Parameter-free Small Variance Asymptotics for Dictionary Learning

Hong-Phuong Dang[(1)] and Clément Elvira[(2)]
[(1)] CREST - UMR 9194, Ensai, France
[(2)] Univ. Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France
Email: firstname.lastname@{ensai.fr,inria.fr}
Authors contributed equally.

Abstract—**Learning redundant dictionaries for sparse representation from sets of patches has proven its efficiency in solving inverse problems. However, the optimization process often calls for the prior knowledge of the noise level or the regularization parameters for sparse encoding. In a Bayesian framework, these parameters are integrated within the probabilistic model through the choice of prior distributions. Although efficient, these methods come with numerical disadvantages for large-scale data. Small-variance asymptotic (SVA) approaches pave the way to much cheaper though approximate methods for inference by taking advantage from a fruitful interaction between Bayesian models and optimization algorithms. We propose such a SVA analysis of a Bayesian dictionary learning (DL) model where the noise level and regularization level are jointly estimated so that nearly no parameter tuning is needed. We analyze this algorithm and demonstrate its efficiency on real data to illustrate the relevance of the resulting dictionaries.**

Index Terms—**Bayesian model, small variance asymptotic, sparse representations, dictionary learning, inverse problems.**

1. Introduction

Inverse problems in signal or image processing (*e.g.* denoising or inpainting) are most often ill-posed so that prior information, namely *regularization*, becomes necessary to reduce the set of potential solutions. *Sparse promoting* regularizers have sparked a surge of interest since the notable discovery that many natural signals are approximated well by linear combinations of a few elements from some overcomplete family, called a *dictionary*.

More formally, let $\mathbf{X} \in \mathbb{R}^{L\times N}$ be the patches from the initial clean image. Each column vector $\mathbf{x}_n \in \mathbb{R}^L$ represents a square patch (e.g. 8×8 so $L = 64$) in lexicographic order. In presence of additive noise ε, the acquisition process is often modeled as $\mathbf{Y} = \mathbf{X} + \varepsilon$ where $\mathbf{Y} \in \mathbb{R}^{L\times N}$ is the observation matrix. The goal is to reconstruct $\mathbf{X}$ as a linear combination of a few elements of the dictionary $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_K] \in \mathbb{R}^{L\times K}$ following $\mathbf{X} = \mathbf{DW}$ and such that each column vector $\mathbf{w}_n$ of the encoding coefficients matrix $\mathbf{W} \in \mathbb{R}^{K\times N}$ satisfies $\|\mathbf{w}_n\|_0 \ll L$.

Although the dictionary can originate from mathematical functions (DCT, wavelets...), dictionaries directly learned from the data often outperform the non-adaptive ones [1]. The latter task is referred to as dictionary learning (DL). Alternate procedures have been proposed, solving iteratively sparse approximations problems followed by Least Squares updates of the atoms [2] or rank-1 approximations of the residual error [3]. Although numerically efficient, both methods require to invert high dimensional matrices. Several works have been proposed to tackle this problem, updating the atoms using either gradient descent [4] or online approaches [5, 6]. Assuming i.i.d. Gaussian Noise (*i.e.*, quadratic discrepancy), most methods reduce to minimizing a functional of the form

$$(\widehat{\mathbf{D}}, \widehat{\mathbf{W}}) = \underset{(\mathbf{D},\mathbf{W})}{\operatorname{argmin}} \ \tfrac{1}{2}\|\mathbf{Y} - \mathbf{DW}\|_2^2 + \lambda\|\mathbf{W}\|_p \qquad (1)$$

where sparsity is imposed through the ℓ_0 pseudo-norm or its convex relaxation. In (1), the regularization parameter λ has to be adjusted with the (unknown) noise level and strongly impacts performances. Among the few proposed strategies to choose λ, we mention cross-validation [7], Homotopy based algorithms [8] or probabilistic models.

In a Bayesian framework, Problem (1) is translated in a Gaussian likelihood plus a penalty term originating from a prior distribution $\mathrm{p}(\mathbf{D}, \mathbf{W}, \lambda)$. Designing efficient yet 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 algorithms to approximate Bayesian estimators.

Small-Variance asymptotics (SVA) analyses were recently introduced as a computationally efficient framework to approximate the MAP estimators of a Bayesian model. Initially designed to perform inference with Bayesian non parametric priors [9, 10], its scope now goes beyond machine learning tasks [11]. Coupling hyper-parameters with the noise level, the rationale of SVA is to take the limiting behavior of the MAP estimator as the noise tends to zero. However, such a coupling introduces controlled parameters, losing the advantages of Bayesian model. In this paper, we propose a SVA approach for DL where the controlled parameters are integrated within the Bayesian model, leading to a fully parameter-free Bayesian analysis.

H.-P. Dang is partially supported by AIS Rennes Metropole.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

Up to our knowledge, this is the first parameter-free SVA analysis proposed in the literature. The relevance of the learned dictionaries is illustrated on denoising applications.

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

2. Beta-Bernoulli-Gaussian model

This section introduces the Bernoulli Gaussian model used for DL. We describe the conditional posterior distributions that will be used in a Gibbs sampler for inference. In Sec. 3.1, these posteriors are exploited in the SVA analysis. Note that this model corresponds to the parametric approximation of the Indian Buffet Process used in [12].

$$\begin{array}{rcll}
\mathbf{y}_i & = & \mathbf{D}\mathbf{w}_i + \boldsymbol{\varepsilon}_i & \forall 1 \leq i \leq N \\
\mathbf{d}_k & \sim & \mathcal{N}(0, L^{-1}\mathbb{I}_L) & \forall 1 \leq k \leq K \\
\mathbf{w}_i & = & \mathbf{z}_i \odot \mathbf{s}_i & \\
\mathbf{z}_i & \sim & \prod_{k=1}^{K} \mathrm{Ber}(\pi_k) \quad \text{with} & \boldsymbol{\pi} \sim \prod_{k=1}^{K} \mathrm{Beta}(a_0, b_0) \\
\mathbf{s}_i & \sim & \mathcal{N}(0, \sigma_s^2 \mathbb{I}_K) \quad \text{with} & \sigma_s^2 \sim \mathcal{IG}(c_0, d_0) \\
\boldsymbol{\varepsilon}_i & \sim & \mathcal{N}(0, \sigma_{\boldsymbol{\varepsilon}}^2 \mathbb{I}_L) \quad \text{with} & \sigma_{\boldsymbol{\varepsilon}}^2 \sim \mathcal{IG}(e_0, f_0)
\end{array}$$

Except for σ_D^2 that is fixed to $1/L$ to avoid a multiplicative factor indeterminacy, vague conjugate priors are used for $\boldsymbol{\theta}=(\sigma_S^2, \sigma_{\boldsymbol{\varepsilon}}^2)$, *i.e.*, inverse Gamma with small hyperparameters ($c_0, d_0, e_0, f_0 = 10^{-6}$). We set $a_0 = b_0 = 1$ which is equivalent to choosing a uniform distribution on $[0, 1]$ since we have no prior information on the use of each atom.

Using the Bayes rule, the joint posterior distribution of the unknown parameters $\mathbf{D}, \mathbf{W} = \mathbf{Z} \odot \mathbf{S}, \boldsymbol{\pi}, \boldsymbol{\theta}$ writes

$$\begin{aligned}
&\mathrm{p}\big(\mathbf{D}, \mathbf{Z}, \mathbf{S}, \boldsymbol{\pi}, \boldsymbol{\theta} | \mathbf{Y}\big) \propto \mathrm{p}(\mathbf{Y} \mid \mathbf{D}, \mathbf{W}, \sigma_{\boldsymbol{\varepsilon}})\mathrm{p}(\mathbf{D}, \mathbf{W}, \boldsymbol{\pi}, \boldsymbol{\theta}) \\
&\propto \Big(\frac{1}{2\pi\sigma_{\boldsymbol{\varepsilon}}^2}\Big)^{\frac{NL}{2}} \exp\Big(-\frac{1}{2\sigma_{\boldsymbol{\varepsilon}}^2}\mathbf{tr}[(\mathbf{Y} - \mathbf{D}\mathbf{W})^{\mathrm{T}}(\mathbf{Y} - \mathbf{D}\mathbf{W})]\Big) \\
&\times \Big(\tfrac{L}{2\pi}\Big)^{\frac{LK}{2}} \exp\big(-\tfrac{L}{2}\mathbf{tr}[\mathbf{D}^{\mathrm{T}}\mathbf{D}]\big) \times \Big(\tfrac{1}{2\pi\sigma_S^2}\Big)^{\frac{NK}{2}} \exp\Big(-\tfrac{1}{2\sigma_S^2}\mathbf{tr}[\mathbf{S}^{\mathrm{T}}\mathbf{S}]\Big) \\
&\times \prod_{k=1}^{K} \pi_k^{\sum_{i=1}^{N} \mathbf{Z}(k,i)} (1 - \pi_k)^{N - \sum_{i=1}^{N} \mathbf{Z}(k,i)} \times \prod_{k=1}^{K} \mathbb{1}_{[0,1]}(\pi_k)\mathrm{p}(\boldsymbol{\theta}) \quad (2)
\end{aligned}$$

where $\mathbf{Z}(k, i)$ denotes the (k, i) entry of $\mathbf{Z}$. Bayesian estimators can be approximated by resorting to a Gibbs sampler. It consists in drawing from (2) by sampling alternately $\mathbf{D}, \mathbf{Z}, \mathbf{S}, \boldsymbol{\pi}, \boldsymbol{\theta}$ according to

Posterior distribution of each atom $\mathbf{d}_k$

$$\mathbf{d}_k | \mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{D}_{-k}, \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{d}_k}, \boldsymbol{\Sigma}_{\mathbf{d}_k}) \quad (3)$$

$$\begin{cases}
\boldsymbol{\Sigma}_{\mathbf{d}_k} = (L\mathbb{I}_L + \sigma_{\boldsymbol{\varepsilon}}^{-2}\mathbb{I}_L \sum_{i=1}^{N} w_{ki}^2)^{-1} \\
\boldsymbol{\mu}_{\mathbf{d}_k} = \sigma_{\boldsymbol{\varepsilon}}^{-2}\boldsymbol{\Sigma}_{\mathbf{d}_k} \sum_{i=1}^{N} w_{ki}(\mathbf{y}_i - \sum_{j \neq k}^{K} \mathbf{d}_j w_{ji}).
\end{cases}$$

Note that when an atom $\mathbf{d}_k$ is unused, its posterior distribution reduces to its prior. The same phenomenon will occur for the coefficients w_{ki} (cf. (5)).

Posterior distribution of z_{ki}

$$z_{ki} | \mathbf{Y}, \mathbf{Z}_{-ki}, \mathbf{S}, \mathbf{D}, \boldsymbol{\theta} \sim \mathrm{Ber}\Big(\frac{p_1}{p_1 + p_0}\Big) \quad (4)$$

$$p_1 = \pi_k \exp\Big[\frac{-1}{2\sigma_{\boldsymbol{\varepsilon}}^2}(s_{ki}^2 \mathbf{d}_k^{\mathrm{T}}\mathbf{d}_k - 2s_{ki}\mathbf{d}_k^{\mathrm{T}}(\mathbf{y}_i - \sum_{j \neq k}^{K} w_{ji}\mathbf{d}_j)\Big]$$

$$p_0 = (1 - \pi_k)\mathrm{e}^0 = (1 - \pi_k).$$

Posterior distribution of s_{ki}

$$s_{ki} | \mathbf{Y}, \mathbf{D}, \mathbf{Z}, \mathbf{S}_{-(k,i)}, \boldsymbol{\theta}) \sim \mathcal{N}(\mu_{s_{ki}}, \Sigma_{s_{ki}}) \quad (5)$$

$$z_{ki} = 1 \Rightarrow \begin{cases}
\Sigma_{s_{ki}} = & (\sigma_{\boldsymbol{\varepsilon}}^{-2}\mathbf{d}_k^{\mathrm{T}}\mathbf{d}_k + \sigma_S^{-2})^{-1} \\
\mu_{s_{ki}} = & \sigma_{\boldsymbol{\varepsilon}}^{-2}\Sigma_{s_{ki}}\mathbf{d}_k^{\mathrm{T}}(\mathbf{y}_i - \sum_{j \neq k}^{K} \mathbf{d}_j w_{ji})
\end{cases}$$

$$z_{ki} = 0 \Rightarrow \quad \Sigma_{s_{ki}} = \sigma_S^2, \quad \mu_{s_{ki}} = 0$$

Posterior distribution of π_k

$$\pi_k | \mathbf{Z}(k, :) \sim \mathrm{Beta}\Big(a_0 + \sum_{i=1}^{N} z_{ik}, b_0 + N - \sum_{i=1}^{N} z_{ik}\Big). \quad (6)$$

Posterior distribution of $\sigma_{\boldsymbol{\varepsilon}}^2$ and σ_S^2

$$\sigma_{\boldsymbol{\varepsilon}}^2 | \mathbf{Y}, \mathbf{D}, \mathbf{W} \sim \mathcal{IG}\Big(e_0 + \tfrac{NL}{2}, f_0 + \tfrac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{W}\|_F^2\Big) \quad (7)$$

$$\sigma_S^2 | \mathbf{S} \sim \mathcal{IG}\Big(c_0 + \frac{KN}{2}, d_0 + \tfrac{1}{2}\sum_{i=1}^{N} \mathbf{s}_i^{\mathrm{T}}\mathbf{s}_i\Big). \quad (8)$$

3. Proposed method

We study the limiting behavior of the Gibbs sampler described in Section 2. The main novelty is that our analysis yields a parameter-free method.

3.1. Small Variance Asymptotic (SVA) Analysis

In Bayesian inference, the posterior (2) is the cornerstone of the definition of estimators. We focus here on the MAP point estimator which is obtained by maximizing (2) or, equivalently, its neg-log posterior. Since the exact solution is difficult to obtain, one often resorts to Monte Carlo integration or variational approaches to solve the solution. In a manner akin to [10, 13, 14], we propose instead to approximate the solution by analyzing the limiting behavior of (2) as $\sigma_{\boldsymbol{\varepsilon}}^2$ tends to zero.

Since the dictionary $\mathbf{D}$ is assumed over-complete, we can reasonably assume that taking this limit will promote a small reconstruction error, resulting in a non-sparse code $\mathbf{W}$. Therefore, it is necessary to scale the weight associated to each factor with ${\sigma_{\boldsymbol{\varepsilon}}}^2$. For all $k \in [\![1, K]\!]$, we consider the parametrization $\pi_k = \exp\Big(-\frac{\lambda_k}{2\sigma_{\boldsymbol{\varepsilon}}^2}\Big)$ with $\lambda_k > 0$, so that $\pi_k \to 0$ as $\sigma_{\boldsymbol{\varepsilon}}^2 \to 0$, hence promoting a parsimonious use of the factor. With this dependence in mind, we consider the following approximation of the MAP estimator

$$\widehat{\mathbf{D}}, \widehat{\mathbf{Z}}, \widehat{\mathbf{S}}, \widehat{\boldsymbol{\pi}} \simeq \underset{\mathbf{D}, \mathbf{Z}, \mathbf{S}, \boldsymbol{\pi}}{\arg\max} \lim_{\sigma_{\boldsymbol{\varepsilon}}^2 \to 0} -2\sigma_{\boldsymbol{\varepsilon}}^2 \, \mathrm{logp}(\mathbf{D}, \mathbf{Z}, \mathbf{S}, \boldsymbol{\pi}, \boldsymbol{\theta} | \mathbf{Y}) \quad (9a)$$

$$\widehat{\sigma_{\boldsymbol{\varepsilon}}^2}, \widehat{\sigma_S^2} \simeq \underset{\sigma_{\boldsymbol{\varepsilon}}^2, \sigma_S^2}{\arg\max} - \log \mathrm{p}(\sigma_{\boldsymbol{\varepsilon}}^2, \sigma_S^2 | \mathbf{Y}, \widehat{\mathbf{D}}, \widehat{\mathbf{Z}}, \widehat{\mathbf{S}}, \widehat{\boldsymbol{\pi}}). \quad (9b)$$

The discussion on the potential theoretical difficulties in the inversion between the limit and the maximization is postponed to the extended version of this work. Note that the value of σ_S^2 in (9a) has no effect on the estimation. Indeed, letting $\sigma_{\varepsilon}^2 \to 0$, we see that the r.h.s. of (9a) becomes

$$\mathbf{tr}\big[(\mathbf{Y}-\mathbf{D}\mathbf{W})^{\mathrm{T}}(\mathbf{Y}-\mathbf{D}\mathbf{W})\big] + \sum_{k=1}^{K} \lambda_k \sum_{i=1}^{N} \mathbf{Z}[k,i]. \quad (10)$$

The trace originates from the exponential function in the Gaussian likelihood, and the penalty term - reminiscent of ℓ_0-penalty - originates from the Bernoulli prior. It follows that finding the MAP estimate for the dictionary learning problem is asymptotically equivalent to solve

$$\underset{\mathbf{D},\mathbf{W},\boldsymbol{\lambda}}{\operatorname{argmin}} \quad \big\|\mathbf{Y}-\mathbf{D}\mathbf{W}\big\|_F^2 + \sum_{k=1}^{K} \lambda_k \big\|\mathbf{W}[k,:]\big\|_0 \quad (11)$$

The solution of (11) is referred to as asymptotic MAP (aMAP). Note that the choice of the regularization parameters λ_k is of importance since they should decrease as the noise level σ_ε increases. Moreover, unlike the standard optimization problem (1), Eq. (11) penalizes the use of each atom $\mathbf{d}_k$. The novelty of our approach is to benefit from the Bayesian model in order to jointly estimate the hyperparameters $\sigma_{\varepsilon}^2, \sigma_S^2$ within the SVA framework. Indeed, once the optimal value of $\widehat{\mathbf{D}}, \widehat{\mathbf{W}}, \widehat{\boldsymbol{\pi}}$ is known, estimating $\boldsymbol{\theta}$ is equivalent to solve (9b). Next section describes the proposed strategy to approximate the aMAP estimator.

3.2. BBG-SVA algorithm

We formulate the proposed BBG-SVA algorithm (cf. Alg. 1) to solve the optimization problem (11). Letting the noise variance tend to 0, BBG-SVA is deduced from the limiting behavior of the Gibbs sampler described in Section 2. In some cases, this analysis returns to taking the mode of the posterior distribution.

Update $\mathbf{d}_k$. Define $m_k \triangleq \|\mathbf{W}[k,:]\|_0$. When $m_k \neq 0$, we deduce from (3) that the posterior distribution of the atoms reduces to a degenerated Gaussian. Hence

$$\mathbf{d}_k = \frac{1}{\sum\limits_{i=1}^{N} w_{ki}^2} \sum_{i=1}^{N} w_{ki}(\mathbf{y}_i - \sum_{j\neq k}^{K} \mathbf{d}_j w_{ji}). \quad (12)$$

When $m_k = 0$, as emphasized below (3), we sample $\mathbf{d}_k \sim \mathcal{N}(\mathbf{0}, \sigma_S^2 \mathbb{I}_L)$. Finally, the updated vector $\mathbf{d}_k$ is normalized to avoid multiplicative indeterminacy.

Update z_{ki}. Let $\rho = s_{ki}^2 - s_{ki}\mathbf{d}_k{}^{\mathrm{T}}(\mathbf{y}_i - \sum_{\ell\neq k}^{N} s_{\ell i}\mathbf{d}_\ell) + \lambda_k$. Remembering that $\pi_k = \exp(-\frac{\lambda_k}{2\sigma_\varepsilon^2})$, we deduce from (4) that $p_1 = \exp(-\frac{\rho}{2\sigma_\varepsilon^2})$ so $\lim\limits_{\sigma_\varepsilon\to 0} p_1 = +\infty$ if $\rho < 0$ and 0 if $\rho > 0$. Since $\lim\limits_{\sigma_\varepsilon\to 0} p_0 = 1$, we obtain

$$\begin{cases} z_{ki} = 1 & \text{if } \rho < 0 \\ z_{ki} = 0 & \text{if } \rho > 0. \end{cases} \quad (13)$$

```
Input: $\mathbf{Y}$, $\mathbf{D}$, $\mathbf{W}$, $\boldsymbol{\lambda}$, $\sigma_S^2 = 1$
$\mathbf{E} \leftarrow \mathbf{Y} - \mathbf{D}\mathbf{W}$ ;
for each iteration $t$
    for each $k \in [\![1, K]\!]$
        \\ Remove influence of atom $k$
            $\mathbf{E}_{-k} \leftarrow \mathbf{E} + \mathbf{D}[:,k]\,\mathbf{W}[k,:]$ ;
        \\ Update $\mathbf{d}_k$ according to (12)
            if $\|\mathbf{W}[k,:]\|_0 = 0$ then
                $\mathbf{d}_k \sim \mathcal{N}(0, L^{-1}\mathbb{I}_L)$
            else
                $\mathbf{d}_k \leftarrow \mathbf{E}_{-k}\mathbf{W}[k,:]^{\mathrm{T}}$;
            $\mathbf{d}_k \leftarrow \frac{\mathbf{d}_k}{\|\mathbf{d}_k\|_2}$;
        \\ Update $\mathbf{Z}[k,:]$ acc. to (13)
            $\mathbf{s}_{tmp} \leftarrow \mathbf{W}[k,:]$ ;
            $\mathbf{s}_{tmp}[\mathbf{W}[k,:] = 0] \sim \mathcal{N}(0, \sigma_S^2)$ ;
            $p \leftarrow \mathbf{s}_{tmp}^{\odot 2} - 2\mathbf{s}_{tmp} \odot (\mathbf{d}_k^{\mathrm{T}}\mathbf{E}_{-k}) + \boldsymbol{\lambda}[k]$ ;
            $\mathbf{W}\big(k, p \geq 0\big) \leftarrow 0$ ;
        \\ Update $\mathbf{S}[k,:]$ acc. to (14)
            $\ell_0 \leftarrow \mathbf{W}[k,:] \neq 0$ ;
            $\mathbf{W}[k,\ell_0] \leftarrow \mathbf{d}_k^{\mathrm{T}}\mathbf{E}_{-k}[:,\ell_0]$ ;
        \\ Restore influence of atom $k$
            $\mathbf{E} \leftarrow \mathbf{E}_{-k} - \mathbf{D}[:,k]\,\mathbf{W}[k,:]$ ;
        \\ Update $\pi_k$ acc. to (15)
            if $\|\mathbf{W}[k,:]\|_0 = 0$ then
                $\boldsymbol{\pi}[k] \leftarrow 1/(N+2)$;
            else
                $\boldsymbol{\pi}[k] \leftarrow \frac{1}{N}\|\mathbf{W}[k,:]\|_0$ ;
    \\ Update $\sigma_\varepsilon$ acc. to (17)
        $\sigma_\varepsilon^2 \leftarrow \frac{f_0 + 0.5\|\mathbf{E}\|_F^2}{e_0 + 0.5NL + 1}$ ;
    \\ Update $\boldsymbol{\lambda}$ acc. to (16)
        $\boldsymbol{\lambda} \leftarrow -2\sigma_\varepsilon^2 \log(\boldsymbol{\pi})$ ;
    \\ Update $\sigma_S$ acc. to (18)
        $\sigma_S^2 \leftarrow \frac{d_0 + \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{1}{2}(KN - \|\mathbf{W}\|_0)\sigma_S^2}{c_0 + \frac{1}{2}KN + 1}$
Output: dictionary $\mathbf{D}$, code $\mathbf{W}$, noise level $\sigma_\varepsilon^2$
```

Algorithm 1: Proposed BBG-SVA algorithm.

When $\rho{=}0$, $\lim\limits_{\sigma_\varepsilon\to 0} p_1 {=} \lim\limits_{\sigma_\varepsilon\to 0} p_0 {=} 1$ so we sample $z_{ki} {\sim} \mathrm{Ber}(\frac{1}{2})$. In practice, this case is seldom met.

Update s_{ki}. According to (5), the posterior distribution of s_{ki} is normally distributed. However, when $z_{ki} = 1$, the distribution degenerates as $\sigma_\varepsilon^2 \to 0$. Hence

$$\begin{cases} s_{ki} = \frac{1}{\mathbf{d}_k^{\mathrm{T}}\mathbf{d}_k}\mathbf{d}_k^{\mathrm{T}}(\mathbf{y}_i - \sum\limits_{j\neq k}^{K} \mathbf{d}_j w_{ji}) & \text{if } z_{ki} = 0 \\ s_{ki} \sim \mathcal{N}(0, \sigma_S^2) & \text{otherwise.} \end{cases} \quad (14)$$

Update π_k then λ_k. According to (6), the posterior distribution of π_k is beta distributed. Since a parsimonious use of each atom is desired, we expect $m_k \triangleq \|\mathbf{W}[k,:]\|_0 \ll N$ for large values of N. The variance of the posterior then writes

$$\mathrm{var}[\pi_k|\mathbf{Z}] \simeq \frac{m_k(N-m_k)}{(m_k+N)^2(m_k+N+1)} \simeq \frac{1}{N^2} \ll 1$$

To avoid resorting to randomness, we choose to reduce the distribution to its mode. Hence, for $a_0 = b_0 = 1$, the update

$$\pi_k = \frac{a_0 + m_k - 1}{a_0 + b_0 + N - 2} = \frac{m_k}{N}. \tag{15}$$

Note that when $m_k = 0$ and since $a_0 = b_0 = 1$, the mode of the beta distribution is 0. In this case, we have found that setting π_k to a small value, *e.g.*, $\pi_k = 1/N \simeq \mathbb{E}[\pi_k|\mathbf{Z}]$, leads to better performances. Finally, since $\pi_k = \exp(-\frac{\lambda_k}{\sigma_{\boldsymbol{\varepsilon}}^2})$, we deduce

$$\lambda_k = -2\sigma_{\boldsymbol{\varepsilon}}^2 \log(\pi_k). \tag{16}$$

Update $\sigma_{\boldsymbol{\varepsilon}}^2$ and σ_S^2. Using the same arguments as for the update of π_k, one can show that the variance of the posterior distributions of both $\sigma_{\boldsymbol{\varepsilon}}^2$ and σ_S^2 tends to be small for large values of N. Again, we choose to update with respect to the mode of the distribution, *i.e.*,

$$\sigma_{\boldsymbol{\varepsilon}}^2 = \mathrm{Mode}(\sigma_{\boldsymbol{\varepsilon}}^2|-) = \frac{f_0 + \frac{1}{2}\|\mathbf{Y} - \mathbf{DW}\|_F^2}{e_0 + \frac{1}{2}NL + 1} \tag{17}$$

$$\sigma_S^2 = \mathrm{Mode}(\sigma_S^2|-) = \frac{d_0 + \frac{1}{2}\|\mathbf{S}\|_F^2}{c_0 + \frac{1}{2}KN + 1}. \tag{18}$$

The originality of the approach is to mix deterministic / random approaches. Future work aims at showing that each deterministic move decreases the cost function. However, since the objective function (11) is not convex, there is no guaranties that one converges to a global minimizer. Similarly to MCMC, we expect that such random moves will allow for escaping from spurious minimizers.

4. Numerical experiments

In this section we illustrate the relevance of the dictionary learnt with BBG-SVA on a denoising task.

4.1. Experiments set-up

A set of 5 images of size 512×512 is considered - Barbara, Hill, Mandrill, Lena, Peppers - for 2 noise levels $\sigma_{\boldsymbol{\varepsilon}} = 25$ and 40. Considering patches of size 8×8 (*i.e.* dimension $L = 64$), there are $N = (512-7)^2 = 255025$ overlapping patches for each image. In image processing, when working on image patches of size 8×8, a dictionary of size K=256 or 512 atoms is typically learnt [1, 3, 12]. In this experiment, we choose rather $K = 300$ for the proposed BBG-SVA algorithm. This choice will be motived in Section 4.2. Finally, all results are averaged over 10 Monte Carlo simulations.

Simulations are run on a personal laptop with a Python implementation. Here, according to the model detailed in Section 2, a random initialization is used for $\mathbf{D}$ and $\mathbf{W} = \mathbf{Z} \odot \mathbf{S}$. Only the first atom (column) of $\mathbf{D}$ and row of $\mathbf{W}$ are initialized with the empirical mean and a vector of ones, respectively. Denoting σ^2 the empirical variance of the residual error, the vectors $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ are initialized with $\frac{1}{K}\mathbf{1}_K$ and $-2\sigma_{\boldsymbol{\varepsilon}}^2 \log(\boldsymbol{\pi})$, respectively. Recall that none of the parameters of eq (11) have been fixed. We emphasize that the BBG-SVA approach infers all these parameters: $\mathbf{D}$ and $\mathbf{W}$, sparsity levels $\boldsymbol{\lambda}$ - originates from the parameter $\boldsymbol{\pi}$ of the Bernoulli prior - and noise level $\sigma_{\boldsymbol{\varepsilon}}^2$.

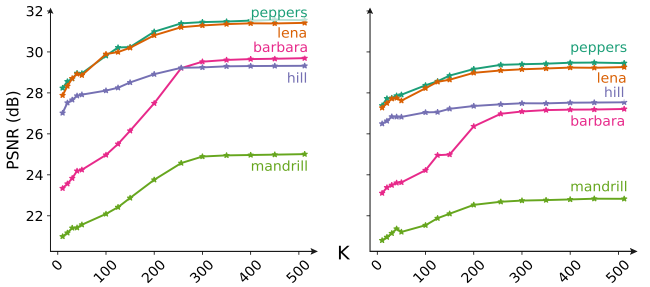

Figure 1. Evolution of BBG-SVA performances (PSNR of the reconstructed image) seen as function of the dictionary size K for several images and $\sigma_{\boldsymbol{\varepsilon}} = 25$ (left) and 40 (right).

4.2. Denoising results

Choosing the optimal size K. As stated in Section 4.1, most methods select $K = 256$ or 512. Before comparing BBG-SVA to other methods, we first motivate our choice of dictionary size K. Figure 1 shows the evolution of the PSNR obtained with BBG-SVA seen as a function of K for two noise levels and several images. In all situations, we observe that performances stabilize for some value of K. As already observed in [15], these results suggest the existence of an optimal size that depends on both the image and the noise level. Moreover, overestimating K only affects the computational cost of the method. We also observed that BBG-SVA was able to perform some pruning, *i.e.*, that some atoms are not used when K is too big. For these reasons, we found that K =300 was a reasonable choice.

Denoising results. We compare BBG-SVA denoising results with: 1/ the original K-SVD [3], 2/ DLENE [16] - an adaptive approach to learn overcomplete dictionaries with an efficient number of atoms, 3/ the state-of-the-art method for denoising BM3D [17] - block matching with 3D filtering and 4/ BPFA [12] - a Bayesian model that resorts to a Beta-Bernoulli process. Results from BM3D are recalled for information only since we do not expect to perform better. The four algorithms assume that the noise variance and / or sparsity level are known, while the proposed model automatically estimates both of them.

Table 1 gathers all the numerical results. Although parameter-free, BBG-SVA outperforms K-SVD and achieves performances comparable to DLENE and BPFA. BBG-SVA even outperforms BM3D on Pepper; we made the same observation for a similar - yet supervised - model [14]. Figure 2 displays typical denoising results obtained by using

	$\sigma_\varepsilon = 25$ PSNR $\approx$ 20.14 dB			$\sigma_\varepsilon = 40$ PSNR $\approx$ 16.06 dB		
Barbara	29.60 $\widehat{\sigma} = 25.5$	29.88 28.82	29.60 *30.72*	27.12 $\widehat{\sigma} = 40.5$	27.72 25.60	26.65 *27.99*
Hill	29.28 $\widehat{\sigma} = 25.8$	29.57 28.58	29.18 *29.85*	27.50 $\widehat{\sigma} = 40.5$	27.60 26.29	27.30 *27.99*
Mandrill	24.99 $\widehat{\sigma} = 27.5$	25.30 24.88	24.38 *27.85*	22.78 $\widehat{\sigma} = 42.7$	23.18 22.43	22.26 *25.37*
Lena	31.37 $\widehat{\sigma} = 25.4$	31.63 30.45	31.32 *32.08*	29.18 $\widehat{\sigma} = 40.2$	29.30 27.58	28.90 *29.86*
Peppers	31.47 $\widehat{\sigma} = 25.3$	30.00 30.23	29.73 *30.16*	29.38 $\widehat{\sigma} = 40.2$	27.57 27.27	27.36 *27.70*

Table 1. DENOISING RESULTS (IN DB) FOR NOISE LEVELS $\sigma = 25$ AND 40 ON 5 IMAGES. LEFT ARE BBG-SVA PSNR (TOP) AND ESTIMATED NOISE LEVEL σ_ε (BOTTOM). CENTER ARE PSNR USING BPFA (TOP), DLENE (BOTTOM). RIGHT ARE K-SVD (TOP), BM3D (BOTTOM).

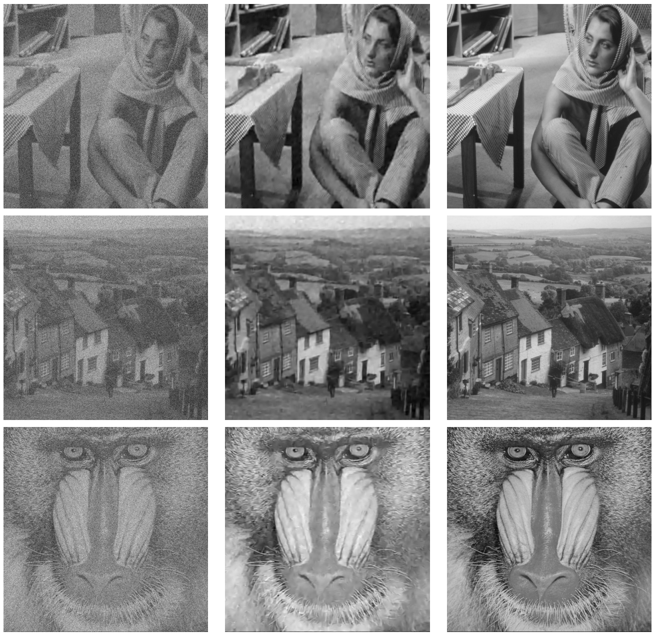

Figure 2. Denoising ($\sigma = 40$) results obtained by using BBG-SVA. From left to right are the noisy, the denoised and the original images.

BBG-SVA on several examples of Table 1. Finally, note that the noise level σ_ε is inferred as well with good accuracy. Except for Mandrill, the estimation error varies from 1.5 to 4.% for σ_ε =25 and from 0.5 to 1.5% when $\sigma_\varepsilon = 40$. This accurate estimate is a benefit of this approach. We conclude that BBG-SVA was able to automatically select relevant values of the hyper-parameters in the objective function (11).

5. Conclusion

This paper presents a parameter-free yet computationally efficient approach for dictionary learning (DL). The methods results from a Small Variance Asymptotic (SVA) analysis of a Bernoulli Gaussian Bayesian model for DL. Such an analysis yields an objective function which is minimized through a SVA analysis of the corresponding Gibbs sampler. The main novelty is that the coupling parameters resulting from the SVA analysis are also estimated within the Bayesian framework. Therefore, the proposed approach gathers both the flexibility of Bayesian modeling and the numerical efficiency of optimization methods. The relevance of the inferred dictionary has been assessed on a denoising task, results are comparable to supervised methods. Future work will investigate the computational cost of the method. Finally, the SVA analysis proposed in [14] can be revisited along the same lines.

References

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

[2] K. Engan, S. Aase, and J. H. Husoy, “Method of optimal directions for frame design,” in *ICASSP*, 1999.

[3] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Trans. Sig. Process.*, 2006.

[4] B. Mailhé and M. D. Plumbley, “Dictionary learning with large step gradient descent for sparse representations,” in *Latent Variable Analysis and Signal Separation*, 2012.

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online learning for matrix factorization and sparse coding,” *J. Mach. Learning Research*, 2010.

[6] N. Rao and F. Porikli, “A clustering approach to optimize online dictionary learning,” in *ICASSP*, 2012.

[7] J. Mairal, F. Bach, and J. Ponce, “Task-driven dictionary learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[8] C. Soussen, J. Idier, J. Duan, and D. Brie, “Homotopy based algorithms for ℓ_0-regularized least-squares,” *IEEE Trans. Sig. Process.*, 2015.

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

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

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

[12] M. Zhou *et al*, “Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images,” *IEEE Trans. Image Process.*, 2012.

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

[14] C. Elvira, H.-P. Dang, and P. Chainais, “Small variance asymptotics and bayesian nonparametrics for dictionary learning,” in *EUSIPCO*, 2018.

[15] H.-P. Dang and P. Chainais, “Indian buffet process dictionary learning : algorithms and applications to image processing,” *Int. J. of Approx. Reasoning*, 2017.

[16] M. Marsousi, K. Abhari, P. Babyn, and J. Alirezaie, “An adaptive approach to learn overcomplete dictionaries with efficient numbers of elements,” *IEEE Trans. Sig. Process.*, 2014.

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