Small variance asymptotics and bayesian nonparametrics for dictionary learning

Clément Elvira, Hong-Phuong Dang, Pierre Chainais

Small variance asymptotics and bayesian nonparametrics for dictionary learning

Clément Elvira[(1)*], Hong-Phuong Dang[(2)*] and Pierre Chainais[(3)]
[(1)] Univ. Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France
[(2)] National School for Statistics and Information Analysis, UMR 9194 - CREST, France
[(3)] Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
Email: firstname.lastname@{inria.fr, ensai.fr,ec-lille.fr}
[*] Authors contributed equally.

***Abstract*—Bayesian nonparametric (BNP) is an appealing framework to infer the complexity of a model along with the parameters. To this aim, sampling or variational methods are often used for inference. However, these methods come with numerical disadvantages for large-scale data. An alternative approach is to relax the probabilistic model into a non-probabilistic formulation which yields a scalable algorithm. One limitation of BNP approaches can be the cost of Monte-Carlo sampling for inference. Small-variance asymptotic (SVA) approaches paves the way to much cheaper though approximate methods for inference by taking benefit from a fruitful interaction between Bayesian models and optimization algorithms. In brief, SVA lets the variance of the noise (or residual error) distribution tend to zero in the optimization problem corresponding to a MAP estimator with finite noise variance for instance. We propose such an SVA analysis of a BNP dictionary learning (DL) approach that automatically adapts the size of the dictionary or the subspace dimension in an efficient way. Numerical experiments illustrate the efficiency of the proposed method.**

***Index Terms*—Bayesian nonparametrics, small variance asymptotic, Indian Buffet Process, sparse representations, dictionary learning, inverse problems.**

1. Introduction

Designing efficient and scalable algorithms for Bayesian inference has become a tremendous topic. Two noteworthy lines of research have arisen, namely fast sampling of MCMC chains with strong space exploration potential and deterministic algorithm to approximate Bayesian estimators. Yet, few methods have been proposed for *bayesian non parametric* (BNP) models.

The BNP framework permits to build models with an adaptive number of degrees of freedom. In BNP approaches, the number of parameters is potentially infinite but the effective number is controlled by the complexity of the data, without going through a further model selection step, hence the name 'nonparametric'. For instance, BNP allows for latent feature models with a potentially infinite number of features, e.g. using the Indian buffet process. Two families of algorithms have been usually proposed for inference, namely MCMC methods and variational approximations. The main drawback of MCMC methods is their high computational cost and variational analysis still relies on parametric approximations only.

Small-Variance asymptotics (SVA) have been recently proposed to derive efficient yet scalable optimization algorithms [1], [2] for inference from probabilistic models. They have been successfully applied to signal processing problems [3]. This approach takes benefits from both worlds: the flexibility of BNP models and the numerical advantage of optimization methods. In this paper, a SVA approach is proposed in a *dictionary learning* (DL) problem [4], a latent feature model where each observation can be associated with several latent features.

The BNP model *Indian buffet process for dictionary learning* (IBP-DL) has been proposed in [5]. The *Indian Buffet Process* (IBP) [6] is a BNP prior that permits to estimate a dictionary of adaptive size. Numerical experiments on inverse problems in image processing (e.g. denoising, inpainting) have shown the relevance of this approach. A Gibbs sampler has been proposed for inference at the price of a prohibitive computational cost. This paper first proposes a generalized version of the IBP-DL model using the two-parameter IBP presented in [7]. Then a SVA analysis is carried out that connects the proposed BNP model with well known optimization problems. We propose a new algorithm that takes benefits from the optimization methods for inference while conserving the desired regularizing properties of the probabilistic BNP approach. The relevance of the resulting dictionaries is illustrated by denoising experiments.

Section 2 recalls the dictionary learning problem. Section 3 describes the proposed Bayesian model. The SVA analysis as well as the proposed algorithm are describes in Section 4. Section 5 illustrates the relevance of the approach on numerical experiments on a denoising problem. Section 6 gathers conclusions and prospects.

2. Dictionary Learning (DL)

Dictionary learning for sparse representation is known as an efficient approach to resolve ill-posed inverse problems in image processing [4]. The problem is often modeled as :

$$\mathbf{Y} = \mathbf{H}(\mathbf{X} + \boldsymbol{\varepsilon}) \quad \text{where} \quad \mathbf{X} = \mathbf{DW} \tag{1}$$

$\mathbf{Y} \in \mathbb{R}^{L \times N}$ is a set of N observations $\mathbf{y}_i$. Each column vector $\mathbf{y} \in \mathbb{R}^L$ represents a square patch (e.g. 8×8, then L=64), in lexicographic order. $\mathbf{X} \in \mathbb{R}^{L \times N}$ represents patches from the initial image which is disturbed by an observation known linear operator $\mathbf{H}$ and an noise $\boldsymbol{\varepsilon}$.For this, patches are represented by the encoding coefficients $\mathbf{W}=[\mathbf{w}_1, ..., \mathbf{w}_N] \in \mathbb{R}^{K \times N}$ of their representation in a dictionary $\mathbf{D}=[\mathbf{d}_1, ..., \mathbf{d}_K] \in \mathbb{R}^{L \times K}$ with K atoms. Each $\mathbf{x}_i$ is described by $\mathbf{x}_i=\mathbf{D}\mathbf{w}_i$ where $\mathbf{w}_i$ is sparse. The recovery of $\mathbf{X}$ is equivalent to finding an optimal pair ($\mathbf{D}$, $\mathbf{W}$) from $\mathbf{Y}$.

Sparsity is typically imposed through a ℓ_0 or its convex relaxation ℓ_1-penalty in the mixed optimization problem (other formulations are possible)

$$(\mathbf{D}, \mathbf{W}) = \underset{(\mathbf{D},\mathbf{W})}{\operatorname{argmin}} \quad \frac{1}{2}\|\mathbf{Y} - \mathbf{H}(\mathbf{DW})\|_2^2 + \lambda \|\mathbf{W}\|_p \tag{2}$$

Various approaches have been proposed to solve this problem by an alternate optimization on $\mathbf{D}$ and $\mathbf{W}$ for fixed heuristic size dictionaries, e.g. with 256 or 512 atoms [4], [8].

In the Bayesian framework, the problem is translated in a Gaussian likelihood according to the model (1). The prior $p(\mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}})$ acts as a regularization and the joint posterior distribution writes

$$p(\mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}} \mid \mathbf{Y}, \mathbf{H}) \propto p(\mathbf{Y} \mid \mathbf{H}, \mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}) p(\mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}) \tag{3}$$

Using Gibbs sampling for inference, the problem is solved by sampling alternately $\mathbf{D}$, $\mathbf{W}$ and $\sigma_{\boldsymbol{\varepsilon}}$. In a BNP framework, the dictionary is learned without setting the size in advance and no parameter tuning is necessary. The model IBP-DL [5] has used Indian buffet process to deal both with the sparsity constraint and the desirable adaptive number of atoms.

3. Extension of IBP-DL

3.1. Extension of Indian Buffet Process (IBP)

The IBP was introduced in [6] to deal with problems of latent feature analysis that naturally promotes sparsity. The IBP is a BNP prior on infinite binary feature-assignment matrices $\mathbf{Z}$: $\mathbf{Z}(k,i)$=1 indicates whether the observation $\mathbf{y}_i$ uses feature $\mathbf{d}_k$ (0 otherwise). Above all, it allows the number of latent features K to be not a priori fixed and simultaneously penalizes large values of K. The IBP emulates an exchangeable distribution over *sparse* binary matrices with a *potentially infinite* number of lines.

An extension of the two-parameter IBP is presented in [7]. Its generative process is as follows. N customers (observations) taste dishes (features) in a potentially infinite (Indian) buffet. The first customer tries Poisson(α) dishes. Recursively, the customer $i+1$ takes portions from previously-selected dish k with probability $m_k/(i+c)$, where m_k is the number of previous customers who selected dish k before him; then he also tries Poisson($\alpha c/(i+c)$) new dishes, which corresponds to adding new lines to matrix $\mathbf{Z}$. The corresponding probability distribution over equivalence classes $[\mathbf{Z}]$, when the number of features tends to infinity, is:

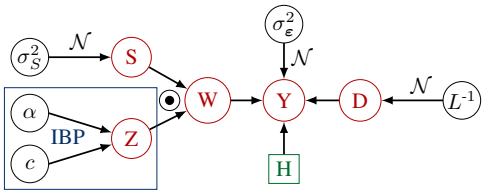

Figure 1. Graphical model of IBP-DL.

$$\mathrm{P}([\mathbf{Z}]) = \frac{(\alpha c)^K}{\prod_{h=1}^{2^N-1} K_h!} \exp(-\alpha c H_N) \prod_{k=1}^{K} \beta(m_k, N - m_k + c) \tag{4}$$

where $H_N = \sum_{j=1}^{N} \frac{1}{c+j-1}$, β denotes the Beta function, K is the number of "active" features for which $m_k > 0$ is the number of customers that have chosen dish k. The mass parameter $\alpha > 0$ controls the total number of dishes tried by the customers. The concentration parameter $c > 0$ controls the number of customers that will try each dish. When $c = 1$, the process reduces to the usual IBP with one parameter [6]. The resulting prior mean of latent features behaves as $\mathbb{E}[K] = \alpha c H_N \approx \alpha c \log(N)$.

3.2. Generalization of IBP-DL model

In [5], an BNP model for dictionary learning (IBP-DL) has been proposed by using the one parameter IBP (*i.e.* $c = 1$) as a BNP prior to promote sparcity on the latent features. In this paper, we prospect the advantage of the generalization of the IBP-DL model with two parameters. Fig. 1 shows the graphical model which may be expressed as : $\forall i \in [\![1, N]\!]$

$$\mathbf{y}_i = \mathbf{H}_i[(\mathbf{D}\mathbf{w}_i) + \boldsymbol{\varepsilon}_i], \text{ with } \mathbf{w}_i = \mathbf{z}_i \odot \mathbf{s}_i, \tag{5}$$

$$\mathbf{d}_k \sim \mathcal{N}(0, \sigma_D^2 \mathbb{I}_L), \forall k \in \mathbb{N}, \tag{6}$$

$$\mathbf{Z} \sim \mathrm{IBP}(\alpha, c), \quad s_{ki} \sim \mathcal{N}(0, \sigma_s^2), \forall k \in \mathbb{N}, \tag{7}$$

$$\mathbf{H}_i \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma_{\boldsymbol{\varepsilon}}^2 \mathbb{I}_L). \tag{8}$$

$\odot$ denotes the Hadamard product. Recall that the observation matrix $\mathbf{Y}$ contains N column vectors $\mathbf{y}_i \in \mathbb{R}^L$. The representation coefficients are defined as $\mathbf{w}_i = \mathbf{z}_i \odot \mathbf{s}_i$, in the spirit of a parametric Bernoulli-Gaussian model. The sparsity of $\mathbf{W}$ is induced by the sparsity of $\mathbf{Z}$ thanks to the IBP prior. Except for $\sigma_D^2 = L^{-1}$ due to the problem of indeterminacy of the pair ($\mathbf{D}, \mathbf{W}$) to a multiplicative factor, conjugate priors are used for others parameters : Gamma and inverse Gamma distributions, see [5] for details. Note that setting σ_D^2 to L^{-1} amounts to writing that the energy contained in each atom

k is approximately $\mathbb{E}[\mathbf{d}_k^{\mathrm{T}}\mathbf{d}_k]$=1. It is therefore a gentle way to normalize. The vector $\mathbf{z}_i \in \{0,1\}^K$ denotes which of the K atoms of $\mathbf{D}$ are used for representation of $\mathbf{y}_i$, and the vector $\mathbf{s}_i$ gathers the coefficients: z_{ki}=1 then w_{ki}=s_{ki} and z_{ki}=0 then w_{ki}=0. We consider in this paper the denoising problem, *i.e.*, $\mathbf{H}_i$=$\mathbb{I}_L$.

4. Proposed method

This section presents a computationally efficient approach for approximating Bayesian estimators based on a small-variance asymptotic (SVA) approximation referred to as IBPDL-SVA. Conceptual links between IBPDL-SVA and non-probabilistic approaches will be established.

4.1. Small Variance Asymptotic (SVA)

In a manner akin to [5], a metropolis within Gibbs sampler could be designed to sample according to the posterior

$$f(\mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}^2, \dots | \mathbf{Y}) \propto \mathrm{p}(\mathbf{Y}|\mathbf{D}, \mathbf{W}, \sigma^2)\mathrm{p}(\mathbf{D}), \mathrm{p}(\mathbf{W})\mathrm{p}(\sigma_{\boldsymbol{\varepsilon}}^2) \quad (9)$$

and approximate Bayesian estimators. Whereas efficient, such a Gibbs sampler remains computationally costly. We alternatively propose to conduct a SVA analysis to derive the *asymptotic MAP* (aMAP) estimator of (9) defined as

$$\widehat{\mathbf{D}}, \widehat{\mathbf{W}} = \underset{\mathbf{D},\mathbf{W}}{\operatorname{argmin}} \lim_{\sigma_{\boldsymbol{\varepsilon}}^2 \to 0} \; -2\sigma^2 \log f(\mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}}^2, \dots | \mathbf{Y}). \quad (10)$$

Without further improvement, the aMAP is the maximum likelihood estimator. As pointed out by [9], it is necessary to couple the model's hyperparameters to make them scale with $\sigma_{\boldsymbol{\varepsilon}}^2$ and preserve the desired regularization property of the Bayesian model.

Let $\alpha = \exp\left(\frac{\sigma_{\boldsymbol{\varepsilon}}^2}{\lambda_1} - \frac{\lambda_1}{2\sigma_{\boldsymbol{\varepsilon}}^2}\right), c = \exp\left(\frac{\lambda_2}{2\sigma_{\boldsymbol{\varepsilon}}^2} - \frac{\sigma_{\boldsymbol{\varepsilon}}^2}{\lambda_2}\right)$ and $\lambda_1, \lambda_2 > 0$. As $\sigma_{\boldsymbol{\varepsilon}}^2 \longrightarrow 0$, one find asymptotically

$$\begin{aligned} -2\sigma_{\boldsymbol{\varepsilon}}^2 \log \mathrm{p}(\mathbf{Y}, \mathbf{D}, \mathbf{W}) \sim \mathbf{tr}\left[(\mathbf{Y} - \mathbf{D}\mathbf{W})^{\mathrm{T}}(\mathbf{Y} - \mathbf{D}\mathbf{W})\right] \\ + \lambda_2 \sum_k^K m_k + (\lambda_1 - \lambda_2)(K+1) \end{aligned} \quad (11)$$

where m_k is the number of observations that use the k^{th} atom. The trace originates from the exponential function in the Gaussian likelihood, and the penalty term originates from the IBP prior. Note that the trace operator returns to the Frobenius norm $\|.\|_F$ and $\sum_k^K m_k = \|\mathbf{W}\|_0$.

We see from Eq. (11) that finding the asymptotic MAP estimate of the DL problem is asymptotically equivalent to solving the following optimization problem

$$\underset{K,D,W}{\operatorname{argmin}} \quad \|\mathbf{Y} - \mathbf{D}\mathbf{W}\|_F^2 + \lambda_2 \|\mathbf{W}\|_0 + (\lambda_1 - \lambda_2)(K+1). \quad (12)$$

In [2], a functional similar to Eq. (12) without the ℓ_0 term is referred to as the BP-means objective. The parametrization of α is also different, but the discussion is reported to Section 4.3. Note also that Eq. (12) contains a regularization on the size K of dictionary compared to Eq. (2) of the standard optimization methods. The next section describes the proposed strategy to approximate the aMAP estimator.

4.2. A greedy within alternate approach

Input: $\mathbf{Y}$, $n_{\mathrm{it}}, \lambda_1, \lambda_2$

$\boldsymbol{R} = [\boldsymbol{r}_1, \dots, \boldsymbol{r}_N] \longleftarrow \mathbf{Y}$;
for $t \leftarrow 1$ **to** n_{it} **do**
  `// Encoding`
  for $n \leftarrow 1$ **to** N **do**
    $k \longleftarrow 0$;
    while *True* **do**
      $k \longleftarrow k+1$
      $\boldsymbol{r}_n \leftarrow \mathbf{y}_n, \mathbf{w}_n \leftarrow [0, \dots, 0], \mathcal{S}_n \leftarrow \emptyset$
      $k^* = \operatorname{argmax} |\langle \boldsymbol{r}_n, \mathbf{d}_k \rangle|$ $\mathcal{O}(K-k+1)$;
      find $\mathbf{w}^*$ with LS $\mathcal{O}(k^2L + k^3)$;
      if $\boldsymbol{w}^*$ *increases Eq.* (12) **then**
        break ;
      end
    end
    `// Add atoms`
    $\mathbf{d}_{\mathrm{new}} = N^{-1} \sum_{n=1}^N \boldsymbol{r}_n$ (Eq. (16));
    if $\boldsymbol{d}_{\mathrm{new}}$ *lowers Eq.* (12) **then**
      $\mathbf{D} = [\mathbf{D}, \mathbf{d}_{\mathrm{new}}]$;
      Recompute $\mathbf{w}_n$ $\mathcal{O}(K^2L + K^3)$;
    end
  end
  `// Dictionary update`
  Remove unused atoms ;
  Update $\mathbf{D}$ according to (17) $\mathcal{O}(L^3 + NLK)$;
end

Output: dictionary $\mathbf{D}$ code $\mathbf{W}$.

Algorithm 1: Proposed IBPDL-SVA algorithm.

The SVA framework also suggests designing deterministic algorithm based on the asymptotic behavior of a Gibbs sampler [2], [9]. Such a sampler would requires Metropolis within Gibbs steps that are described in [5]. The resulting method, described below is an alternate optimization on $\mathbf{D}$ and $\mathbf{W}$ and is summed up in Alg. 1.

Updating $\boldsymbol{w}_n$. Let $\boldsymbol{r}_{k,n} = \mathbf{y}_n - \sum_{j \neq k} s_{j,n}\mathbf{d}_j$ be the residual vector. The conditional posterior distribution of $\mathbf{w}_{k,n}$ $\mathrm{p}(\mathbf{w}_{k,n}|\mathbf{Y}) \propto \sum_{z \in \{0,1\}} \mathrm{p}(s_{k,n}|\mathbf{Y}, z_{k,n} = z)\mathrm{P}[z_{k,n} = z|\mathbf{Y}]$ $\propto p_0 + p_1$ can be marginalized w.r.t. $s_{k,n}$. When $\sigma_\varepsilon^2 \to 0$, one has

$$\log p_0 = \|\boldsymbol{r}_{k,n}\|_2^2 + \lambda_2 m_k \quad (13)$$

$$\log p_1 = \|\boldsymbol{r}_{k,n} - \mathbf{d}_k^T \boldsymbol{r}_{k,n}\|_2^2 + \lambda_2(m_k + 1). \quad (14)$$

The resulting Bernoulli random variable $z_{k,n}|\mathbf{Y}$ of parameters $p_{k,n}$ indicates whether observation $\mathbf{y}_n$ is described by $\mathbf{d}_k$. This suggests setting $z_{k,n} = 1$ if $p_{k,n} \geq .5$ and 0 otherwise. One immediately see from (13) and (14) that $p_{k,n} \geq .5$ *iff* setting $w_{k,n} = \mathbf{d}_k^T \boldsymbol{r}_{k,n}$ decreases the cost function (12) compared to $w_{k,n} = 0$.
In a Gibbs sampler, one would loop over a random ordering of $k \in \{1, \dots, K\}$. To get rid of randomness, we rather propose to set $\mathbf{w}_{1:K,n} = 0$ and start by the atoms the more

correlated with the residual error. This procedure resumes to a Matching Pursuit (MP) [10] algorithm with the cost function as stopping criteria. We choose to replace MP by Orthonormal Matching Pursuit (OMP) [11] since it is known to perform significantly better at a reasonably higher cost.

Adding new atoms. After the coding vector stage, one tests whether adding a new atom permits a better reconstruction. For technical reasons, this stage is performed is performed using a Metropolis Hasting move [12]. The practitioner is free to design the best strategy to propose new atoms since the proposal is corrected by a ratio of probability distribution. We simply choose here to explore the space around residual vector, *i.e.* $\mathrm{q}(\mathbf{d}_{\mathrm{new}}|\mathbf{Y},\mathbf{W},\mathbf{D})$ of a new atom $\mathbf{d}_{\mathrm{new}}$ is normally distributed with mean vector

$$\boldsymbol{\mu}_{\mathbf{d}_{\mathrm{new}}} = \left(\frac{\sigma_{\varepsilon}^2}{\sigma_D^2} + \sum_{i=1}^{N} s_{\mathrm{new},i}^2\right)^{-1} \sum_{i=1}^{N} s_{\mathrm{new},i}(\mathbf{y}_i - \sum_{k=1}^{K} \mathbf{d}_k s_{ki}) \tag{15}$$

and known covariance matrix. When the $s_{k,n}$ are marginalized out and $\sigma_{\boldsymbol{\varepsilon}}^2 \to 0$, the Gaussian distribution reduces to a Dirac, suggesting the proposal

$$\mathbf{d}_{\mathrm{new}} = \sum_{i=1}^{N} \frac{1}{N}(\mathbf{y}_i - \sum_{k=1}^{K} \mathbf{d}_j \mathbf{w}_{ji}). \tag{16}$$

Eq. (16) corresponds to a rank-one approximation of the residual error. Note that the optimal rank-one approximation in the least-square sense is the eigenvector associated to the highest eigenvalue, corresponding to a K-SVD like update [8]. This choice has been discarded here, but we report the discussion in Section 4.3.

The proposed vector is accepted if the ratio of the Metropolis Hastings p^* is higher than .5 and refused otherwise. Setting $\sigma_{\varepsilon}^2 \to 0$, p^* is anew higher than .5 when adding $\mathbf{d}_{\mathrm{new}}$ decreases the cost function (12).

*Updating **D**.* The conditional posterior $\mathbf{D}|\mathbf{Y},\mathbf{W}$ is normally distributed and reduces to its expectation when $\sigma_{\boldsymbol{\varepsilon}}^2 \to 0$. We have chosen to keep the noise correction part rising from the prior term for the sake of numerical stability. This choice leads to

$$\mathbf{D}^{(t+1)} = \mathbf{Y}\mathbf{W}^{\mathrm{T}}(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \frac{\bar{\sigma}_{\varepsilon}^2}{\sigma_{\mathbf{D}}^2}\mathbb{I}_K)^{-1}, \tag{17}$$

where $\bar{\sigma}_{\varepsilon}^2$ is the residual variance at iteration t.

4.3. Comments about the algorithm

We analyze in this section the properties of IBPDL-SVA. Interestingly, the underlying objective function in eq. (12) referred to as BP-mean objective reduces to (2) plus an additive term penalizing the number of atoms. The sparsity is controlled by λ_2 while λ_1 drives the number of latent features. We emphasize that compared to [2], our proposed SVA approximations truly promotes sparsity because of the ℓ_0 penalization. In addition, the change of variable $\alpha = f(\lambda_1)$ is different to match the whole domain of α.

The algorithm described section 4.2 uses rank-1 approximation of the residual error to add new atom. Even tough the approximation has been chosen non optimal for sake of numerical complexity, performing the SVA limiting argument in a simple Gibbs sampler has naturally led to an algorithm that can be interpreted as a non parametric version of K-SVD [8]. In [5], the new atom $\mathbf{d}_{\mathrm{new}}$ is marginalized out instead of the weights $w_{k,n}$ resulting in an estimator with fewer variance. Such a strategy has no equivalent in the optimization framework but could be imported by mean of a SVA analysis. This work is under investigation.

Another important contribution of the proposed approach is to reduce the numerical complexity by a factor KN. In particular, the complexity of the sparse coding stage is reduced to $\mathcal{O}(N(K^2L + K^3))$. In [5], even though accelerated sampling is used, the complexity scales as $\mathcal{O}(NK(NK^2 + KL)$. For now, this acceleration comes at the cost of the knowledge of 2 regularizing parameters λ_1 and λ_2. We believe that these parameters could be jointly inferred by also exploiting Bayesian approaches.

5. Numerical experiments

This section describes a brief experiment to show that IBPDL-SVA can enjoy some properties of Bayesian techniques while featuring the speed and scalability of deterministic methods. 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 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 perform better.

A set of 5 images of size 512×512 is considered - Barbara, Hill, Mandrill, Lena, Peppers - for 2 noise levels $\sigma_{\boldsymbol{\varepsilon}} = 25$ or 40. There are $(512-7)^2 = 255025$ overlapping patches in each image, but the proposed approach works with $N = 16129$, *i.e.* 50% overlapping. The hypeparameters (λ_1, λ_2) are tuned with cross validation on Pepper and reused for the four other images. We have found $(0.12, 0.08)$ for $\sigma_{\varepsilon} = 25$ and $(0.4, 0.2)$ for $\sigma_{\varepsilon} = 40$. Simulations are run on a personal laptop and a Python implementation.

Fig. 2 shows the evolution of the size K of the dictionary across iterations for pepper (*i.e.*, the training set) with $\sigma_{\boldsymbol{\varepsilon}} = 25$ for several couples of (λ_1, λ_2). The blue curve denotes the chosen couple (λ_1, λ_2) while the orange ones stands for the rejected couples. One can obverse that a pattern applies to all chains : the method starts by adding too much atoms before stabilizing after a pruning stage. Learning the dictionary related to the blue curve costs about 30 minutes for 150 iterations. As a comparison, IBP-DL needs 1 hour for 30 iterations on a smaller dataset with a Matlab implementation. Note that such behavior does not apply to all images, since time computation depends on the size of the dictionary. For instance, the chosen (λ_1, λ_2) requires a couples of hours for Mandrill since the size of the dictionary barely stabilizes.

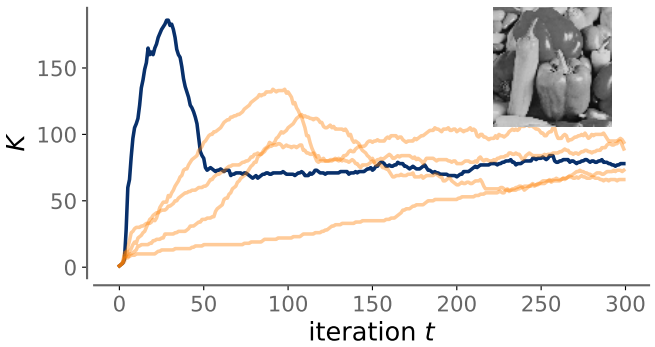

Figure 2. Evolution of the dictionary size K across iterations for several couples of (λ_1, λ_2). The orange lines stand for discarded couples and the blue curve is the retained one.

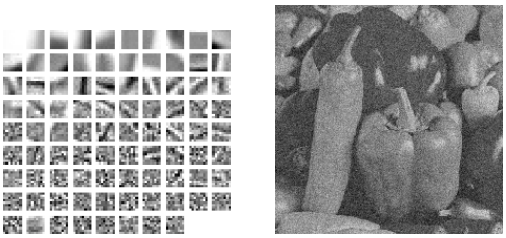

Figure 3. Dictionary of 88 atoms learn on Peppers image with σ_ε=25

Table 1 illustrates the relevance of IBPDL-SVA by comparing its denoising performances with 1) IBP-DL [5] 2) DLENE [14] 3) 4) K-SVD with K=256 [8] and 5) BM3D [13] as state of the art reference. DLENE is an approach to learn overcomplete dictionaries with an efficient numbers of elements that targets a compromise between reconstruction error and sparsity. One can see that IBPDL-SVA achieve performances similar to IBP-DL and DLENE but lower to BM3D which is state of the art. In addition, the same conclusions can be drawn compared to IBP-DL : the inferred dictionary size K is lower that 256, the value used for K-SVD (except for Mandrill) and tends to increase with small noise. IBPDL-SVA even outperforms BM3D on Pepper but recall that this image has served as trained set. Note also that using different values of (λ_1, λ_2) for these images (obtained also by grid search) permit better performances. This motivates our ongoing work about jointly inferring these hyperparameters using Bayesian approaches.

6. Conclusion

This paper presents a new computationally efficient approach for dictionary learning (DL) resulting from a Small Variance Asymptotic (SVA) analysis of a Bayesian non parametric (BNP) model. The proposed approach gathers some of the benefits of BNP such as inferring a dictionary of unknown size and the lower computational cost of optimization algorithms. It also outlines connections arising between the asymptotic behavior of MCMC methods and well known algorithms for dictionary learning (DL). The relevance of the inferred dictionary has been assessed on a denoising task; results are promising. Future work aims at taking benefit from the Bayesian framework to also infer the hyperparameters and yield a swift unsupervised approach. The non parametric PCA proposed in [15] can be revisited along the same lines; this is the subject of ongoing work.

	$\sigma_\varepsilon = 25$ PSNR $\approx$ 20.14 dB			$\sigma_\varepsilon = 40$ PSNR $\approx$ 16.06 dB		
Barbara	28.28 K=80	29.06 28.82	27.84 30.72	25.76 K=71	26.34 25.60	25.17 27.99
Hill	28.65 K=63	28.80 28.58	28.51 29.85	27.29 K=14	26.93 26.29	26.80 27.99
Mandrill	24.29 K= 148	24.59 24.88	23.58 27.85	22.25 K= 61	22.29 22.43	21.71 25.37
Lena	30.49 K=74	31.12 30.45	28.86 32.08	28.81 K=24	28.78 27.58	26.74 29.86
Peppers	30.25 K= 88	29.64 30.23	28.87 30.16	28.23 K= 13	27.06 27.27	26.66 27.70

Table 1. DENOISING RESULTS FOR 2 NOISE LEVELS σ = 25 AND 40 FOR 5 IMAGES. LEFT ARE IBPDL-SVA PSNR (TOP) AND DICTIONARY SIZE (BOTTOM). CENTER ARE PSNR USING IBPDL-GIBBS (TOP), DLENE (BOTTOM). RIGHT ARE K-SVD (TOP), BM3D (BOTTOM).

References

[1] K. Jiang, B. Kulis, and M. I. Jordan, "Small-variance asymptotics for exponential family dirichlet process mixture models," in *NIPS*, 2012.

[2] T. Broderick, B. Kulis, and M. Jordan, "Mad-bayes: Map-based asymptotic derivations from bayes," in *ICML*, 2013.

[3] M. Pereyra and S. McLaughlin, "Fast unsupervised bayesian image segmentation with adaptive spatial regularisation," *IEEE Trans. on Image Process.*, 2017.

[4] I. Tosic and P. Frossard, "Dictionary learning: What is the right representation for my signal," *IEEE Signal Process. Magazine*, 2011.

[5] H.-P. Dang and P. Chainais, "Indian buffet process dictionary learning : algorithms and applications to image processing," *International Journal of Approximate Reasoning*, 2017.

[6] T. L. Griffiths and Z. Ghahramani, "Infinite latent feature models and the indian buffet process," in *NIPS*, 2006.

[7] Z. Ghahramani, T. L. Griffiths, and P. Sollich, "Bayesian nonparametric latent feature models," *Bayesian Statistic*, 2007.

[8] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Process.*, 2006.

[9] B. Kulis and M. I. Jordan, "Revisiting k-means: New Algorithms via Bayesian Nonparametrics," in *ICML*, 2012.

[10] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on Sig. Process.*, 1993.

[11] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," *Asilomar*, 1993.

[12] D. Knowles and Z. Ghahramani, "Nonparametric Bayesian sparse factor models with application to gene expression modeling," *Ann. Appl. Stat.*, vol. 5, 2011.

[13] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Trans. on Image Process.*, 2007.

[14] M. Marsousi, K. Abhari, P. Babyn, and J. Alirezaie, "An adaptive approach to learn overcomplete dictionaries with efficient numbers of elements," *IEEE Trans. on Signal Process.*, 2014.

[15] C. Elvira, P. Chainais, and N. Dobigeon, "Bayesian nonparametric principal component analysis," *preprint arXiv*, 2017.