Sampling theorems for signals from the union of finite-dimensional linear subspaces

Citation for published version:
Blumensath, T & Davies, ME 2009, 'Sampling theorems for signals from the union of finite-dimensional linear subspaces', *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1872-1882. https://doi.org/10.1109/TIT.2009.2013003

Digital Object Identifier (DOI):
10.1109/TIT.2009.2013003

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

Published In:
IEEE Transactions on Information Theory

Sampling Theorems for Signals from the Union of Finite-Dimensional Linear Subspaces

Thomas Blumensath, *Member, IEEE,* Mike E. Davies, *Member, IEEE*

Abstract

Compressed sensing is an emerging signal acquisition technique that enables signals to be sampled well below the Nyquist rate, given that the signal has a sparse representation in an orthonormal basis. In fact, sparsity in an orthonormal basis is only one possible signal model that allows for sampling strategies below the Nyquist rate. In this paper we consider a more general signal model and assume signals that live on or close to the union of linear subspaces of low dimension. We present sampling theorems for this model that are in the same spirit as the Nyquist-Shannon sampling theorem in that they connect the number of required samples to certain model parameters.

Contrary to the Nyquist-Shannon sampling theorem, which gives a necessary and sufficient condition for the number of required samples as well as a simple linear algorithm for signal reconstruction, the model studied here is more complex. We therefore concentrate on two aspects of the signal model, the existence of *one to one* maps to lower dimensional observation spaces and the smoothness of the inverse map. We show that *almost all* linear maps are *one to one* when the observation space is at least of the same dimension as the largest dimension of the convex hull of the union of *any two* subspaces in the model. However, we also show that in order for the inverse map to have certain smoothness properties such as a given finite Lipschitz constant, the required observation dimension necessarily depends logarithmically

The authors are with the Institute for Digital Communication and with the Joint Research Institute for Signal and Image Processing, Edinburgh University, King's Buildings, Mayfield Road, Edinburgh EH9 3JL, UK (Tel.: +44(0)131 6505659, Fax.: +44(0)131 6506554, e-mail: thomas.blumensath@ed.ac.uk, mike.davies@ed.ac.uk).

This research was supported by EPSRC grants D000246/1 and D002184/1. MED acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the Heriot-Watt University as a component part of the Edinburgh Research Partnership.

This is a corrected version of the papers in which a few small errors have been corrected. Importantly, the dependence on δ in Theorem 3.3 and its corollaries has been corrected.

on the number of subspaces in the signal model. In other words, whilst unique linear sampling schemes require a small number of samples depending only on the dimension of the subspaces involved, in order to have stable sampling methods, the number of samples depends necessarily logarithmically on the number of subspaces in the model. These results are then applied to two examples, the standard compressed sensing signal model in which the signal has a sparse representation in an orthonormal basis and to a sparse signal model with additional tree structure.

Index Terms

Compressed sensing, unions of linear subspaces, sampling theorems, embedding and restricted isometry

I. INTRODUCTION

A. *Compressed sensing*

Since Nyquist [1] and Shannon [2] we are used to sampling continuous signals at a rate that is twice the bandwidth of the signal. However, during the last decades, the focus has shifted and the problem of recovering signals from fewer measurements than would be required by the Nyquist rate has been posed [3]. Over the last few years interest in this problem has dramatically increased, fuelled by several recent publications, including the work by Vetterli et al. [4] and by Eldar [5] on sampling of continuous signals and the seminal papers by Candes, Romberg and Tao [6], [7], [8] and by Donoho [9] on sampling of signals with finite dimensional discrete representations. Many of these publications have given theoretical justification for many of the previously proposed approaches. In particular, it is now known that finite dimensional signals with certain structures (to be made more concrete below) can be sampled at a lower rate without incurring any loss of information. While the sampling operation is a simple linear mapping, the reconstruction becomes non-trivial. The papers by Candes, Romberg and Tao [6], [7], [8] and by Donoho [9] have shown that under certain conditions on the signal structure and the sampling operator (which are often satisfied by certain random matrices), the original signal can be reconstructed using weakly polynomial time algorithms.

The problem can be formulated as follows. A continuous or discrete signal f from an $N < \infty$ dimensional Hilbert space is to be sampled. This is done by using M linear measurements $\{\langle f, \phi_n \rangle\}_n$, where $\langle \cdot, \cdot \rangle$ is the inner product and where $\{\phi_n\}$ is a set of N dimensional vectors from the Hilbert space under consideration. Let $\mathbf{x}$ be the vector of elements x_i such that $f = \sum_{i=1}^{N} \psi_i x_i$ for some orthonormal basis ψ_i of the signal space. As f and $\mathbf{x}$ are equivalent, we will from now on assume that $\mathbf{x}$ is the signal.

Let $\Phi \in \mathbb{R}^{M \times N}$ be the matrix with entries $\langle \psi_i, \phi_j \rangle$. The observation can then be written as

$$\mathbf{y} = \mathbf{\Phi} \mathbf{x}. \tag{1}$$

In compressed sensing it is paramount to consider signals $\mathbf{x}$ that are highly structured and in the original papers, $\mathbf{x}$ was assumed to be an *exact k-sparse* vector, i.e. a vector with not more than k non-zero entries. Related models of $\mathbf{x}$ were considered in [9], [10], [11], [12], [13] and [14] and extensions to noisy observations were presented in [10], [15] and [16].

B. Unions of Linear Subspaces

In this paper we consider a quite general signal model and assume the signal $\mathbf{x}$ to be an element from a union of linear subspaces $\mathcal{A}$, defined formally as

$$\mathcal{A} = \bigcup_j^L S_j, \;\; S_j = \{\mathbf{x} = \Omega_j \mathbf{a}, \Omega_j \in \mathbb{R}^{N \times k_j}, \mathbf{a} \in \mathbb{R}^{k_j}\}, \tag{2}$$

where the Ω_j are bases for linear subspaces and where $k_j \leq k < \infty$. This model is a special case of the union of subspaces model introduced in [17] with the difference that we will restrict the discussion to mixtures of finitely many finite dimensional subspaces, i.e we assume $L < \infty$.

If $S_{i,j}$ is the convex hull of the set $S_i \bigcup S_j$, we define

$$k_{\max} = \max_{i \neq j} \dim(S_{i,j}), \tag{3}$$

that is, the maximum dimension of the convex hull of the unions of two distinct subspaces in the model. In this paper we follow a similar approach to that in [17] and rely heavily on the difference between any two vectors $\mathbf{x}_i - \mathbf{x}_j$, both from $\mathcal{A}$. The vectors $\mathbf{x}_i - \mathbf{x}_j$ lie in a union of subspaces $S_{i,j}$ and $k_{\max}$ gives the largest dimensions of any of these subspaces.

This quite general signal model incorporates many, though not all, previously considered compressed sensing settings. It includes, for example

- The 'traditional', noiseless k-sparse model as considered by Candes et al. in for example [6] and by Donoho in [9], in which $\mathbf{x}$ is assumed to have no more than k non-zero elements.
- The set of vectors with a k-sparse representation in a general, possibly overcomplete and non-orthogonal dictionary Ψ as considered in [12] and [13], i.e. $\mathbf{x} = \Psi \mathbf{z}$, where Ψ is a general matrix with unit norm columns and possibly more columns than rows and where $\mathbf{z}$ is a vector with no more than k non-zero elements.
- The set of k-sparse signals in which the non-zero elements form a tree as considered in [18].

- The simultaneous sparse approximation problem [19] [20] [21] [22] [23], where a number of observations $\mathbf{y}_i$ is assumed to follow the model $\mathbf{y}_i = \mathbf{\Phi}_i \mathbf{x}_i$ where the $\mathbf{x}_i$ are constrained to have the same non-zero elements.
- The set consisting of the union of statistically independent k-dimensional subspaces S_i as considered by Fletcher et al. in [14].

C. Contribution

Many of the previous papers in compressed sensing have addressed two important aspects, namely, the specification of conditions that guarantee an efficient reconstruction of the original signal from the measurement samples and practical constructions of measurement ensembles. In this paper we will study the problem of compressed sensing of signals which are known to lie in $\mathcal{A}$. In particular we address two fundamental aspects, for each of which, the primary question is the relationship between the required observation space dimension as a function of both, the maximum dimension of any of the subspaces as well as the total number of subspaces.

The first aspect studied is that of characterising linear maps that map each $\mathbf{x} \in \mathcal{A}$ to a unique observation $\mathbf{y}$. We here study necessary and sufficient conditions for the existence of such *one to one* maps. This *one to one* property of $\mathbf{\Phi}$ for elements of $\mathcal{A}$ is clearly an important aspect in sampling as it specifies a 'minimal' requirement that allows us to sample a signal without loss of information. Whilst this property has previously been studied in [17], our first main contribution is to show that under appropriate conditions *almost all* linear maps are *one to one*. For the mixture model with finitely many finite dimensional subspaces, our results are therefore stronger than those derived in [17], which only states that the set of *one to one* maps is a dense subset of all linear maps.

Fletcher et al. also studied a similar formulation of the problem but assumed statistically independent subspaces. They have shown that for this probabilistic model and for certain probability models on $\mathbf{x}$, *almost all* $\mathbf{x}$ can be mapped *one to one* under even milder conditions. In this paper we show that a similar result also holds for the more general union of subspaces model considered here.

The second important aspect addressed is a theoretical characterisation of the inverse map. Here we are particularly interested in the Lipschitz property of this inverse map and we derive conditions for the existence of a bi-Lipschitz embedding from $\mathcal{A}$ into a subset of $\mathbb{R}^N$. This Lipschitz property is an important aspect of the map which ensures stability and controls the behaviour of any reconstruction[1] to perturbations of the observation, i.e. it controls the amount by which small perturbations are amplified in the reconstruction. This in effect specifies the robustness of compressed sensing against noise, quantisation errors and perturbations of the signal from the exact subspace model. Furthermore, in the k-sparse model, the Lipschitz property is also an important aspect for the existence of efficient and robust reconstruction algorithms.

Whilst sufficient conditions for the existence of Lipschitz inverses have been extensively studied (however, under a different name) in the context of k-sparse signal models, necessary conditions have to our knowledge not been reported (see however the discussion below). The derivation of such conditions, for the general model considered here, constitute the second main contribution of this paper. In particular, we derive novel sufficient conditions for the existence of maps whose inverse has a specific Lipschitz constant. In the special case of k-sparse signals, the theorem reduces to well known results.

D. Notation

The set $\mathcal{A}$ will denote the union of L subspaces in an N dimensional *ambient space*. Each subspace will have dimension not more than k and is often denoted by S_i. When talking about dimension in this paper, we in general mean the box counting dimension. Let $N(\epsilon)$ be the minimum number of boxes, each of side length ϵ, required to cover a given set. The box counting dimension is then defined as [24, p. 185]

$$\mathrm{dim}_{\mathrm{box}} := \lim_{\epsilon \to 0} \frac{\log(N(\epsilon))}{\log(1/\epsilon)}. \tag{4}$$

For linear subspaces, this is equivalent to the normal Euclidean notion of dimension. The set of signals $\mathbf{x}$ will be assumed to be taken from the set $\mathcal{A}$. The linear map $\mathbf{\Phi}$ will map any element from $\mathcal{A}$ into an M dimensional *observation space*, elements of which will be denoted by $\mathbf{y}$. We will often use the notation $\mathcal{B}_\rho^N(p)$ to refer to the ball in $\mathbb{R}^N$, i.e. the set of points $\{\mathbf{x} : \|p - \mathbf{x}\|_2 \leq \rho, \mathbf{x} \in \mathbb{R}^N\}$. If $p = \mathbf{0}$ we will write $\mathcal{B}_\rho^N$. A similar notation is used for the sphere which is denoted by $\mathcal{S}_\rho^{N-1}(p)$, which will be the sphere living in $\mathbb{R}^N$.

[1] It is important to stress that this is a property of the inverse map itself, which is uniquely (but indirectly) defined for any *one to one* map. This property has therefore nothing to do with any particular algorithm one might design to (approximately) calculate this inverse map.

E. Paper Overview

The first main section of this paper, section II, derives two theorems that give conditions under which the map $\mathbf{\Phi}$ is *one to one* for elements from $\mathcal{A}$. The developed theory is strongly inspired by work on embedding theory, some of the relevant results of which are reviewed in subsection II-A. This is followed by two subsections stating the main results for the existance of *one to one* maps for the two conditions $M \geq k_{\max}$ (subsection II-B) and $k < M \leq k_{\max} - 1$ (subsection II-C), where $k_{\max}$ is defined in (3). In section III we tighten the requirements on $\mathbf{\Phi}$. Not only do we require $\mathbf{\Phi}$ to be *one to one* for elements of $\mathcal{A}$, we further assume that Φ and its inverse have certain properties such as a given Lipschitz constant. This requirement leads to stricter necessary as well as to more stringent sufficient conditions on the number of observations to be taken. To demonstrate the generality of the results of section III, section IV looks at two particular cases that fit the union of subspace model studied. The first case is the standard k-sparse signal model traditionally considered in compressed sensing (subsection IV-A) while the second example is a k-sparse signal model in which non-zero coefficients are constrained to form a tree structure (subsection IV-B). Most of the proofs are stated in the appendices.

II. Existence of a Unique Inverse Map

One quite natural property to be required from any signal acquisition or sampling system is that the system preserves (at least most) of the information contained in the signal. In compressed sensing it is therefore often required that the system maps any $\mathbf{x}$ from the set under consideration to a unique observation $\mathbf{y}$. Under this condition, knowledge of $\mathbf{y}$ is, at least in theory, equivalent to knowledge of $\mathbf{x}$.

A map $\mathbf{\Phi}$ that maps different points $\mathbf{x}$ to unique vectors $\mathbf{y}$ is said to be *one to one*. In this section we derive conditions under which $\mathbf{\Phi}$ is *one to one* for all $\mathbf{x} \in \mathcal{A}$.

This section considers sufficient conditions relating M and k. We here distinguish two cases, $M \geq k_{\max}$ and $k < M \leq k_{\max} - 1$. We prove that *almost all* linear maps are *one to one* whenever $M \geq k_{\max}$, whilst for $k < M \leq k_{\max} - 1$ it can be shown that *almost all* linear maps are *one to one* for *almost all* $\mathbf{x} \in \mathcal{A}$ (where *almost all* requires the definition of a smooth measure on $\mathcal{A}$). However, before stating the main results, the next subsection recalls some motivating results from embedding theory, which form the basis of the main theorems.

A. Embedding of low dimensional compact sets

Whitney's Embedding Theorem [25, chapter 10] states that "Every compact metrizable k-dimensional topological space (or alternatively every smooth k-manifold) can be embedded into $\mathbb{R}^M$ if $M > 2k$".

Dimension here generally refers to the (Lebesgue) covering dimension as defined in for example [24, p. 96].

An extension of this theorem by Sauer et al. [26] is

Theorem 2.1 (Theorem 2.3 [26]): Assume a compact subset $\mathcal{C}$ of R^N with box counting dimension k and let $M > 2k$, then *almost all* smooth maps F from $\mathbb{R}^N$ to $\mathbb{R}^M$ have the property that F is *one to one* on $\mathcal{C}$.

As pointed out by Sauer et al., the space of smooth maps is infinite-dimensional and there is no Lebesgue measure on such a space. The term “almost all” in the above theorem has therefore to be understood in terms of prevalence [27]. As we are dealing with finite dimensional spaces of linear maps in this paper, we will from now on use the term “almost surly” to mean that the complement will have Lebesgue measure zero. The distinction between this definition and that in terms of prevalence is therefore not required here and we refer the interested reader to the original literature cited above.

These results can be seen in terms of a quite general compressed sensing problem. Assume that the data lives on a k dimensional compact subset of the data-space. It is then clear that we would only need $M > 2k$ observations to exactly specify the data. This suggests an extension of compressed sensing to more general low dimensional data structures. An example of this, where the data was assumed to lie on a smooth manifold was already considered by Baranuick and Wakin in [28]. The above theorem further suggests the use of non-linear measurements, i.e. the use of smooth maps. To our knowledge, such maps have not been considered for compressed sensing so far.

It is important to note that Whitney’s result and Sauer’s extension hold for general low dimensional compact manifolds as well as general smooth embeddings (not necessarily linear). We show below that, in the case of a unions of k-dimensional linear subspaces and for linear embeddings, we can actually get an embedding into $\mathbb{R}^M$ if $M \geq 2k$ (rather than the strict inequality of the above theorems).

B. The case $M \geq k_{\max}$

The theorem by Sauer et al. sheds new light on the more traditional compressed sensing problem that considers signals that are well approximated as lying on the union of linear subspaces and mappings that are assumed to be linear. In this context, the paper of Lu and Do [17] derived results related to those of Sauer et al. In particular [17] shows that the set of maps from the union of linear subspaces into R^M with $M \geq k_{\max}$ is dense. For the k sparse model, [3] (see also Corollary 4 in [29]) presented the following result

Theorem 2.2: A linear system $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$ for which all possible combinations of M of the N columns of $\mathbf{\Phi}$ are linearly independent and for which $\mathbf{x}$ has $k < M/2$ non-zero elements does not admit any other solution with less than $M - k + 1$ non-zero elements.

The property[2] that all combinations of M of the N columns of $\mathbf{\Phi}$ are linearly independent was termed the Unique Representation Property (URP) in [3]. Combining this theorem with the fact that *almost all* linear maps have the URP gives a similar result to the one we derive below, but for the special case of k-sparse signals. See also lemma 2.1 in [30] for an alternative statement of the same result. A related result has also been presented in [31, Theorem 2.1] in which sparse signals are considered and in which the measurement ensemble was a matrix with Gaussian distributed entries.

We derive our result based on the analysis of Sauer et al. and show that for signals from finite unions of low dimensional linear subspaces, *almost all* linear maps are *one to one* whenever $M \geq k_{\max}$. This result is a restriction of the more general results of Sauer et al. to unions of linear subspaces and linear mappings and allows us to reduce the required observation dimension by one and therefore extends the results from [17] and [31]. In particular, we can use a slight variation of the proof used by Sauer et al. [26] to derive the following theorem

Theorem 2.3: Almost all linear maps $\mathbf{\Phi} : \mathcal{A} \to R^M$ are *one to one* if $M \geq k_{\max}$.

The somewhat lengthy and rather involved proof can be found in Appendix I. In [17] a very similar result was derived for a countably infinite union of subspaces. The difference to our result is that for the finite union of subspaces, our theorem states that *almost all* maps have the desired property, whilst in [17] it was shown that the set of maps with the property is dense, which is a slightly weaker statement (though derived for a more general model) as density of a set does not imply anything about the measure of the set.

This theorem tells us that, not only is there a map that will map the union of subspaces of interest *one to one* into an $M \geq k_{\max}$ dimensional observation space, but also, that *almost all* linear maps will do. Therefore, if we chose the maps at random (as is often advertised in compressed sensing) we will find such a map with probability one. We are then guaranteed that there also exists a *unique* inverse map that will get us back to the original signal. However, the theorem does not give any insight into the behaviour of this inverse map, which is the topic of the next section.

It is also important to note that the above theorem is tight as is shown by the following necessary condition.

Theorem 2.4: A necessary condition for the map $\mathbf{\Phi} : \mathcal{A} \to R^M$ to be *one to one* is that $M \geq k_{\max}$.

[2] Note that this property is equivalent to $spark(\mathbf{\Phi}) = M + 1$, where $spark$ is defined as in [29].

This is Proposition 3 in [17]. For completeness, a proof can also be found in Appendix II.

C. The case $k < M \leq k_{\max} - 1$

The paper [26] sports a further theorem that is of interest for compressed sensing.

Theorem 2.5 (Self-Intersection Theorem [26]): For any compact subset $\mathcal{C}$ of a metric space R^N, let $\mathcal{C}$ have box-counting dimension k and let $M \leq 2k$ be an integer. For any δ and *almost all* smooth maps $F : \mathcal{A} \rightarrow R^M$, the set

$$\{\mathbf{x}_1 \in A : \exists \mathbf{x}_2 \in A, F(\mathbf{x}_1) = F(\mathbf{x}_2), \|\mathbf{x}_1 - \mathbf{x}_2\|_2 \geq \delta\} \tag{5}$$

has box-counting dimension at most $2k - M$.

The self-intersection theorem of Sauer et al. hints at a further possible reduction in the number of required observations if we can specify a smooth measure on each of the subspaces. We can decompose $\mathcal{A} = \bigcup_{i=1}^{k} \mathcal{A}_i$, where the sets $\mathcal{A}_i$ are unions of subspaces of dimension exactly i. We define a measure on $\mathcal{A}$ by assuming a measure defined on each of the $\mathcal{A}_i$ in one of the possible decomposition of $\mathcal{A}$. The measures on $\mathcal{A}_i$ are assumed to be such that all subsets of $\mathcal{A}_i$ of dimension less than i are of measure zero[3].

The argument given in Appendix III then proves the following theorem.

Theorem 2.6: Assume a measure defined on $\mathcal{A}$ as outlined above, then *almost all* linear maps $\mathbf{\Phi} : R^N \rightarrow R^M$ are *one to one* on *almost all* elements of $\mathcal{A}$ whenever $k < M \leq k_{\max} - 1$.

While the theorem in the previous subsection was valid for all $\mathbf{x} \in \mathcal{A}$, the theorem in this subsection assumes that the elements $\mathbf{x}$ are drawn randomly from $\mathcal{A}$.

Again note that a similar result for sparse signals and Gaussian measurement ensembles has been presented in [31, Theorem 2.1]. Our results again extend these results to more general linear subspaces and to *almost all* linear measurements. In this context, it is also interesting to note the result by Fletcher et al. [14] who considered a mixture of statistically independent k dimensional subspaces with a Gaussian measure on each subspace. Using information theoretic arguments, they have shown that with probability one compressed sensing does not lose any information for their model, whenever $M > k$. Our results are a more general version of this observation.

[3] For example, this models includes the standard sparse signal model in which one randomly selects the number of non-zero coefficients (constrained to be no more than k), their position and their value.

III. Properties of the Inverse Map

While the existence of a *one to one* map is important as it guarantees the existence of a unique inverse map, in practical applications more stringent requirements are often called for. Two such requirements are that the observations are robust to noise and the existence of efficient algorithms for the recovery of the original signal. Both of these properties can be shown to be strongly related to distance preserving properties of the map. For example, it is not only required that two distinct signals are mapped to distinct observations, it is also important that the distance between distinct signals is not changed too much in the observation domain [6], [10], [16], [32], [16]. A mathematical tool to measure this property for k-sparse signals is the k-restricted isometry constant δ_k [33] defined as follows

Definition: (k-restricted isometry) For any matrix Φ and integer k we define the k-restricted isometry constant $\delta_k(\Phi)$ to be the smallest quantity such that

$$(1-\delta_k(\Phi)) \leq \frac{\|\Phi\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} \leq (1+\delta_k(\Phi)), \tag{6}$$

holds for all $\mathbf{x}$ with no more than k non-zero elements.

The restricted isometry constant δ_k bounds the amount by which the length of any K-sparse vector is changed by $\mathbf{\Phi}$. If we normalise $\mathbf{\Phi}$, such that both inequalities in (6) are tight, then, if $\delta_{2K}=0$, then there are vectors for which no change takes place. In order for each K-sparse signal to admit a one to one map, it is therefore necessary that there exist a normalisation of $\mathbf{\Phi}$ such that $\delta_{2K}<1$ for otherwise there would be a $2k$ sparse vector $\mathbf{x}$ that is the linear combination of two k-sparse vectors $\mathbf{x}=\mathbf{x}_1-\mathbf{x}_2$ for which $\mathbf{\Phi}\mathbf{x}=0$, which would imply that $\mathbf{\Phi}$ cannot be *one to one* for all k-sparse signals, i.e. we would have $\mathbf{\Phi}\mathbf{x}_1=\mathbf{\Phi}\mathbf{x}_2$. We therefore have the following corollary to Theorem 2.3:

Corollary 3.1: If $M>2k-1$, then *almost all* linear maps $\mathbf{\Phi}$ can be normalised such that $\delta_{2k}(\mathbf{\Phi})<1$.

A natural extension of the k-restricted isometry for the more general union of subspaces model would be

Definition: ($\mathcal{A}$-restricted isometry) For any matrix Φ and any subset $\mathcal{A}\subset\mathbb{R}^N$ we define the $\mathcal{A}$-restricted isometry constant $\delta_{\mathcal{A}}(\Phi)$ to be the smallest quantity such that

$$(1-\delta_{\mathcal{A}}(\Phi)) \leq \frac{\|\Phi\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} \leq (1+\delta_{\mathcal{A}}(\Phi)), \tag{7}$$

holds for all $\mathbf{x}\in\mathcal{A}$.

If we define the set $\bar{\mathcal{A}}=\{\mathbf{x}=\mathbf{x}_1+\mathbf{x}_2:\mathbf{x}_1,\mathbf{x}_2\in\mathcal{A}\}$ (Note that $\delta_{\bar{\mathcal{A}}}(\Phi)\geq\delta_{\mathcal{A}}(\Phi)$.), then a similar argument to the one above shows that a necessary condition for the existence of a *one to one* map would require that $\delta_{\bar{\mathcal{A}}} < 1$ for otherwise there would exist $\mathbf{x}_1 \neq \mathbf{x}_2$ such that $\mathbf{\Phi}\mathbf{x}_1 = \mathbf{\Phi}\mathbf{x}_2$, i.e. such that $\mathbf{\Phi}\mathbf{x}_1 - \mathbf{\Phi}\mathbf{x}_2 = 0$. We again get a corollary to Theorem 2.3.

Corollary 3.2: If $M \geq k_{\max}$, then *almost all* linear maps $\mathbf{\Phi}$ can be normalised such that $\delta_{\bar{\mathcal{A}}}(\Phi) < 1$.

Interestingly, this result does not depend on the ambient dimension of the signal space nor on the number of subspaces in the signal model. However, the corollary only guarantees that *almost all* (suitably normalised) linear maps of correct dimension satisfy $\delta_{\bar{\mathcal{A}}}(\Phi) < 1$, but allows $\delta_{\bar{\mathcal{A}}}(\Phi)$ to get arbitrarily close to one. We therefore turn now to a more stringent requirement on $\delta_{\bar{\mathcal{A}}}(\Phi)$, namely we require $\delta_{\bar{\mathcal{A}}}(\Phi) \leq c < 1$ for some given constant c. The theorems of this section show that the existence of a fixed restricted isometry constant requires a logarithmic dependence on the number of subspaces.

One important interpretation of the restricted isometry constant is in terms of the "smoothness" of the inverse map from $\mathbf{y}$ to $\mathbf{x}$. Denote the inverse map by $f(\mathbf{y})$ say. This "smoothness" can, for example, be measured by the Lipschitz constant K_I (I for inverse), defined as the smallest number K_I such that

$$\|f(\mathbf{y}_1) - f(\mathbf{y}_2)\|_2 \leq K_I \|\mathbf{y}_1 - \mathbf{y}_2\|_2. \tag{8}$$

If $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{A}$ and if $\delta_{\bar{\mathcal{A}}} < 1$, then we can write

$$\|\mathbf{x}_1 - \mathbf{x}_2\|_2 = \|f(\mathbf{y}_1) - f(\mathbf{y}_2)\|_2 \leq \frac{1}{\sqrt{1-\delta_{\bar{\mathcal{A}}}}} \|\mathbf{y}_1 - \mathbf{y}_2\|_2, \tag{9}$$

i.e. $\frac{1}{\sqrt{1-\delta_{\bar{\mathcal{A}}}}}$ is a bound on the Lipschitz constant K_I of the inverse map from the observations to the signal space. Similarly, the Lipschitz constant K_F (F for forward) of the map $\mathbf{\Phi}$ (defined as the map from $\mathcal{A}$ to $\mathbf{\Phi}(\mathcal{A})$) can be bound by $\sqrt{1+\delta_{\bar{\mathcal{A}}}}$. It is important to note that in the definition of the Lipschitz constant K_F used throughout this paper, we consider $\mathbf{\Phi}$ to be restricted to elements from $\mathcal{A}$. One should therefore think of K_F as a *restricted* Lipschitz constant.

A. A sufficient condition for the existence of a $\mathbf{\Phi}$ with required $\delta_{\mathcal{A}}$

We first state a sufficient condition that guarantees the existence of a fixed $\delta_{\mathcal{A}} < 1$. The proof can be found in Appendix IV

Theorem 3.3: For any $t > 0$, let

$$M \geq \frac{1}{c(\delta_{\mathcal{A}}/6)} \left(\ln(2L) + k \ln\left(\frac{36}{\delta_{\mathcal{A}}}\right) + t \right), \tag{10}$$

then there exist a matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ and a function $c(\delta) > 0$ depending only on δ such that

$$(1-\delta_{\mathcal{A}})\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Phi}\mathbf{x}\|_2^2 \leq (1+\delta_{\mathcal{A}})\|\mathbf{x}\|_2^2 \tag{11}$$

holds for all $\mathbf{x}$ from the union of L arbitrary k dimensional subspaces $\mathcal{A}$. What is more, if $\mathbf{\Phi}$ is generated by randomly drawing i.i.d. entries from an appropriately scaled subgaussian distribution[4], then this matrix satisfies equation (11) with probability at least

$$1 - e^{-t}. \tag{12}$$

The function $c(\delta)$ then only depends on the distribution of the entries in $\mathbf{\Phi}$ and is $c(\delta) = \delta^2/4 - \delta^3/6$ if the entries of $\mathbf{\Phi}$ are i.i.d. normal or if the entries are either $1/\sqrt{N}$ or $-1/\sqrt{N}$ with equal probability. Note that, in contrast to the results of the previous section, this sufficient condition is logarithmic in the number of subspaces considered. In the next subsection we show that this logarithmic dependence is in fact necessary.

We have here stated the results for a general union of subspace model $\mathcal{A}$. By choosing the appropriate values for L and k, results for $\mathcal{A}$, $\bar{\mathcal{A}}$ or other unions of subspaces can be derived. For example, assume $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{A}$. The vectors $\mathbf{x}_i - \mathbf{x}_j$ lie in $\bar{\mathcal{A}}$, which is the union of $\bar{L} = (L^2 - L)/2$ subspaces of dimension no more than $k_{\max}$. We therefore get the following corollary

Corollary 3.4: For any $t > 0$, let

$$M \geq \frac{1}{c(\delta_{\bar{\mathcal{A}}}/6)} \left(\ln(2\bar{L}) + k_{\max} \ln\left(\frac{36}{\delta_{\bar{\mathcal{A}}}}\right) + t \right), \tag{13}$$

then there exists a matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ and a function $c(\delta) > 0$ such that

$$K_I \leq \frac{1}{\sqrt{1 - \delta_{\bar{\mathcal{A}}}}}, \tag{14}$$

and

$$K_F \leq \sqrt{1 + \delta_{\bar{\mathcal{A}}}}. \tag{15}$$

If $\mathbf{\Phi}$ is generated by randomly drawing i.i.d. entries from an appropriately scaled subgaussian distribution, the matrix will satisfy the conditions with probability at least $1 - e^{-t}$.

B. A necessary condition for the existence of a $\mathbf{\Phi}$ with required $\delta_{\bar{\mathcal{A}}}$

In this subsection we study a necessary condition for the existence of a map $\mathbf{\Phi}$ with a required Lipschitz constant for points in $\mathcal{A}$ and with an inverse, also with a given Lipschitz constant. To derive this result, we relate the distance between p unit norm vectors packed optimally into $\mathcal{A}$ to the number of required

[4] Examples of these distributions include the Gaussian distribution and random variables that are $\pm\frac{1}{\sqrt{N}}$ with equal probability [34] [12].

observations. To formalise the notion of optimal packing, we define the *maximum packing distance* of a set $\mathcal{A}$ as follows

Definition: (maximum packing distance $\omega_p(\mathcal{A})$) Let P_p be a set of p unit norm vectors from $\mathcal{A}$. For any set $\mathcal{A}$, the maximum packing distance is defined as

$$\omega_p(\mathcal{A}) = \max_{P_p} \left[\min_{\substack{\mathbf{x}_i, \mathbf{x}_j \in P_p \\ \mathbf{x}_i \neq \mathbf{x}_j}} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \right], \tag{16}$$

where the maximum is taken over all sets of p unit norm vectors taken from the set $\mathcal{A}$ and where the minimum is taken over all combinations of distinct elements from that set of p vectors.

In words, we are looking for a set of p vectors in $\mathcal{A}$, for which the closest vectors are as far apart from each other as possible. In this set of vectors, the maximum packing distance is the smallest distance between any two vectors.

We are now able to state a necessary condition for the existence of a map $\mathbf{\Phi}$ with a prescribed $\delta_{\bar{\mathcal{A}}}$. In fact, we derive a slightly more general theorem in terms of the Lipschitz constants K_F of $\mathbf{\Phi}$ (for values in $\mathcal{A}$) and K_I of the inverse map. This automatically gives the necessary condition for the existence of a restricted isometry constant $\delta_{\bar{\mathcal{A}}}$ by using $K_F = \sqrt{(1+\delta_{\bar{\mathcal{A}}})}$ and $K_I = \sqrt{\frac{1}{(1-\delta_{\bar{\mathcal{A}}})}}$. We have the following theorem proven in Appendix V.

Theorem 3.5: Let $\mathcal{A}$ be the union of L subspaces of dimension no more than k. In order for a linear map $\mathbf{\Phi} : \mathcal{A} \mapsto \mathbb{R}^N$ to exist such that it has a Lipschitz constant K_F and such that its inverse map $f : \mathbf{\Phi}(\mathcal{A}) \mapsto \mathcal{A}$ has a Lipschitz constant K_I, it is necessary that for all integers $p > 0$

$$M \geq \frac{\ln(p)}{\ln\left(\frac{4K_F K_I}{\omega_p(\mathcal{A})}\right)}. \tag{17}$$

The above inequality must hold for all integer $p > 0$. A tighter bound would therefore require the study of the bound for varying p. This requires a detailed analysis of $\omega_p(\mathcal{A})$. Instead of following this route further here, we use a simpler approach based on a different geometric property of $\mathcal{A}$ defined as follows.

Definition: ($\Delta(\mathcal{A})$ subspace separation) Let $\mathcal{A} = \bigcup_i S_i$ be the union of L subspaces S_i and let $P = \{\mathbf{x}_i \in S_i : \|\mathbf{x}_i\|_2 = 1\}_{i \in \{1,2,\dots,L\}}$, that is, P is a set of L unit length vectors, each being an element of a different subspace. The subspace separation of $\mathcal{A}$ is defined as

$$\Delta(\mathcal{A}) = \max_{P} \left[\min_{\substack{\mathbf{x}_i, \mathbf{x}_j \in P \\ \mathbf{x}_i \neq \mathbf{x}_j}} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \right]. \tag{18}$$

In words, we are looking for a set of L vectors in $\mathcal{A}$, where each vector is in a different subspace, for which the closest vectors are as far apart from each other as possible. The subspace separation $\Delta(\mathcal{A})$ is the smallest such distance and measures the separation of subspaces in the model. It depends on $\mathcal{A}$ and in particular on the number of subspaces in $\mathcal{A}$. However, for a given model $\mathcal{A}$ one might be able to add additional subspaces without changing $\Delta(\mathcal{A})$. There is however a limit to this, depending on N. For example, in order for $\mathcal{A}$ to have a separation $\Delta(\mathcal{A})$ it is necessary that there are L vectors that can be packed into $\mathbb{R}^N$ with an ℓ_2 distance between any two vectors given by $\Delta(\mathcal{A})$.

As the set of L vectors over which we maximise in the definition of $\omega_L(\mathcal{A})$ includes all the sets over which we maximise in the definition of $\Delta(\mathcal{A})$, we have

$$\omega_L(\mathcal{A}) \geq \Delta(\mathcal{A}). \tag{19}$$

This bound can then be used in Theorem 3.5 to derive the following necessary condition relating the number of subspaces L to the number of required observations.

Corollary 3.6: Let $\mathcal{A}$ be the union of L subspaces of dimension no more than k. In order for a linear map $\mathbf{\Phi} : \mathcal{A} \mapsto \mathbb{R}^N$ to exist such that it has a Lipschitz constant K_F and such that its inverse map $f : \mathbf{\Phi}(\mathcal{A}) \mapsto \mathcal{A}$ has a Lipschitz constant K_I, it is necessary that

$$M \geq \frac{\ln(L)}{\ln\left(\frac{4K_F K_I}{\Delta(\mathcal{A})}\right)}. \tag{20}$$

Therefore, if we keep the subspace separation of the model fixed, increasing the number of subspaces (and possibly increasing the ambient dimension), whilst keeping the dimension of the subspaces fixed, we see that it is *necessary* for the number of samples to grow logarithmically with the number of subspaces.

It is again worth stressing the difference between the results of this section and those of the previous section. In the previous section we have analysed the existence of *one to one* maps. While this *one to one* property implies the existence of an inverse map, it does not give any guarantees on the Lipschitz constant of this map, which might be arbitrary large. The theorems in this section differ in that they are asking for conditions that give a fixed Lipschitz constant. Under this additional constraint we see that it is not only necessary that $M \geq k_{\max}$, but that M necessarily has to depend logarithmically on the number of subspaces considered.

IV. Examples

A. k-Sparse Signals

As a specific example, we now return to the "traditional" compressed sensing model in which $\mathbf{x}$ is assumed to have k non-zero elements. Using Stirling's formula, we can bound the number of subspaces in the k-sparse model by

$$(N/k)^k \leq L = \left(\begin{array}{c} N \\ k \end{array} \right) \leq (eN/k)^k. \tag{21}$$

With this bound, the sufficient condition from the previous section reduces immediately to the results similar to those first presented in [33] for the Gaussian case. In particular we see that with probability at least

$$1 - e^{-t} \tag{22}$$

a matrix with i.i.d. Gaussian (or Bernoulli) entries satisfies the $\delta_k(\Phi)$-restricted isometry condition whenever

$$M \geq \frac{1}{c(\delta_k/6)} \left(k \ln(eN/k) + k \ln\left(\frac{36}{\delta_k}\right) + \ln(2) + t \right), \tag{23}$$

where c is as in Theorem 3.3.

In the k-sparse case, $\Delta(\mathcal{A}) \geq \sqrt{2/k}$, because we can always choose L k-sparse vectors which differ in their support (i.e. which lie on a different subspace) and for which all the non-zero elements have magnitude $1/\sqrt{k}$. Two such vectors are closest if they differ only in two elements, from which we get the bound.

Therefore, the following corollary is a simple consequence of Theorem 3.6.

Corollary 4.1: Let $\mathcal{A}$ be the union of subspaces spanned by all k-sparse vectors. In order for a linear map $\mathbf{\Phi} : \mathcal{A} \mapsto \mathbb{R}^N$ to exist such that it has a Lipschitz constant K_F and such that its inverse map $f : \mathbf{\Phi}(\mathcal{A}) \mapsto \mathcal{A}$ has a Lipschitz constant K_I, it is necessary that

$$M \geq \frac{k \ln(N/k)}{\ln\left(\sqrt{8k} K_F K_I\right)}. \tag{24}$$

It is interesting to note the following argument [35][5] showing the necessity of the logarithmic dependence on the signal space dimension for signals with bounded l_p "norm" ($p \leq 1$). It has been shown that the restricted isometry constant implies the recovery of these signals to within a given error bound (See for example [8]). On the other hand, it is also known that the best attainable error bound is related to the Gelfand width of the signal classes. This relationship shows that the given error bound is only attainable if the observation dimension depends logarithmically on the signal space dimension. See [11] for details on Gelfand width and the relationship to the restricted isometry condition.

[5] We would also like to thank Joel Tropp for pointing this argument out to us.

B. k-Sparse Rooted Sub-Trees

Let us consider another example. For many images, it is known that significant non-zero wavelet coefficients have certain tree structures. A more powerful signal model will take such structure into account. One recent example is the model considered by La and Do [18] who also presented a practical algorithm to recover the non-zero coefficients in such a model.

Assume that $\mathbf{x}$ has k non-zero elements as in the k-sparse model. In addition, the elements in $\mathbf{x}$ are assumed to form a *binary tree* and we restrict the k non-zero coefficients of $\mathbf{x}$ to form a rooted sub-tree. The number of sub-trees with k elements is in general much smaller than the number of all k-sparse vectors. Each sub-tree with k nodes defines a subspace and we need to bound the total number of these subspaces to use the theory developed above. The number of different sub-trees with k nodes is clearly bounded by the total number of different trees with k nodes. The number of different trees with k nodes is known to be the Catalan number

$$C_k = \frac{1}{k+1}\left(\begin{array}{c} 2k \\ k \end{array}\right), \tag{25}$$

which we can bound using Stirling's formula

$$C_k \leq \frac{(2e)^k}{k+1}. \tag{26}$$

Therefore, we have a bound on the number of k-dimensional subspaces in the k-sparse tree model

$$L \leq \frac{(2e)^k}{k+1}. \tag{27}$$

Importantly, and in contrast to the bound (21) for the k sparse model, L does not depend on the ambient dimension N.

Using this bound in theorems 3.3 gives the following corollary

Corollary 4.2: Let $\mathcal{A}$ be the set of signals with k non-zero elements that form a rooted sub-tree. There exists a matrix $\mathbf{\Phi}$ such that with probability at least

$$1 - e^{-t}. \tag{28}$$

$$(1 - \delta_{\mathcal{A}}(\mathbf{\Phi}))\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Phi}\mathbf{x}\|_2^2 \leq (1 + \delta_{\mathcal{A}}(\mathbf{\Phi}))\|\mathbf{x}\|_2^2 \tag{29}$$

for all $\mathbf{x} \in \mathcal{A}$, whenever

$$M \geq \frac{1}{c(\delta_{\mathcal{A}}/6)}\left(k\ln\left(\frac{72e}{\delta_{\mathcal{A}}}\right) - \ln\left(\frac{k+1}{2}\right) + t\right). \tag{30}$$

Again, c is as in Theorem 3.3.

A general lower bound for L is not available, but if we assume that the height of the tree is larger than k (note that this implies that $k \leq \log_2(N-1)-1$), then the Catalan number gives the exact number of k element rooted sub-trees. Under this assumption we have

$$\frac{2^k}{k+1} \leq L. \tag{31}$$

Therefore, (again using $\Delta(\mathcal{A}) \geq \sqrt{2/k}$) we get the necessary condition.

Corollary 4.3: Let $\mathcal{A}$ be the set of signals with k non-zero elements that form a rooted sub-tree and assume that $2 \leq k \leq \log_2(N-1)-1$. In order for a linear map $\mathbf{\Phi} : \mathcal{A} \mapsto \mathbb{R}^N$ to exist such that it has a Lipschitz constant K_F and such that its inverse map $f : \mathbf{\Phi}(\mathcal{A}) \mapsto \mathcal{A}$ has a Lipschitz constant K_I, it is necessary that

$$M \geq \frac{k\ln(2) - \ln(k+1)}{\ln\left(\sqrt{8k}K_F K_I\right)}. \tag{32}$$

V. Discussion and conclusion

The union of linear subspaces model considered in this paper is a general signal model that includes many of the signal models previously studied in compressed sensing. Results derived for this model have therefore a wide applicability and can provide new insight into the traditional sparse coding problem. In this paper we have studied two aspects of this general model, the existence of *one to one* maps into low dimensional observation spaces and the properties of the inverse maps. We were particularly interested in the behaviour of these properties in terms of the observation dimension, thereby deriving theorems that are in the same spirit to the Nyquist-Shannon sampling theorem.

The first new result presented was that *almost all* linear maps are *one to one*, whenever the observation dimension is at least twice the largest subspace dimension in the model considered. This is an interesting result similar to that given in [17][6] showing that *one to one* maps are relatively "easy" to come by and do not depend on either, the signal ambient dimension nor on the number of subspaces in the model.

While the *one to one* property is clearly desirable for signal acquisition as it guarantees that the observation contains the same information as the original signal, in practical applications two other important properties have to be taken into account, the robustness to noise and the existence of efficient algorithms to calculate the inverse map. The second part of this paper therefore concentrated on properties of the inverse map and in particular its smoothness. In order to guarantee a given smoothness of the

[6] The differenc being that our results show that *almost all* maps have the desired property, whilst in [17] it was shown that for the countably infinite union of subspaces, the set of maps having the *one to one* property is dense.

inverse map, we showed that the observation dimension had to depend at least logarithmically on the number of subspaces in the signal model and we showed that this logarithmic dependence was sufficient for the existence of a smooth inverse. This is a much stricter condition than the one required for the existence of *one to one* maps, however, for these stricter conditions we get the additional guarantee of the existence of an inverse map with a fixed Lipschitz constant.

An important realisation is that linear *one to one* embeddings only require twice as many observations as the dimension of the largest subspace in the model, however, these sampling schemes can be arbitrarily unstable. In order to control the stability of such sampling schemes, the number of samples must depend logarithmically on the number of subspaces in the model. Similar results are known for the sparse signal model in compressed sensing and we have here shown that this behaviour is valid in a much more general settings. In fact we could show that for the k sparse signal, we directly recovered know sufficient conditions, though the derived necessary conditions are novel. In particular, we showed that the logarithmic dependence of the observation space dimension on the signal ambient dimension is necessary for the k-sparse model. The dependence on the ambient dimension is however a result of the growth of the number of subspaces in the k-sparse model when the ambient space dimension is increased. Interestingly, if we further constrain the model and assume the non-zero coefficients to have tree structures, the dependence on the signal ambient space dimension disappears.

We have here studied theoretical properties of a compressed sensing approach for signals from the union of linear subspaces and specified theoretical properties of the inverse map $f(\mathbf{y})$. Whilst a small Lipschitz constant of $f(\mathbf{y})$ clearly specifies a robustness against noise, we also believe that this constant controls other important aspects of $f(\mathbf{y})$. It is for example well known that in the k-sparse signal model, $f(\mathbf{y})$ can be implemented efficiently using linear programming algorithms whenever the restricted isometry constant is sufficiently small. The problem of calculating $f(\mathbf{y})$ for the more general unions of subspace model has not been addressed in this paper, however, we believe that the ability to implement $f(\mathbf{y})$ efficiently might well be related to the Lipschitz constant of $f(\mathbf{y})$ or to the restricted isometry constant and we hope that the developments of this paper are instrumental in such a theory, but a detailed study of these issues has to be relegated to future publications.

We have here concentrated on the union of finitely many finite dimensional subspaces. Extensions to unions of infinitely many finite dimensional subspaces were considered previously in [17], where the existence of *one to one* maps for this model was studied. The question now arises whether our results are extendable to this setting. In particular the conditions for the existence of stable embeddings is of interest. For a model with infinitely many subspaces, a result like that of Theorem 3.3 would clearly not be desirable, in that it would imply the requirement of infinitely many observations. However, the necessary condition in Theorem 3.5 also depends on how well we can pack L vectors onto the subspaces. In order to get a finite result, one important issue therefore seems to be the necessity to control this geometric property.

APPENDIX I

PROOF OF THEROEM 2.3

For the proof we need to consider the set $\mathcal{C}$ defined as the union of hyper-spheres of dimension $k_j - 1$

$$\mathcal{C} = \bigcup_{j}^{L} \mathcal{C}_j, \ \mathcal{C}_j = \{\mathbf{x} = \Omega_j \mathbf{a}, \Omega_j \in \mathbb{R}^{N \times k_j}, \mathbf{a} \in \mathbb{R}^{k_j}, \|\mathbf{x}\| = 1\}, \tag{33}$$

which can be used as an alternative definition of $\mathcal{A}$

$$\mathcal{A} = \{\hat{\mathbf{x}} : \hat{\mathbf{x}} = \alpha \mathbf{x}, \mathbf{x} \in \mathcal{C}, \alpha \in \mathbb{R}\}. \tag{34}$$

Note that the set $\mathcal{C}$ has box counting dimension $k-1$, where k is the maximum of the k_j.

The goal is then to show that, for the set defined below, almost all linear maps $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ are *one to one* under appropriate conditions on M. We proceed as follows. First we state a lemma that bounds the volume of the subset of any bounded set, that maps into a ball centred at zero. Later, shrinking the diameter of the ball in this lemma, it can be shown that the set that exactly maps to zero must have a bounded dimension. This dimension depends on M and the dimension of the set. To prove that this holds for *almost all* maps, we also need to consider a neighbourhood of an arbitrary linear map and show that the property holds for *almost all* maps in the neighbourhood of *any* map. This approach mirrors closely the derivation in [26] from where we borrow the first lemma

Lemma 1.1 (Lemma 4.2 from [26]): Let $\theta_{\mathbf{\Theta}}(\alpha) = \mathbf{\Theta}\alpha + \mathbf{d}$ be an affine map from R^T to R^M, where $\mathbf{\Theta}$ is a $T \times M$ matrix and $\mathbf{d} \in \mathbb{R}^M$. Given a positive integer r, let σ denote the r^{th} largest singular value of the matrix $\mathbf{\Theta}$. Let $\mathcal{B}_t^T$ be the ball centred at the origin in R^T space with radius t. Similarly, let $\mathcal{B}_m^M$ be the ball centred at the origin in R^M space with radius m. Then the portion of the volume of $\mathcal{B}_t^T$ that overlaps with the set $\theta_{\mathbf{\Theta}}(\alpha)^{-1}(\mathcal{B}_m^M)$ can be upper bounded by

$$\frac{Vol(\mathcal{B}_t^T \cap \theta_{\mathbf{\Theta}}(\alpha)^{-1}(\mathcal{B}_m^M))}{Vol(\mathcal{B}_t^T)} < 2^{\frac{T}{2}} \left(\frac{m}{\sigma t}\right)^r. \tag{35}$$

The second lemma also follows closely the development in [26], with the main difference that we here concentrate on linear maps. As this lemma does most of the work in proving the main result, it is given here for completeness. Let us, however, first introduce the set

$$B = \{\mathbf{b} : \mathbf{b} = \mathbf{x}_1 + \mathbf{x}_2; \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{A}, \|\mathbf{b}\| = 1\}. \tag{36}$$

Note that B has dimension no more than $k_{\max} - 1$.

Lemma 1.2: Let the set $\{\mathbf{\Phi}_i\}_{i \in \{0,1,\ldots,T\}}$ be a basis for linear maps from R^N to R^M. For any $\mathbf{\Phi}_0$ we define $\mathbf{\Phi}_\alpha = \mathbf{\Phi}_0 + \sum_i^T \alpha_i \mathbf{\Phi}_i$ for $\alpha \in \mathbb{R}^t$, or in the language of the previous lemma, $\mathbf{\Phi}\mathbf{b} = \theta_{\mathbf{\Theta}_b}(\alpha) = \mathbf{\Theta}_b \alpha + \mathbf{d} = \sum_i^T \alpha_i \mathbf{\Phi}_i \mathbf{b} + \mathbf{\Phi}_0 \mathbf{b}$. Let $k_b = k_{\max} - 1$ be the box counting dimension of B. Then for *almost all* $\alpha \in \mathbb{R}^T$, the set $\mathbf{\Phi}_\alpha^{-1}(0)$, i.e. the set of $\mathbf{b} \in B$ for which $\mathbf{\Phi}_\alpha \mathbf{b} = 0$, has lower box counting dimension at most $k_b - M$ and in particular, if $M > k_b = k_{\max} - 1$, then $\mathbf{\Phi}_\alpha^{-1}(0) = \emptyset$.

Proof: $\mathbf{\Phi}_\alpha(\mathbf{b})$ is a bilinear map. In order to prove the lemma, we are considering the set $\mathbf{\Phi}_\alpha(\mathbf{b})$ for both, α and $\mathbf{b}$ taking values in some ball in their respective spaces. Lemma 1.1 can then be used to bound the probability for the set $\mathbf{\Phi}_\alpha(\mathbf{b})$ to overlap with a ball centred at zero. This probability can be made arbitrary small, from which the lemma will follow. We proceed in several steps.

1) Because of the linearity of $\mathbf{\Phi}_\alpha$ as a function of α, it suffices to prove the result for $\alpha \in \mathcal{B}_t^T$ for some $t > 0$. This defines a neighbourhood of linear maps and we need to show that *almost all* maps in this neighbourhood do not map any $\mathbf{b}$ into zero.
2) Let $B_r \subseteq B$ be the set of $\mathbf{b} \in B, \mathbf{b} \neq 0$ for which the matrix

$$\mathbf{\Theta}_{\mathbf{b}} = \{\mathbf{\Phi}_1 \mathbf{b}, \mathbf{\Phi}_2 \mathbf{b}, \ldots, \mathbf{\Phi}_t \mathbf{b}\} \tag{37}$$

has rank M with the M^{th} largest singular value of $\mathbf{\Theta}_{\mathbf{b}}$ greater than σ. Let k_r be the lower box counting dimension of B_r. Note that $\mathbf{\Theta}_{\mathbf{b}}$ is of rank M almost surely.
3) We next consider the ball $\mathcal{B}_m^M \subset R^M$ of radius m and centred at zero. Lemma 1.1 shows that for a fixed $\mathbf{b}$ the maximum overlap of the set $\mathcal{B}_t^T$ and the set $\mathbf{\Phi}_\alpha^{-1}(\mathcal{B}_m^M)$ is less than $2^{\frac{T}{2}}(\frac{m}{\sigma t})^r$, i.e. for a fixed $\mathbf{b}$ the probability of any of the $\mathbf{\Phi}_\alpha, \alpha \in \mathcal{B}_t^T$ mapping $\mathbf{b}$ into a ball centred at zero is bounded. In other words, for fixed $\mathbf{b}$, $\mathbf{\Phi}_\alpha(\mathbf{b})$ is more than t away from zero except with probability $2^{\frac{T}{2}}(\frac{m}{\sigma t})^r$.
4) Next we consider a third ball, this time in R^N, centred at an arbitrary $\mathbf{b}$ and with radius n, say $\mathcal{B}_n^N(\mathbf{b}) \subset R^k$. Because of the linearity of $\mathbf{\Phi}_\alpha$, the image of any $\mathcal{B}_n^N(\mathbf{b})$ under $\mathbf{\Phi}_\alpha, \alpha \in \mathcal{B}_t^T$ is a subset of some ball in R^M with bounded radius Cn for some C. If $\mathbf{\Phi}_\alpha$ maps $\mathbf{b}$ (the centre of our ball) further away from zero than Cn, i.e. if $|\mathbf{\Phi}_\alpha(\mathbf{b})| > Cn$, then the image of the ball $\mathcal{B}_n^N(\mathbf{b})$ does not contain zero. We can therefore bound the probability that the image of the ball $\mathcal{B}_n^N(\mathbf{b})$ under $\mathbf{\Phi}_\alpha, \alpha \in \mathcal{B}_t^T$ contains zero with the probability that $\mathbf{\Phi}_\alpha, \alpha \in \mathcal{B}_t^T$ maps $\mathbf{b}$ to a point within the ball with radius Cn. By the argument above, this probability is bounded by $2^{\frac{T}{2}}(\frac{Cn}{\sigma t})^r$. For $\sigma > 0$ and $t > 0$ fixed, we have a bound of $C_1 n^r$.
5) Because B_r is bounded, we can cover B_r with say p balls of radius n. Furthermore, we can bound p using $p \leq n^{-k}$, where we can make k arbitrary close to the box counting dimension k_r of B_r

by choosing n small enough. Therefore, the probability that q of the images of the p balls contain zero can be bounded by $C_1 n^{r-k}/q$. Writing $q = n^{-k}$, i.e. writing the probability as $C_1 n^{d-(k-r)}$, we see that if $d > k - r$ we can make the probability arbitrary small by decreasing n.

6) We can therefore decrease n such that fewer than n^{-d} balls cover the subset of B that maps into zero except with a probability approaching zero as $n \rightarrow 0$. By the definition of box counting dimension, we realise that b is a bound on the box counting dimension of the set $\mathbf{\Phi}_\alpha^{-1}(0) \subset B$. The argument holds for all $d > k_r - r$, which therefore gives the bound in the lemma.
7) The above argument holds for all $\sigma > 0$ and therefore holds for all $\mathbf{\Theta_b}$ almost surely.

■

Proof: [Proof of Theorem 2.3] Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$, then $\alpha\mathbf{x}_1, \beta\mathbf{x}_2 \in \mathcal{A}$ if $\alpha, \beta \in \mathbb{R}$. Assume there were $\mathbf{x}_1, \mathbf{x}_2, \alpha, \beta$ such that $\alpha\mathbf{x}_1 \neq \beta\mathbf{x}_2$ and $\mathbf{\Phi}\alpha\mathbf{x}_1 = \mathbf{\Phi}\beta\mathbf{x}_2$. W.l.o.g. $\alpha \neq 0$ so that $c = \frac{\beta}{\alpha}$. Then $\mathbf{\Phi}\mathbf{x}_1 = \mathbf{\Phi}c\mathbf{x}_2$. We can write $c\mathbf{x}_2 = \mathbf{x}_1 + \gamma\mathbf{b}$ for some γ and some $\mathbf{b} \in B$. $\mathbf{\Phi}\alpha\mathbf{x}_1 = \mathbf{\Phi}\beta\mathbf{x}_2$ implies therefore that $\gamma\mathbf{\Phi}\mathbf{b} = 0$. If $\gamma = 0$, then we would have $\gamma\mathbf{x}_2 = \mathbf{x}_1$ which is impossible by the assumption that $\alpha\mathbf{x}_1 \neq \beta\mathbf{x}_2$, therefore $\gamma \neq 0$ which means that $\mathbf{\Phi}\mathbf{b} = 0$. By definition $\mathbf{b}$ is from a bounded set of box counting dimension $k_{\max} - 1$. But by the previous lemma $\mathbf{\Phi}\mathbf{b} \neq 0$ for *almost all* $\mathbf{\Phi}$ if $M > 2d - 1$, from which the theorem follows. ■

APPENDIX II
PROOF OF THEOREM 2.4

Proof: [Proof of Theorem 2.4] The *one to one* property implies that

$$\mathbf{\Phi}(\mathbf{x}_i - \mathbf{x}_j) \neq \mathbf{0}, \tag{38}$$

for all $\mathbf{x}_i \neq \mathbf{x}_j$ where $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{A}$.

Let $B = \{\mathbf{b} = \mathbf{x}_i - \mathbf{x}_j : \mathbf{x}_i \in S_i, \mathbf{x}_j \in S_j, \mathbf{b} \neq \mathbf{0}\}$, where the subsets S_i and S_j are chosen such that the dimension of B is $k_{\max}$. $\mathbf{\Phi}(\mathbf{x}_i - \mathbf{x}_j) \neq \mathbf{0}$ then implies that $\mathbf{\Phi}(\mathbf{b}) \neq \mathbf{0}$. For any set of $k_{\max}$ linearly independent vectors $\mathbf{b}_1 \in B$ we know that $\sum_i \alpha_i \mathbf{b}_i \neq 0$ holds for all $\alpha_i \neq 0$. This implies that the set of $k_{\max}$ vectors $\mathbf{\Phi}\mathbf{b}_i$ is also linearly independent. This can be shown by contradiction. Assume $\mathbf{\Phi}\mathbf{b}_i$ is linearly dependent, i.e. there exist $\alpha_i \neq 0$ such that $\sum_i \alpha_i(\mathbf{\Phi}\mathbf{b}_i) = 0$. By linearity, this would imply that $\mathbf{\Phi}\sum_i \alpha_i(\mathbf{b}_i) = 0$. Now $\sum_i \alpha_i(\mathbf{b}_i) \in B$. But none of the elements of B are in the null space of $\mathbf{\Phi}$, so that $\mathbf{\Phi}\mathbf{b} = 0$ would imply $\sum_i \alpha_i(\mathbf{b}_i) = \mathbf{b} = 0$. But by the linear independence of the $\mathbf{b}_i$ this is impossible if $\alpha_i \neq 0$. Therefore, the $\mathbf{\Phi}\mathbf{b}_i$ are linearly independent and span a $k_{\max}$ dimensional subspace of $\mathbb{R}^M$, which is only possible if $M \geq k_{\max}$. ■

APPENDIX III
PROOF OF THEOREM 2.6

For the $\mathcal{A}_i$ in the definition of the measure in the theorem, we define $\mathcal{C}_i$ similarly to the definitions of $\mathcal{C}$. Let the sets

$$B_{i,j} = \{\mathbf{b}_{i,j} : \mathbf{b}_{i,j} = \mathbf{y}_i + \mathbf{y}_j, \mathbf{y}_i \in \mathcal{C}_i, \mathbf{y}_j \in \mathcal{A}_i\}. \tag{39}$$

The sets $B_{i,j}$ are then of dimension $i + j - 1$.

Proof: [Proof of Theorem 2.6] If $M > i+j-1$, we can apply Theorem 2.3 and *almost all* $\mathbf{\Phi}$ do not map any subset $B_{i,j}$ to zero. If $i+j-1 \geq M > k > i$ and assume that without loss of generality $i \geq j$, we have from Lemma 1.2 that the set for which $\mathbf{\Phi b} = 0$, with $b \in B_{i,j}$ has dimension $i+j-1-M < i$, which by definition of the measure on $\mathcal{A}_i$ has measure zero. Taking the union of these measure zero events is also a measure zero event showing that for *almost all* $\mathbf{\Phi}$, and *almost all* $\mathbf{b}$, $\mathbf{\Phi b} \neq \mathbf{0}$. ■

APPENDIX IV
PROOF OF THEOREM 3.3

The proof of Theorem 3.3 follows closely the approach set out in [34] and [12].

Proof: [of Theorem 3.3] For a fixed $\mathbf{x}$, any matrix $\mathbf{\Phi}$ with entries drawn i.i.d. form subgaussian distribution satisfies [12]

$$P(|\|\mathbf{\Phi x}\|_2^2 - \|\mathbf{x}\|_2^2| \geq \epsilon\|\mathbf{x}\|_2^2) \leq 2e^{-\frac{M}{2}c(\epsilon)}. \tag{40}$$

From lemma 5.1 in [34] we know that if

$$P\{|\|\mathbf{\Phi y}\|_2^2 - \|\mathbf{y}\|_2^2| > \delta\|\mathbf{y}\|_2^2\} \leq 2e^{-c(\delta/2)N}, \tag{41}$$

then

$$(1-\delta)\|\mathbf{x}\|_2 \leq \|\mathbf{\Phi x}\|_2 \leq (1+\delta)\|\mathbf{x}\|_2 \tag{42}$$

holds with probability more than

$$1 - 2\left(\frac{12}{\delta}\right)^k e^{-c(\delta/2)M}. \tag{43}$$

Let $\delta_A = 3\delta$, then

$$(1-\delta_A)\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Phi x}\|_2^2 \leq (1+\delta_A)\|\mathbf{x}\|_2^2 \tag{44}$$

holds with probability more than

$$1 - 2\left(\frac{36}{\delta_A}\right)^k e^{-c(\delta_A/6)M}. \tag{45}$$

A union bound tells us that for L subspaces, the probability of failure will be bounded by

$$1 - L2\left(\frac{36}{\delta_A}\right)^k e^{-c(\delta_A/6)M}. \tag{46}$$

from which the theorem follows. ■

Appendix V
Proof of Theorem 3.5

Proof: [of Theorem 3.5]

The proof of this theorem is based on the following inequality which we will show to hold for a particular subset of p vectors from $\mathcal{A}$.

$$\frac{1}{K_I^2} \leq \frac{\|\mathbf{\Phi}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2}{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2} \leq \frac{1}{\omega_p^2(\mathcal{A})}\|\mathbf{\Phi}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \tag{47}$$

The first inequality is due to the definition of the Lipschitz constants. The second inequality follows from the definition of $\omega_p(\mathcal{A})$, which guarantees the existence of p vectors taken from $\mathcal{A}$ that satisfy the required bound.

The vectors $\mathbf{x}_i$ are unit length. We require that $\|\mathbf{\Phi}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \leq K_F^2$. Therefore, the vectors $\mathbf{y}_i = \mathbf{\Phi}\mathbf{x}_i$ all lie within a ball of radius K_F and the vectors $\hat{\mathbf{y}}_i = \frac{\mathbf{y}_i}{K_F}$ lie in a ball of unit radius. This, with inequality (47), shows that we require $\|\hat{\mathbf{y}}_i - \hat{\mathbf{y}}_j\|_2^2 \geq \frac{\omega_p^2(\mathcal{A})}{K_I^2 K_F^2}$ for $i \neq j$. We therefore see that it is necessary for the vectors $\hat{\mathbf{y}}_i$ to be separated by at least $d = \sqrt{\frac{\omega_p^2(\mathcal{A})}{K_I^2 K_F^2}}$ and we can use a packing argument to complete the proof.

Lemma 5.1: We can pack p points within the unit ball in $\mathbb{R}^M$ with a separation of d *only if*

$$M \geq \frac{1}{\ln(2/d + 1)} \ln(p). \tag{48}$$

Proof: [of Lemma 5.1] A set of points within the unit ball which constitutes such an d separated set, also constitutes a $d/2$ packing. Consider the ball with radius $1 + d/2$. We need a necessary condition on the number of $d/2$ balls which can be packed into such a ball. We know that the volume of all of the $d/2$ balls in such a packing must be smaller than the total volume of the ball.

$$\begin{aligned}
p \ \mathrm{Vol}(\mathcal{B}(d/2)) &\leq \mathrm{Vol}(\mathcal{B}_{1+d/2}) & (49)\\
p &\leq \frac{\mathrm{Vol}(\mathcal{B}_{1+d/2})}{\mathrm{Vol}(\mathcal{B}_{d/2})} \\
p &\leq \frac{(1 + d/2)^M}{(d/2)^M} \\
p &\leq (2/d + 1)^M. & (50)
\end{aligned}$$

Therefore, a necessary condition to be able to find such a packing is that we have $p \leq (2/d+1)^M$ balls, or, equivalently that

$$M \geq \frac{1}{\ln(2/d+1)} \ln(p). \tag{51}$$

■

Using this lemma with $d = \frac{\omega_p(\mathcal{A})}{K_I K_F}$ we get the necessary condition that

$$\frac{\ln(p)}{M} \leq \ln\left(\frac{2K_I K_F}{\omega_p(\mathcal{A})} + 1\right). \tag{52}$$

Theorem 3.5 follows then by recognising that $1 < \frac{2K_I K_F}{\omega_p(\mathcal{A})}$. ■

References

[1] H. Nyquist, "Certain topics in telegraph transmission theory," *Transactions of the A. I. E. E.*, pp. 617–644, Feb. 1928.

[2] C. A. Shannon and W. Weaver, *The mathematical theory of communication.* University of Illinois Press, 1949.

[3] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm," *IEEE Transactions on Signal Processing*, vol. 45, pp. 600–616, March 1997.

[4] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Transactions on Signal Processing*, vol. 50, no. 6, pp. 1417–1428, 2002.

[5] Y. C. Eldar, "Compressed sensing of analog signals," *arXiv:0806.3332v1*, 2008.

[6] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information.," *IEEE Transactions on information theory*, vol. 52, pp. 489–509, Feb 2006.

[7] E. Candès and J. Romberg, "Quantitative robust uncertainty principles and optimally sparse decompositions," *Foundations of Comput. Math*, vol. 6, no. 2, pp. 227 – 254, 2006.

[8] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. on Information Theory*, vol. 52, no. 12, pp. 5406 – 5425, 2006.

[9] D. Donoho, "Compressed sensing," *IEEE Trans. on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[10] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[11] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," *Journal of the american mathematical society*, vol. 22, no. 1, pp. 211–231, 2009.

[12] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *IEEE Transactions on Information Theory*, vol. 54, pp. 2210–2219, May 2007.

[13] T. Blumensath and M. Davies, "Compressed sensing and source separation," in *Independent Component Analysis and Blind Source Separation*, Lecture Notes on Computer Science 4666, M.E. Davies, C.J. James, S. Abdallah and M.D. Plumbley, Ed. Springer, pp. 341-348, 2007.

[14] A. K. Fletcher, S. Rangan, and V. K. Goyal, "On the rate-distortion performance of compressed sensing," in *Proc. IEEE conf. Acoustics, Speech and Signal Processing*, Vol. 3, pp. 885–888, 2007.

[15] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. on Information Theory*, vol. 52, no. 9, pp. 4036–4048, 2006.

[16] E. Candès and T. Tao, “The dantzig selector: statistical estimation when p is larger than n,” *Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[17] Y. Lu and M. Do, “A theory for sampling signals from a union of subspaces,” *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2334–2345, 2008.

[18] C. La and M. Do, “Signal reconstruction using sparse tree representations,” in *Proc. SPIE Conf, Wavelet Applications in Signal and Image Processing XI*, (San Diego, California), Vol. 5914, Sep 2005.

[19] S. F. Cotter, B. D. Rao, K. Engan, and K. K-Delgado, “Sparse solutions to linear inverse problems with multiple measurement vectors,” *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, 2005.

[20] C. J and X. Huo, “Theoretical results on sparse representations of multiple-measurement vectors,” *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4634–4643, 2006.

[21] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, “Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit,” *Signal Processing*, vol. 86, pp. 572–588, 2006.

[22] J. A. Tropp, “Algorithms for simultaneous sparse approximation. Part II: Convex relaxation,” *Signal Processing*, vol. 86, pp. 589–602, 2006.

[23] K. S. R. Gribonval, H. Rauhut and P. Vandergheynst, “Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms,” *Journal of Fourier analysis and applications*, Published online, DOI:10.1007/s00041-008-9044-y, October, 2008.

[24] G. A. Edgar, *Measure, Topology, and Fractal Geometry*. Springer-Verlag New York Inc., 1990.

[25] J. M. Lee, *Introduction to smooth manifolds*. Springer, 2000.

[26] T. Sauer, J. A. Yorke, and M. Casdagli, “Embedology,” *Journal of Statistical Physics*, vol. 65, no. 3/4, pp. 579–616, 1991.

[27] R. Hunt, T. Sauer, and A. Yorke, “Prevalence: a translation-invariant “almost every” on infinite-dimensional spaces,” *Bull. Amer. Math. Soc.*, vol. 99, no. 11, pp. 217–238, 1992.

[28] R. Baraniuk and M. Wakin, “Random projections of smooth manifolds,” *Foundations of Computational Mathematics*, Published online, DOI:10.1007/s10208-007-9011-z, September, 2007.

[29] D. L. Donoho and M. Elad, “Optimally-sparse representation in general (non-orthogonal) dictionaries via l1 minimization,” *Proc. Nat. Aca. Sci.*, vol. 100, pp. 2197–2202, 2003.

[30] D. L. Donoho, “For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution,” *Comm. Pure and Applied Math.*, vol. 59, no. 6, pp. 797–829, 2006.

[31] M. B. Wakin, *The Geometry of Low-Dimensional Signal Manifolds*. PhD thesis, Rice University, Texas, USA, 2006.

[32] E. Candès and J. Romberg, “Practical signal recovery from random projections,” in *Proc. SPIE Conf, Wavelet Applications in Signal and Image Processing XI*, Vol. 5914, Jan. 2005.

[33] E. J. Candès and T. Tao, “Decoding by linear programming,” *IEEE Transactions on Information Theory*, vol. 51, pp. 4203–4215, 2004.

[34] R. Baraniuk, M. Davenport, R. De Vore, and M. Wakin, “A simple proof of the restricted isometry property for random matrices,” *Constructive Approximation*, Published online, DOI:10.1007/s00365-007-9003-x, February, 2008.

[35] E. Candès, “Compressive sampling,” in *Proceedings of the International Congress of Mathematics*, (Madrid, Spain), Vol. 3, pp. 1433-1452, 2006.