

## Context

### Sparse representation problem

Sparse representations (SR) aim at describing a signal as the combination of a small number of atoms chosen from an overcomplete dictionary.

Let  $\mathbf{D} \in \mathbb{R}^{N \times M}$  be a rank- $N$  matrix whose columns are normed to 1 and  $\mathbf{y} \in \mathbb{R}^N$  an observed signal. A standard formulation of the sparse representation problem writes

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \leq \epsilon,$$

where  $\|\mathbf{x}\|_0$  denotes the  $l_0$ -norm, *i.e.*, the number of nonzero coefficients in  $\mathbf{x}$  and  $\epsilon$  is a given constant.

### State of the art: SR algorithms

NP-hard problem  $\rightsquigarrow$  suboptimal (but tractable) algorithms:

- the **greedy algorithms** build up the sparse vector  $\mathbf{x}$  by making a succession of locally-optimal decisions,
- the **algorithms based on a problem relaxation** approximate the SR problem by relaxed problems that can be solved efficiently by standard optimization procedures,
- the **Bayesian algorithms** express the sparse representation problem as the solution of a Bayesian inference problem and apply statistical tools to solve it.

### Contributions of this paper

In this paper we address the problem of sparse representation within a Bayesian framework. We assume that the observations are generated from a **Bernoulli-Gaussian process** and consider the corresponding Bayesian inference problem. Tractable solutions are then proposed based on the “mean-field” approximation and the **variational Bayes EM algorithm**. The resulting SR algorithms are shown to have a tractable complexity and very good performance over a wide range of sparsity levels.

## Bernoulli-Gaussian formulation of the SR problem

### Probabilistic model

Assume that the observed vector  $\mathbf{y}$  has a Gaussian distribution with mean  $\mathbf{D}\mathbf{x}$  and covariance  $\sigma_n^2 \mathbf{I}_N$ , *i.e.*,

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{D}\mathbf{x}, \sigma_n^2 \mathbf{I}_N),$$

where  $\mathbf{I}_N$  is the  $N \times N$  identity matrix.

Suppose moreover that  $\mathbf{x}$  obeys the following probabilistic model:

$$\begin{aligned} p(\mathbf{x}) &= \sum_{\mathbf{s}} p(\mathbf{x}, \mathbf{s}), \\ &= \prod_{i=1}^M \sum_{s_i} p(x_i | s_i) p(s_i), \end{aligned}$$

where

$$\begin{aligned} p(x_i | s_i) &= \mathcal{N}(0, \sigma^2(s_i)), \\ p(s_i) &= \text{Ber}(p_i), \end{aligned}$$

and  $\text{Ber}(p_i)$  denotes a Bernoulli distribution with parameter  $p_i$ .

$\rightsquigarrow$  If  $\sigma^2(s_i = 0) \ll \sigma^2(s_i = 1)$  and  $p_i \ll 1 \forall i$ , only a small fraction of the  $x_i$ 's will have an amplitude significantly larger than the others.

### Optimization problem

Sparse solutions for  $\mathbf{x}$  can be found as the maximum or the mean of posterior distribution  $p(\mathbf{x}|\mathbf{y})$ :

$$\begin{aligned} \hat{\mathbf{x}} &= \arg \max_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}), \\ \hat{\mathbf{x}} &= \int \mathbf{x} p(\mathbf{x}|\mathbf{y}) d\mathbf{x}, \end{aligned}$$

where

$$p(\mathbf{x}|\mathbf{y}) = \sum_{\mathbf{s}} p(\mathbf{x}, \mathbf{s}|\mathbf{y}).$$

$\rightsquigarrow$  Intractable problems but solutions can be found using variational approximations of  $p(\mathbf{x}, \mathbf{s}|\mathbf{y})$ . One of them: “mean-field” (MF) approximation.

## Mean-field approximations

### Basics

Let  $\boldsymbol{\theta}$  denote a vector of random variables (*e.g.*,  $\boldsymbol{\theta} = [\mathbf{x}^T \mathbf{s}^T]^T$ ) and  $p(\boldsymbol{\theta}|\mathbf{y})$  its a posteriori probability. We define  $q(\boldsymbol{\theta})$  a probability distribution such that:

$$q(\boldsymbol{\theta}) \triangleq q(\boldsymbol{\theta}_1) q(\boldsymbol{\theta}_2),$$

where  $\boldsymbol{\theta}_1$  and  $\boldsymbol{\theta}_2$  are such that  $\boldsymbol{\theta}^T = [\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T]$  and  $\int q(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i = 1$ .

Based on this definition, the **mean-field approximation** of  $p(\boldsymbol{\theta}|\mathbf{y})$ , say  $q^*(\boldsymbol{\theta})$ , can be expressed as:

$$q^*(\boldsymbol{\theta}) = \arg \min_{q(\boldsymbol{\theta})} \text{KL}(q(\boldsymbol{\theta}); p(\boldsymbol{\theta}|\mathbf{y})),$$

where  $\text{KL}(q(\boldsymbol{\theta}); p(\boldsymbol{\theta}|\mathbf{y}))$  is the Kullback-Leibler distance between  $q(\boldsymbol{\theta})$  and  $p(\boldsymbol{\theta}|\mathbf{y})$ , *i.e.*,

$$\text{KL}(q(\boldsymbol{\theta}); p(\boldsymbol{\theta}|\mathbf{y})) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})} d\boldsymbol{\theta}.$$

The solution of this problem can be computed by the “**variational Bayes EM**” (VB-EM) algorithm which is ensured to converge and iterates between estimations of the distributions:

$$\begin{aligned} q^{(n+1)}(\boldsymbol{\theta}_1) &\propto \exp \left\{ \langle \log p(\mathbf{y}, \boldsymbol{\theta}) \rangle_{q^{(n)}(\boldsymbol{\theta}_2)} \right\}, \\ q^{(n+1)}(\boldsymbol{\theta}_2) &\propto \exp \left\{ \langle \log p(\mathbf{y}, \boldsymbol{\theta}) \rangle_{q^{(n+1)}(\boldsymbol{\theta}_1)} \right\}, \end{aligned}$$

where  $\propto$  denotes equality up to a normalization factor and

$$\langle \log p(\mathbf{y}, \boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta}_i)} \triangleq \int q(\boldsymbol{\theta}_i) \log p(\mathbf{y}, \boldsymbol{\theta}) d\boldsymbol{\theta}_i.$$

### Algorithm VBSR1: $p(\mathbf{x}, \mathbf{s}|\mathbf{y}) \simeq q(\mathbf{x})q(\mathbf{s})$

Let  $\Sigma_{\mathbf{s}}$  be a diagonal matrix whose  $i$ th diagonal element is defined as  $(\Sigma_{\mathbf{s}})_{ii} \triangleq \sigma^2(s_i)$ . The MF approximation  $q(\mathbf{x}, \mathbf{s}) = q(\mathbf{x})q(\mathbf{s})$  leads to the following expressions:

$$\begin{aligned} q(\mathbf{s}) &\propto (\sqrt{\det \Sigma_{\mathbf{s}}})^{-1} \exp \left\{ -\frac{1}{2} \langle \mathbf{x}^T \Sigma_{\mathbf{s}}^{-1} \mathbf{x} \rangle_{q(\mathbf{x})} \right\} p(\mathbf{s}), \\ &\propto \prod_i \frac{1}{\sqrt{\sigma^2(s_i)}} \exp \left\{ -\frac{\langle x_i^2 \rangle_{q(x_i)}}{2\sigma^2(s_i)} \right\} p(s_i), \\ q(\mathbf{x}) &= \mathcal{N}(m, \Gamma), \end{aligned}$$

where

$$\Gamma = \left( \frac{\mathbf{D}^T \mathbf{D}}{\sigma_n^2} + \langle \Sigma_{\mathbf{s}}^{-1} \rangle_{q(\mathbf{s})} \right)^{-1}, \quad m = \frac{1}{\sigma_n^2} \Gamma \mathbf{D}^T \mathbf{y}.$$

Back to the optimization problem:

- An approximation of  $p(\mathbf{x}|\mathbf{y})$  can be computed as:

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \int p(\mathbf{x}, \mathbf{s}|\mathbf{y}) d\mathbf{s}, \\ &\simeq \int q^*(\mathbf{x}, \mathbf{s}) d\mathbf{s} = \int q^*(\mathbf{x}) q^*(\mathbf{s}) d\mathbf{s}, \\ &= q^*(\mathbf{x}), \end{aligned}$$

- So that the posterior estimate is:

$$\begin{aligned} \hat{\mathbf{x}} &= \arg \max_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}), \\ &\simeq \arg \max_{\mathbf{x}} \log q(\mathbf{x}) = m = \int \mathbf{x} q(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

Mean and posterior estimates are in this case (Gaussian distribution) equal.

- Complexity order similar to the one of the algorithms based on problem relaxation (such as BP or FOCUS).

### Algorithm VBSR2: $p(\mathbf{x}, \mathbf{s}|\mathbf{y}) \simeq \prod_i q(x_i, s_i)$

The MF approximation  $q(\mathbf{x}, \mathbf{s}) = \prod_i q(x_i, s_i)$  leads to the following expressions:

$$q(x_i, s_i) = q(x_i | s_i) q(s_i) \quad \forall i,$$

where  $q(x_i | s_i)$  and  $q(s_i)$  are defined as follows:

$$\begin{aligned} q(x_i | s_i) &= \mathcal{N}(m(s_i), \Gamma(s_i)), \\ q(s_i) &\propto \frac{1}{\sqrt{\sigma_n^2 + \sigma^2(s_i)}} \exp \left\{ -\frac{1}{2} \frac{(\mathbf{r}_i - m(s_i))^2}{\sigma_n^2 + \sigma^2(s_i)} \right\} p(s_i), \end{aligned}$$

and

$$\Gamma(s_i) = \frac{\sigma^2(s_i) \sigma_n^2}{\sigma^2(s_i) + \sigma_n^2}, \quad m(s_i) = \frac{\sigma^2(s_i)}{\sigma_n^2 + \sigma^2(s_i)} \mathbf{d}_i^T \mathbf{r}_i, \quad \mathbf{r}_i = \mathbf{y} - \sum_{j \neq i} \langle x_j \rangle_{q(x_j, s_j)} \mathbf{d}_j.$$

Back to the optimization problem:

- An approximation of  $p(\mathbf{x}|\mathbf{y})$  can be computed as:

$$p(\mathbf{x}|\mathbf{y}) \simeq \prod_i \sum_{s_i} q(x_i, s_i),$$

- So that the mean estimate is:

$$\hat{x}_i \simeq \langle x_i \rangle_{q(x_i)} = \sum_{s_i} m(s_i) q(s_i) \quad \forall i.$$

- Complexity order similar to the one of the greedy algorithms (such as MP or GP).

### Algorithm VBSR3: Combination of mean-field approximations

- **Motivations:** reduce the loss of information due to the MF factorizations by taking benefit from both decompositions.

- **In practice:** definition of an algorithm as the combination of VBSR1 and VBSR2 updates:

1.  $q^{(n)}(\mathbf{x}) = \mathcal{N}(m, \Gamma)$  where  $m$  and  $\Gamma$  are defined in VBSR1.
2.  $q^{(n)}(\mathbf{s}) = \prod_i q(s_i)$  where  $q(s_i)$  are computed from VBSR2 by using  $\langle x_j \rangle_{q(x_j, s_j)} = \langle x_j \rangle_{q^{(n)}(\mathbf{x})}$ .

- Complexity similar to the one of the algorithms based on problem relaxation (like VBSR1).

- Convergence not theoretically ensured since update equations do not define a VB-EM algorithm.

## Performance analysis

- **Experiments setup:**  $N = 128$ ,  $M = 256$ ,  $\sigma_n^2 = 10^{-5}$ . Elements of the dictionary are *i.i.d* realizations of a zero-mean Gaussian distribution with variance  $N^{-1}$ . Positions of the non-zero coefficients are drawn uniformly at random. Amplitudes of the active (resp. inactive) coefficients are generated from a zero-mean Gaussian with variance  $\sigma^2(s_i = 1) = 10$  (resp.  $\sigma^2(s_i = 0) = 10^{-8}$ ). For each point of simulation, we run 200 trials.

- **Performance measurements:** We compare the performance achieved by VBSR1, VBSR2 and VBSR3 with other standard SR algorithms (MP, OMP, BP, SP and RVM).

Performance is evaluated via the empirical frequency of correct reconstruction versus the number of non-zero coefficients in  $\mathbf{x}$ , say  $K$ .

- **Stopping criterion:** MP and OMP are run until the  $l_2$ -norm of the residual drops below  $\sqrt{N\sigma_n^2}$ . The probabilities of activity used by the VBSR algorithms are set to  $p(s_i = 1) = K/M \forall i$ . We noticed that the performance of VBSR1 and VBSR3 can be greatly improved by progressively decreasing the variance on the inactive coefficients. We used the following strategy:

$$(\sigma^2(s_i = 0))^{(n)} = 0.8 \sigma^2(s_i = 1) \alpha^n + \sigma^2(s_i = 0) \quad \forall i,$$

where  $n$  is the iteration number and  $\alpha < 1$ .

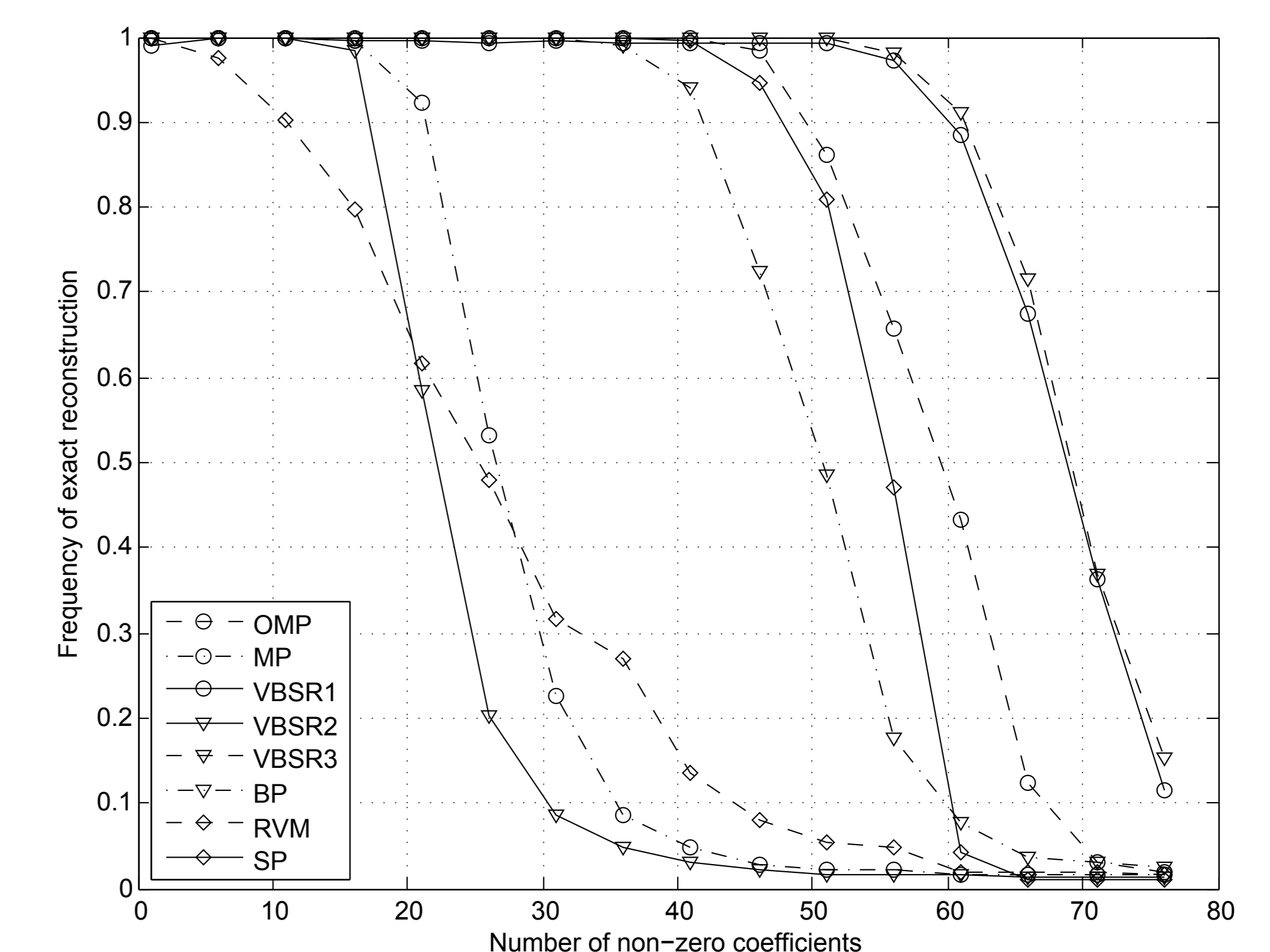


FIGURE 1: Frequency of exact reconstruction versus number of non-zero coefficients;  $N = 128$ ,  $M = 256$ ,  $\sigma_n^2 = 10^{-5}$ ,  $\sigma^2(s_i = 0) = 10^{-8}$ ,  $\sigma^2(s_i = 1) = 10 \forall i$ .