$$

$$J_S(S_{prop}) = \mathcal{N}(S_{prop}; 0, \sigma_S^2) \text{ then } a_S = 1 \tag{21}$$

Then the acceptance threshold is simply governed by the collapsed likelihood ratio. The proposal is accepted, that is $\zeta_{new} = \zeta_{prop}$, if a uniform random variable $u \in (0, 1)$ verifies

$$u \leq \min\left(1, \frac{p(\mathbf{Y}|\zeta_{prop}, \text{rest})}{p(\mathbf{Y}|\zeta_{sing}, \text{rest})}\right) \tag{22}$$

Note that when we integrate out a variable somewhere and later this variable occurs in another posterior, it must be sampled before reusing [21]. Therefore $\mathbf{D}$ must be sampled immediately after $\mathbf{Z}$. Sampling $\mathbf{D}$, $\mathbf{S}$ and $\boldsymbol{\theta} = (\sigma_S^2, \sigma_{\boldsymbol{\varepsilon}}^2, \alpha)$ are done according to

$$\mathbf{D} \begin{cases} p(\mathbf{d}_k|\mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{D}_{-k}, \boldsymbol{\theta}) \propto \mathcal{N}(\boldsymbol{\mu}_{\mathbf{d}_k}, \boldsymbol{\Sigma}_{\mathbf{d}_k}) \\ \boldsymbol{\Sigma}_{\mathbf{d}_k} = (\sigma_D^{-2} \mathbb{I}_P + \sigma_{\boldsymbol{\varepsilon}}^{-2} \mathbb{I}_P \sum_{i=1}^{N} w_{ki}^2)^{-1} \\ \boldsymbol{\mu}_{\mathbf{d}_k} = \sigma_{\boldsymbol{\varepsilon}}^{-2} \boldsymbol{\Sigma}_{\mathbf{d}_k} \sum_{i=1}^{N} w_{ki} (\mathbf{y}_i - \sum_{j \neq k}^{K} \mathbf{d}_j w_{ji}) \end{cases} \tag{23}$$

$$\mathbf{S} \begin{cases} p(s_{ki}|\mathbf{Y}, \mathbf{D}, \mathbf{Z}, \mathbf{S}_{-(k,i)}, \boldsymbol{\theta}) \propto \mathcal{N}(\mu_{s_{ki}}, \Sigma_{s_{ki}}) \\ z_{ki} = 1 \Rightarrow \begin{cases} \Sigma_{s_{ki}} = (\sigma_{\boldsymbol{\varepsilon}}^{-2} \mathbf{d}_k^T \mathbf{d}_k + \sigma_S^{-2})^{-1} \\ \mu_{s_{ki}} = \sigma_{\boldsymbol{\varepsilon}}^{-2} \Sigma_{s_{ki}} \mathbf{d}_k^T (\mathbf{y}_i - \sum_{j \neq k}^{K} \mathbf{d}_j w_{ji}) \end{cases} \\ z_{ki} = 0 \Rightarrow \begin{cases} \Sigma_{s_{ki}} = \sigma_S^2 \\ \mu_{s_{ki}} = 0 \end{cases} \end{cases} \tag{24}$$

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 $\boldsymbol{Z} \sim$ *IBP*(α)
    for *data i=1:N* **do**
      for *atom k=1:K* **do**
        Sample $\mathbf{Z}(k,i)$ according to (16)
      end
      Sample k_{new} (# of new atoms) acc. to (22)
      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
  Sample $\boldsymbol{D}$ *and* $\boldsymbol{S}$
    for *atoms k=1:K* **do**
      Sample $\mathbf{d}_k \sim \mathcal{N}(\boldsymbol{\mu}_{dk}, \boldsymbol{\Sigma}_{dk})$ (23)
      Sample $\mathbf{S}(k, \mathbf{z}_k \neq 0) \sim \mathcal{N}(\mu_{sk}, \Sigma_{sk})$ (24)
    end
  Sample $\boldsymbol{\theta} = (\sigma_S^2, \sigma_{\boldsymbol{\varepsilon}}^2, \alpha)$
    Sample σ_S according to (25)
    Sample $\sigma_{\boldsymbol{\varepsilon}}$ according to (26)
    Sample α according to (27)
end

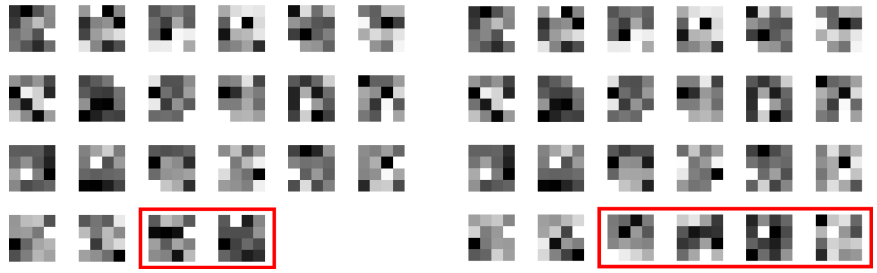
Algorithm 1: Pseudo-algorithm of the IBP-DL method.

$$\frac{1}{\sigma_S^2} \sim \mathcal{G}\left(c_0 + \frac{KN}{2}, d_0 + \frac{1}{2}\sum_{i=1}^{N} \mathbf{s}_i^T \mathbf{s}_i\right) \tag{25}$$

$$\frac{1}{\sigma_{\boldsymbol{\varepsilon}}^2} \sim \mathcal{G}\left(e_0 + \frac{NP}{2}, f_0 + \frac{1}{2}\sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{D}\mathbf{w}_i]\|_2^2\right) \tag{26}$$

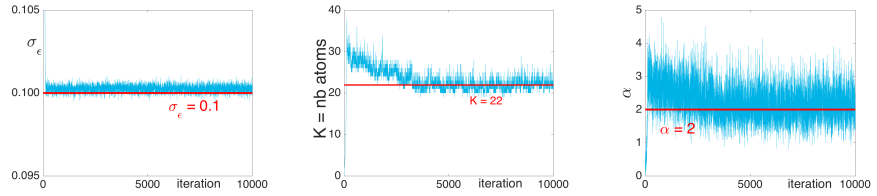
$$\alpha \sim \mathcal{G}\left(1 + K, 1 + \sum_{j=1}^{N} 1/j\right) \tag{27}$$

One limitation of our algorithm is its computational cost because of Gibbs sampling. Indeed, the complexity per-iteration of the IBP sampler is $O(N^3(K^2+KP))$ due to the matrix in the exponent of the collapsed likelihood (15). The accelerated Gibbs sampling [19] proposes to maintain the posterior over $\mathbf{D}$ instead of integrating out $\mathbf{D}$ entirely. The observations $\mathbf{Y}$ and the features assignement matrix $\mathbf{W}$ can split into two parts : one for the observation i, and one for the rest. Sampling of $\mathbf{z}_i$ will consist of *removing* the influence of a single observation $\mathbf{y}_i$ from the posterior over $\mathbf{D}$. Once $\mathbf{z}_i$ is sampled, we *restore* the influence of $\mathbf{y}_i$ into this posterior. The accelerated sampling [19] can reduce the complexity to $O(N(K^2+KP))$. In pratice, sufficient statistics (information form) are used to remove and to restore easily the influence of a single observation [20].



(a) (b)

Fig. 5 Comparison between (a) the original dictionary of $K = 22$ atoms for synthesis and (b) the dictionary of $K' = 24$ atoms estimated by IBP-DL. Red rectangles point on atoms for which no correlation > 0.55 was found between original and estimated atoms : 18 atoms are retrieved at a level of 0.55 correlation.



(a) (b) (c)

Fig. 6 Evolution of parameters' samples through iterations of IBP-DL: (a) the noise level $\sigma_{\boldsymbol{\epsilon}}$, (b) the number of atoms K, (c) the parameter α of the IBP prior.

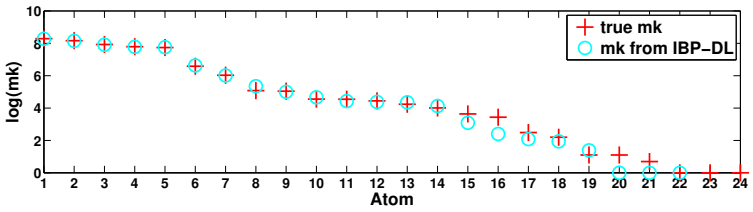
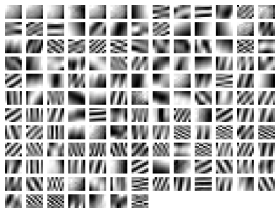


Fig. 7 Comparaison the distribution of the number m_k of observations using atom k beetween the true m_k of synthetic dataset and m_k from IBP-DL.

4 Numerical experiments

4.1 Toy example

As a first experiment, we prepare a synthetic dataset of $N = 10000$ samples from the proposed generative model in dimension $P = 16$. Each sample can be seen as a small 4×4 image. First, a realization $\mathbf{Z}$ of an IBP(α) with $\alpha = 2$ is drawn, leading to a total number of $K = 22$ features ($\mathbf{Z}$ is $K \times N$). Note that $K = 22$ is close to $E[K] \simeq \alpha \log N \simeq 18$. Then a dictionary of K atoms is built according to a normal law with variance $\sigma_D^2 = 1/P$. Coefficients $\mathbf{S}$ are i.i.d.


(a)Dictionary : 150 atoms


(b) Initial image


(c) Restored image , PSNR=44.94dB

Fig. 8 Illustration of noiseless restoration obtained by using IBP-DL on a segment of barbara image

Gaussian variables with $\sigma_S^2 = 1$. The dataset is corrupted by some additive white Gaussian noise with $\sigma_{\boldsymbol{\varepsilon}} = 0.1$. Finally, the dataset is built from :

$$\mathbf{Y} = \mathbf{D}(\mathbf{Z} \odot \mathbf{S}) + \boldsymbol{\varepsilon} \tag{28}$$

Fig. 5 shows the comparison between the original dictionary, used to synthesize the dataset, and the dictionary sampled at the last iteration of IBP-DL, here after 10000 iterations. Atoms have been reordered to make the correspondence more visible. One can observe that 20 atoms over 22 are retrieved at a 0.55 level of correlation (18 at 0.9 level, 14 at 0.99 level). Only 2 atoms of the original dictionary, which are much correlated to her atoms within the original dictionary, are not identified. This shows that the algorithm behaves very well on this toy model.

Fig. 6 illustrates the behaviour of parameters' samples across iterations. One can see that the noise level $\sigma_{\boldsymbol{\varepsilon}}$, the number of atoms K and the parameter α of the IBP prior soon fluctuate around their expected values after about 3000 iterations corresponding to the burn in time of the sampler.

Fig. 7 shows the distribution in descending order of the number m_k of observations using atom k in the synthetic dataset compared to the m_k inferred by IBP-DL. They are very similar, above all the first 14 atoms that reach a 0.99 level of correlation. The less frequently atoms are used, the more difficult they are to retrieve. In the synthetic dataset, 4 atoms are used by less than 10 observations over 10000, corresponding to 4 atoms which are not identified at a 0.55 level of correlation.

4.2 Noiseless restoration example

Now we consider dictionary learning from an image to evaluate the performance of the IBP-DL. A dictionary is learnt from a segment of size 256×256 of Barbara image without noise. The full data set of 62001 overlapping patches is used to restore this segment. IBP-DL method yields an adapted dictionary with 150 atoms and the reconstruction is very accurate since PSNR=44.94dB. Fig. 8 displays the result of image restoration without noise by using IBP-DL. For comparison, K-SVD produces a dictionary of fixed size 256 and a larger

Table 1 Results of IBP-DL for denoising applied to 9 images. Averaged estimated noise level are 25.97 using $\sigma_{init} = 51$, resp. 40.87 using $\sigma_{init} = 76.5$, when the true level was 25, resp. 40.

	PSNR $\simeq$ 20.17dB, $\sigma_\epsilon = 25$			PSNR $\simeq$ 16.08dB, $\sigma_\epsilon = 40$		
	PSNR [dB]	# atoms	$\hat{\sigma}_\epsilon$	PSNR [dB]	# atoms	$\hat{\sigma}_\epsilon$
Barbara	29.06	100	25.86	26.34	58	40.76
Boat	28.92	91	25.82	26.75	44	40.64
Carmeraman	28.57	413	26.10	26.24	121	41.16
Fingerprint	26.72	34	25.79	23.99	20	40.90
GoldHill	28.80	54	25.89	26.93	19	40.70
House	31.55	60	25.53	29.11	28	40.46
Lena	31.12	62	25.45	28.78	31	40.24
Mandrill	24.59	169	27.57	22.29	80	42.50
Peppers	29.46	116	25.76	27.06	55	40.58

reconstruction error since PSNR=43.97dB. The Bayesian method proposed in [15] with a dictionary of size 256 as well obtains PSNR=42.92dB. IBP-DL restores the image with an adapted yet smaller number of atoms and a better quality of approximation.

4.3 Denoising example

Dictionary learning (DL) provides an adapted representation to solve inverse problems. Even though there exist potentially better state of the art methods for denoising, e.g. BM3D [22], one simple and usual way to compare the relevance of different dictionary learning methods is to compare their denoising performances. Present experiments aim at checking the relevance of the dictionaries obtained from the proposed IBP-DL.

Table 1 gathers numerical denoising performances of IBP-DL as well as the dictionary size and the estimated noise level for 9 images of size 512×512 (8 bits) for 2 noise levels $\sigma_{\boldsymbol{\epsilon}}$=25 or 40 correspond respectively to PSNR=20.17dB and 16.08dB. There are $(512-7)^2 = 255025$ overlapping patches in each image. Here IBP-DL learns from 16129 50%-overlapping patches only (for sake of limited numerical complexity). The initial value of $\widehat{\sigma}_{\boldsymbol{\epsilon}}$ is set to a crude estimate of twice the true one in Algo. 1. Using a really non-parametric approach like IBP-DL, it appears that the size of the dictionary can considerably vary from one image to another, for instance from dozens to hundreds at the same level of noise, see Table 1. Note that the noise level $\sigma_{\boldsymbol{\epsilon}}$ is inferred with good accuracy.

Fig. 9 displays typical denoising results obtained by using IBP-DL on several examples of Table 1. The denoising images have a good quality. IBP-DL learns from a reduced set of 50%-overlapping patches. The DC component (the mean value) is kept apart: it is associated to the constant atom $\mathbf{d}_0 = (1, ..., 1)$. The resulting IBP-DL dictionary and estimated noise level $\hat{\sigma}_{\boldsymbol{\epsilon}}$ are then used to

(a) (b) (c)

Fig. 9 Illustration of typical denoising results obtained by using IBP-DL on images. From top to bottom are the IBP-DL dictionary, the noisy, the denoised and the original images; (a) Lena, from a PSNR of 20.17 dB to 31.12 dB, (b) Boat from a PSNR of 20.17 dB to 28.92 dB, (c) Barbara from a PSNR of 16.08 dB to 26.34 dB.

denoise images. To be consistent with the denoising method in [5,6][3], the images are denoised by averaging pixel estimates from overlapping patches reconstructed by Orthogonal Matching Pursuit (OMP) with a maximum tolerance of representation error of $1.15\sigma_{\boldsymbol{\epsilon}}$ and a Lagrangian multiplier $\lambda = 30/\sigma_{\boldsymbol{\epsilon}}$.

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

We illustrate the relevance of IBP-DL by first comparing denoising results with BM3D (state of the art as a top reference) and several K-SVD based methods [5]. In the following, we compare with DLENE [14], an adaptive approach to learn overcomplete dictionaries with efficient numbers of elements and BPFA [15], a bayesian approach that can be seen as a parametric approximation of the IBP :

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 with a compromise between reconstruction error and sparsity by adapting the number of atoms.
5. BPFA with an initial number K of atoms depending on the size of the image.

Fig. 10 compares the denoising performance of IBP-DL, see Table 1, with BM3D [22] and these K-SVD based methods [5]. Table 2 shows numerical comparisons of IBP-DL with BPFA: dictionary size of IBP-DL, denoising performances and estimated noise level. Note that, the comparison between IBP-DL and BPFA method is realized on the full training set. Results from BM3D [22] are used as a reference only since we do not expect to obtain better results here. 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 at least as relevant as K-SVD ones. Note that the results using K-SVD [5, 6] 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 noise level is available. We emphasize that in IBP-DL the noise level is part of the 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. Here, to reduce computational time, IBP-DL works with a reduced set of 50% overlapping patches (16129 patches). Fig. 10 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 as the noise level gets higher: the adaptive sized dictionary is not always redundant, see Table 1 & 2. 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. The dictionary size tends to increase for a smaller noise level. This is expected 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 averaged to reduce noise leading to a smaller number of atoms.

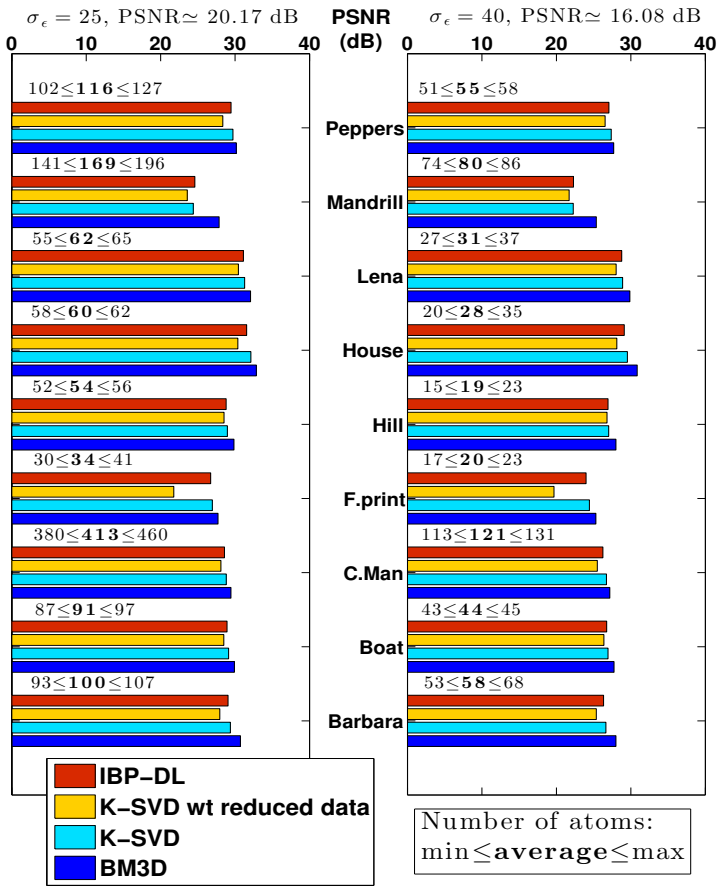


Fig. 10 Denoising results and sizes of IBP-DL dictionaries for noise level (right) $\sigma_{\boldsymbol{\varepsilon}} = 25$, (left) $\sigma_{\boldsymbol{\varepsilon}} = 40$. The text above each group of bars is the IBP-DL dictionary size. From bottom to top are the PSNR using IBP-DL learnt from the reduced training set, 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.

We now compare our results with DLENE [14], a recent work which also adapts the size of the dictionary. DLENE uses the reduced training set as well. It targets a compromise between reconstruction error and sparsity by adapting the number of atoms. For Peppers with $\sigma_{\boldsymbol{\varepsilon}}$=40, $\text{PSNR}_{\text{IBP-DL}}$=27.07dB and DLENE yields $\text{PSNR}_{\text{DLENE}}$= 27.27dB. For Barbara with $\sigma_{\boldsymbol{\varepsilon}}$=25, we get $\text{PSNR}_{\text{IBP-DL}}$=29.06 dB and $\text{PSNR}_{\text{DLENE}}$= 28.82 dB, see results in [14] for other comparisons. In general, IBP-DL performs as well as DLENE for denoising. Again this supports the relevance of the IBP-DL dictionaries.

We also compare the results of IBP-DL to those of BPFA [15], a bayesian method implemented using Gibbs sampling. Despite a connection with the Indian Buffet Process, this approach is not really a non-parametric approach

Table 2 The results of IBP-DL and BPFA approaches when the true noise level was 25 and 40. For each noise level, from left to right are the IBP-DL dictionary size K, the denoising PSNR (dB) and the estimated noise level then the denoising PSNR (dB) and the estimated noise level using BPFA.

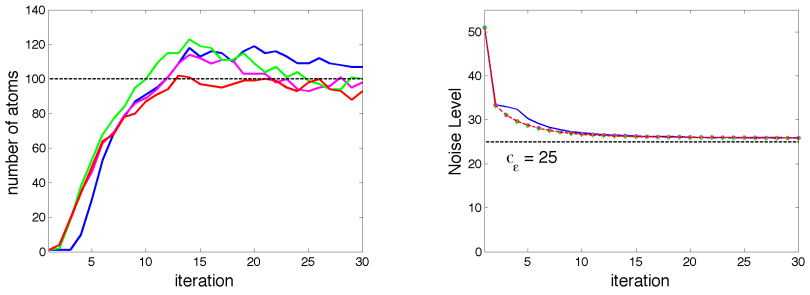
	PSNR $\simeq$ 20.17dB, $\sigma_\epsilon = 25$					PSNR $\simeq$ 16.08dB, $\sigma_\epsilon = 40$				
	IBP-DL			BPFA		IBP-DL			BPFA	
	# atoms	PSNR	$\hat{\sigma}_\epsilon$	PSNR	$\hat{\sigma}_\epsilon$	# atoms	PSNR	$\hat{\sigma}_\epsilon$	PSNR	$\hat{\sigma}_\epsilon$
House	57	31.95	25.37	32.14	25.43	40	29.47	40.43	29.73	40.54
Peppers	191	29.40	25.48	29.88	25.50	163	27.15	40.34	27.06	40.67
Barbara	134	29.31	25.66	29.79	25.45	107	26.81	40.33	26.34	40.15
Lena	147	31.40	25.20	31.58	25.32	111	29.15	40.04	29.27	40.19

and is a parametric approximation of the IBP because it works with a fixed number of atoms in advance. The initial size (K=256 or 512) of the dictionary of BPFA depends on the size of the image. Then, a subset of atoms is used that is slightly smaller than the initial size. This times, both IBP-DL and BPFA approaches train from the full set of available patches : 62001 overlapping patches for House and Peppers images and 255025 overlapping for Barbara, Lena images. BPFA approach initializes $K = 256$ for House and Peppers images and $K = 512$ for Barbara, Lena images. Table 2 illustrates the results of IBP-DL and BPFA with 2 noise levels $\sigma_{\boldsymbol{\varepsilon}}$=25 or 40. The image restoration method[4] is the same as in [15].

For House with $\sigma_{\boldsymbol{\varepsilon}}$=25, $\text{PSNR}_{\text{IBP-DL}}$=31.95dB with 57 atoms and BPFA yields $\text{PSNR}_{\text{BPFA}}$= 32.14dB. For Barbara with $\sigma_{\boldsymbol{\varepsilon}}$=40, $\text{PSNR}_{\text{BPFA}}$=26.34 dB while we get $\text{PSNR}_{\text{IBP-DL}}$= 26.81 dB with 107 atoms. The IBP-DL performances are comparable to BPFA [15] while the adapted size of IBP-DL dictionaries are often relatively smaller than the BPFA method. To this respect, IBP-DL improves on the previous method [15] and our observations support the interest of a non parametric approach that is more adaptive to the actual content of the image. Again, we note that the dictionary size is increased for a smaller noise level.

One important limitation of IBP-DL however is that sampling from an IBP is expensive even though the accelerated sampling [19] is implemented, reducing the complexity from $O(N^3(K^2 + KP))$ to $O(N(K^2 + KP))$; that is still larger than the $O(K(N + P))$ complexity of BPFA with an initial number K of atoms. For example, training on a reduced dataset of Barbara image, IBP-DL costs about 1 hour for 30 iterations using Matlab_R2013b on a recent personal computer. There is room for a significant improvement to this respect either by using a more efficient implementation or by proposing other inference methods. In this case, Gibbs sampling for non parametric methods is close to prohibitive.

[4] Matlab code by M. Zhou is available at `http://mingyuanzhou.github.io/Code.html`



(a) (b)

Fig. 11 (a) Evolution of the number of atoms in the dictionary across iterations of IBP-DL on Barbara for 4 different noise realizations with $\sigma_{\boldsymbol{\varepsilon}} = 25$. (b) Evolution of the noise level sampled over iterations of IBP-DL on Barbara when $\sigma_{\boldsymbol{\varepsilon}} = 25$ and $\sigma_{init} = 51$.

Fig. 11(a) shows the evolution of the size K of the dictionary across iterations. The final size (around 100 in this example) is reached after about 15 iterations only; implicitly it means that α converges to about $K/\log N \simeq 10$. Fig. 11(b) shows the evolution of the sampled σ_ε with iterations on an example. After 15 iterations, the sampled value has converged very close to the true value. The estimation error is at most of a few percents only 2% - 10% when $\sigma_{\boldsymbol{\varepsilon}}$=25 and 1% - 6% when $\sigma_{\boldsymbol{\varepsilon}}$=40. This accurate estimate is an essential benefit of this approach.

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 [5] for fixed size, DLENE [14] with an adaptive size of the dictionary learnt from a reduced training set or BPFA [15] for an initial number K of atoms.

Starting from an empty dictionary apart from the DC atom, IBP-DL yields an efficient dictionary. It simultaneously infers the size of the dictionary as well as all the parameters of the model such as the noise level that is a crucial input to later solve 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 than Gibbs sampling for scalability and a more general model for other cases of inverse problems in image processing.

References

1. I. Tosic and P. Frossard, Dictionary learning: What is the right representation for my signal, *IEEE Signal Processing Magazine*, vol. 28, pp. 27–38 (2011).
2. N. Ahmed, T. Natarajan, and K.R. Rao, Discrete cosine transform, *IEEE Transactions on Computers*, vol. C-23, pp. 90–93 (1974).
3. S.Mallat, A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way, *Academic Press* (2008).
4. B.A. Olshausen and D.J. Field, Emergence of simple-cell receptive properties by learning a sparse code for natural images, *Nature*, vol. 381, pp. 607–609 (1996).
5. 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, pp. 4311–4322 (2006).
6. M. Elad and M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, *IEEE Transactions on Image Processing*, vol. 15, pp. 3736 –3745 (2006).
7. Julien Mairal, Michael Elad, and Guillermo Sapiro, Sparse representation for color image restoration, *IEEE Transactions on Image Processing*, vol. 17, pp. 53–69 (2008).
8. J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, Non-local sparse models for image restoration, *IEEE International Conference on Computer Vision*, pp. 2272–2279, (2009)
9. N. Rao and F. Porikli, A clustering approach to optimize online dictionary learning, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1293–1296 (2012).
10. 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, (2010).
11. R. Mazhar and P.D. Gader, Ek-svd: Optimized dictionary design for sparse representations, *International Conference on Pattern Recognition*, pp. 1–4 (2008).
12. J. Feng, L. Song, X. Yang, and W. Zhang, Sub clustering k-svd: Size variable dictionary learning for sparse representations, *IEEE International Conference on Image Processing*, pp. 2149–2152 (2009).
13. C. Rusu and B. Dumitrescu, Stagewise k-svd to design efficient dictionaries for sparse representations, *IEEE Signal Processing Letters*, vol. 19, pp. 631–634 (2012).
14. 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, pp. 3272–3283 (2014).
15. 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).
16. T. Griffiths and Z. Ghahramani, Infinite latent feature models and the indian buffet process, *Advances in NIPS 18*, MIT Press, pp. 475–482 (2006).
17. 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).
18. D A Knowles and Z Ghahramani, Nonparametric Bayesian sparse factor models with application togene expression modeling, *The Annals of Applied Statistics*, vol. 5, pp. 1534–1552 (2011).
19. F. Doshi-Velez and Z. Ghahramani, Accelerated sampling for the indian buffet process, *International Conference on Machine Learning*, pp. 273–280 (2009).
20. David Andrzejewski, Accelerated Gibbs Sampling for Infinite Sparse Factor Analysis, *LLNL Technical Report* (LLNL-TR-499647) .
21. David A van Dyk and Taeyoung Park, Partially collapsed gibbs samplers, *Journal of the American Statistical Association*, vol. 103, pp. 790–796 (2008).
22. 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, pp. 2080–2095 (2007).