

Context

Sparse representation problem

Let $\mathbf{D} \in \mathbb{R}^{N \times M}$ be a dictionary with $N \leq M$ and $\mathbf{y} \in \mathbb{R}^N$ an observed signal. Find the vector $\mathbf{x} \in \mathbb{R}^M$ such that:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 \quad \text{subject to} \quad \|\mathbf{x}\|_0 \leq L,$$

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

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_0,$$

where λ is a Lagrangian multiplier.

Design of dictionaries adapted to sparse representations

Given a training set $\{\mathbf{y}_j\}_{j=1}^K$, find the dictionary \mathbf{D}^* which leads to the best distortion-sparsity compromise, *i.e.*,

$$\mathbf{D}^* = \arg \min_{\mathbf{D}} \left\{ \sum_j \min_{\mathbf{x}_j} \|\mathbf{y}_j - \mathbf{D}\mathbf{x}_j\|_2^2 + \lambda \|\mathbf{x}_j\|_0 \right\}.$$

Low bit rate compression and sparsity in context of orthonormal basis [1]

Dependency of distortion and rate on the number of nonzero quantized transform coefficients, say L

$$D = \varphi(L), \\ R = \gamma L,$$

where $\varphi(L)$ and γ depend on the basis.

↪ At low bit rates and in context of orthonormal basis, the rate-distortion performance depends on the ability of the basis to provide a good approximation of the signal with few coefficients.

Contributions of this paper

In this paper, we focus on the learning of dictionary made up of the union of orthonormal bases. This topic was the object of some recent contributions ([2],[3]). We propose here a probabilistic interpretation of one of them and suggest a novel optimization procedure based on the expectation-maximization (EM) algorithm ([4]).

A probabilistic framework

Let $\{\mathbf{y}_j\}_{j=1}^K$ be a set of training signals for the optimization of an overcomplete dictionary \mathbf{D} . We suppose that \mathbf{D} is made up of P orthonormal bases, *i.e.*,

$$\mathbf{D} \triangleq [\mathbf{D}_1, \dots, \mathbf{D}_i, \dots, \mathbf{D}_P], \quad \mathbf{D}_i^T \mathbf{D}_i = \mathbf{I}_N,$$

where \mathbf{I}_N is the N -dimensional identity matrix. Let finally \mathbf{x}_j be the vector made up of the \mathbf{x}_{ji} 's which correspond to the sparse representations of \mathbf{y}_j in bases \mathbf{D}_i 's, *i.e.*,

$$\mathbf{x}_j^T \triangleq [\mathbf{x}_{j1}^T, \dots, \mathbf{x}_{ji}^T, \dots, \mathbf{x}_{jP}^T]^T.$$

We consider the following model for \mathbf{y}_j :

$$p(\mathbf{y}_j | \mathbf{D}) = \int_{\mathbb{R}^M} \sum_{c_j=1}^P p(\mathbf{y}_j | \mathbf{x}_j, \mathbf{D}, c_j) p(\mathbf{x}_j | c_j) p(c_j) d\mathbf{x}_j,$$

with

$$p(\mathbf{y}_j | \mathbf{x}_j, \mathbf{D}, c_j = i) = \mathcal{N}(\mathbf{D}_i \mathbf{x}_{ji}, \sigma^2 \mathbf{I}_N),$$

where $\mathcal{N}(\mu, \Gamma)$ denotes a Gaussian distribution with mean μ and covariance Γ , and

$$p(\mathbf{x}_j | c_j = i) \propto \exp\{-\lambda \|\mathbf{x}_{ji}\|_0\},$$

where $\lambda > 0$.

Learning algorithms

Sezer's algorithm [3]

0. Initialization

Set $\mathbf{D}^{(0)} = [\mathbf{D}_1^{(0)}, \dots, \mathbf{D}_i^{(0)}, \dots, \mathbf{D}_P^{(0)}]$, and

$$\forall i \in \{1, \dots, P\}, \forall j \in \{1, \dots, K\}, \mathbf{x}_{ji}^{(0)} = \arg \min_{\mathbf{x}_{ji}} \left\{ \|\mathbf{y}_j - \mathbf{D}_i^{(0)} \mathbf{x}_{ji}\|_2^2 + \lambda' \|\mathbf{x}_{ji}\|_0 \right\}.$$

1. Classification

$\forall i \in \{1, \dots, P\}$, compute $\mathcal{S}_i^{(k)} = \{j \in \{1, \dots, K\} \mid c_j^{(k)} = i\}$,

where $c_j^{(k)} = \arg \min_{i \in \{1, \dots, P\}} \left\{ \|\mathbf{y}_j - \mathbf{D}_i^{(k-1)} \mathbf{x}_{ji}^{(k-1)}\|_2^2 + \lambda' \|\mathbf{x}_{ji}^{(k-1)}\|_0 \right\}$.

2. Basis update

$\forall i \in \{1, \dots, P\}, \forall j \in \{1, \dots, K\}$, update \mathbf{D}_i and \mathbf{x}_{ji} as follows:

$$\mathbf{D}_i^{(k)} = \arg \min_{\mathbf{D}_i} \left\{ \sum_{j \in \mathcal{S}_i^{(k)}} \min_{\mathbf{x}_{ji}} \left\{ \|\mathbf{y}_j - \mathbf{D}_i \mathbf{x}_{ji}\|_2^2 + \lambda' \|\mathbf{x}_{ji}\|_0 \right\} \right\} \quad \text{subject to} \quad \mathbf{D}_i^T \mathbf{D}_i = \mathbf{I}_N,$$

$$\mathbf{x}_{ji}^{(k)} = \arg \min_{\mathbf{x}_{ji}} \left\{ \|\mathbf{y}_j - \mathbf{D}_i^{(k)} \mathbf{x}_{ji}\|_2^2 + \lambda' \|\mathbf{x}_{ji}\|_0 \right\}.$$

...Revisited

With assumptions $p(c_j) = \frac{1}{P}$, $\forall c_j, \forall j$, and $\lambda' = 2\lambda\sigma^2$, Sezer's algorithm is equivalent to the MAP estimation problem:

$$(\mathbf{D}^*, \mathbf{X}^*, \mathbf{c}^*) = \arg \max_{(\mathbf{D}, \mathbf{X}, \mathbf{c})} \sum_{j=1}^K \log p(\mathbf{y}_j, \mathbf{x}_j, \mathbf{D}, c_j).$$

We can indeed recognize the two steps:

$$\mathbf{c}^{(k)} = \arg \max_{\mathbf{c}} \sum_{j=1}^K \log p(\mathbf{y}_j, \mathbf{x}_j^{(k-1)}, \mathbf{D}^{(k-1)}, c_j), \\ \text{with } \mathbf{c} = [c_1, \dots, c_K]^T,$$

$$(\mathbf{D}^{(k)}, \mathbf{X}^{(k)}) = \arg \max_{(\mathbf{D}, \mathbf{X})} \sum_{j=1}^K \log p(\mathbf{y}_j, \mathbf{x}_j, \mathbf{D}, c_j^{(k)}).$$

Alternative approach

We consider instead the marginalized MAP estimation problem:

$$(\mathbf{D}^*, \mathbf{X}^*) = \arg \max_{(\mathbf{D}, \mathbf{X})} \sum_{j=1}^K \log p(\mathbf{y}_j, \mathbf{x}_j, \mathbf{D}),$$

where $p(\mathbf{y}_j, \mathbf{x}_j, \mathbf{D}) = \sum_{c_j=1}^P p(\mathbf{y}_j, \mathbf{x}_j, \mathbf{D}, c_j)$.

This problem is solved by the EM algorithm:

- E-step computes a lower bound on the log likelihood with respect to the current estimates of the model parameters,
- M-step estimates the model parameters which maximize the function evaluated in the E-step.

EM-based algorithm

0. Initialization

Set $\mathbf{D}^{(0)} = [\mathbf{D}_1^{(0)}, \dots, \mathbf{D}_i^{(0)}, \dots, \mathbf{D}_P^{(0)}]$, $\lambda' \triangleq 2\lambda\sigma^2$, and

$$\forall i \in \{1, \dots, P\}, \forall j \in \{1, \dots, K\}, \mathbf{x}_{ji}^{(0)} = \arg \min_{\mathbf{x}_{ji}} \left\{ \|\mathbf{y}_j - \mathbf{D}_i^{(0)} \mathbf{x}_{ji}\|_2^2 + \lambda' \|\mathbf{x}_{ji}\|_0 \right\}.$$

1. E-step

$\forall i \in \{1, \dots, P\}, \forall j \in \{1, \dots, K\}$, compute

$$w_{ji}^{(k)} \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y}_j - \mathbf{D}_i^{(k-1)} \mathbf{x}_{ji}^{(k-1)}\|_2^2 - \lambda \|\mathbf{x}_{ji}^{(k-1)}\|_0\right) p(c_j).$$

2. M-step

$\forall i \in \{1, \dots, P\}, \forall j \in \{1, \dots, K\}$, update \mathbf{D}_i and \mathbf{x}_{ji} as follows:

$$\mathbf{D}_i^{(k)} = \arg \min_{\mathbf{D}_i} \left\{ \sum_{j=1}^K w_{ji}^{(k)} \min_{\mathbf{x}_{ji}} \left\{ \|\mathbf{y}_j - \mathbf{D}_i \mathbf{x}_{ji}\|_2^2 + \lambda' \|\mathbf{x}_{ji}\|_0 \right\} \right\} \quad \text{subject to} \quad \mathbf{D}_i^T \mathbf{D}_i = \mathbf{I}_N,$$

$$\mathbf{x}_{ji}^{(k)} = \arg \min_{\mathbf{x}_{ji}} \left\{ \|\mathbf{y}_j - \mathbf{D}_i^{(k)} \mathbf{x}_{ji}\|_2^2 + \lambda' \|\mathbf{x}_{ji}\|_0 \right\}.$$

Estimation of the noise variance

Noise variance included as a new unknown variable in the MAP problem

$$(\mathbf{D}^*, \mathbf{X}^*, (\sigma^2)^*) = \arg \max_{(\mathbf{D}, \mathbf{X}, \sigma^2)} \sum_{j=1}^K \log p(\mathbf{y}_j, \mathbf{x}_j, \mathbf{D}, \sigma^2).$$

Addition of the estimation of the noise variance in the M-step

$$(\sigma^2)^{(k)} = \frac{1}{NK} \sum_{j=1}^K \sum_{i=1}^P w_{ji}^{(k)} \|\mathbf{y}_j - \mathbf{D}_i^{(k)} \mathbf{x}_{ji}^{(k)}\|_2^2.$$

Performance analysis

- **Synthetic signals:** 500 16-dimensional signals are generated as the noisy combination of 4 atoms of one single basis taken from a set of 6 bases. The amplitude of the nonzero coefficients are drawn from a zero-mean Gaussian distribution with variance $\sigma_a^2 = 16$. Finally, the dictionary is initialized from the original dictionary as:

$$\forall i \in \{1, \dots, P\} \quad \mathbf{D}_i^{(0)} = \mathbf{D}_i \mathbf{M}^T$$

where $\mathbf{M} = GS(\mathbf{I}_6 + N(a))$, GS represents the Gram-Schmidt orthogonalization process and $N(a)$ represents a 16×16 -matrix whose elements are *i.i.d.* realizations of a uniform law on $[-a, a]$.

- **Performance measurements:** We evaluate and compare the performance of three algorithms:

- “Sezer”: learning algorithm proposed in [3],
- “EM”: proposed algorithm where the noise variance estimation is also implemented,
- “EM thresholded”: similar to “EM” where the E-step is approximated by the thresholded decision:

$$c_j^{(k)} = \arg \max_{c_j} p(c_j = i | \mathbf{y}_j, \mathbf{x}_j^{(k-1)}, \mathbf{D}^{(k-1)}).$$

Performance is evaluated via the missed-detection rate versus the signal-to-noise ratio (SNR).

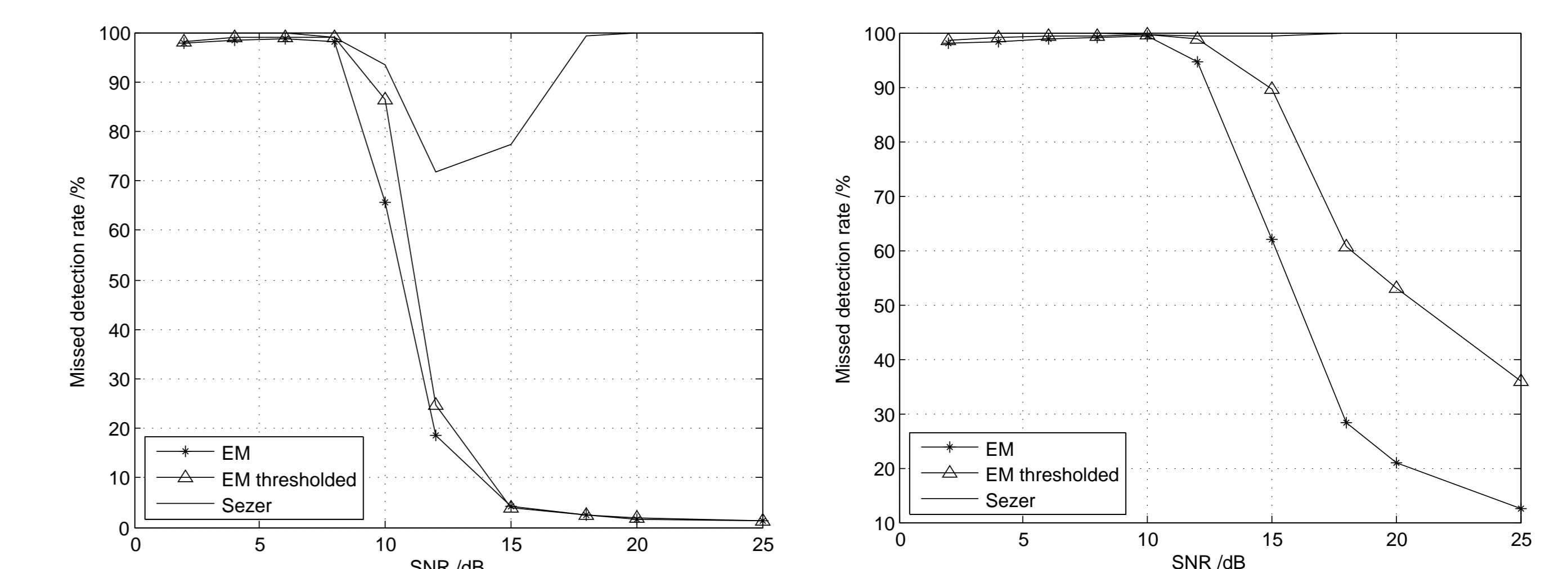


FIGURE 1: Comparison between Sezer's, EM and EM-thresholded algorithms for different dictionary initializations (left: $a=0.3$, right: $a=0.4$)

References

- [1] S. Mallat and F. Falzon, “Analysis of low bit rate image transform coding,” *IEEE Trans. On Signal Processing*, vol. 46, no. 4, pp. 1027–1042, April 1998.
- [2] S. Lesage, R. Gribonval, F. Bimbot, and L. Benaroya, “Learning unions of orthonormal bases with thresholded singular value decomposition,” in *Proc. IEEE Int'l Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 18-23 March 2005, vol. 5, pp. v293–v296.
- [3] O. G. Sezer, O. Harmanci, and O. G. Guleryuz, “Sparse orthonormal transforms for image compression,” in *Proc. IEEE Int'l Conference on Image Processing (ICIP)*, San Diego, CA., October 2008.
- [4] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.