# AVERAGE CASE ANALYSIS OF MULTICHANNEL THRESHOLDING

*Rémi Gribonval, Boris Mailhe, Holger Rauhut, Karin Schnass and Pierre Vandergheynst*

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Institute - ITS
CH- 1015 Lausanne, Switzerland

## ABSTRACT

This paper introduces $p$-thresholding, an algorithm to compute simultaneous sparse approximations of multichannel signals over redundant dictionaries. We work out both worst case and average case recovery analyses of this algorithm and show that the latter results in much weaker conditions on the dictionary. Numerical simulations confirm our theoretical findings and show that $p$-thresholding is an interesting low complexity alternative to simultaneous greedy or convex relaxation algorithms for processing sparse multichannel signals with balanced coefficients.

## 1. INTRODUCTION

Transform coding is one of the most successful paradigms in signal processing. Generally speaking, it asserts that many signals can be efficiently compressed because they have a sparse representation in some fixed basis. A simple transform coder would then decompose the signal over this optimal basis and threshold all projections to locate and keep only the $m$ strongest ones. This simple algorithm is at the core of the success of modern image and video coders such as JPEG2000 where a wavelet basis is used.

Recently though, new problems have come to challenge that paradigm. Restricting our models to decompositions over fixed bases drastically narrows the class of signals that can be efficiently processed. A lively strand of research advocates richer models based on redundant dictionaries, which can capture a much broader range of signals. A dictionary $\mathbf{\Phi}$ is a large collection of unit norm vectors $\|\varphi_k\|_2 = 1$, $k = 1$, ..., $K$ in $\mathbb{R}^d$, usually with $K \gg d$. Handling arbitrary dictionaries is no easy task, though. First, uniqueness of a signal representation is not guaranteed anymore. Second, even computing a decomposition becomes a complicated issue: several algorithms, most notably greedy algorithms and convex relaxation techniques can be used, but until recently analyzing their performance remained a daunting challenge. The situation unlocked with the realization that sparse models solve these problems. To illustrate the role of sparsity, let us introduce the *coherence* of the dictionary, i.e. the strongest correlation between any two distinct vectors in $\mathbf{\Phi}$: $\mu = \max_{i \neq j} |\langle \varphi_i, \varphi_j \rangle|$. Schematically, if our dictionary is not too coherent and the signal is an arbitrary superposition of less than $\mathcal{O}(\sqrt{d})$ elements of $\mathbf{\Phi}$, this representation is unique and can be recovered by standard algorithms [1, 2, 3].

In parallel to developments in sparse signal models, various application scenarios motivated renewed interest in processing not just a single signal, but many signals or channels at the same time. A striking example is sensor networks, where signals are monitored by low complexity devices whose observations are transfered to a central collector [4]. This central node thus faces the task of analyzing many, possibly high-dimensional, signals. Moreover, signals measured in sensor networks are typically not uncorrelated: there are global trends or components that appear in all signals, possibly in slightly altered forms. Modeling multichannel signals by means of redundant dictionaries, generalizing existing mono-channel algorithms and understanding their properties are thus important challenges.

In this paper we analyze the theoretical performance of $p$-thresholding, a simple algorithm for recovering simultaneous sparse approximations of multichannel signals. Our analysis is based on studying the average in addition to the worst case, and the spirit of our results is the following: given a not too coherent dictionary and signals with coefficients sufficiently large and balanced over the number of channels, $p$-thresholding can recover superpositions of up to $\mathcal{O}(d)$ atoms *with overwhelming probability*. Our conditions on $\mathbf{\Phi}$ are thus much less restrictive than in the worst case.

## 2. SIGNAL MODEL

Suppose we are to design a network of $N$ sensors monitoring a common phenomenon. Each of our sensors observes a $d$-dimensional signal $y_n \in \mathbb{R}^d$, $n = 1, ..., N$. As explained in the previous section, a sparsity hypothesis will be the central assumption of our model. Moreover, we will assume that each signal $y_n$ admits a sparse approximation over a single

R. Gribonval and B. Mailhe are with IRISA, Rennes, and H. Rauhut with the University of Vienna. This work was completed while they were visiting EPFL and partly supported by the HASSIP network, HPRN-CT-2002-00285.

Contact author: karin.schnass@epfl.ch

dictionary $\mathbf{\Phi}$:

$$y_n = \mathbf{\Phi} x_n + e_n,\ n = 1, ..., N.$$

In order to model correlations between signals, we will refine this model by imposing that all signals share a common sparse support, i.e.

$$y_n = \mathbf{\Phi}_\Lambda x_n + e_n,$$

where $\mathbf{\Phi}_\Lambda$ is the restriction of the synthesis matrix $\mathbf{\Phi}$ to the columns listed in the set $\Lambda$. This model is inspired by a recent series of papers on distributed sensing, see [5] and references therein. It describes a network of sensors monitoring a signal with a strong global component that appears at each node. Localized effects are modeled by letting synthesis coefficients $x_n \in \mathbb{R}^S$, $S := |\Lambda|$, vary across nodes and through the noise $e_n$. As an illustrative example, imagine sensors measuring the chemical composition of the atmosphere at some locations of a geographical area. There is a common component modeled by the fixed support $\Lambda$. Slight changes from node to node due to different sensor locations are modeled by varying amplitudes $x_n$ of components from node to node. Localized effects can drastically alter the signal and are captured by noise $e_n$. Let us now turn to describing a generative model for the synthesis coefficients $x_n$. In order to obtain a sufficiently general model, we will assume that the components $x_n(k)$ of the random vector $x_n$ are independent Gaussian variables of variance $\alpha_k$. This model is fairly general to accommodate various practical problems: the Gaussian assumption is one of the most widely used in signal processing, while incorporating different variances allows us to shape the synthesis coefficients, imposing statistical decay for example on the $x_n(k)$.

In order to simplify our analysis we will adopt a global matrix notation. We will collect all signals $y_n$ on the columns of the $d \times N$ matrix $Y$ and the synthesis coefficients $x_n$ on the columns of the $S \times N$ matrix $X$. Let $U$ be a $S \times N$ random matrix with independent standard Gaussian entries and let $D$ be a $S \times S$ diagonal matrix whose entries are positive real numbers $\alpha_k$. Our model can then be written in compact form

$$Y = \mathbf{\Phi}_\Lambda X + E = \mathbf{\Phi}_\Lambda D U + E, \tag{1}$$

where $E$ is a $d \times N$ matrix collecting noise signals $e_n$ on its columns.

## 3. ALGORITHM

### 3.1. Principle

Let us now describe more precisely our sensing algorithm. The observed signals $y_n$ are sent to a central processing unit that tries to recover the common sparse support $\Lambda$. The problem thus boils down to estimating the joint sparse support of a set of signals generated from a redundant dictionary $\mathbf{\Phi}$. A number of algorithms have been proposed lately to jointly process sparse signals, most of them based on multichannel generalizations of greedy algorithms [6] or convex relaxation algorithms. A common weakness to all these techniques is a high computational complexity. To overcome this problem, we would like to resort here to one of the simplest possible algorithms: thresholding. More precisely, our algorithm computes the $p$-norm of the correlation of the multichannel signal $Y$ with the atoms $\psi_k$ of a sensing dictionary $\mathbf{\Psi}$:

$$\|\psi_k^\star Y\|_p^p := \sum_{n=1}^{N} |\langle \psi_k, y_n \rangle|^p.$$

The sensing dictionary $\mathbf{\Psi}$ has the same cardinality as $\mathbf{\Phi}$, so the atoms in both dictionaries are in a one-to-one relationship. We could set $\mathbf{\Psi} \equiv \mathbf{\Phi}$, but we voluntarily keep the possibility of optimizing both dictionaries in the spirit of [7].

### 3.2. Recovery conditions

Define $\Lambda_S$, the set of indices $k$ with the $S$ largest $p$-norms. This algorithm is successful if for $S = \sharp\Lambda$ we have $\Lambda_S = \Lambda$. Since $\mathbf{\Psi}^\star Y = \mathbf{\Psi}^\star \mathbf{\Phi}_\Lambda X + \mathbf{\Psi}^\star E$, the strongest $p$-norm of projections on the set $\overline{\Lambda}$ of bad atoms is

$$\|\mathbf{\Psi}_{\overline{\Lambda}}^\star Y\|_{p,\infty} \leq \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{\Phi}_\Lambda X\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty},$$

where the $(p,\infty)$-norm of a matrix $\|\mathbf{M}\|_{p,\infty}$ is defined as the maximum of the $p$-norms of its rows. Conversely, the smallest $p$-norm of projections on the set of good atoms reads

$$\min_{i\in\Lambda} \|\psi_i^\star Y\|_p \geq \min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda X\|_p - \|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty}.$$

and the algorithm will thus succeed as soon as

$$\min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda X\|_p - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{\Phi}_\Lambda X\|_{p,\infty} > \|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty}. \tag{2}$$

### 3.3. Worst case analysis

This condition can be checked based on simple characteristics of the multichannel signals and the dictionaries. To capture the requirements on the dictionary we need to define $\beta := \min_{i\in\Lambda} |\langle \psi_i, \varphi_i \rangle|$ the minimum correlation between sensing and synthesis atoms, and to adapt the definition of the standard cumulative coherence [1]:

$$\mu_q(\mathbf{\Psi}, \mathbf{\Phi}, \Lambda) := \sup_{l\notin\Lambda} \|\mathbf{\Phi}_\Lambda^\star \psi_l\|_q = \sup_{l\notin\Lambda} \left( \sum_{i\in\Lambda} |\langle \psi_l, \varphi_i \rangle|^q \right)^{1/q}. \tag{3}$$

As for properties of the signal we need to define the p-Peak SNR and the dynamic range $R_p$:

$$\begin{aligned} \text{PSNR}_p &:= \frac{\|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty} + \|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty}}{\|X\|_{p,\infty}}, \\ R_p &:= \frac{\min_{i\in\Lambda} \|X(i)\|_p}{\|X\|_{p,\infty}}, \end{aligned}$$

where we denote $\|X(i)\|_p = (\sum_{n=1}^N |x_n(i)|^p)^{1/p}$ the $p$-norm of the $i$-th row of $X$. Following the analysis in [8], it is easy to check [9] that the following condition implies (2):

$$\begin{aligned} \mu_1(\mathbf{\Psi}, \mathbf{\Phi}, \Lambda) + \sup_{i\in\Lambda} \mu_1(\mathbf{\Psi}_\Lambda, \mathbf{\Phi}_\Lambda, \Lambda/\{i\}) \\ < \beta \cdot R_p - \text{PSNR}_p. \end{aligned} \tag{4}$$

The success of $p$-thresholding is thus governed by the condition that the dynamic range of the signal should be bigger than the noise level and the sum of correlations among atoms on the support and between the support and the remaining of $\mathbf{\Phi}$. We note that $\mu_1$ can be very big even for reasonably small $\Lambda$. For example, when $\mathbf{\Psi} = \mathbf{\Phi}$, the quantity $\mu_1(\mathbf{\Psi}, \mathbf{\Phi}, \Lambda) + \mu_1(\mathbf{\Psi}_\Lambda, \mathbf{\Phi}_\Lambda, \Lambda/\{i\})$ is often replaced by its upper estimate $(2S-1)\mu$. The r.h.s in (4) is at most one, so the resulting condition can only be satisfied when $S < (1+\mu^{-1})/2$. In the next sections, we develop an average case analysis of $p$-thresholding and show that the *typical* recovery conditions are much less restrictive.

## 4. AVERAGE CASE ANALYSIS

To state our central theoretical result for the average case we need to define a probabilistic PSNR and dynamic range, remember we had $Y = \mathbf{\Phi}_\Lambda DU + E$ where $D = \mathbf{diag}(\alpha_i)$,

$$\begin{aligned} \overline{\text{PSNR}}_p &:= \frac{\|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} + \|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty}}{\max_{i\in\Lambda}|\alpha_i|}, \\ \overline{R} &:= \frac{\min_{i\in\Lambda}|\alpha_i|}{\max_{i\in\Lambda}|\alpha_i|}. \end{aligned}$$

**Theorem 1.** *Assume that the noise level and the dynamic range are sufficiently small (respectively large), that is to say*

$$\mu_2(\mathbf{\Phi}, \mathbf{\Psi}, \Lambda) < \min_{i\in\Lambda} \|\mathbf{\Phi}_\Lambda^\star \psi_i\|_2 \cdot \overline{R} - \overline{\text{PSNR}}_p / C_p(N). \tag{5}$$

*where $C_p(N)$ is a constant depending only on $p$ and the number of channels $N$, see Theorem 2. Then, under signal model (1), the probability that $p$-thresholding fails to recover the indices of the atoms in $\Lambda$ does not exceed*

$$\mathbb{P}(p-\textit{thresholding fails}) \le K \cdot \exp\left(-AN\gamma^2\right)$$

*with*

$$\gamma = \frac{\overline{R}\cdot \min_{i\in\Lambda}\|\mathbf{\Phi}_\Lambda^\star\psi_i\|_2 - \overline{\text{PSNR}}_p/C_p(N) - \mu_2(\mathbf{\Phi},\mathbf{\Psi},\Lambda)}{\overline{R}\cdot \min_{i\in\Lambda}\|\mathbf{\Phi}_\Lambda^\star\psi_i\|_2 + \mu_2(\mathbf{\Phi},\mathbf{\Psi},\Lambda)}.$$

This result has unique features compared to the worst case, see (4). First, the condition on $\mathbf{\Phi}$ is expressed in terms of the cumulative coherence of order 2 which is much smaller than that of order one. For example assuming that there is no noise and that the variances $\alpha_i$ are constant the r.h.s in (5) is larger than one. If additionally $\mathbf{\Psi} = \mathbf{\Phi}$, an upper estimate of $\mu_2(\mathbf{\Phi}, \mathbf{\Psi}, \Lambda)$ is $\mu\sqrt{S}$ and we see that typically thresholding can be successful even when $S \approx \mu^{-2} \gg \mu^{-1}$. Second, due to typicality, we see that the probability of failure quickly diminishes as the number of channel grows, suggesting that we should use $N \sim \log K$ channels in practice.

**Ingredients and Flavour of the Proof**

As explained in the previous section, thresholding works by collecting statistics on the signals through projections. The intuition is that, when there are sufficiently many channels, typical behavior will emerge and allow us to detect the meaningful components. The following classical result of measure concentration will be our main tool [10].

**Theorem 2.** *Let $1 \le p \le \infty$. Suppose $Z = (Z_1, \ldots, Z_N)$ is a vector of independent standard Gaussian variables. Then there exist constants $C_p(N), A_p(N)$ such that*

$$\mathbb{P}(\|Z\|_p \ge (1+\epsilon)C_p(N)) \le \exp\left(-\epsilon^2 A_p(N)\right) \tag{6}$$

*and*

$$\mathbb{P}(\|Z\|_p \le (1-\epsilon)C_p(N)) \le \exp\left(-\epsilon^2 A_p(N)\right), \tag{7}$$

*For the important cases $p = 1, 2$ we have $C_1(N) = \sqrt{\frac{2}{\pi}}N$, $C_2(N) \sim \sqrt{N}$ and $A_2(N) \ge A_1(N) = N/\pi$.*

This theorem highlights the emergence of typicality in high-dimensional random Gaussian vectors: the probability that the $p$-norm of $Z$ differs significantly from $C_p(N)$ decreases exponentially with $N$.

The main idea for the proof of Theorem 1 is that when the number of channels is sufficiently large, each $p$-correlation $\|\psi_k^\star \mathbf{\Phi}_\Lambda DU\|_p$ of the noiseless multichannel signal with a sensing atom is, with very large probability, almost equal to $C_p(N)\cdot \|\psi_k^\star \mathbf{\Phi}_\Lambda D\|_2 = C_p(N)\cdot\|D\mathbf{\Phi}_\Lambda^\star\psi_k\|_2$, where $C_p(N)$ grows with the number of channels. Therefore, if

$$\begin{aligned} C_p(N)\cdot\left(\min_{i\in\Lambda}\|D\mathbf{\Phi}_\Lambda^\star\psi_i\|_2 - \max_{\ell\notin\Lambda}\|D\mathbf{\Phi}_\Lambda^\star\psi_\ell\|_2\right) \gtrsim \\ \|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} + \|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty}, \end{aligned}$$

the recovery condition (2) will be met with high probability. The easy part is that inserting the following estimates into the above and simplifying a bit we arrive at condition (5) of the theorem,

$$\begin{aligned} \min_{i\in\Lambda}\|D\mathbf{\Phi}_\Lambda^\star\psi_i\|_2 &\ge \min_{j\in\Lambda}|\alpha_j|\cdot\|\mathbf{\Phi}_\Lambda^\star\psi_j\|_2, \\ \max_{\ell\notin\Lambda}\|D\mathbf{\Phi}_\Lambda^\star\psi_\ell\|_2 &\le \max_{j\in\Lambda}|\alpha_j|\cdot\mu_2(\mathbf{\Phi},\mathbf{\Psi},\Lambda). \end{aligned}$$

The hard part of the proof is to make precise estimates of the typicality and precision of the approximation $\|\psi_k^\star\mathbf{\Phi}_\Lambda DU\|_p \approx C_p(N)\|D\mathbf{\Phi}_\Lambda^\star\psi_k\|_2$ using Theorem 2. Although this is clearly out of the scope of this paper, let us summarize the main steps. Observe that $\psi_k^\star\mathbf{\Phi}_\Lambda DU =: v_k^* U$ is a vector of independent Gaussian random variables whose components have variance

$\|v_k\|_2$. Therefore, applying Theorem 2 and the union bound we arrive at

$$\mathbb{P}\left(\max_{\ell\in\overline{\Lambda}} \|v_\ell^\star U\|_p \geq (1+\epsilon_1)C_p(N)\max_{\ell\in\overline{\Lambda}}\|v_\ell\|_2\right) \leq |\overline{\Lambda}|\cdot\exp(-\epsilon_1^2 A_p(N)).$$

Similar arguments are used to estimate the probability that $\min_{i\in\Lambda}\|\psi^\star\mathbf{\Phi}_\Lambda DU\|_p$ takes a small value. The exact result is then obtained by carefully choosing the constants $\epsilon_i$.

## 5. EXPERIMENTAL RESULTS

In this section we compare our theoretical findings with simulations of the performance of 2-thresholding with $\mathbf{\Psi} = \mathbf{\Phi}$. As dictionary we chose a combination of the Dirac and Fourier basis, $\mathbf{\Phi} = (\mathbf{I}_d, \mathcal{F}_d)$, in dimension $d = 1024$, which has coherence $\mu = 1/\sqrt{d}$. For each number of channels $N$, varying from 1 to 128, and support size, varying from 1 to 1024 in steps of 16, we created 180 signals by choosing a support $\Lambda$ uniformly at random and independent Gaussian coefficients with variances $\alpha_i = 1$ and calculated the percentage of thresholding being able to recover the full support. The results can be seen in Figure 1.

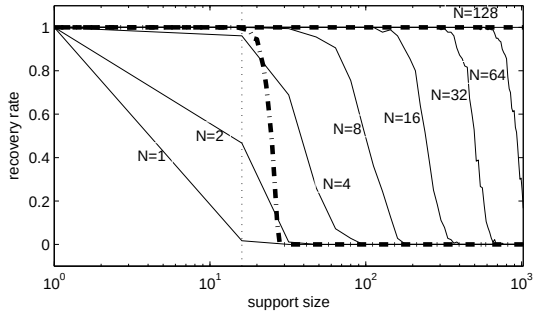


**Fig. 1**. Comparison of Recovery Rates for Different Support Sizes and Number of Channels.

As reference we also calculated how many out of 200 randomly chosen supports of a given size satisfy the worst case recovery condition $\mu_1(\lambda) + \sup_{i\in\Lambda}\mu_1(\Lambda/\{i\}) < 1$. This is indicated by the dash dotted line and can be seen to drop rapidly once the theoretical limit $|\Lambda| = 16$ is reached. Since $\mu = 1/\sqrt{d}$ the average recovery condition $\mu_2\Lambda < 1$, indicated by the dashed line, is always satisfied. We can see that as predicted by Theorem 1 with an increasing number of channels we get closer to the average case bound, which is actually attained once $N = 128$.

## 6. CONCLUSIONS

Thresholding is a computationally inexpensive algorithm for simultaneous sparse signal approximation. We have shown that, in a probabilistic multichannel setting, it shares good recovery properties with much more complex alternatives such as greedy algorithms and convex relaxation algorithms. The worst case recovery condition is reminiscent of Tropp's recovery condition, but the typical behaviour is instead driven by a much less restrictive condition and improves with the numbers of channels. This is clearly confirmed by our simulation results.

One of the main drawbacks of thresholding is that its performance relies heavily on the assumption that the signal coefficients are well balanced, in addition to the Gaussian model. Orthogonal Matching Pursuit is a natural candidate for dealing with signals that do not have balanced coefficients. Preliminary results [9] indicate that its typical performance in a multi-channel probabilistic setup is also driven by much less restrictive conditions on the dictionary than the worst case ones . Last but not least, since the characterization of what drives the average performance of thresholding work involves the mutual coherence of order 2 between a sensing dictionary and a synthesis dictionary, an interesting new perspective is the design of a sensing dictionary to optimize the recovery performance for a given signal model.

## 7. REFERENCES


[1] J.A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. on Information Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.

[2] J.A. Tropp, "Just relax: Convex programming methods for subset selection and sparse approximation," *IEEE Trans. Information Theory*, vol. 51, no. 3, pp. 1030–1051, March 2006.

[3] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. on Inform.Theory*, vol. 52, no. 1, January 2006.

[4] Z. Luo, M. Gaspar, J. Liu, and A. Swami, "Distributed signal processing in sensor networks," *IEEE Signal processing magazine*, vol. 23, no. 4, pp. 14–15, July 2006.

[5] D. Baron, M.F. Duarte, S. Sarvotham, M.B. Wakin, and R.G. Baraniuk, "An information-theoretic approach to distributed compressed sensing," in *Proc. 45rd Conference on Communication, Control, and Computing*, 2005.

[6] J.A. Tropp, A.C. Gilbert, and M.J. Strauss, "Algorithms for simultaneous sparse approximations. Part I: Greedy pursuit," *Signal Processing, special issue "Sparse approximations in signal and image processing"*, vol. 86, pp. 572–588, 2006.

[7] K. Schnass and P. Vandergheynst, "Deterministic measurement ensembles for greedy algorithms," *Submitted to IEEE Transactions on Signal Processing*, May 2006.

[8] R. Gribonval, M. Nielsen, and P. Vandergheynst, "Towards an adaptive computational strategy for sparse signal approximation," IRISA preprint, Rennes, France, 2006.

[9] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Average case analysis of multichannel thresholding and greedy algorithms," in preparation.

[10] M. Ledoux, *The Concentration of Measure Phenomenon*, AMS, 2001.