MMSE Approximation for Denoising Using Several Sparse Representations

Irad Yavneh and Michael Elad
The Computer Science Department – The Technion, Israel Institute of Technology
Haifa 32000 Israel, E-mail: [irad,elad]@cs.technion.ac.il

Keywords: Sparse representation, Denoising, Greedy Algorithm, MMSE.

Abstract

Cleaning of noise from signals is a classical and long-studied problem in signal processing. For signals that admit sparse representations over a known dictionary, MAP-based denoising seeks the sparsest representation that synthesizes a signal close to the corrupted one. While this task is NP-hard, it can usually be approximated quite well by a greedy method, such as the Orthogonal Matching Pursuit (OMP). In this work we consider a Minimum-Mean-Squared-Error (MMSE) denoising algorithm, superior to the above MAP approach. We show that this estimator amounts to a weighted averaging of many sparse representation solutions. As its deployment is also NP-hard, we propose a practical randomized version of the OMP algorithm for generating such a group of representations. Simulations of the proposed algorithm are provided and its superiority over plain OMP is demonstrated.

1 Introduction

Cleaning of additive noise from signals is a classical and long-studied problem in signal processing. This task, known as denoising, considers a given measurement signal $\mathbf{y} \in \mathbb{R}^n$ obtained from the clean signal $\mathbf{x} \in \mathbb{R}^n$ by a contamination of the form $\mathbf{y} = \mathbf{x} + \mathbf{v}$. In this paper we shall restrict our discussion to noise vectors $\mathbf{v} \in \mathbb{R}^n$, assumed to be zero mean i.i.d. Gaussian, with entries drawn at random from the normal distribution $\mathcal{N}(0, \sigma)$. The denoising goal is to recover $\mathbf{x}$ from $\mathbf{y}$.

In order to design an effective denoising algorithm, we must have at our disposal two pieces of information: The first is a knowledge about the noise characteristics, as described above. Along with it, we must also introduce some knowledge about the class of signals that $\mathbf{x}$ belongs to. Only with these two can one design a scheme to decompose $\mathbf{y}$ into its original components, $\mathbf{x}$ and $\mathbf{v}$. There are numerous algorithms for denoising, as there are numerous ways to describe the a-priori knowledge about the signal characteristics. Among these, a recently emerging model for signals that attracts much attention is one that relies on sparse and redundant representations [Mallat (1998), Bruckstein-Donoho-Elad (2008)]. This model will be the focus of the work presented here.

A signal $\mathbf{x}$ is said to have a sparse representation over a known dictionary $\mathbf{D} \in \mathbb{R}^{n \times m}$ (we typically assume that $m > n$, implying that this is a redundant representation), if there exists a sparse vector $\alpha \in \mathbb{R}^m$ such that $\mathbf{x} = \mathbf{D}\alpha$. The vector α is said to be the representation of $\mathbf{x}$. Referring to the columns of $\mathbf{D}$ as prototype signals or *atoms*, α describes how to construct $\mathbf{x}$ from a few such atoms by a linear combination. The representation is sparse – the number of non-zeros in it, $k = \|\alpha\|_0$, is expected to be much smaller than n. Also, this is a redundant representation – it is longer than the original signal it represents. In this paper we consider the family of signals that admit sparse representations over a known dictionary $\mathbf{D}$ and discuss ways to denoise them.

Assuming that $\mathbf{x} = \mathbf{D}\alpha$ with a sparse representation α, how can we denoise a corrupted version of it, $\mathbf{y}$? A commonly used denoising technique is to seek the sparsest representation that synthesizes a signal close enough to the corrupted one [Bruckstein-Donoho-Elad (2008)]. Put formally, one way to define our task is given by

$$\hat{\alpha} = \arg\min_{\alpha} \; \|\alpha\|_0 + \lambda \|\mathbf{y} - \mathbf{D}\alpha\|_2^2. \tag{1}$$

The first penalty directs the minimization task towards the sparsest possible representation, exploiting our a-priori knowledge about the formation of the signal. The second penalty manifests our knowledge about the noise being white and Gaussian. This overall expression is inversely proportional to the posterior probability, $p(\alpha|\mathbf{y})$, and as such, its minimization forms the Maximum A-posteriori Probability

(MAP) estimate. The parameter λ should be chosen based on σ and the fine details that model how the representations are generated. Once $\hat{\alpha}$ is found, the denoising result is obtained by $\hat{\mathbf{x}} = \mathbf{D}\hat{\alpha}$.

The problem posed in Equation (1) is too complex in general, requiring a combinatorial search that explores all possible sparse supports [Natarajan (1995)]. Approximation methods are therefore often employed, with the understanding that their result may deviate from the true solution. One such approximation technique is the Orthogonal Matching Pursuit (OMP), a greedy algorithm that accumulates one atom at a time in forming $\hat{\alpha}$, aiming at each step to minimize the representation error $\|\mathbf{y} - \mathbf{D}\alpha\|_2^2$ [Mallat-Zhang (1993), Bruckstein-Donoho-Elad (2008)]. When this error falls below some predetermined threshold, or when the number of atoms reaches a destination value, this process stops. While crude, this technique works very fast and can guarantee near-optimal results in some cases.

How good is the denoising obtained by the above approach? Past work mostly concentrated on the accuracy with which one can approximate the true representation (rather than the signal itself), adopting a worst-case point of view. The only work that targets the theoretical question of denoising performance head-on is reported in [Fletcher-Rangan-Goyal-Ramchandran (2006)], providing asymptotic assessments of the denoising performance for very low and very high noise powers, assuming that the original combinatorial problem can be solved exactly.

In this paper we consider the following question: Suppose we are served with a group of competing sparse representations, each claiming to explain the signal differently. Can those be fused somehow to lead to a better result? Surprisingly, the answer to this question is positive. In this paper we propose a practical way to generate a set of sparse representations for a given signal by randomizing the OMP algorithm. We demonstrate the gain in using such a set of representations by a plain averaging. Most important of all, we develop analytical expressions for the MAP and the Minimum Mean-Squared-Error (MMSE) estimators for the model discussed and show that while the MAP estimator aims to find and use the sparsest representation, the MMSE estimator fuses a collection of representations to form its result.

This paper is organized as follows. In Section 2 we build a case for the use of several sparse representations, leaning on intuition and some preliminary experiments that suggest that this idea is worth a closer look. Section 3 contains the analytic part of this paper, which develops the MAP and the MMSE exact estimators and their expected errors, showing how they relate to the use of several representations. We conclude in Section 4 by highlighting the main contribution of this paper, and drawing attention to important open questions to which our analysis points. We should note that this paper is a (much) shorter version of [Elad-Yavneh (2008)].

2 A Mixture of Representations

2.1 The RandOMP – Creation of a Set of Representations

Here is a clear definition of our goal at the moment: Given a dictionary $\mathbf{D}$ and a signal $\mathbf{y}$, we aim to find a group of sparse representations α_i, such that each satisfies $\|\mathbf{D}\alpha_i - \mathbf{y}\|_2 \leq T$, and all aim to be as sparse as possible yet different from each other. Alternatively, we may desire to find this set such that each has the same pre-specified number of non-zeros, k, and all aim to get residuals, $\|\mathbf{D}\alpha_i - \mathbf{y}\|_2$, that are as low as possible. We shall work in this section with the former option, since it is more relevant to denoising in cases when the noise power is fixed and known, as in the case studied here.

Figure 1 presents the OMP algorithm with a stopping rule that depends on the residual energy [Mallat-Zhang (1993)]. At each iteration, the set $\{\epsilon(j)\}_{j=1}^m$ is computed, whose jth term indicates the error that would remain if atom j is added to the current solution. The atom chosen is the one yielding the smallest error. Note that if there are several candidate atoms that show a relatively small residual energy, the smallest one is chosen regardless of the proximity of the others to it. This brings us naturally to the randomization approach we intend to apply.

In order to use this algorithm to generate a set of (probably) distinct sparse representations, all that we need to do is to randomize the choice of the next atom to be added. For example, rather than choose the atom that minimizes $\epsilon(j)$, we can choose it at random with a probability inversely proportional to these error values, or proportional to $|\mathbf{d}_j^T\mathbf{r}^{k-1}|^2/\|\mathbf{d}_j\|_2^2$ (since $\epsilon(j) = \|\mathbf{r}^{k-1}\|_2^2 - |\mathbf{d}_j^T\mathbf{r}^{k-1}|^2/\|\mathbf{d}_j\|_2^2$). For reasons to be explained in detail in the next section, the specific way we choose to draw the next atom is with probability linearly proportional to $\exp\{\frac{c^2}{2\sigma^2} \cdot |\mathbf{d}_j^T\mathbf{r}^{k-1}|^2/\|\mathbf{d}_j\|_2^2\}$, with $c^2 = \sigma_x^2/(\sigma_x^2 + \sigma^2)$. Here σ_x is the variance of the non-zero entries of the representation of the original signal. Figure 2 presents this algorithm, concentrating on the part that is different from Figure 1.

Task: Approximate the solution of $\min_{\alpha} \ \|\alpha\|_0$ s.t. $\|\mathbf{D}\alpha - \mathbf{y}\|_2 \leq T$.

Parameters: The matrix $\mathbf{D}$, the signal $\mathbf{y}$, and the error threshold T.

Initialization: Initialize $k = 0$, and set

- The initial solution $\alpha^0 = 0$.
- The initial residual $\mathbf{r}^0 = \mathbf{y} - \mathbf{D}\alpha^0 = \mathbf{y}$.
- The initial solution support $\mathcal{S}^0 = Support\{\alpha^0\} = \emptyset$.

Main Iteration: Increment k by 1 and perform the following steps:

- **Sweep:** Compute the errors $\epsilon(j) = \min_{z_j} \|\mathbf{d}_j z_j - \mathbf{r}^{k-1}\|_2^2$ for all j using the optimal choice $z_j^* = \mathbf{d}_j^T \mathbf{r}^{k-1} / \|\mathbf{d}_j\|_2^2$.
- **Update Support:** Find j_0–the minimizer of $\epsilon(j)$, and update the support, $\mathcal{S}^k = \mathcal{S}^{k-1} \cup \{j_0\}$.
- **Update Solution:** Compute α^k, the minimizer of $\|\mathbf{D}\alpha - \mathbf{y}\|_2^2$ subject to $Support\{\alpha\} = \mathcal{S}^k$.
- **Update Residual:** Compute $\mathbf{r}^k = \mathbf{y} - \mathbf{D}\alpha^k$.
- **Stopping Rule:** If $\|\mathbf{r}^k\|_2 < T$, stop. Otherwise, apply another iteration.

Output: The proposed solution is α^k obtained after k iterations.

Figure 1: The OMP – a greedy algorithm.

2.2 Experimental Study

By running this algorithm J_0 times, this randomization leads to J_0 solutions $\{\alpha_i\}_{i=1}^{J_0}$, as desired. Common to all these representations are the facts that (i) their representation error $\|\mathbf{D}\alpha_i - \mathbf{y}\|_2$ is below T due to the stopping rule enforced; and (ii) all of them tend to be relatively sparse due to the greedy nature of this algorithm that aims to decrease the residual energy, giving preference to those atoms that serve this goal better.

$\vdots$

Main Iteration: ...

- **Sweep:** ...
- **Update Support:** Draw j_0 at random with probability proportional to $\exp\{\frac{c^2}{2\sigma^2} \cdot |\mathbf{d}_j^T \mathbf{r}^{k-1}|^2 / \|\mathbf{d}_j\|_2^2\}$, and update the support, $\mathcal{S}^k = \mathcal{S}^{k-1} \cup \{j_0\}$.

$\vdots$

Figure 2: RandOMP – generating random sparse representations.

We demonstrate the behavior of this algorithm by performing the following test. We build a random dictionary $\mathbf{D}$ of size 100×200 by drawing its entries at random from the normal distribution $\mathcal{N}(0, 1)$, and then ℓ_2 normalizing its columns. We then generate a random representation α_0 with $k = 10$ non-zeros chosen at random and with values drawn from $\mathcal{N}(0, \sigma_x)$ with $\sigma_x = 1$. The clean signal is obtained by $\mathbf{x} = \mathbf{D}\alpha$, and its noisy version $\mathbf{y}$ is obtained by adding white Gaussian noise with entries drawn from $\mathcal{N}(0, \sigma)$ with σ in the range $[0, 2]$.

Armed with the dictionary $\mathbf{D}$, the corrupted signal $\mathbf{y}$ and the noise threshold $T = n\sigma^2 = 100$, we can run the plain OMP, and obtain a representation α^{OMP}. The denoising effect obtained can be evaluated by the expression $\|\mathbf{D}\alpha^{OMP} - \mathbf{x}\|_2^2 / \|\mathbf{y} - \mathbf{x}\|_2^2$.

We also run the RandOMP with $J_0 = 40$, obtaining $\{\alpha_j^{RandOMP}\}_{j=1}^{40}$. Instead of trying to pinpoint the representation that performs best among those, we simply compute their average. This experiment is repeated $1,000$ times in order to average over different noise realizations, sharing the same dictionary but generating different signals α, $\mathbf{x}$ and $\mathbf{y}$ using the same parameters ($\sigma_x = 1$ and $k = 10$). Figure 3 presents the denoising performance of the averaging as a function of σ, and as can be seen, our method is better [1] for all the choices of σ.

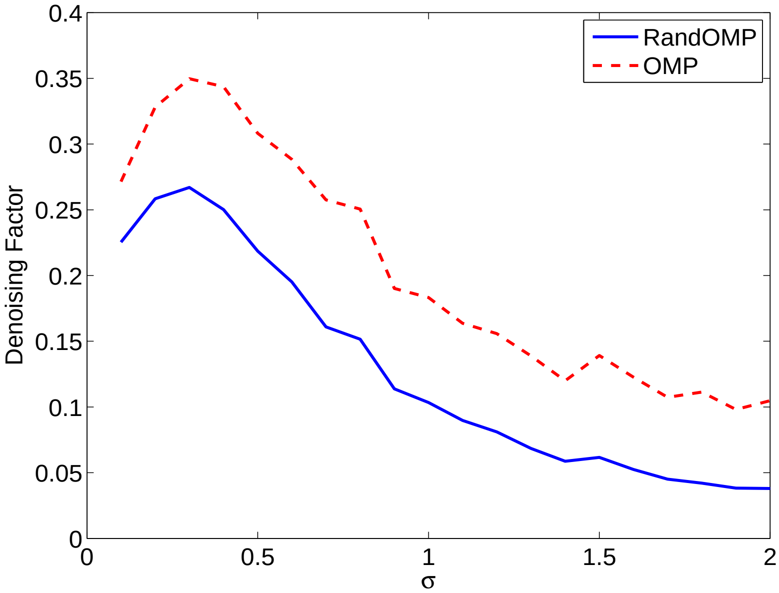

Figure 3: The performance of the OMP and RandOMP algorithm for various noise powers.

We now turn to explain these results, tying the Rand-OMP algorithm and its averaged result to the MMSE estimator.

3 Relation to the MMSE Estimator

We start by modelling the signal source in a complete manner, define the denoising goal in terms of the MSE, and derive several estimators for it. We start with a very general setting of the problem, and then narrow it down to the case discussed above on sparse representations. Our main goal in this section is to show that the MMSE estimator can be written as a weighted averaging of various sparse representations, which explains the results of the previous section. Beyond this, the analysis derives exact expressions for the MSE for various estimators, enabling us to assess analytically their behavior and relative performance, and to explain results that were obtained empirically in Section 2. Towards the end of this section we tie the empirical and the theoretical parts of this work – we again perform simulations and show how the actual denoising results obtained by OMP and RandOMP compare to the analytic expressions developed here.

3.1 Preliminaries

Given a dictionary $\mathbf{D} \in \mathbb{R}^{n \times m}$, let Ω denote the set of all 2^m sub-dictionaries, where a sub-dictionary, $\mathbf{S}$, will interchangeably be considered as a subset of the columns of $\mathbf{D}$ or as a matrix comprised of such columns. We assume that a random signal, $\mathbf{x} \in \mathbb{R}^n$, is selected by the following process. With each sub-dictionary, $\mathbf{S} \in \Omega$, we associate a non-negative probability, $P(\mathbf{S})$, with $\sum_{\mathbf{S} \in \Omega} P(\mathbf{S}) = 1$. Furthermore, with each signal $\mathbf{x}$ in the range of $\mathbf{S}$ (that is, such that there exists a vector $\mathbf{z} \in \mathbb{R}^k$ satisfying $\mathbf{S}\mathbf{z} = \mathbf{x}$,) denoted $\mathbf{x} \in \mathcal{R}(\mathbf{S})$, we associate a conditional PDF, $p(\mathbf{x}|\mathbf{S})$. Then, the clean signal $\mathbf{x}$ is assumed to be

[1] The gain provided by the RandOMP is higher for lower SNR in this experiment. This differs from results to appear later in the paper in Figure 4. An explanation of this gap in the performances is provided there.

generated by first randomly selecting $\mathbf{S}$ according to $P(\mathbf{S})$, and then randomly choosing $\mathbf{x} \in \mathbf{S}$ according to $p(\mathbf{x}|\mathbf{S})$. After the signal is generated, an additive random noise term, $\mathbf{v}$, with PDF $p_v(\mathbf{v})$, is introduced, yielding a noisy signal $\mathbf{y} = \mathbf{x} + \mathbf{v}$.

Note that $P(\mathbf{S})$ can be used to represent a tendency towards sparsity. For example, we can choose $P(\mathbf{S})$ to be a strongly decreasing function of the number of elements in $\mathbf{S}$, or we can choose $P(\mathbf{S})$ to be zero for all $\mathbf{S}$'s except those with a particular (small) number of elements, etc.

Given $\mathbf{y}$, and assuming we know $p_v(\mathbf{v})$, $P(\mathbf{S})$ and $p(\mathbf{x}|\mathbf{S})$, our objective is to find an estimator, $\hat{\mathbf{x}}$, that will be as close as possible to the clean signal $\mathbf{x}$ in some sense. In this work we will mainly strive to minimize the conditional mean square error (MSE),

$$\mathrm{MSE}_y = \mathcal{E}\left(\|\hat{\mathbf{x}} - \mathbf{x}\|^2 \,|\mathbf{y}\right). \tag{2}$$

We write the conditional MSE as the sum

$$\mathrm{MSE}_y = \sum_{\mathbf{S}\in\Omega} \mathrm{MSE}_{S,y} P(\mathbf{S}|\mathbf{y}), \tag{3}$$

with $\mathrm{MSE}_{S,y}$ defined as

$$\mathrm{MSE}_{S,y} = \mathcal{E}\left(\|\hat{\mathbf{x}} - \mathbf{x}\|^2 \,|\mathbf{S}, \mathbf{y}\right). \tag{4}$$

The first factor of the summation in (3) is the MSE subject to a noisy signal $\mathbf{y}$ and a given sub-dictionary $\mathbf{S}$, and the second factor is the probability of $\mathbf{S}$ given a noisy signal $\mathbf{y}$. By Bayes's formula, the latter is given by

$$P(\mathbf{S}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{S})P(\mathbf{S})}{p(\mathbf{y})}, \tag{5}$$

where

$$p(\mathbf{y}|\mathbf{S}) = \int_{\mathbf{x}\in\mathcal{R}(\mathbf{S})} p_v(\mathbf{y} - \mathbf{x}) p(\mathbf{x}|\mathbf{S})\, d\mathbf{x} \tag{6}$$

is the PDF of $\mathbf{y}$ given the sub-dictionary $\mathbf{S}$.

Next, we consider the first factor of the summation in (3), $\mathrm{MSE}_{S,y}$, the MSE for a given $\mathbf{y}$ and sub-dictionary $\mathbf{S}$. Defining $\mathcal{M}_{S,y}(\mathbf{x}) = \mathcal{E}\left(\mathbf{x}\,|\mathbf{S}, \mathbf{y}\right)$, we have

$$\begin{aligned} \mathcal{E}\left(\|\mathbf{x}\|^2 \,|\mathbf{S}, \mathbf{y}\right) &= \mathcal{E}\left(\|\mathcal{M}_{S,y}(\mathbf{x}) + \mathbf{x} - \mathcal{M}_{S,y}(\mathbf{x})\|^2 \,|\mathbf{S}, \mathbf{y}\right) \\ &= \|\mathcal{M}_{S,y}(\mathbf{x})\|^2 + \mathcal{E}\left(\|\mathbf{x} - \mathcal{M}_{S,y}(\mathbf{x})\|^2 \,|\mathbf{S}, \mathbf{y}\right) \\ &= \|\mathcal{M}_{S,y}(\mathbf{x})\|^2 + \mathcal{V}_{S,y}(\mathbf{x}). \end{aligned} \tag{7}$$

This property, along with the linearity of the expectation, can be used to rewrite the first factor of the summation in (3) as follows:

$$\begin{aligned} \mathrm{MSE}_{S,y} = \mathcal{E}\left(\|\hat{\mathbf{x}} - \mathbf{x}\|^2 \,|\mathbf{S}, \mathbf{y}\right) &= \mathcal{E}\left(\|\hat{\mathbf{x}}\|^2 - 2\hat{\mathbf{x}}^T\mathbf{x} + \|\mathbf{x}\|^2 \,|\mathbf{S}, \mathbf{y}\right) \\ &= \|\hat{\mathbf{x}}\|^2 - 2\hat{\mathbf{x}}^T\mathcal{M}_{S,y}(\mathbf{x}) + \|\mathcal{M}_{S,y}(\mathbf{x})\|^2 + \mathcal{V}_{S,y}(\mathbf{x}) \\ &= \|\hat{\mathbf{x}} - \mathcal{M}_{S,y}(\mathbf{x})\|^2 + \mathcal{V}_{S,y}(\mathbf{x}). \end{aligned} \tag{8}$$

Finally, plugging this into (3) we obtain

$$\mathrm{MSE}_y \;=\; \sum_{\mathbf{S}\in\Omega} \left[\|\hat{\mathbf{x}} - \mathcal{M}_{S,y}(\mathbf{x})\|^2 + \mathcal{V}_{S,y}(\mathbf{x})\right] P(\mathbf{S}|\mathbf{y}) = \mathcal{E}\left(\|\hat{\mathbf{x}} - \mathcal{M}_{S,y}(\mathbf{x})\|^2|\mathbf{y}\right) + \mathcal{E}\left(\mathcal{V}_{S,y}(\mathbf{x})|\mathbf{y}\right), \tag{9}$$

with $P(\mathbf{S}|\mathbf{y})$ given by (5).

3.2 Various Estimators – A General Form

By (9), the optimal $\hat{\mathbf{x}}$ that minimizes MSE_y is, not surprisingly, given by

$$\hat{\mathbf{x}}^{MMSE} = \mathcal{E}\left(\mathcal{M}_{S,y}(\mathbf{x})|\mathbf{y}\right), \tag{10}$$

and, plugged to Equation (9), the resulting optimal conditional MSE is given by

$$\mathrm{MSE}_y^{MMSE} = \mathcal{E}\left(\|\mathcal{M}_{S,y}(\mathbf{x}) - \mathcal{E}\left(\mathcal{M}_{S,y}(\mathbf{x})|\mathbf{y}\right)\|^2|\mathbf{y}\right) + \mathcal{E}\left(\mathcal{V}_{S,y}(\mathbf{x})|\mathbf{y}\right). \tag{11}$$

Finally, from (9) and (10) we obtain for an arbitrary estimator $\hat{\mathbf{x}}$ the conditional MSE

$$\mathrm{MSE}_y = \mathrm{MSE}_y^{MMSE} + \|\hat{\mathbf{x}} - \hat{\mathbf{x}}^{MMSE}\|^2. \tag{12}$$

This can be used to determine how much better the optimal estimator does compared to any other estimator.

The MAP estimator is obtained by maximizing the probability of $\mathbf{x}$ given $\mathbf{y}$,

$$\hat{\mathbf{x}}^{MAP} = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) = \arg\max_{\mathbf{x}} \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} = \arg\max_{\mathbf{x}} p_v(\mathbf{y}-\mathbf{x})p(\mathbf{x}), \tag{13}$$

with

$$p(\mathbf{x}) = \sum_{\mathbf{S}\in\Omega:\ \mathbf{x}\in\mathcal{R}(\mathbf{S})} p(\mathbf{x}|\mathbf{S})P(\mathbf{S})\,.$$

At the moment these expressions remain vague, but as we turn to use the specific signal and noise models discussed in Section 3.1.2, these will assume an explicit form.

Suppose that the sub-dictionary $\mathbf{S}$ that was chosen in the generation of $\mathbf{x}$ is revealed to us. Given this information, we clearly minimize MSE_y by setting $\hat{\mathbf{x}} = \mathcal{M}_{S,y}(\mathbf{x})$ for the given $\mathbf{S}$. We call this the oracle estimator. The resulting conditional MSE is evidently given by the last term of (9),

$$\mathrm{MSE}_y^{oracle} = \mathcal{E}\left(\mathcal{V}_{S,y}(\mathbf{x})|\mathbf{y}\right). \tag{14}$$

We shall use this estimator to assess the performance of the various alternatives and see how close we get to this “ideal” performance.

3.3 Back to Our Story – Sparse Representations

Our aim now is to harness the general derivation to the development of a practical algorithm for the sparse representation and white Gaussian noise. Motivated by the sparse-representation paradigm, we concentrate on the case where $P(\mathbf{S})$ depends only on the number of atoms (columns) in $\mathbf{S}$, denoted $|\mathbf{S}|$. We start with the basic case where $P(\mathbf{S})$ vanishes unless $|\mathbf{S}|$ is exactly equal to some particular $0 \le k \le \min(n,m)$, and $\mathbf{S}$ has column rank k. We denote the set of such $\mathbf{S}$'s by Ω_k, and define the uniform distribution

$$P(\mathbf{S}) = \begin{cases} \frac{1}{|\Omega_k|} & \mathbf{S}\in\Omega_k\,, \\ 0 & \text{otherwise.} \end{cases}$$

We assume throughout that the columns of $\mathbf{D}$ are normalized, $\|\mathbf{d}_j\| = 1$, for $j = 1,\ldots,n$. This assumption comes only to simplify the expressions we are about to obtain. Next, we recall that the noise is modelled via a Gaussian distribution with zero mean and variance σ^2, and thus

$$p\left(\mathbf{y}|\mathbf{x}\right) = p_v\left(\mathbf{y}-\mathbf{x}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \cdot \exp\left\{\frac{-\|\mathbf{y}-\mathbf{x}\|^2}{2\sigma^2}\right\}. \tag{15}$$

Similarly, given the sub-dictionary $\mathbf{S}$ from which $\mathbf{x}$ is drawn, the signal $\mathbf{x}$ is assumed to be generated via a Gaussian distribution with mean zero and variance σ_x^2, thus $p(\mathbf{x}|\mathbf{S})$ is given by

$$p\left(\mathbf{x}|\mathbf{S}\right) = \begin{cases} \frac{1}{(2\pi\sigma_x^2)^{k/2}} \cdot \exp\left\{\frac{-\|\mathbf{x}\|^2}{2\sigma_x^2}\right\} & \mathbf{x}\in\mathcal{R}(\mathbf{S}) \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

For convenience, we introduce the notation $c^2 = \sigma_x^2/(\sigma^2+\sigma_x^2)$. Also, we denote the orthogonal projection of any vector $\mathbf{a}$ onto the subspace spanned by the columns of $\mathbf{S}$ by

$$\mathbf{a}_S = \mathbf{S}\left(\mathbf{S}^T\mathbf{S}\right)^{-1}\mathbf{S}^T\mathbf{a}.$$

We now follow the general derivation given above. From Equation (6) we can develop a closed-form expression for $p(\mathbf{y}|\mathbf{S})$. By integration and rearrangement we obtain

$$\begin{aligned} p(\mathbf{y}|\mathbf{S}) &= \int_{\mathbf{x}\in\mathcal{R}(\mathbf{S})} \frac{1}{\left(2\pi\sigma^2\right)^{n/2}\cdot\left(2\pi\sigma_x^2\right)^{k/2}} \cdot \exp\left\{\frac{-\|\mathbf{y}-\mathbf{x}\|^2}{2\sigma^2} + \frac{-\|\mathbf{x}\|^2}{2\sigma_x^2}\right\} d\mathbf{x} \\ &= \frac{(1-c^2)^{k/2}}{|\Omega_k|\left(2\pi\sigma^2\right)^{n/2}} \cdot \exp\left\{\frac{-(1-c^2)\|\mathbf{y}\|^2}{2\sigma^2}\right\} \cdot \exp\left\{\frac{-c^2\|\mathbf{y}-\mathbf{y}_S\|^2}{2\sigma^2}\right\}. \end{aligned} \tag{17}$$

Since the only dependence of $p(\mathbf{y}|\mathbf{S})$ on $\mathbf{S}$ is through the right-most factor, we immediately obtain by (5) the simple formula

$$P(\mathbf{S}|\mathbf{y}) = \frac{\exp\left\{-\frac{c^2\|\mathbf{y}-\mathbf{y}_S\|^2}{2\sigma^2}\right\}}{\sum_{\mathbf{S}'\in\Omega_k}\exp\left\{\frac{c^2\|\mathbf{y}-\mathbf{y}_{S'}\|^2}{2\sigma^2}\right\}}. \tag{18}$$

The denominator here is just a normalization. The numerator implies that, given a noisy signal $\mathbf{y}$, the probability that the clean signal was selected from the subspace $\mathbf{S}$ decays at a Gaussian rate with the distance between $\mathbf{y}$ and $\mathbf{S}$, i.e., $\|\mathbf{y}-\mathbf{y}_S\|$. This result is expected, given the Gaussian noise distribution.

Continuing to follow the general analysis, we compute the conditional mean, $\mathcal{M}_{S,y}(\mathbf{x})$, for which we require the conditional probability

$$\begin{aligned} p(\mathbf{x}|\mathbf{S},\mathbf{y}) &= \frac{p(\mathbf{y}|\mathbf{S},\mathbf{x})\,p(\mathbf{x}|\mathbf{S})}{p(\mathbf{y}|\mathbf{S})} \\ &= \frac{1}{p(\mathbf{y}|\mathbf{S})}\cdot\frac{1}{(2\pi\sigma^2)^{n/2}}\cdot\exp\left\{\frac{-\|\mathbf{y}-\mathbf{x}\|^2}{2\sigma^2}\right\}\cdot\frac{1}{(2\pi\sigma_x^2)^{k/2}}\cdot\exp\left\{\frac{-\|\mathbf{x}\|^2}{2\sigma_x^2}\right\}. \end{aligned} \tag{19}$$

By integration, we then obtain the simple result,

$$\mathcal{M}_{S,y}(\mathbf{x}) = \int_{\mathbf{x}\in\mathcal{R}(\mathbf{S})}\mathbf{x}p(\mathbf{x}|\mathbf{S},\mathbf{y})d\mathbf{x} = c^2\mathbf{y}_S. \tag{20}$$

Now the conditional variance can be computed, yielding

$$\mathcal{V}_{S,y}(\mathbf{x}) = \int_{\mathbf{x}\in\mathcal{R}(\mathbf{S})}\|\mathbf{x}-c^2\mathbf{y}_S\|^2p(\mathbf{x}|\mathbf{S},\mathbf{y})d\mathbf{x} = kc^2\sigma^2, \tag{21}$$

which is independent of $\mathbf{S}$ and $\mathbf{y}$. Thus, the oracle MSE_y in this case is simply

$$\mathrm{MSE}_y^{oracle} = kc^2\sigma^2. \tag{22}$$

The optimal estimator is given by Equation (10),

$$\hat{\mathbf{x}}^{MMSE} = c^2\sum_{\mathbf{S}\in\Omega_k}\mathbf{y}_SP(\mathbf{S}|\mathbf{y}) = \frac{c^2}{\sum_{\mathbf{S}'\in\Omega_k}\exp\left\{-\frac{c^2\|\mathbf{y}-\mathbf{y}_{S'}\|^2}{2\sigma^2}\right\}}\cdot\sum_{\mathbf{S}\in\Omega_k}\exp\left\{\frac{-c^2\|\mathbf{y}-\mathbf{y}_S\|^2}{2\sigma^2}\right\}\mathbf{y}_S, \tag{23}$$

with $P(\mathbf{S}|\mathbf{y})$ taken from (18). This MMSE estimate is a weighted average of the projections of $\mathbf{y}$ onto all the possible sub-spaces $\mathbf{S}\in\Omega_k$, as claimed. The MSE of this estimate is given by

$$\mathrm{MSE}_y^{MMSE} = kc^2\sigma^2 + \sum_{\mathbf{S}\in\Omega_k}\|\hat{\mathbf{x}}^{MMSE} - c^2\mathbf{y}_S\|^2P(\mathbf{S}|\mathbf{y}). \tag{24}$$

The latter can also be written as

$$\mathrm{MSE}_y^{MMSE} = kc^2\sigma^2 - \|\hat{\mathbf{x}}^{MMSE}\|^2 + \sum_{\mathbf{S}\in\Omega_k}\|c^2\mathbf{y}_S\|^2P(\mathbf{S}|\mathbf{y}). \tag{25}$$

We remark that *any* spherically symmetric $p_v(\mathbf{v})$ and $p(\mathbf{x}|\mathbf{S})$ produce a conditional mean, $\mathcal{M}_{S,y}(\mathbf{x})$, that is equal to $\mathbf{y}_S$ times some scalar coefficient. The choice of Gaussian distributions makes the result in (20) particularly simple in that the coefficient, c^2, is independent of $\mathbf{y}$ and $\mathbf{S}$.

Next, we consider the Maximum a Posterior (MAP) estimator, using (13). For simplicity, we shall neglect the fact that some $\mathbf{x}$'s may lie on intersections of two or more sub-dictionaries in Ω_k, and therefore their PDF is higher according to our model. This is a set of measure zero, and it therefore does not influence the MMSE solution, but it does influence somewhat the MAP solution for $\mathbf{y}$'s that are close to such $\mathbf{x}$'s. We can overcome this technical difficulty by modifying our model slightly so as to eliminate the favoring of such $\mathbf{x}$'s. Noting that $P(\mathbf{S})$ is a constant for all $\mathbf{S}\in\Omega_k$, we obtain from (13)

$$\hat{\mathbf{x}}^{MAP} = \arg\max_{\mathbf{x}\in\mathcal{R}(\Omega_k)}\exp\left\{\frac{-\|\mathbf{y}-\mathbf{x}\|^2}{2\sigma^2}\right\}\cdot\exp\left\{\frac{-\|\mathbf{x}\|^2}{2\sigma_x^2}\right\}, \tag{26}$$

where $\mathcal{R}(\Omega_k)$ is defined as the union of the ranges of all $\mathbf{S} \in \Omega_k$. Multiplying through by $\exp(2c^2\sigma^2)$, we find that the maximum is obtained by minimizing $c^2\|\mathbf{y} - \mathbf{x}\|^2 + (1 - c^2)\|\mathbf{x}\|^2$, subject to the constraint that $\mathbf{x}$ belongs to some $\mathbf{S} \in \Omega_k$. The resulting estimator is readily found to be given by

$$\hat{\mathbf{x}}^{MAP} = c^2 \mathbf{y}_{S_{MAP}} , \tag{27}$$

where $\mathbf{S}_{MAP}$ is the sub-space $\mathbf{S} \in \Omega_k$ which is closest to $\mathbf{y}$, i.e., for which $\|\mathbf{y} - \mathbf{y}_S\|^2$ is the smallest. The resulting MSE_y is given by substituting $\hat{\mathbf{x}}^{MAP}$ for $\hat{\mathbf{x}}$ in (12).

Note that in all the estimators we derive, the oracle, the MMSE, and the MAP, there is a factor of c^2 that performs a shrinking of the estimate. For the model of $\mathbf{x}$ chosen, this is a mandatory step that was omitted in Section 2.

3.4 Combining It All

It is now time to combine the theoretical analysis of section and the estimators we tested in Section 2, and we achieve this by a controlled experiment. We start by building a random dictionary of size 20×30 with ℓ^2-normalized columns. We generate signals following the model described above, by randomly choosing a support with $k = 3$ columns, orthogonalizing the chosen columns, and multiplying them by a random i.i.d. vector with entries drawn from $N(0, 1)$ (i.e. $\sigma_x = 1$). We add noise to these signals with σ in the range $[0, 2]$ and evaluate the following values:

1. **Empirical Oracle** estimation and the MSE it induces. This estimator is simply the projection of $\mathbf{y}$ on the correct support, followed by a multiplication by c^2, as described in Equation (20) .

2. **Theoretical Oracle** estimation error, as given in Equation (22).

3. **Empirical MMSE** estimation and its MSE. We use the formula in Equation (23) in order to compute the estimation, and then assess its error empirically. Note that in applying this formula we gather all the $\binom{30}{k}$ possible supports, compute the projection of $\mathbf{y}$ onto them, and weight them according to the formula. This explains why in the experiment reported here we have restricted the sizes involved.

4. **Theoretical MMSE** estimation error, using Equation (25) directly.

5. **Empirical MAP** estimation and its MSE. We use the analytic solution to (26) as described above, by sweeping through all the possible supports, and searching the one with the smallest projection error. This gives us the MAP estimation, and its error is evaluated empirically.

6. **Theoretical MAP** estimation error, as given in Equation (12), when plugging in the MAP estimation.

7. **OMP** estimation and its MSE. The OMP is the same as described in Section 2, but the stopping rule is based on the knowledge of k, rather than on representation error. Following the MAP analysis done in Section 3, the result is multiplied by c^2 as well.

8. **Averaged RandOMP** estimation and its MSE. The algorithm generates $J_0 = 100$ representations and averages them. As in the OMP, the stopping rule for those is the number of atoms k, and the result is also multiplied by c^2.

The above process is averaged over $1,000$ signal generations, and the resulting values are shown in Figures 4 for $k = 3$. First we draw attention to several general observations. As expected, we see in all these graphs that there is a good alignment between the theoretical and the empirical evaluation of the MSE for the oracle, the MMSE, and the MAP estimators. In fact, since the analysis is exact for this experiment, the differences are only due to the finite number of tests per σ. We also see that the denoising performance weakens as k grows. This is different from the behavior observed in Figure 3, and the reason for this is the lack of the shrinkage (multiplication by c^2) force in the early tests. A third and intriguing observation that we will not explore here is the fact that there appears to be a critical input noise power ($\sigma \approx 0.4$) for which the MAP and the MMSE estimators (and their approximations) give their worst denoising performance, as exhibited by the hump in all the MMSE/MAP cases.

The OMP algorithm is an attempt to approximate the MAP estimation, replacing the need for sweeping through all the possible supports by a greedy detection of the involved atoms. As such, we expect it to be competitive and close to the MAP results we get (either analytically or empirically). In fact, for $k = 1$ it aligns perfectly with the empirical MAP, since both are going through the same computational stages. As k grows, there are some differences between the empirical MAP and the OMP, especially for low noise, but for the cases studied here these differences are relatively small.

Just as OMP is an attempt to approximate the MAP estimation, the RandOMP averaging is approximating the MMSE estimator, thereby yielding much better denoising than OMP. The core idea is to replace the summation over all possible supports with a much smaller selected group of representations that are sampled from the distribution governed by the weights in Equation (23). Indeed, the representations chosen by RandOMP are those that correspond to large weights, since they are built in a way that leads to small projection error $\|\mathbf{y} - \mathbf{y}_P\|^2$ for the k atoms chosen. Since the sampling already mimics approximately the required distribution, all that remains is a simple averaging, as indeed we do in practice. What is required is to tune the sampling to be faithful, and for that we revisit the case of $k = 1$.

Considering the case of $k = 1$, we see from Equation (23) that an atom should be chosen as a candidate representation with a probability proportional to $\exp\{-c^2\|\mathbf{y} - \mathbf{y}_P\|^2/2\sigma^2\}$. This in turn implies that this probability is also proportional to[2] $\exp\{c^2|\mathbf{y}^T\mathbf{d}_i|^2/2\sigma^2\}$. Thus, RandOMP as described in Section 2 is with perfect agreement with this probability. We see that RandOMP remains close to the empirical MMSE for $k = 3$, implying that while our sampling strategy is not perfect, it is fair enough. Further investigation is required to better sample the representations in order to get closer to the MSE estimate.

Finally, we note an additional advantage of RandOMP: the MMSE estimator varies continuously with y, whereas the MAP estimator does not, possibly leading to artifacts.

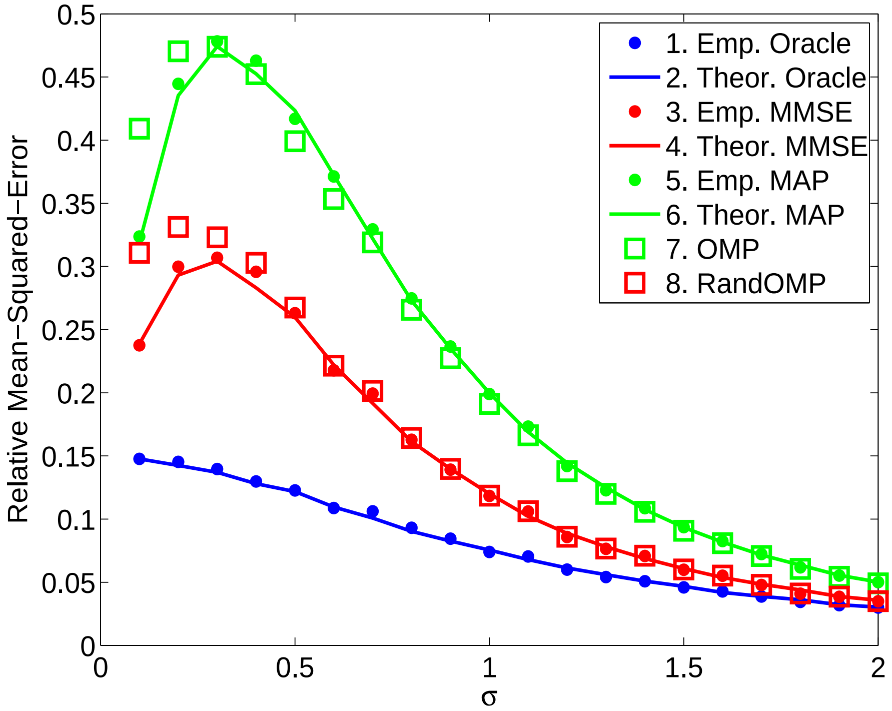

Figure 4: Empirical and theoretical evaluations of the MSE as a function of the input noise for $k = 3$.

3.5 Summarizing This Section

Under the assumptions of this section, we obtain simple explicit expressions for the optimal (MMSE) estimator and its resulting MSE_y. The optimal estimator turns out to be a weighted average of the orthogonal projections of the noisy signal on the feasible subspaces, multiplied by a "shrinkage factor"

[2] Since the columns of the dictionary are normalized, the projection is given by $\mathbf{y}_P = (\mathbf{y}^T\mathbf{d_i}) \cdot \mathbf{d}_i$. Thus, $\|\mathbf{y} - \mathbf{y}_P\|^2 = \|\mathbf{y}\|^2 - (\mathbf{y}^T\mathbf{d}_i)^2$. The term $\exp\{-c^2\|\mathbf{y}\|^2\}$ is therefore a constant that cancels-out in the normalization.

c^2, which tends to zero when the noise variance, σ^2, is large compared to the signal variance, σ_x^2, and to 1 when the opposite is true. The weights in the weighted average depend on the distances between $\mathbf{y}$ and the subspaces, favoring short distances of course, especially when $c^2\|\mathbf{y}\|^2/\sigma^2$ is large.

While the expressions obtained are indeed simple, they involve either an intolerable summations over $\binom{m}{k}$ (for the MMSE estimate), or searching over this amount of sub-spaces (for the MAP). Thus, these formulas are impractical for a direct use. In that sense, one should consider the RandOMP approach in Section 2 as a sampler from this huge set of subspaces over which we average. Roughly speaking, since the RandOMP algorithm tends to find near-by sub-spaces that lead to sparse representations, it gives priority to elements in the summation in Equation (23) that are assigned higher weights. We see experimentally that RandOMP samples well from the representations, judging by the proximity of its results to the MMSE error (both empirical and theoretical).

The results of this section can easily be extended to the case where we allow a range of values of k with given probabilities. That is, we can extend these results for the case where

$$P(\mathbf{S}) = f(|\mathbf{S}|), \tag{28}$$

for general non-negative functions f.

4 Summary and Conclusions

The Orthogonal Matching Pursuit is a simple and fast algorithm for approximating the sparse representation for a given signal. It can be used for denoising of signals, as a way to approximate the MAP estimation. In this work we have shown that by running this algorithm several times in a slightly modified version that randomizes its outcome, one can obtain a collection of competing representations, and those can be averaged to lead to far better denoising performance. This work starts by showing how to obtain a set of such representations to merge, how to combine them wisely, and what kind of results to expect. The analytic part of this paper explains this averaging as a way to approximate the MMSE estimate as a sampler of the summation required. Future work on this topic should consider better sampling strategies for better approximation of the MMSE result, an analytical and numerical study of the required number of samples, an assessment of the robustness of this approach with respect to non-Gaussian distribution of signals and limited accuracy in determining their variance, and exploration of special cases for which practical deterministic algorithms are within reach.

References

S. Mallat, *A Wavelet Tour of Signal Processing*, Academic-Press, 1998.

A.M. Bruckstein, D.L. Donoho, and M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, to appear in *SIAM Review.*

B.K. Natarajan, Sparse approximate solutions to linear systems, *SIAM Journal on Computing*, 24:227–234, 1995.

S. Mallat and Z. Zhang, Matching pursuits with time-frequency dictionaries, *IEEE Trans. Signal Processing*, 41(12):3397–3415, 1993.

A.K. Fletcher, S. Rangan, V.K. Goyal, and K. Ramchandran, Denoising by sparse approximation: error bounds based on rate-distortion theory, *EURASIP Journal on Applied Signal Processing*, Paper No. 26318, 2006.

M. Elad and I. Yavneh, A weighted average of sparse representations is better than the sparsest one alone, Submitted to *IEEE Trans. on Information Theory.*