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., $8 \times 8$ 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 $8 \times 8$ 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 initialization 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.
2. DICTIONARY LEARNING (DL)

Here is a brief introduction to the DL optimization problem, see [11] for a review. Let matrix $Y = [y_1, \ldots, y_N] \in \mathbb{R}^{P \times N}$ a set of $N$ image patches of size $\sqrt{P} \times \sqrt{P}$, ordered lexicographically as column vectors $y \in \mathbb{R}^P$. Let dictionary of $K$ atoms $D = [d_1, \ldots, d_K] \in \mathbb{R}^{P \times K}$. In presence of some additive noise $\varepsilon \in \mathbb{R}^{P \times N}$, the data is modeled by

$$ Y = DW + \varepsilon \quad (1) $$

where $W = [w_1, \ldots, w_N] \in \mathbb{R}^{K \times N}$ are the encoding coefficients. Each observation $y_i$ should be described by a sparse set of coefficients $w_i$. Usually, when working on image patches of size $8 \times 8$ (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):

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

Various approaches have been proposed to solve this problem by an alternate optimization on $D$ and $W$, including K-SVD (batch DL) [1, 11] and ODL (online DL) [6]. Note that the choice of the regularization parameter $\lambda$ is of importance and should be decreasing with the noise level $\sigma_\varepsilon$. 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.

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

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-assignment matrices $Z: Z(k,i) = 1$ 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$(\alpha/i)$ 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]:

$$ P[Z] = \frac{1}{2^{N-1}} \prod_{k=1}^{K} \frac{(N - m_k)! (m_k - 1)!}{N!} \quad (3) $$

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’ $Z(k, \cdot) = h$. Parameter $\alpha > 0$ controls the expected total number of features that $K_+ \sim \text{Poisson}(\alpha H_N)$ hence $E[K_+] = \alpha H_N \simeq \alpha \log N$. The IBP permits to both deal with a variable sized dictionary (potentially infinite 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

$$ y_i = Dw_i + \varepsilon_i, \forall 1 \leq i \leq N \quad (4) $$

$$ w_i = z_i \circ s_i, \forall 1 \leq i \leq N \quad (5) $$

$$ d_k \sim \mathcal{N}(0, P^{-1}P) , \forall k \in \mathbb{N} \quad (6) $$

$$ Z \sim \text{IBP}(\alpha) \quad (7) $$

$$ s_i \sim \mathcal{N}(0, \sigma_2^2I_K) , \forall 1 \leq i \leq N \quad (8) $$

$$ \varepsilon_i \sim \mathcal{N}(0, \sigma_2^2P) , \forall 1 \leq i \leq N \quad (9) $$

where $y_i$ is a column vector of dimension $P$, $\circ$ represents the Hadamard product. We place priors on $D$, $W$ and $\varepsilon$. The vector $z_i \in \{0, 1\}^K$ denotes wich of the K columns of $D$ are used for representation of $y_i$; $s_i \in \mathbb{R}^K$ represents the coefficients used for this representation. The representation...
In [10], from Bayes' rule:

\[ w \]

the properties of a parametric Bernoulli-Gaussian model. Hence, the sparsity regularization term that tends to penalize the number \( d \) of rows, but only non-zero rows are kept in memory. Let \( \epsilon \) be the noise variance estimated as well during inference, making the approach very close to truly 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 \( P(\mathbf{D}, \mathbf{S}, \mathbf{Z}, \theta | \mathbf{Y}) \).

**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 \( \mathbf{z}_{ki} = \mathbf{Z}(k, i) \) for ‘active’ atoms \( k \) such that \( m_{k,-i} \geq 0 \) (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 \( p(z_{ki} = 1 | Z_{k,-i}) = m_{k,-i}/N \). The likelihood \( p(Y | \mathbf{D}, \mathbf{S}, \mathbf{Z}, \theta \) 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

\[
p(Y | \mathbf{Z}, \mathbf{S}, \sigma^2_S, \sigma^2_D) = p(Y | \mathbf{W}, \sigma^2_S, \sigma^2_D) = 
\exp \left\{ -\frac{1}{2 \sigma^2_S} \| Y - \mathbf{W}^T \mathbf{W} \|_F^2 \right\} 
\exp \left\{ -\frac{1}{2 \sigma^2_D} \| \mathbf{W} \|_F^2 \right\}
\exp \left\{ -\frac{1}{2 \sigma^2_D} \| \mathbf{W} \|_F^2 \right\}
\]

In [10], from Bayes’ rule:

\[
p(z_{ki} | Y_{k,-i}, S, \sigma^2_S, \sigma^2_D) \propto p(Y | \mathbf{W}, \sigma^2_S, \sigma^2_D) p(z_{ki} | Z_{k,-i})
\]

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

If row \( Z(k, \cdot) = 0 \), 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 number of new atoms according to the same distribution as the prior on \( k_{sing} \) that is a Poisson law with parameter \( \lambda \) in the IBP model:

\[
J(k_{prop}) = \mathcal{P}(k_{prop}; \alpha/N)
\]

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

\[
u \leq \min \left\{ 1, \frac{p(Y | k_{prop}, \text{rest})}{p(Y | k_{sing}, \text{rest})} \right\}
\]

Let \( \gamma_S = 1/\sigma^2_S, \gamma_D = 1/\sigma^2_D, \gamma_S = 1/\sigma^2_S \). Sampling \( \mathbf{D} \) and \( \mathbf{S} \) and \( \theta = (\sigma^2_S, \sigma^2_D, \alpha) \) are done according to...
$$p(d_k|Y, Z, S, D_{-k}, \theta) \propto N(\mu_{d_k}, \Sigma_{d_k})$$  

$$\Sigma_{d_k} = \left(\gamma_D I_p + \gamma_\varepsilon \sum_{i=1}^N w_{ki}^2\right)$$  

$$\mu_{d_k} = \gamma_\varepsilon \Sigma_{d_k} \sum_{i=1}^N w_{ki}(y_i - D w_i + d_k w_{ki})$$  

$$p(s_{ki}|Y, D, Z, S, k, i, \theta) \propto N(\mu_{s_{ki}}, \Sigma_{s_{ki}})$$  

$$z_{ki} = 1 \Rightarrow \begin{cases} 
\Sigma_{s_{ki}} = (\gamma_d d_k^T d_k + \gamma_s)^{-1} \\
\mu_{s_{ki}} = \gamma_\varepsilon \Sigma_{s_{ki}} d_k^T (y_i - D w_i + d_k s_{ki}) 
\end{cases}$$  

$$z_{ki} = 0 \Rightarrow \begin{cases} 
\Sigma_{s_{ki}} = \sigma^2_S \\
\mu_{s_{ki}} = 0 
\end{cases}$$

An optional intern loop of F iterations can be used to sample $(D, S, \sigma_S, \sigma_\varepsilon, Z)$ at fixed support $Z$ (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 $512 \times 512$ (8 bits) are processed for 2 noise levels $\sigma_\varepsilon = 25$ or 40. The initial value of $\gamma_S$ 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 $d_0 = (1, ..., 1)$. The denoising method [14] uses 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 $1.15\sigma_\varepsilon$ and Lagrangian multiplier $\lambda = 30/\sigma_\varepsilon$. But the K-SVD dictionary is here replaced by IBP-DL and the noise level $\sigma_\varepsilon$ is its IBP-DL estimate $\hat{\sigma}_\varepsilon$.

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.

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

<table>
<thead>
<tr>
<th>Image</th>
<th>Case $\sigma_\varepsilon = 25$</th>
<th>Case $\sigma_\varepsilon = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial value</td>
<td>$\sigma_{init} = 51$</td>
<td>$\sigma_{init} = 76.5$</td>
</tr>
<tr>
<td>Barbara</td>
<td>25.86</td>
<td>40.76</td>
</tr>
<tr>
<td>Boat</td>
<td>25.82</td>
<td>40.64</td>
</tr>
<tr>
<td>Cameraman</td>
<td>26.10</td>
<td>41.16</td>
</tr>
<tr>
<td>Fingerprint</td>
<td>25.79</td>
<td>40.90</td>
</tr>
<tr>
<td>GoldHill</td>
<td>25.89</td>
<td>40.70</td>
</tr>
<tr>
<td>House</td>
<td>25.53</td>
<td>40.36</td>
</tr>
<tr>
<td>Lena</td>
<td>25.45</td>
<td>40.24</td>
</tr>
<tr>
<td>Mandrill</td>
<td>27.57</td>
<td>42.50</td>
</tr>
<tr>
<td>Peppers</td>
<td>25.76</td>
<td>40.58</td>
</tr>
</tbody>
</table>

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

\[1\] Matlab code by R. Rubinstein is available at http://www.cs.technion.ac.il/~ronrubin/software.html
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 $\alpha$ converges to about $K / \log N \approx 10$. 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 $\sigma_\delta=25$, DLENE yields PSNR$_{\text{DLENE}}=28.82$ dB while we get PSNR$_{\text{IBP-DL}}=29.06$ dB; for Peppers with $\sigma_\delta=40$, PSNR$_{\text{DLENE}}=27.27$ dB and PSNR$_{\text{IBP-DL}}=27.07$ dB; 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 $\sigma_\delta$ is inferred with good accuracy. Fig. 5 shows the evolution of the sampled $\sigma_\delta$ 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 $\sigma_\delta = 25$ and from 1% to 6% when $\sigma_\delta = 40$. This accurate estimate is an essential profit 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) [1] for fixed size or DLNE [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