# Dictionary Identification Results for K-SVD with Sparsity Parameter 1

Karin Schnass
Computer Vision Laboratory
University of Sassari
Porto Conte Ricerche, 07041 Alghero, Italy
kschnass@uniss.it

***Abstract*—In this paper we summarise part of the results from our recent work [1]. We give theoretical insights into the performance of K-SVD, a dictionary learning algorithm that has gained significant popularity in practical applications, by answering the question when a dictionary $\Phi$ can be recovered as local minimum of the minimisation criterion underlying K-SVD from a set of training signals $y_n = \Phi x_n$. Assuming the training signals are generated from a tight frame with coefficients drawn from a random symmetric distribution, then in expectation the generating dictionary can be recovered as a local minimum of the K-SVD criterion if the coefficient distribution exhibits sufficient decay. This decay can be characterised by the coherence of the dictionary and the $\ell_1$-norm of the coefficients. Further it is demonstrated that given a finite number of training samples $N$ with probability $O(\exp(-N^{1-4q}))$ there is a local minimum of the K-SVD criterion within a radius $O(N^{-q})$ of the generating dictionary.**

***Index Terms*—dictionary learning, sparse coding, finite samples, K-SVD, sampling complexity, dictionary identification, minimisation criterion, sparse representation**

## I. Introduction

Research in the last decade has proven that sparsity provides an efficient way of dealing with high-dimensional data, since sparse signals are easily compressed, are robust to corruption and can therefore easily be restored from incomplete information. Triggered by this success an increasingly important research direction is how to learn dictionaries providing sparse representations for the data at hand, known as dictionary learning or sparse coding. The problem under investigation is usually formulated as follows. Given $N$ signals $y_n \in \mathbb{R}^d$, stored as columns in a matrix $Y = (y_1, \dots y_N)$ find a decomposition,

$$Y \approx \Phi X,$$

into a $d \times K$ dictionary matrix $\Phi$ with unit norm columns and a $K \times N$ coefficient matrix with sparse columns.

So far research has provided several dictionary learning algorithms, which are efficient in practice and therefore popular in applications, but there exists only a handful of dictionary learning schemes, for which theoretical results available, [3], [4], [5], [6], [7]. Unfortunately, however, these then tend to be rather cumbersome in practice. In this talk we start bridging the gap between practically efficient and provably efficient dictionary learning schemes, by shedding some light on the theoretical performance of K-SVD, one of the most widely applied dictionary algorithms.

K-SVD was introduced by Aharon, Elad and Bruckstein in [8] as an algorithm to solve the following minimisation problem. Given some signals $Y = (y_1, \dots, y_N)$, $y_n \in \mathbb{R}^d$, find

$$\min_{\Phi \in \mathcal{D}, X \in \mathcal{X}_S} \|Y - \Phi X\|_F^2, \tag{1}$$

for $\mathcal{D} := \{\Phi = (\phi_1, \dots, \phi_K), \phi_i \in \mathbb{R}^d, \|\phi_i\|_2 = 1\}$ and $\mathcal{X}_S := \{X = (x_1, \dots, x_N), x_n \in \mathbb{R}^K, \|x_n\|_0 \leq S\}$, where $\|x\|_0$ counts the number of non-zero entries of x, and $\|\cdot\|_F$ denotes the Frobenius norm. In short we are looking for the dictionary $\Phi$ that provides on average the best $S$-term approximation to the signals in $Y$.

Since for a signal $y_n$ the best $S$-term approximation using $\Phi$ is given by the largest projection onto a set of $S$ atoms $\Phi_I = (\phi_{i_1} \dots \phi_{i_S})$, ie. $P_I(\Phi) = \Phi_I \Phi_I^\dagger$ where $\Phi_I^\dagger$ denotes the Moore-Penrose pseudo inverse of $\Phi_I$, instead of (1) we can equivalently consider the following maximisation problem,

$$\max_{\Phi \in \mathcal{D}} \sum_i \max_{|I| \leq S} \|P_I(\Phi) y_n\|_2^2. \tag{2}$$

Let us assume that the training signals are all created from an admissible generating dictionary $\bar{\Phi} \in \mathcal{D}$, and coefficients drawn at random from a distribution $\nu$ of sparse or rapidly decaying coefficient, ie.

$$y_n = \bar{\Phi} \bar{x_n}. \tag{3}$$

The goal of dictionary identification is to give conditions under which an algorithm can locally identify the generating dictionary from the training signals. To achieve this we will first study when $\bar{\Phi}$ is exactly at a local maximum in the limiting case, ie. when we replace the sum in (2) with the expectation,

$$\max_{\Phi \in \mathcal{D}} \mathbb{E}_y \left( \max_{|I| \leq S} \|P_I(\Phi) y\|_2^2 \right). \tag{4}$$

In the next section we will provide identification results for the case when in (4) we have $S = 1$, ie. $\mathcal{X}_S = \mathcal{X}_1$, assuming first a simple (discrete, noise-free) signal model and then progressing

This work was supported by the Austrian Science Fund (FWF) under Grant no. Y432 an J3335.

to a noisy, continuous signal model. In Section III we will extend these asymptotic results to the case of a finite number of samples and finally we will discuss the implications of our results for practical applications and compare them to related dictionary identification results.

## II. Asymptotic identification results

### *A. The problem for $S = 1$*

In case $S = 1$ the objective function in (4) can be radically simplified and the maximisation problem we want to analyse reduces to,

$$\max_{\Phi \in \mathcal{D}} \mathbb{E}_y \left( \|\Phi^\star y\|_\infty^2 \right). \tag{5}$$

Clearly if the signals $y$ are all 1-sparse in a dictionary $\bar{\Phi}$ then $\bar{\Phi}$ is a global maximiser of (5). However what happens if we do not have perfect sparsity? Let us start with a very simple negative example.

*Example 2.1:* Let $U$ be an orthonormal basis and $x$ be randomly 2-sparse with 'flat' coeffcients, ie. pick two indices $i, j$ uniformly at random, choose $\sigma_{i/j} = \pm 1$ uniformly at random and set $x_k = \sigma_k$ for $k = i, j$ and zero else. Then $U$ is not a local maximum of (5), since we can construct an ascent direction. Choose $U_\varepsilon = (u_1, \ldots, u_{d-1}, (u_d + \varepsilon u_1)/\sqrt{1+\varepsilon^2})$, then we have

$$\begin{aligned} \mathbb{E}_y \left( \|U_\varepsilon^\star y\|_\infty^2 \right) &= \mathbb{E}_x \left( \|(x_1, \ldots, x_{d-1}, \tfrac{x_d + \varepsilon x_1}{\sqrt{1+\varepsilon^2}})\|_\infty^2 \right) \\ &= 1 + \tfrac{1}{d(d-1)} \tfrac{\varepsilon}{1+\varepsilon^2} > 1 = \mathbb{E}_y \left( \|U^\star y\|_\infty^2 \right). \end{aligned}$$

From the above example we see that in order to have a local maximum at the original dictionary we need a signal/coefficient model where the coefficients show some type of decay.

### *B. A simple model of decaying coefficients*

We first consider a very simple coefficient model, constructed from a non-negative, non-increasing sequence $c \in \mathbb{R}^K$ with $\|c\|_2 = 1$, which we permute uniformly at random and provide with random ± signs. To be precise for a permutation $p : \{1, ..., K\} \to \{1, ..., K\}$ and a sign sequence $\sigma$, $\sigma_i = \pm 1$, we define the sequence $c_{p,\sigma}$ component-wise as $c_{p,\sigma}(i) := \sigma_i c_{p(i)}$, and set $y = \Phi x$ where $x = c_{p,\sigma}$ with probability $(2^K K!)^{-1}$.

The normalisation $\|c\|_2 = 1$ has the advantage that for dictionaries, which are an orthonormal basis, the resulting signals also have unit norm and for general dictionaries the signals have unit square norm in expectation, ie. $\mathbb{E}(\|y\|_2^2) = 1$. This reflects the situation in practical application, where we would normalise the signals in order to equally weight their importance.

Armed with this model we can now prove a first dictionary identification result for (5).

*Theorem 2.1:* Let $\Phi$ be a unit norm tight frame with frame constant $A = K/d$ and coherence $\mu$. Let $x \in \mathbb{R}^d$ be a random permutation of a sequence $c$, where $c_1 \geq c_2 \geq c_3 \ldots \geq c_K \geq 0$ and $\|c\|_2 = 1$, provided with random ± signs, i.e. $x = c_{p,\sigma}$ with probability $\mathbb{P}(p, \sigma) = (2^K K!)^{-1}$. If $c$ satisfies $c_1 > c_2 + 2\mu\|c\|_1$, then there is a local maximum of (5) at $\Phi$. Moreover we have the following quantitative estimate for the basin of attraction around $\Phi$. For all perturbations $\Psi = (\psi_1 \ldots \psi_K)$ of $\Phi = (\phi_1 \ldots \phi_K)$ with $0 < \max_i \|\psi_i - \phi_i\|_2 \leq \varepsilon$ we have $\mathbb{E}_x \|\Psi^\star \Phi x\|_\infty^2 < \mathbb{E}_x \|\Phi^\star \Phi x\|_\infty^2$ as soon as $\varepsilon < 1/5$ and

$$\varepsilon \leq \frac{\left(1 - 2\frac{c_2 + \mu\|c\|_1}{c_2 + c_1}\right)^2}{2A \log\left(2AK/(c_1^2 - \frac{1-c_1^2}{K-1})\right)}. \tag{6}$$

*Proof:* We briefly sketch the proof. The condition $c_1 > c_2 + 2\mu\|c\|_1$ ensures that the maximal inner product $|\langle \phi_i, \Phi c_{p,\sigma} \rangle|$ is always attained by $i_p = p^{-1}(1)$, leading to

$$\mathbb{E}_x \|\Phi^\star \Phi x\|_\infty^2 = c_1^2 + \frac{(1 - c_1^2)}{K - 1}(A - 1).$$

The main idea now is that for small perturbations and most sign patterns $\sigma$ the maximal inner product is still attained by $i_p$. For an $\varepsilon$-perturbations $\Psi$ of the original dictionary $\Phi$ where $\psi_i = (1 - \varepsilon_i^2/2)\phi_i + (\varepsilon_i^2 - \varepsilon_i^4/4)^{\frac{1}{2}} z_i$, for some $z_i$ with $\langle \phi_i, z_i \rangle = 0$, $\|z_i\|_2 = 1$ and $\varepsilon_1 \leq \varepsilon$, we have

$$\max_{i=1 \ldots K} |\langle \psi_i, \Phi c_{p,\sigma} \rangle| = |\langle \psi_{i_p}, \Phi c_{p,\sigma} \rangle|,$$

except with probability

$$\eta := 2 \sum_{i | \varepsilon_i \neq 0} \exp\left( -\frac{\left(1 - \frac{\varepsilon^2}{2} - 2\frac{c_2 + \mu\|c\|_1}{c_2 + c_1}\right)^2}{2A\varepsilon_i^2} \right),$$

which leads to the following bound

$$\begin{aligned} \mathbb{E}_x \|\Psi^\star \Phi x\|_\infty^2 &\leq 2A\eta + \frac{c_1^2}{K} \sum_{i=1}^K (1 - \varepsilon_i^2/2)^2 \\ &\quad + \frac{1 - c_1^2}{K - 1} \left( A - \frac{1}{K} \sum_{i=1}^K (1 - \varepsilon_i^2/2)^2 \right). \end{aligned}$$

Since $e^{-c/\varepsilon^2}$ and therefore $\eta$ decays much faster than $\varepsilon^2$ as $\varepsilon$ goes to zero we have $\mathbb{E}_x \|\Psi^\star \Phi x\|_\infty^2 < \mathbb{E}_x \|\Phi^\star \Phi x\|_\infty^2$, as soon as $\varepsilon$ is small enough. ■

*Remark 2.2:* (i) Note that in some sense Theorem 2.1 is sharp. Assume that $\Phi$ is an orthonormal basis (ONB) then $\mu = 0$ and the condition to be a local maximum reduces to $c_1 > c_2$. However from Example 2.1 we see that if $c_1 = c_2$ we can again construct an ascent direction and so $\Phi$ is not a local maximum.

(ii) Similarly the condition that $\Phi$ is a tight frame is almost necessary in the non-trivial case where $|c_1| < 1$, as otherwise the candidate local maximiser at the generating dictionary may be distorted towards the maximal eigenvector of the frame.

### *C. A continuous model of decaying coefficients*

Next we would like to extend the result from the last subsection to a wider range of coefficient distributions, especially continuous ones. To characterise suitable distributions we make the following definitions.

*Definition 2.1:* A probability measure $\nu$ on the unit sphere $S^{d-1} \subset \mathbb{R}^d$ is called symmetric if for all measurable sets $\mathcal{X} \subseteq S^{d-1}$, for all sign sequences $\sigma \in \{-1,1\}^d$ and all permutations $p$ we have

$$\begin{aligned} \nu(\sigma\mathcal{X}) = \nu(\mathcal{X}), \quad & \sigma\mathcal{X} := \{(\sigma_1 x_1, \dots, \sigma_d x_d) : x \in \mathcal{X}\} \\ \nu(p(\mathcal{X})) = \nu(\mathcal{X}), \quad & p(\mathcal{X}) := \{(x_{p(1)}, \dots, x_{p(d)}) : x \in \mathcal{X}\} \end{aligned}$$

*Definition 2.2:* A probability distribution $\nu$ on the unit sphere $S^{K-1} \subset \mathbb{R}^K$ is called $(\beta,\mu)$-decaying if there exists a $\beta < 1/2$ such that for $c_1(x) \geq c_2(x) \geq \ldots \geq c_d(x) \geq 0$ a non increasing rearrangement of the absolute values of the components of $x$ we have,

$$\nu\left(\frac{c_2(x) + \mu\|c(x)\|_1}{c_2(x) + c_1(x)} \leq \beta\right) = 1 \tag{7}$$

For the case $\mu = 0$ it will also be useful to define the following notion. A probability distribution $\nu$ on the unit sphere $S^{d-1} \subset \mathbb{R}^d$ is called $f$-decaying if there exists a function $f$ such that

$$\begin{aligned} & \exp\left(-\frac{f(\varepsilon)^2}{8\varepsilon^2}\right) = o(\varepsilon^2) \\ \text{and} \qquad & \nu\left(\frac{c_2(x)}{c_1(x)} \geq 1 - f(\varepsilon)\right) = o(\varepsilon^2). \end{aligned}$$

Note that $(\beta, 0)$-decaying is a special case of $f$-decaying, ie. $f(\varepsilon)$ can be chosen constant $\beta$. To illustrate both concepts we give simple examples for $(\beta,\mu)$- and $f$-decaying distributions on $S^1$.

*Example 2.3:*

- Let $\nu$ be the symmetric distribution on $S^1$ defined by $c_2(x)$ being uniformly distributed on $[0, \frac{1}{\sqrt{2}} - \theta]$ for $\theta > 0$ (and accordingly $c_1(x) = \sqrt{1 - c_2^2(x)}$), then $\nu$ is $(\beta,\mu)$-decaying for all $\mu < \frac{\theta}{\sqrt{2}}$.
- Let $\nu$ be the symmetric distribution on $S^1$ defined by $c_2(x)$ being distributed on $[0, \frac{1}{\sqrt{2}}]$ with density $20\sqrt{2}(\frac{1}{\sqrt{2}} - x)^4$, then $\nu$ is $f$-decaying for e.g. $f(\varepsilon) = \sqrt{\varepsilon}$.
- Let $\nu$ be the symmetric distribution on $S^1$ defined by $c_2(x)$ being distributed on $[0, \frac{1}{\sqrt{2}}]$ with density $4(\frac{1}{\sqrt{2}} - x)$, then $\nu$ is not $f$-decaying.

With these examples of suitable probability distributions in mind we can now give a continuous version of Theorem 2.1.

*Theorem 2.2:* (a) Let $\Phi$ be a unit norm tight frame with frame constant $A = K/d$ and coherence $\mu$. If $x$ is drawn from a symmetric $(\beta,\mu)$-decaying probability distribution $\nu$ on the unit sphere $S^{K-1}$, then there is a local maximum of (5) at $\Phi$.

(b) If $\Phi$ is an orthonormal basis, there is a local maximum of (5) at $\Phi$ whenever $x$ is drawn from a symmetric $f$-decaying probability distribution $\nu$ on the unit sphere $S^{d-1}$.

### *D. Bounded white noise*

With the tools used to prove the two noiseless identification results it is also possible to analyse the case of (very small) bounded white noise.

*Theorem 2.3:* Let $\Phi$ be a unit norm tight frame with frame constant $A = K/d$ and coherence $\mu$. Assume that the signals $y$ are generated from the following model

$$y = \Phi x + r, \tag{8}$$

where $r$ is a bounded random white noise vector, ie. there exist two constants $\rho, \rho_{\max}$ such that $\|r\|_2 \leq \rho_{\max}$ almost surely, $\mathbb{E}(r) = 0$ and $\mathbb{E}(rr^\star) = \rho^2 I$. If $x$ is drawn from a symmetric decaying probability distribution $\nu$ on the unit sphere $S^{K-1}$ with $\mathbb{E}_x\|x\|_\infty^2 = \bar{c_1}^2$ and the maximal size of the noise is small compared to the size and decay of the coefficients $c_1, c_2$, meaning there exists $\beta < 1/2$, such that

$$\nu\left(\frac{c_2(x) + \mu\|c(x)\|_1 + \rho_{\max}}{c_1(x) - c_2(x)} \leq \beta\right) = 1 \tag{9}$$

then there is a local maximum of (5) at $\Phi$.

## III. Finite sample size

We are now ready to analyse the local maxima of the non-asymptotic maximisation problem for $S = 1$. For simplicity we choose a normalised version of (2).

$$\max_{\Phi \in \mathcal{D}} \frac{1}{N} \sum_{n=1}^{N} \|\Phi^\star y_n\|_\infty^2. \tag{10}$$

*Theorem 3.1:* Let $\Phi$ be a unit norm tight frame with frame constant $A = K/d$ and coherence $\mu$. Assume that the signals $y_n$ are generated as $y_n = \Phi x_n + r_n$, where $r_n$ is a bounded random white noise vector, ie. there exist two constants $\rho, \rho_{\max}$ such that $\|r_n\|_2 \leq \rho_{\max}$ almost surely, $\mathbb{E}(r_n) = 0$ and $\mathbb{E}(r_n r_n^\star) = \rho^2 I$. Further let $x_n$ be drawn from a symmetric decaying probability distribution $\nu$ on the unit sphere $S^{K-1}$ with $\mathbb{E}_x\|x\|_\infty^2 = \bar{c_1}^2$ and the maximal size of the noise be small compared to the size and decay of the coefficients $c_1, c_2$, meaning there exists $\beta < 1/2$, such that

$$\nu\left(\frac{c_2(x) + \mu\|c(x)\|_1 + \rho_{\max}}{c_1(x) - c_2(x)} \leq \beta\right) = 1. \tag{11}$$

Abbreviate $\gamma := \bar{c_1}^2 - \frac{1 - \bar{c_1}^2}{K-1}$ and $C_L = (\sqrt{A} + \rho_{\max})^2$. If for some $0 < q < 1/4$ the number of samples $N$ satisfies

$$N^{-q} + N^{-2q}/K \leq \frac{(1 - 2\beta)^2}{4A\log(4AK/\gamma)} \tag{12}$$

then except with probability

$$\exp\left(\frac{-N^{1-4q}\gamma^2}{4K^2 C_L^2} + Kd\log(NKC_L/\gamma)\right),$$

there is a local maximum of (10) resp. local minimum of (1) within distance at most $2N^{-q}$ to $\Phi$, ie. for the local maximum $\tilde{\Psi}$ we have $\max_k \|\tilde{\psi_k} - \phi_k\|_2 \leq 2N^{-q}$.

*Proof:* We again give a brief sketch of the proof. From the last section we know that for any $\varepsilon$-perturbation we have

$$\mathbb{E}_y\|\Phi^\star y\|_\infty^2 - \mathbb{E}_y\|\Psi^\star y\|_\infty^2 \approx \varepsilon^2\gamma/K.$$

Hoeffding's inequality lets us estimate the probability that for a fixed perturbation the finite sample sum deviates from its expectation as

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N} \|\Psi^\star y_n\|_\infty^2 - \mathbb{E}_y\|\Psi^\star y\|_\infty^2\right| > t\right) \leq e^{-Nt^2/C_L^2}.$$

Using a union bound this leads to an estimate for the probability that the above holds for a $\delta$-net $\mathcal{N}$ for the set of all $\varepsilon$-perturbations with $\varepsilon \leq \varepsilon_{\max}$. Since this set is the product of $K$ $(d-1)$-dimensional balls with radius $\varepsilon_{max}$ we have

$$\sharp\mathcal{N} \leq (3\varepsilon_{\max}/\delta)^{K(d-1)}.$$

Choosing $\delta$ and $t$ to be $O(N^{-q})$ the final result then follows from a triangular inequality argument and the fact that

$$\left| \|\Psi^\star y_n\|_\infty^2 - \|\bar{\Psi}^\star y_n\|_\infty^2 \right| \leq 3C_L \max_k \|\psi_k - \bar{\psi}_k\|_2.$$

■

## IV. Discussion

We have shown that the K-SVD minimisation principle with sparsity parameter 1 can correctly identify a tight frame from signals generated from a wide class of decaying coefficients distributions. Since any simple greedy algorithm will always find the best 1-term approximation for any signal in any dictionary our results give conditions under which the K-SVD algorithm can identify the underlying dictionary given a good initialisation.

Before turning to a comparison of our results to other dictionary learning schemes we illustrate the limitations of the K-SVD principle for learning non-tight frames. We generated 1000 coefficients by drawing $c_2(x)$ uniformly at random from the interval $[0, 0.6]$, setting $c_1(x) = \sqrt{1 - c_2^2(x)}$, randomly permuting the resulting vector and providing it with random ± signs. We then generated two sets of signals, using an orthogonal and an oblique basis with the same coefficients, and for both sets of signals found the minimiser of the K-SVD criterion (1) with $S = 1$. Figure 1 shows the two signal sets, the generating bases and the recovered bases. As predicted by our theoretical results when the generating basis is orthogonal it is also the minimiser of the K-SVD criterion, while for the oblique generating basis the minimiser is distorted towards the maximal eigenvector of the basis.

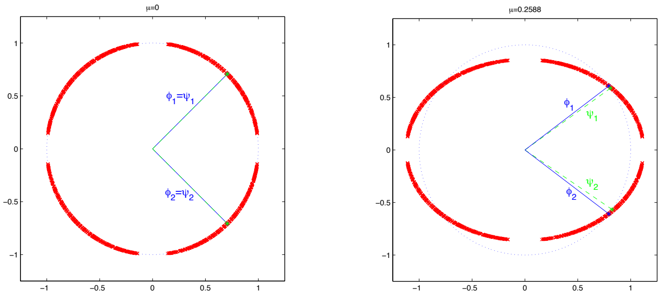

Fig. 1. Signals created from an orthogonal and an oblique basis $\Phi = (\phi_1, \phi_2)$ with decaying coefficients, together with the corresponding minimiser $\Psi = (\psi_1, \psi_2)$ of the K-SVD-criterion for $S = 1$.

Finally let us point out further research directions based on a comparison of our results for the K-SVD-minimisation principle to the identification results for the $\ell_1$-minimisation principle,

$$\min_{\Phi \in \mathcal{D}, X : Y = \Phi X} \sum_{ij} |X_{ij}|, \tag{13}$$

derived in [5], [6]. At first glance it seems that the K-SVD-criterion requires a larger sample size than the $\ell_1$-criterion, ie. $N^{1-4q}/\log N = O(K^3 d)$ as opposed to $O(d^2 \log d)$ reported in [5] for a basis and $O(K^3)$ reported in [6] for an overcomplete dictionary. Also it does not allow for exact identification with high probability but only guarantees stability. However this effect may be due to the more general signal model which assumes decay rather than exact sparsity. Indeed it is very interesting to compare our results to a recent result for a noisy version of the $\ell_1$-minimisation principle, [7], which provides stability results under unbounded white noise and, omitting log factors, also derives a sampling complexity of $O(K^3 d)$.

Another difference, apparently intrinsic to the minimisation criteria is that the K-SVD criterion can only identify tight dictionary frames exactly, while the $\ell_1$-criterion allows identification of arbitrary dictionaries. Thus to support the use of K-SVD for the learning of non-tight dictionaries also theoretically, we plan to study the stability of the K-SVD criterion under non-tightness by analysing the maximal distance between an original, non tight dictionary with condition number $\sqrt{B/A} > 1$ and the closest local maximum, cp. also Figure 1.

The last research direction we want to point out is how much decay of the coefficients is actually necessary. For the asymptotic results we used condition $c_1 > c_2 + 2\mu\|c\|_1$ to ensure that the maximal inner product is always attained at $i_p$. However typically we have $|\langle \phi_i, \Phi c_{p,\sigma} \rangle| \approx c_{p(i)} \pm \mu$. Therefore a condition such as $c_1 > c_2 + O(\mu)$, which allows for outliers, ie. signals for which the maximal inner product is not attained at $i_p$, might be sufficient to prove exact identifiability or - failing that - to again show stability. Together with the inspiring techniques from [7], we expect the tools developed in the course of such an analysis to allow us also to deal with unbounded white noise.

## References

[1] K. Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD," *arXiv:1301.3375*, 2013.

[2] ——, "Sampling complexity of dictionary learning via the k-svd-minimisation criterion," *in preparation*, 2013.

[3] M. Aharon, M. Elad, and A. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Journal of Linear Algebra and Applications*, vol. 416, pp. 48–67, July 2006.

[4] P. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 992–996, 2005.

[5] R. Gribonval and K. Schnass, "Dictionary identifiability - sparse matrix-factorisation via $l_1$-minimisation," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3523–3539, July 2010.

[6] Q. Geng, H. Wang, and J. Wright, "On the local correctness of $\ell^1$-minimization for dictionary learning," *arXiv:1101.5672v1 [cs.IT]*, 2011.

[7] R. Jenatton, F. Bach, and R. Gribonval, "Local stability and robustness of sparse dictionary learning in the presence of noise," *preprint*, 2012.

[8] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing.*, vol. 54, no. 11, pp. 4311–4322, November 2006.