Dictionary Preconditioning for Greedy Algorithms

Karin Schnass* and Pierre Vandergheynst
Signal Processing Institute
Swiss Federal Institute of Technology
Lausanne, Switzerland
{karin.schnass,pierre.vandergheynst}@epfl.ch
EPFL-STI-ITS-LTS2
CH-1015 Lausanne
Tel: +41 21 693 2657
Fax: +41 21 693 7600
EDICS: SPC-CODC

Abstract

This article introduces the concept of *sensing dictionaries*. It presents an alteration of greedy algorithms like thresholding or (Orthogonal) Matching Pursuit which improves their performance in finding sparse signal representations in redundant dictionaries while maintaining the same complexity. These algorithms can be split into a sensing and a reconstruction step, and the former will fail to identify correct atoms if the cumulative coherence of the dictionary is too high. We thus modify the sensing step by introducing a special sensing dictionary. The correct selection of components is then determined by the *cross cumulative coherence* which can be considerably lower than the cumulative coherence. We characterise the optimal sensing matrix and develop a constructive method to approximate it. Finally we compare the performance of thresholding and OMP using the original and modified algorithms.

1 Introduction

In the last years, constructing sparse signal approximations by means of redundant dictionaries has received a lot of attention, see [10, 2, 4, 3] and the references therein for a thorough introduction. In short the reason for this interest is that a sparse signal representation effectively reduces the dimensionality of the signal and thus makes it easier to store or manipulate. The use of redundant dictionaries is then simply a consequence of the fact that the existence of a sparse signal representation becomes more likely as the number of building blocks or atoms in the dictionary increases. Before we can illustrate the topic further by

stating two of the typically investigated problems, we will need to introduce some vocabulary. We will be working with signals $y \in \mathbb{R}^d$. A dictionary $\mathbf{\Phi}$ is assumed to be represented by a $d \times N$ matrix, with $d \ll N$, whose columns are the atoms φ_i, $\|\varphi_i\|_2 = 1$:

$$\mathbf{\Phi} = [\varphi_1 \dots \varphi_N].$$

The ratio $R = N/d$ is called redundancy. A signal is said to have a K-sparse representation in the dictionary $\mathbf{\Phi}$ if there exists a set I with $|I| = K$ such that we can write

$$y = \sum_{i \in I} x_i \varphi_i = \mathbf{\Phi}_I x.$$

With a slight abuse of language we will call both the set I and the atoms with indices in I the support of y and write $\mathbf{\Phi}_I$ for the $d \times K$ matrix of all the atoms in the support. The complement of the support will be denoted by $\overline{I} = \{1 \dots N\}/I$.

Now, having all definitions in place, the first problem, concerned with finding sparse signal approximations, can be more accurately stated as:

Problem 1. *Given a signal y, find its best K-sparse approximation in the dictionary $\mathbf{\Phi}$, i.e.*

$$\min_{I,x} \|y - \mathbf{\Phi}_I x\|_2 \ \textit{s.t.}\ |I| = K.$$

Or the dual problem given y find the sparsest ε-approximation, i.e.

$$\min_{I} |I| \ \textit{s.t.}\ \min_{x} \|y - \mathbf{\Phi}_I x\|_2 \leq \varepsilon.$$

Of course for any signal and dictionary there always exist solutions to the above problems. However, in order to justify the use of the term sparse, we obviously need to have a dictionary in which the signal has a representation where both ε and K are small, i.e. $K \ll d$. This leads to the next question:

Problem 2. *Given a class of signals Y, find a dictionary $\mathbf{\Phi}$ such that all signals $y \in Y$ will have a good sparse approximation in $\mathbf{\Phi}$.*

Without any further assumption on the signal or the dictionary, finding the solution to the first problem is combinatorial. Thus one would have to try the orthogonal projection of the signal on all possible K-sparse supports. To circumvent this problem people started imposing restrictions on the dictionary and/or the coefficients x. By now there exists detailed theory describing under which assumptions suboptimal algorithms like thresholding, (Orthogonal) Matching Pursuit (OMP), or Basis Pursuit (BP), can be proven to recover the true support, see for instance [10, 5, 1]. The property at the base of most theorems for greedy algorithms is slow growth of the cumulative coherence also called Babel function $\mu_1(K, \mathbf{\Phi})$ of the dictionary, which is defined as:

$$\mu_1(K, \mathbf{\Phi}) = \max_i \max_{|J|=K, i \notin J} \sum_{j \in J} |\langle \varphi_i, \varphi_j \rangle|.$$

It gives an indication of how close/far the dictionary is to/from an orthonormal basis. For compactness reasons we will omit the reference to dictionary, i.e. write $\mu_1(K)$, whenever it is clear which dictionary is meant and write μ for the coherence, i.e. $\mu := \mu_1(1)$. Using this definition a typical result for thresholding, cp. [6], and OMP, cp. [10], reads as:

Theorem 1. *If we have a signal exactly K-sparse in $\mathbf{\Phi}$, i.e. $y = \sum_{i \in I} x_i \varphi_i$ and $|I| = K$, then thresholding is able to recover a component φ_i of the true support if*

$$\frac{|x_i|}{\|x\|_\infty} > \mu_1(K) + \mu_1(K-1). \tag{1}$$

OMP is able to recover all components of the true support I if the exact recovery coefficient is smaller than 1, i.e.

$$\|\mathbf{\Phi}_I^\dagger \mathbf{\Phi}_{\overline{I}}\|_{1,1} < 1,$$

where $\mathbf{\Phi}_I^\dagger$ denotes the Moore-Penrose pseudo-inverse. The above condition is always satisfied if

$$\mu_1(K) + \mu_1(K-1) < 1.$$

One deduction from the theorem is that it is desirable to have a dictionary where the cumulative coherence is growing slowly. Dictionaries having minimal coherence μ are called *Grassmannian frames* and are quite well studied, see [9] and references therein, but the next step of trying to minimise the cumulative coherence seems novel. However we can give a lower bound on the cumulative coherence based on results about Grassmannian frames. The following theorem is an extension of Theorem 2.3 in [9].

Theorem 2. *Let $\mathbf{\Phi}$ be a dictionary of N atoms in dimension d. If $K^2 < N - 1$ then*

$$\mu_1(K) \geq K \cdot \sqrt{\frac{N-d}{d(N-1)}}. \tag{2}$$

Equality holds if and only if the dictionary is an equiangular unit norm tight frame.

Since the proof of the theorem is quite technical and not necessary for further developments it is relegated to the appendix, awaiting inspection by the

genuinely interested there. What should be noted though is that *optimal Grassmannian frames* that meet the lower bound for the coherence, i.e.

$$\mu(\mathbf{\Phi}) = \sqrt{\frac{N-d}{d(N-1)}}$$

simultaneously meet the lower bound for the cumulative coherence $\mu_1(K)$ for all K with $K^2 < N-1$.

On the other hand while a dictionary minimising the cumulative coherence might be interesting for communication applications, it will not be ideal for approximation of a specific class of signals, like for instance EEGs or music. For these purposes learned dictionaries are by definition more suited to the task, see [7]. However these learned dictionaries will not show the desired incoherence properties, that enable us to find the approximation with suboptimal algorithms in the same degree as optimal Grassmannian frames. Assume that we have a dictionary that represents a signal class well but is unfortunately so coherent that already $\mu_1(2) + \mu_1(1) > 1$, meaning that we cannot guarantee for OMP to find even a superposition of only two atoms. Thus in order to find good approximations we would have to use a more complex algorithm. Alternatively we could circumvent the problem by trying to find a new dictionary that still represents the class well but retains small minimal cumulative coherence, an interesting research direction in itself.
However in this paper we will introduce the concept of sensing dictionaries and present a small alteration of the suboptimal algorithms such that they can perform well for dictionaries with high cumulative coherence. In Section 2, we will first explain how to separate the thresholding algorithm into a sensing and a reconstruction part. We will then show that sensing with a different dictionary can lower the cumulative cross-coherence and yield better recovery results. Motivated by structural properties of optimal Grassmannian frames we propose an iterative algorithm to construct a sensing dictionary/matrix giving lower cross-coherence. After analysing its convergence properties theoretically we use it to calculate sensing matrices for various dictionaries and compare the performance of thresholding with and without sensing dictionaries in practice. In Section 3 we will introduce sensing dictionaries as well for (O)MP and from a worst case performance analysis derive a characterisation of the ideal sensing dictionary. We will then again do some numerical simulations of how OMP performs with or without sensing matrices using the sensing dictionaries obtained with the algorithm developed in Section 2. In Section 4 we will discuss the theoretical and numerical limitations of the schemes so far, as well as possible extensions.

2 Sensing Dictionaries for Thresholding

As mentioned above thresholding can be formally decomposed into sensing steps, where we try to identify correct atoms of the support, and reconstruction steps.

Sensing:	find I s.t. $\forall i \in I$, $\forall k \notin I$, $\quad |\langle \varphi_i, y \rangle| \geq |\langle \varphi_k, y \rangle|$
Reconstruction:	$a = \mathbf{\Phi}_I \mathbf{\Phi}_I^\dagger y$

$\mathbf{\Phi}_I^\dagger$ again denotes the Moore-Penrose pseudo inverse.
If the dictionary is too coherent the sensing part will fail to identify correct atoms. Our idea is to change the sensing part and instead of sensing with the dictionary, use a different sensing matrix $\mathbf{\Psi}$ that allows to identify more correct components. This sensing matrix will have as columns the same number of sensing atoms as the original dictionary had atoms, so that we have a one to one correspondence between the sensing and the original atoms. If we denote the sensing atom in $\mathbf{\Psi}$ that corresponds to the atom φ_i in the original dictionary with ψ_i schematically the new algorithm looks like:

Sensing new:	find I s.t. $\forall i \in I$, $\forall k \notin I$, $\quad |\langle \psi_i, y \rangle| \geq |\langle \psi_k, y \rangle|$
Reconstruction:	$a = \mathbf{\Phi}_I \mathbf{\Phi}_I^\dagger y$

This approach can be easily motivated on the following example. Assume for instance that the dictionary $\mathbf{\Phi}$ is a deformed version of a dictionary $\mathbf{\Gamma}$ with low coherence, like an optimal Grassmannian frame or even more simple an orthogonal basis, meaning $\mathbf{\Phi} = A\mathbf{\Gamma}$ where A is an invertible matrix with inverse $A^{-1} = B$. For any K-sparse signal $y = \mathbf{\Phi}x$ by applying the matrix B we can construct a new signal $z = By = B\mathbf{\Phi}x = \mathbf{\Gamma}x$. To find the sparse support I we could equivalently use the original signal and dictionary or solve this new problem. But since for a Grassmannian frame $\mathbf{\Gamma}$ the cumulative coherence grows more slowly - in the case of $\mathbf{\Gamma}$ being an orthogonal basis it is even zero - the second problem is obviously better conditioned:

$$\begin{array}{rcl} y = \mathbf{\Phi}x & \Leftrightarrow & z = \mathbf{\Gamma}x \\ \mu_K(\mathbf{\Phi}) & \geq & \mu_K(\mathbf{\Gamma}) \end{array}$$

However, if we write down explicitly the sensing of z with $\mathbf{\Gamma}$ ($\mathbf{\Gamma}^\star$ denotes the transpose of $\mathbf{\Gamma}$),

$$\mathbf{\Gamma}^\star z = (B\mathbf{\Phi})^\star By = (\mathbf{\Phi}^\star B^\star B)y,$$

we see that we can actually interpret it as sensing the original signal with a sensing matrix of the form $\mathbf{\Psi} = B^\star B\mathbf{\Phi}$. In the special case where we choose B such that $B^\star B = (\mathbf{\Phi}\mathbf{\Phi}^\star)^{-1}$ we get as sensing matrix the canonical dual frame (pseudo-inverse): $\mathbf{\Psi} = (\mathbf{\Phi}\mathbf{\Phi}^\star)^{-1}\mathbf{\Phi}$, which in the even more special case where the dictionary is a basis is just the biorthogonal basis $(\mathbf{\Phi}^{-1})^\star$.
Now in order to generalise the above idea we can investigate what happens if we do not insist on deriving the sensing matrix from a linear transformation of the problem. Instead of restricting ourselves to using sensing matrices of the form $\mathbf{\Psi} = B^\star B\mathbf{\Phi}$, we will allow any matrix of the same size as the original dictionary. To see explicitly what properties we want to infer for the sensing/measuring matrix $\mathbf{\Psi}$ we do the analogue of the analysis leading to (1).

2.1 Worst Case Analysis of Thresholding with a Sensing Dictionary

Let y be a d-dimensional signal that has a K-sparse representation in the overcomplete dictionary $\mathbf{\Phi}$, $|\mathbf{\Phi}| = N$, i.e.

$$y = \sum_{i \in I} x_i \varphi_i.$$

For thresholding to recover a component φ_i in the support, we need the inner product of signal with the corresponding sensing atom ψ_i to be larger than the inner product with any atom in the sensing matrix whose corresponding partner is not part of the support:

$$i \in I: \quad |\langle \psi_i, y \rangle| \quad \geq \quad |\langle \psi_j, y \rangle|, \quad \forall j \notin I.$$

Writing out the inner product we can estimate:

$$\begin{aligned} i \in I: \ |\langle \psi_i, y \rangle| &\geq |x_i||\langle \psi_i, \varphi_i \rangle| - \sum_{j \in I, j \neq i} |x_j||\langle \psi_i, \varphi_j \rangle| \\ &\geq |x_i||\langle \psi_i, \varphi_i \rangle| - \|x\|_\infty \sum_{j \in I, j \neq i} |\langle \psi_i, \varphi_j \rangle| \\ k \notin I: \ |\langle \psi_k, y \rangle| &\leq \sum_{j \in I} |x_j||\langle \psi_k, \varphi_j \rangle| \\ &\leq \|x\|_\infty \sum_{j \in I} |\langle \psi_k, \varphi_j \rangle|. \end{aligned}$$

The right most terms in the above equations show a strong similarity to the cumulative coherence. By analogy we define the cumulative cross-coherence or cross Babel function of two dictionaries $\tilde{\mu}_1(K, \mathbf{\Phi}, \mathbf{\Psi})$ as well as their minimal coherence $\beta(\mathbf{\Phi}, \mathbf{\Psi})$ as:

$$\tilde{\mu}_1(K, \mathbf{\Phi}, \mathbf{\Psi}) := \max_i \max_{|J|=K, i \notin J} \sum_{j \in J} |\langle \psi_i, \varphi_j \rangle|, \tag{3}$$

$$\beta(\mathbf{\Phi}, \mathbf{\Psi}) := \min_i |\langle \psi_i, \varphi_i \rangle|. \tag{4}$$

As before we will leave out the reference to the dictionaries whenever it is clear which ones are meant. Using these definitions we can further simplify the above estimates to get:

$$\begin{aligned} i \in I: \ |\langle \psi_i, y \rangle| &\geq |x_i|\beta - \|x\|_\infty \tilde{\mu}_1(K-1) \\ k \notin I: \ |\langle \psi_k, y \rangle| &\leq \|x\|_\infty \tilde{\mu}_1(K). \end{aligned}$$

Finally the combination of these two estimates leads to the following theorem.

Theorem 3. *Let y be a signal exactly K-sparse in $\mathbf{\Phi}$, i.e. $y = \sum_{i\in I} x_i\varphi_i$. Thresholding with the sensing matrix $\mathbf{\Psi}$ is able to recover a component φ_i of the true support if*

$$\frac{|x_i|}{\|x\|_\infty} > \frac{1}{\beta}(\tilde{\mu}_1(K) + \tilde{\mu}_1(K-1)) := \nu(K, \mathbf{\Phi}, \mathbf{\Psi}). \tag{5}$$

This is a relaxation over the traditional recovery condition (1) if

$$\frac{1}{\beta}(\tilde{\mu}_1(K) + \tilde{\mu}_1(K-1)) < \mu_1(K) + \mu_1(K-1).$$

The obvious questions now are: Given a dictionary $\mathbf{\Phi}$, do there exist complementary sensing dictionaries that give a relaxed recovery condition and if yes how do we find them or rather how do we find the best. Since we want to have the new recovery condition as relaxed as possible we need to find the dictionary for which the recovery coefficient $\nu(K, \mathbf{\Phi}, \mathbf{\Psi})$ is minimal, i.e.

$$\mathbf{\Psi}_0 = \arg\min_{\mathbf{\Psi}} \nu(K, \mathbf{\Phi}, \mathbf{\Psi}). \tag{6}$$

Consequently, unless the minimum in the above equation is attained by the dictionary itself, there will always exist better sensing dictionaries. The next subsection is dedicated to developing an algorithm for finding one of them.

2.2 An Algorithm for Calculating Sensing Dictionaries

If we wanted to find the optimal sensing dictionary we would have to find the solution to Problem (6). This a daunting task as is more clearly demonstrated by looking at the expansion of the objective function after back-inserting the definitions:

$$\min_{\mathbf{\Psi}} \frac{1}{\min_i |\langle \psi_i, \varphi_i\rangle|}\Big(\max_{|J|=K, i\notin J} \sum_{j\in J} |\langle \psi_i, \varphi_j\rangle| + \max_{|J|=K-1, i\notin J} \sum_{j\in J} |\langle \psi_i, \varphi_j\rangle|\Big).$$

Another complication arises from the fact that we may not know the exact sparsity of our signals as this can vary but only its order of magnitude.
Our approach to solving the problem is inspired by the alternative projection method in [11] for constructing equiangular tight frames. The problem of trying to find a sensing matrix $\mathbf{\Psi}$ for the dictionary $\mathbf{\Phi}$ that gives low cumulative coherence can be reformulated as looking for the gram type matrix $G = \mathbf{\Psi}^\star\mathbf{\Phi}$ closest to the ideal gram matrix, which by Theorem 2 has only ones on the diagonal and all off diagonal entries of absolute value $\mu = \sqrt{\frac{N-d}{d(N-1)}}$. So if we

define

$$\begin{aligned}\mathcal{G} &:= \{G = \mathbf{\Psi}^\star\mathbf{\Phi},\ \mathbf{\Psi} \text{ a } N \times d \text{ matrix}\}\\ \mathcal{H} &:= \{H,\ \text{a } N \times N \text{ matrix with}\\ &\qquad H_{ii} = 1 \text{ and } |H_{ij}| \leq \mu \text{ for } i \neq j\}\end{aligned}$$

and equip the space of all $N \times N$ matrices with the Frobenius norm we can write the problem as

$$\min \|G - H\|_F \ \text{ s.t } \ G \in \mathcal{G}, H \in \mathcal{H}, \tag{7}$$

which can be solved via projection onto convex sets (POCS) since both sets $\mathcal{G}$ and $\mathcal{H}$ are convex, see [11] for details. In our case POCS will do the following. We fix a number of iterations, initialise $G = \mathbf{\Phi}^\star\mathbf{\Phi}$ and then in each iterative step do:

1. find $H \in \mathcal{H}$ that minimises $\|G - H\|_F$
2. find $G \in \mathcal{G}$ that minimises $\|H - G\|_F$

After the last iteration we can extract our sensing dictionary from the matrix H, which by definition is of the form $\mathbf{\Psi}^\star\mathbf{\Phi}$. Let us now find explicit expressions for the projection of a matrix A onto $\mathcal{H}$ and $\mathcal{G}$. By writing out the Frobenius norm explicitely

$$\min_{H\in\mathcal{H}} \|A - H\|_F = \min_{H\in\mathcal{H}} \big(\sum_{ij} |A_{ij} - H_{ij}|^2\big)^{\frac{1}{2}} \tag{8}$$

we see that the minimum is attained for the matrix H with

$$H : \begin{cases} H_{ii} = 1 & \\ H_{ij} = A_{ij} & \text{if } |A_{ij}| \leq \mu \\ H_{ij} = \operatorname{sgn}(A_{ij})\mu & \text{if } |A_{ij}| > \mu \end{cases} .$$

The solution to the second minimisation problem is not much harder to find. If we write $A^\star = (a_1 \dots a_N)$ we can rewrite the problem

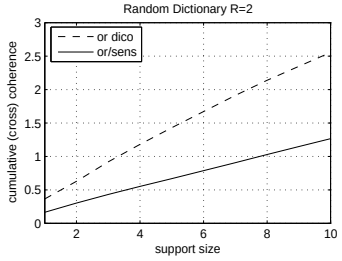
$$\begin{aligned}\min_{G\in\mathcal{G}} \|A - G\|_F &= \min_{\mathbf{\Psi}} \|A - \mathbf{\Psi}^\star\mathbf{\Phi}\|_F\\ &= \min_{\mathbf{\Psi}} \|A^\star - \mathbf{\Phi}^\star\mathbf{\Psi}\|_F\\ &= \min_{\mathbf{\Psi}} (\sum_i \|a_i - \mathbf{\Phi}^\star\psi_i\|_2^2)^{\frac{1}{2}}.\end{aligned}$$

From the last expression it is clear that we should choose $\psi_i = (\mathbf{\Phi}^\star)^\dagger a_i$, leading to $\mathbf{\Psi}^\star = A\mathbf{\Phi}^\dagger$ and $H = A\mathbf{\Phi}^\dagger\mathbf{\Phi}$. Before testing the algorithm numerically note that in case the dictionary was a basis we have $N = d$ resulting in $\mu = 0$. The set $\mathcal{H}$ consequently only contains the identity matrix and so in one iteration the algorithm will find the best sensing dictionary - the biorthogonal basis.

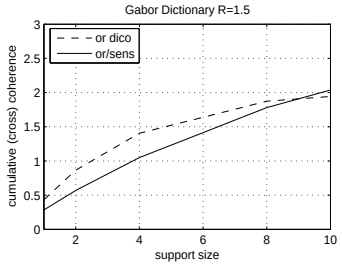
2.3 Simulations

First we calculated sensing dictionaries for three dictionaries of different types to compare the cumulative coherences and cross-coherences. To simplify the comparison we will 'hide' β within the correlations and choose the normalisation of the atoms in $\mathbf{\Psi}$ such that $|\langle \psi_i, \varphi_i \rangle| = \beta = 1$. The first dictionary was a random dictionary, of redundancy $R = 2$ in dimension $d = 128$. So in every atom the entries were drawn independently from a normalised standard Gaussian distribution and then the atom was rescaled to have unit norm. The second dictionary was a Gabor dictionary made up of the time-frequency shifts of one atom φ, i.e. $\mathbf{\Phi} = (\varphi_{n,m})_{n,m}$ where $\varphi_{n,m}(k) = e^{2\pi imbk}\varphi(k - na)$. In our case this atom was a normalised standard Gaussian in dimension $d = 120$ and the time and frequency shift parameters were chosen as $a = 8$, $b = 10$, leading to a redundancy $R = 1.5$. The third dictionary was the union of two orthonormal bases, the Haar-wavelet basis and the Discrete Cosine Transform (DCT) basis in dimension $d = 128$.

(a)



(b)



(c)

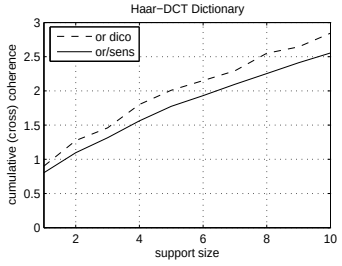


Figure 1: Cumulative coherence (or dico) and cross-coherence (sens dico) for various dictionaries.

Looking at Figure 1 we see that for the random dictionary, (a), the cross coherence is significantly lower than the coherence. We already have $\mu_1(K) > 1$ for $K > 3$ meaning that we can only guarantee to recover super positions of up to two atoms with equal absolute coefficients. On the other hand $\tilde{\mu}_1(4)+\tilde{\mu}_1(3) < 1$ meaning we can recover super-positions of up to 4 atoms. Also for the Gabor dictionary, (b), there is a slight improvement so while $\mu(3) > 1$ we still have $\tilde{\mu}(3) < 1$. For the Haar-DCT dictionary, (c), we still observe the slower growth of the cross-coherence but in this case the difference is not large enough to change the worst case behaviour, i.e. $1 < \tilde{\mu}(2) < \mu(2)$.

As second part of the simulations we tested how the sensing dictionaries performed in average for thresholding. For every support size varying between 1 and 30 we constructed 500 signals by choosing the atoms in the support uniformly at random and coefficients of absolute value one with random signs in the case of the real dictionaries, i.e. the random and the Haar-DCT dictionary, and uniformly random angle $e^{i\theta}$ in case of the complex Gabor dictionary. We ran thresholding using both the original and the sensing dictionary counting how often the full support could be recovered. The results are displayed in Figure 2.

As we can see while for both the random and the Gabor dictionary the recovery rates are higher when using the sensing dictionary there is no improvement for the Haar-DCT dictionary. One of the reasons might be that on average thresholding for the Haar-DCT dictionary is already performing well. So comparing the original recovery rates of the random and the Haar-DCT dictionary, which have about the same redundancy, we observe a performance gap in favour of the Haar-DCT dictionary. However, the gap closes when using the sensing dictionary for the random matrix. Also note that in the above experiment we tested the average performance but used the sensing dictionaries that were designed to give a good worst case performance. Before discussing these issues more thoroughly in Section 4 let us investigate the use of sensing dictionaries for (O)MP.

3 Sensing Dictionaries for (O)MP

Even more clearly than thresholding (O)MP can be decomposed into sensing and reconstruction steps. We initialise $a = 0$, $r = y$, $I = \emptyset$ and then in each step do:

Sensing:	$i = \arg\max_j \lvert\langle \varphi_j, r\rangle\rvert$
Reconstruction:	$a = a + \langle \varphi_i, r\rangle \varphi_i$, $r = y - a$ (MP)
	$I = I \cup i$, $a = \mathbf{\Phi}_I \mathbf{\Phi}_I^\dagger y$, $r = y - a$ (OMP)

As before we can change the sensing step of the algorithm and, instead of trying to identify components of the true support with the dictionary $\mathbf{\Phi}$ itself, use a sensing dictionary $\mathbf{\Psi}$.

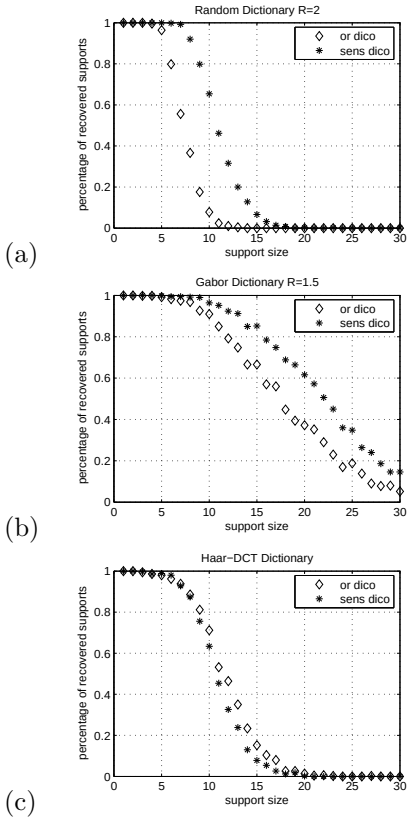


Figure 2: Recovery rates for thresholding using the original dictionary (or dico) and the sensing dictionary (sens dico).

Sensing new:	$i = \arg\max_j \lvert\langle \psi_j, r\rangle\rvert$
Reconstruction:	$a = a + \langle \varphi_i, r\rangle \varphi_i$, $r = y - a$ (MP)
	$I = I \cup i$, $a = \mathbf{\Phi}_I \mathbf{\Phi}_I^\dagger y$, $r = y - a$ (OMP)

To determine which conditions we should impose on the sensing matrix for (O)MP we again do a worst case analysis.

3.1 Worst Case Analysis of (O)MP with a Sensing Dictionary

Theorem 4. *Let y be a signal exactly K-sparse in $\mathbf{\Phi}$, i.e. $y = \sum_{i \in I} x_i \varphi_i$. (Orthogonal) Matching Pursuit using the sensing matrix $\mathbf{\Psi}$ will always select components of the true support I if*

$$\|(\mathbf{\Phi}_I^\star \mathbf{\Psi}_I)^{-1} \mathbf{\Phi}_I^\star \mathbf{\Psi}_{\overline{I}}\|_{1,1} < 1 \tag{9}$$

which is always satisfied if

$$\tilde{\mu}_1(K) + \tilde{\mu}_1(K-1) < \beta. \tag{10}$$

Proof: Basically we just need to rewrite Tropp's proof for *Exact Recovery for OMP* in [10]. As long as we have only selected correct atoms we know that the residual r is still a linear combination of the atoms in the true support, i.e.

$$r = \sum_{i \in I} c_i \varphi_i = \mathbf{\Phi}_I c.$$

(O)MP will again select a correct atom at the next step if the maximal correlation of the residual with an atom in the support $\max_{i \in I} |\langle \psi_i, r \rangle|$ is larger than the maximal correlation with an atom outside the support $\max_{k \in \overline{I}} |\langle \psi_k, r \rangle|$. So we have to make sure that the quotient satisfies

$$\frac{\max_{k \in \overline{I}} |\langle \psi_k, r \rangle|}{\max_{i \in I} |\langle \psi_i, r \rangle|} = \frac{\|\mathbf{\Psi}_{\overline{I}}^\star r\|_\infty}{\|\mathbf{\Psi}_I^\star r\|_\infty} < 1. \tag{11}$$

For further simplification we need to make use of p,q-matrix norms for $1 \leq p, q \leq \infty$, defined as $\|A\|_{p,q} = \max_{\|x\|_p = 1} \|Ax\|_q$. Inserting $r = \mathbf{\Phi}_I c$ into expression (11) and assuming that the matrix $\mathbf{\Psi}_I^\star \mathbf{\Phi}_I$ is invertible so that we can write $z = \mathbf{\Psi}_I^\star \mathbf{\Phi}_I c$, we can bound it as

$$\begin{aligned} \frac{\|\mathbf{\Psi}_{\overline{I}}^\star \mathbf{\Phi}_I c\|_\infty}{\|\mathbf{\Psi}_I^\star \mathbf{\Phi}_I c\|_\infty} &= \frac{\|\mathbf{\Psi}_{\overline{I}}^\star \mathbf{\Phi}_I (\mathbf{\Psi}_I^\star \mathbf{\Phi}_I)^{-1} z\|_\infty}{\|z\|_\infty} \\ &\leq \|\mathbf{\Psi}_{\overline{I}}^\star \mathbf{\Phi}_I (\mathbf{\Psi}_I^\star \mathbf{\Phi}_I)^{-1}\|_{\infty,\infty}. \end{aligned}$$

Finally we note that $\|\mathbf{\Psi}_{\overline{I}}^\star \mathbf{\Phi}_I (\mathbf{\Psi}_I^\star \mathbf{\Phi}_I)^{-1}\|_{\infty,\infty} = \|(\mathbf{\Phi}_I^\star \mathbf{\Psi}_I)^{-1} \mathbf{\Phi}_I^\star \mathbf{\Psi}_{\overline{I}}\|_{1,1}$ which by condition (9) is smaller than one as required.
For the second part of the proof we just have to show that condition (10) implies condition (9). First we can estimate

$$\|(\mathbf{\Phi}_I^\star \mathbf{\Psi}_I)^{-1} \mathbf{\Phi}_I^\star \mathbf{\Psi}_{\overline{I}}\|_{1,1} \leq \|(\mathbf{\Phi}_I^\star \mathbf{\Psi}_I)^{-1}\|_{1,1} \|\mathbf{\Phi}_I^\star \mathbf{\Psi}_{\overline{I}}\|_{1,1}.$$

The second term in the above can easily be bounded with the cross-coherence,

$$\|\mathbf{\Phi}_I^\star \mathbf{\Psi}_{\overline{I}}\|_{1,1} = \max_{k \in \overline{I}} \sum_{i \in I} |\langle \psi_k, \varphi_i \rangle| \leq \tilde{\mu}_1(K).$$

To bound the first term we use the fact that whenever $\|A\|_{1,1} < 1$ we have $\|\mathbf{I} + A\|_{1,1} < (1 - \|A\|_{1,1})^{-1}$. Set $A = \mathbf{\Phi}_I^\star \mathbf{\Psi}_I - \mathbf{I}$, then

$$\begin{aligned} \|A\|_{1,1} &= \max_{i \in I} \left(|\langle \psi_i, \varphi_i \rangle - 1| + \sum_{j \neq i} |\langle \psi_j, \varphi_i \rangle| \right) \\ &\leq 1 - \beta + \tilde{\mu}_1(K-1), \end{aligned} \tag{12}$$

and consequently

$$\begin{aligned}\|(\mathbf{\Phi}_I^\star \mathbf{\Psi}_I)^{-1}\|_{1,1} &\leq (1-(1-\beta+\tilde{\mu}_1(K-1))^{-1} \\ &\leq (\beta-\tilde{\mu}_1(K-1))^{-1}.\end{aligned}$$

If we now combine these two estimates with condition (10) we get the desired bound

$$\|(\mathbf{\Phi}_I^\star \mathbf{\Psi}_I)^{-1}\mathbf{\Phi}_I^\star \mathbf{\Psi}_{\overline{I}}\|_{1,1} \leq \frac{\tilde{\mu}_1(K)}{\beta-\tilde{\mu}_1(K-1)} < 1.$$

□

The theorem above is applicable to both MP and OMP as we only used that in each step the residual is a linear combination of the atoms in the support. Note, however, that picking a correct atom does not mean picking a new correct atom. Indeed since the sensing atoms corresponding to already found atoms are not orthogonal to the residual not even OMP can be guaranteed to find the full support in K steps.
As a consequence to Theorem 4 we get a characterisation of the optimal sensing dictionary for (O)MP. Given a dictionary $\mathbf{\Phi}$ and a sparsity level K, the best sensing dictionary $\mathbf{\Psi}_0$ is the solution to:

$$\mathbf{\Psi}_0 = \arg\min_{\mathbf{\Psi}} \max_{|I|=K} \|(\mathbf{\Phi}_I^\star \mathbf{\Psi}_I)^{-1}\mathbf{\Phi}_I^\star \mathbf{\Psi}_{\overline{I}}\|_{1,1}. \tag{13}$$

Unfortunately solving this problem is even harder than solving the original problem of finding the best sensing dictionary for thresholding in (6), as in addition to the maximum over all subsets of size K we also have to consider the inverse of a pseudo Gram matrix. However we still have the sufficient condition (10) for recovery success in terms of the cross coherence. Thus if we take a sensing dictionary calculated with the algorithm developed in Section 2.2 that has cross-coherence smaller than the coherence we can at least guarantee recovery for signals with higher sparsity. Finally what remains to be done is to check wether these sensing dictionaries also improve the average case performance of OMP.

3.2 Simulations for OMP

For our simulations we used the same three dictionaries and sensing dictionaries as for tresholding and the same set up. So for every support size varying between 10 and 40 we constructed 500 signals in the same way as for thresholding. Then we ran OMP using both the original and the sensing dictionary counting how often the full support could be recovered. The results are displayed in Figure 3.

Surprisingly even though the sensing matrices are derived from optimising only a sufficient worst case condition we can observe the same trends as for thresholding. So for both the random and the Gabor dictionary the recovery rates are higher when using the sensing dictionary but there is no improvement

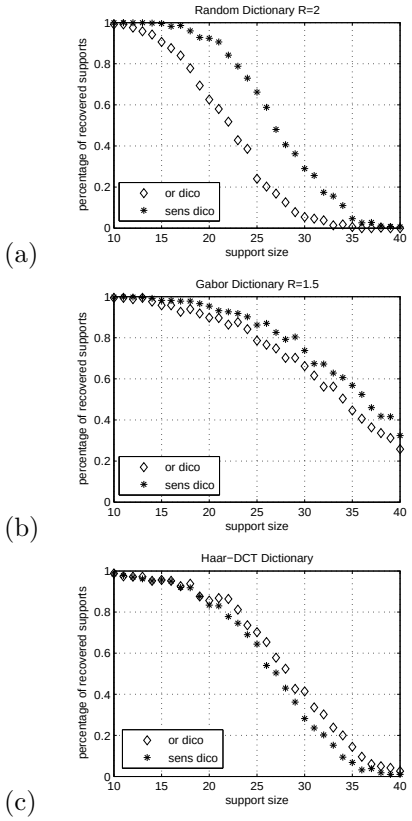


Figure 3: Recovery Rates for OMP using the original dictionary (or dico) and the sensing dictionary (sens dico).

for the Haar-DCT dictionary. Comparing the original recovery rates of the random and the Haar-DCT dictionary we observe the same performance gap in favour of the Haar-DCT dictionary as for thresholding. Again the gap closes when using the sensing dictionary for the random matrix.

4 Discussion, Conclusions & Future Work

In this paper we introduced the concept of sensing dictionaries to improve the performance of thresholding and OMP, while maintaining the same computational complexity. We analysed the worst case behaviour of both algorithms when using a sensing dictionary and from the results derived characterisations of the optimal sensing dictionaries for worst case performance. We developed an approximative algorithm to find good sensing dictionaries and showed that it works in practice, i.e. we get a sensing dictionary with lower cumulative cross coherence than coherence, even though this difference is not always sufficiently

large to guarantee a higher recovery rate. We did some numerical simulations to test the average performance of both algorithms and found out that in some cases the sensing dictionaries for good worst case performance also improve the average performance. There is a simple heuristic argument why the recovery rates increased for the random and the Gabor dictionary but not for the Haar-DCT dictionary. So for the random and the Gabor dictionary lowering the extreme correlations that are contributing to the cumulative coherence went together with lowering all the correlations, while for the Haar-DCT dictionary lowering the extremal correlations came at the price of increasing some of the a priori small correlations. Figure 4 showing the Gram matrices $\mathbf{\Phi}^\star\mathbf{\Phi}$ and pseudo Gram matrices $\mathbf{\Psi}^\star\mathbf{\Phi}$ nicely illustrates this effect.

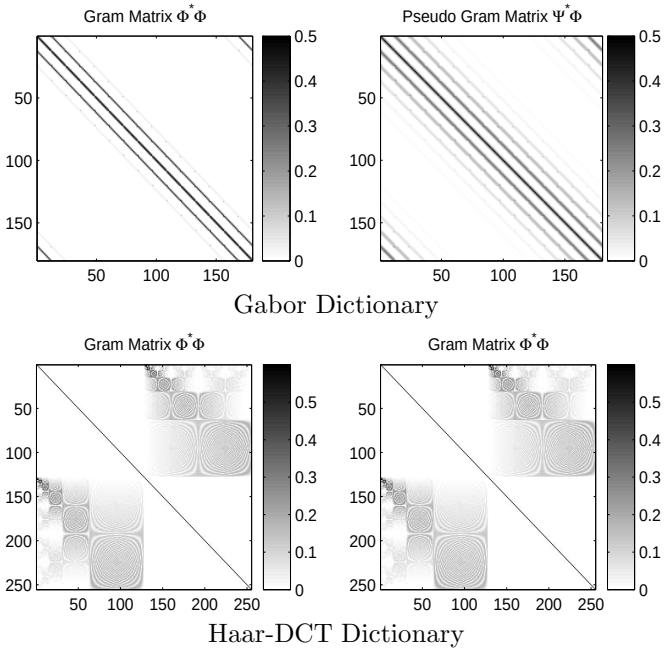


Figure 4: Gram and Pseudo Gram Matrices.

For the future there are plenty of interesting directions to explore. We would like to precisely analyse the average behaviour of OMP, as has already been done for BP, [12], and thresholding, [8]. Given a probabilistic model for our sparse signals, we want to derive design criteria for the sensing dictionaries to be able to recover a high percentage of these signals and consequently develop constructive algorithms for calculating the sensing dictionaries. Similarly we want to characterise good sensing matrices for recovery in the presence of noise and for OMP as an approximation algorithm.
As a parallel direction we would like to improve the algorithm for calculating the worst case sensing dictionaries. For now in order to find a sensing matrix that is close to optimal we use the Frobenius norm as distance measure. This

has the advantage of resulting in simple formulas for the alternate projections. However there is no guarantee that the best sensing matrix is the one having a pseudo Gram matrix that minimises the Euclidean distance to the ideal Gram matrix. A promising idea in that direction is to use a different distance measure like the 1-norm of the Gram matrices when considered as vectors. This would mean that instead of the least square problem in (7) we would have to solve an iteration of weighted least square problem with adaptive weights.
Finally it would be interesting to investigate whether the concept of dictionary preconditioning can be extended to Basis Pursuit, which is the other main approach for finding sparse signal representations.

A Proof of Theorem 2

We will do all the hard work for the proof in the following lemma.

Lemma 1. *Let α_i be a non increasing sequence of n positive real numbers with $\sum_{i=1}^{n} \alpha_i^2 = c$. If $K^2 \leq n$ then*

$$\sum_{i=1}^{K} \alpha_i \geq K\sqrt{\frac{c}{n}} \tag{14}$$

Equality holds if and only if the sequence is constant, i.e. $\alpha_i = \sqrt{\frac{c}{n}}$ for all i.

Proof: We have to show that for any non constant sequence the sum over the first K elements is larger than $K\sqrt{\frac{c}{n}}$. Assume first that the $\alpha_K > \sqrt{\frac{c}{N}}$. Then all the first K summands also have to be since the sequence has to be non increasing and so their sum is larger than $K\sqrt{\frac{c}{n}}$. On the other hand if we assume that $\alpha_K = \beta\sqrt{\frac{c}{n}}$ for a $0 < \beta < 1$, then

$$\sum_{i=1}^{K} \alpha_i^2 = c - \sum_{i=K+1}^{n} \alpha_i^2 \geq c(1 - \beta^2 + \frac{K}{n}\beta^2).$$

Under these circumstances the best possible choice for α_i, $i = 1 \dots K$, is the solution to the following minimisation problem

$$\min \sum_{i=1}^{K} \alpha_i \qquad \text{s.t } \sum_{i=1}^{K} \alpha_i^2 \geq c(1 - \beta^2 + \frac{K}{n}\beta^2)$$
$$\text{and } \alpha_i \geq \beta\sqrt{\frac{c}{n}}.$$

Figure 5 sketches the problem in two dimensions. The grey shaded area shows the region of all possible sequences α. The level curves of the objective function are the parallel translates of the dashed line. It is easy to see that the minimum

is attained in one of the corner of the feasible region, i.e. for

$$\alpha_2 = \ldots = \alpha_K = \beta\sqrt{\frac{c}{n}}$$
$$\alpha_1^2 = c(1 - \beta^2 + \frac{K}{n}\beta^2 - \frac{K-1}{n}\beta^2) = c(1 - \beta^2 + \frac{\beta^2}{n}).$$

Let us check when this minimum is smaller than $K\sqrt{\frac{c}{n}}$, i.e.

$$(K-1)\beta\sqrt{\frac{c}{n}} + \sqrt{c}\sqrt{1 - \beta^2 + \frac{\beta^2}{n}} < K\sqrt{\frac{c}{n}}$$

To see whether the inequality above is valid we have to do the following manipulations,

$$\begin{aligned}
\Leftrightarrow \quad & (K-1)\beta + \sqrt{n}\sqrt{1 - \beta^2 + \frac{\beta^2}{n}} < K \\
\Leftrightarrow \quad & n(1 - \beta^2 + \frac{\beta^2}{n}) < (K(1-\beta) + \beta)^2 \\
\Leftrightarrow \quad & n(1-\beta^2) - K^2(1-\beta)^2 - 2K(1-\beta)\beta < 0 \\
\Leftrightarrow \quad & n(1+\beta) - K^2(1-\beta) - 2K\beta < 0 \\
\Leftrightarrow \quad & (n - K^2)(1+\beta) + 2\beta(K^2 - K) < 0.
\end{aligned}$$

From the last expression we can finally see that the inequality is never valid as long as $K^2 < N$. Thus the only remaining choice is $\alpha_K = \sqrt{\frac{c}{N}}$ and then the minimum is clearly attained only when the sequence is constant. □

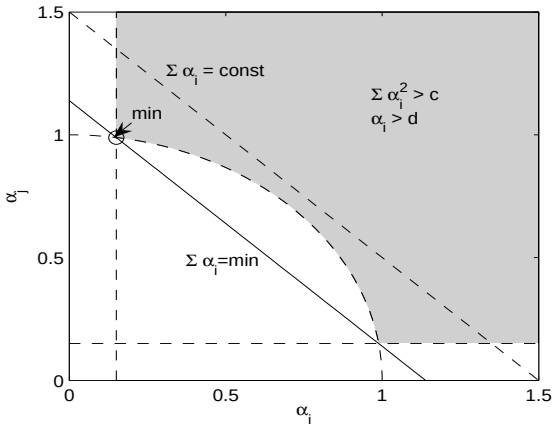


Figure 5: Minimisation problem.

Finally we are able to prove Theorem 2. We know that the energy of all inner products between two atoms, i.e. the squared Frobenius norm of the

Gram matrix satisfies

$$\sum_{i=1}^{N}\sum_{j=1}^{N}|\langle\varphi_i,\varphi_j\rangle|^2 \geq \frac{N^2}{d},$$

where equality holds only if $\mathbf{\Phi}$ is a tight frame, see [9] for a proof. From the inequality above we can deduce that there exists at least one φ_k such that

$$\sum_{i\neq k}|\langle\varphi_i,\varphi_k\rangle|^2 \geq \frac{N}{d}-1.$$

If we now reorder the correlations non increasingly and denote the ith largest correlations with α_i we can apply Lemma 1 with $n = N-1$ and get that

$$\max_{|I|=K}\sum_{i\in I}|\langle\varphi_i,\varphi_k\rangle| = \sum_{i=1}^{K}\alpha_i \geq K\sqrt{\frac{N-d}{d(N-1)}}$$

as long as $K^2 < N-1$. Backtracing the conditions under which all the inequalities are actually equalities we see that this happens if and only if all the correlations are of constant absolute value, i.e. on top of being tight $\mathbf{\Phi}$ is also equiangular.

References

[1] E. Candes, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *preprint*, February 2005.

[2] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive Approximation*, 13:57–98, 1997. Springer-Verlag New York Inc.

[3] D. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization. *Proc. Nat. Aca. Sci.*,, 100(5):2197–2202, March 2003.

[4] J. J. Fuchs. Extension of the pisarenko method to sparse linear arrays. *IEEE Transactions on Signal Processing*, 45(2413-2421), October 1997.

[5] J. J. Fuchs. Detection and estimation of superimposed signals. In *Proc. IEEE ICASSP98*, volume 3, pages 1649–1652, 1998.

[6] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. Technical Report PI-1848, IRISA, 2007.

[7] P. Jost, S. Lesage, P. Vandergheynst, and R. Gribonval. Motif: An efficient algorithm for learning translation invariant dictionaries. In *Proc. IEEE ICASSP06*, May 2006.

[8] K. Schnass and P. Vandergheynst. Average performance analysis for thresholding. *preprint*, 2007.

[9] T. Strohmer and R.W. Heath. Grassmannian frames with applications to coding and communication. *Appl. Comput. Harmon. Anal.*, 14(3):257–275, May 2003.

[10] J. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, October 2004.

[11] J. Tropp, I. Dhillon, R Heath Jr, and T. Strohmer. Designing structured tight frames via an alternating projection method. *IEEE Transactions on Information Theory*, 51(1):188–209, January 2005.

[12] J.A. Tropp. Random subdictionaries of general dictionaries. *preprint*, 2006.