CLASSIFICATION FROM COMPRESSIVE REPRESENTATIONS OF DATA.

Bertrand Coppa, Rodolphe Héliot, Dominique David

CEA-LETI
Minatec Campus
Grenoble, France

Olivier Michel

GIPSA-Lab/DIS
University of Grenoble
France

ABSTRACT

Compressive sensing proposes simple compression of sparse data at the expense of difficult data reconstruction. We focus here on the opportunities in terms of information recovery within the compressed data space, thus avoiding the expensive data reconstruction step. Specifically, we study here how the clustering ability of a dataset is affected by random projections. The proposed result has the advantage to give statistical insights for low dimensions, where traditional results are to no avail. Experiments show that it is possible to achieve high compression rate while preserving clustering abilities, at a low computational cost.

***Index Terms*—** Random embeddings; Compressive Sensing; Clustering

1. INTRODUCTION

It is now well-known that *compressive sensing* (CS) enables the possibility to reconstruct a signal that has been projected on a $m \times N$, $m << N$ matrix under the condition that the signal admits a *sparse* representation on a known dictionary, that is when the number of significant representation coefficients is low [1–5]. Furthermore, this is achievable when the matrix is drawn from a random process [6, 7]. The major drawback of CS is that the reconstruction process requires complex algorithms [8–10] and the a priori knowledge of the sparsity inducing basis.

However, in many applications, the final objective is not signal reconstruction, but information retrieval, such as event detection or signal classification. For example, spikes classification is an essential step of intracortical neural data processing [11]. In this application, the idea of a simple compression system is appealing, since the hardware resources (in terms of size, available power) can be very limited in the case of embedded (or implanted) systems.

CS has shown [1, 6] that a random subspace projection holds all the signal information when the signal is sparse. The next question is: is it possible to maintain classification and clustering ability in the compressed space? Dasgupta [12, 13] has investigated the problem, and found that random projections of highly eccentric Gaussian mixtures make them become more spherical. The way the separability (and thus clustering ability) reacts to the projection is investigated and described in [12]. Others [14, 15] have proposed methods that perform multiple random projections and clustering, and combine the multiple results to enhance the final clustering.

Random projections also have an edge on other dimensionality reduction methods such as PCA [16], as it is a linear, non-adaptive method, has low processing power requirements and makes no assumption on the criterion of interest. Using random binary matrices, the computation is even simpler since there is no floating point operation required, if the signal is sampled as integer values.This can be seen as a low computational cost compression method, and thus is interesting for sensor applications where power is scarce. In these cases, methods like PCA or those presented in [14, 15] are not helpful.

In section 2, we introduce some mathematical concepts, then in section 3 we discuss the possibility of using binary matrices and how it can be applied to cases where m and N are not large, with corresponding experimental results shown in section 4.

2. MATHEMATICAL BACKGROUND

In order to maintain equivalent classification performance in both the original and compressed spaces,it is necessary to preserve the critical relationship used in classification. For many clustering methods, the (Euclidean) distance between points is that criterion: the *k-means* algorithm uses distances, as well as methods based on the graph Laplacian [17, 18]. To characterize how the distances are modified by random projections, we introduce *distortion* that quantifies how close to isometric the projection is. This is very close to the concept of *Restricted Isometry Property* (RIP) used in CS literature [19].

Definition 1 *Given Φ a $m \times N$ matrix representing an embedding from $\mathbb{R}^N$ into $\mathbb{R}^m$, it has a distortion $\delta > 0$ with* *probability P if for all $x \in \mathbb{R}^N$, the following is true with probability at least P:*

$$(1-\delta) \leq \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \leq (1+\delta) \quad (1)$$

The main result about random embeddings distortion is due to Johnson and Lindenstrauss [20] (which can also be used to prove the RIP for random matrices [21]). We will retain the following formulation of the Johnson-Lindenstrauss (JL) lemma as exposed in [22]:

Lemma 1 *Let Q be a finite collection of $\#Q$ points in $\mathbb{R}^N$. Fix $0 < \varepsilon < 1$ and $\beta > 0$. Let Φ be a random orthoprojector from $\mathbb{R}^N$ to $\mathbb{R}^m$ (i.e. a $m \times N$ matrix with orthonormal rows) with*

$$m \geq \left(\frac{4+2\beta}{\varepsilon^2/2+\varepsilon^3/3}\right) \ln \#Q. \quad (2)$$

If $m \leq N$, then with probability exceeding $1-(\#Q)^{-\beta}$, the following statement holds: for every $x, y \in Q$,

$$(1-\varepsilon)\sqrt{\frac{m}{N}} \leq \frac{\|\Phi x - \Phi y\|_2}{\|x-y\|_2} \leq (1+\varepsilon)\sqrt{\frac{m}{N}}. \quad (3)$$

The choice of having an orthoprojector leads to the compaction factor $\sqrt{\frac{m}{N}}$. However, simply renormalizing Φ by a $\sqrt{\frac{N}{m}}$ factor, eq. (3) becomes

$$(1-\varepsilon) \leq \frac{\|\Phi x - \Phi y\|_2}{\|x-y\|_2} \leq (1+\varepsilon), \quad (4)$$

and the compaction is avoided. We will keep to such normalization for the rest of the article.

Note that eq. 4 is not equivalent to eq. 1, which involves squared norms. However, if $\delta < 1$, we have the following relation:

$$(1) \Rightarrow (1-\delta) \leq \frac{\|\Phi x\|_2}{\|x\|_2} \leq (1+\delta) \Leftrightarrow (4). \quad (5)$$

Conversely, if $\varepsilon < 1$, the following result may be easily shown:

$$\begin{aligned}(4) \Leftrightarrow (1-2\varepsilon+\varepsilon^2) &\leq \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \leq (1+2\varepsilon+\varepsilon^2) \\ \Rightarrow (1-\mathbf{3}\varepsilon) &\leq \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \leq (1+\mathbf{3}\varepsilon). \quad (6)\end{aligned}$$

There is no direct equivalence but one can easily switch from one equation to the other. From now on, we will keep using ε and δ in order to make it easier to distinguish between the two equations.

Figure 1 shows how the minimal value for m, as given by lemma 1, evolves with the distortion ε, with probability $P > 0.5$; m_{min} can be very high even for low values of tolerated distortion. Thus in the perspective of constructing a random compressive embedding with low distortion, the result of lemma 1 is granted only for large value of m (and of N as we want dimensionality reduction). The value of m_{min} also depends on $\#Q$. This is expected as the lemma is formulated to characterize the ability of a projector to maintain separation between any two signals (points in $\mathbb{R}^N$) from Q. The larger Q, the lower may be this separation $\|x-y\|$.

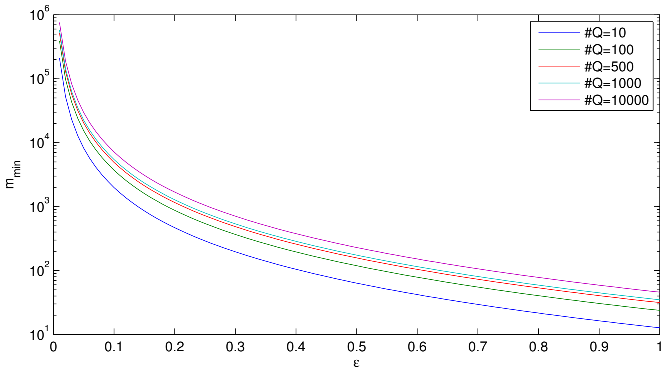

Fig. 1. Variation of the minimal number of projections $m_{min} = \left(\frac{4+2\beta}{\varepsilon^2/2+\varepsilon^3/3}\right) \ln \#Q$ as a function of ε, with such β that $P > 0.5$ and for different values of $\#Q$.

3. DISTORTION PROBABILITY FOR BINARY MATRICES

3.1. Proposed result

We will consider the following system: Φ is a $m \times N$, $m < N$ column-normalized zero-mean binary matrix with matrix element $\phi_{i,j} = \pm\frac{1}{\sqrt{m}}$, $P(\phi_{i,j} = \frac{1}{\sqrt{m}}) = P(\phi_{i,j} = -\frac{1}{\sqrt{m}}) = 0.5$. $x \in \mathbb{R}^N$ is a signal (sampled signals lie in a subset of $\mathbb{R}^N$ thus the result still holds after sampling).

We propose the following result, which is a simplified version of the ExRIP [23], where the matrix is chosen as binary and there is no assumption on the distribution of x or its sparsity:

Proposition 1 *Given Φ as defined above, x from a random process and $\delta > 0$, we have:*

$$P\left\{\left|\frac{\|\Phi x\|^2}{\|x\|^2} - 1\right| \leq \delta\right\} \geq 1 - \frac{2}{m\delta^2}\left(1 - \mathbb{E}\left\{\frac{\sum_{a=1}^{N} x_a^4}{\|x\|^4}\right\}\right). \quad (7)$$

Sketch of proof: *Let $Z = \frac{\|\Phi x\|}{\|x\|}$ then compute $\mathbb{E}\{Z^2\} = 1$ and $\mathbb{E}\{Z^4\} = 1 + \frac{2}{m}\left(1 - \mathbb{E}\left\{\frac{\sum_{a=1}^{N} x_a^4}{\|x\|^4}\right\}\right)$. These results depend on the properties of the binary matrix, namely*

$\mathbb{E}\{\phi_{i,j}\} = 0$ and $\mathbb{E}\{\phi_{i,j}^2\} = 1/m$, as well as the independence between the $\phi_{i,j}$. The final result is then obtained by applying the Bienaymé-Chebyshev inequality:

$$P\{|Z^2 - \mathbb{E}\{Z^2\}| \leq \delta\} \geq 1 - \frac{\mathbb{E}\{Z^4\} - \mathbb{E}\{Z^2\}^2}{\delta^2}.$$

The term $\mathbb{E}\left\{\frac{\sum_{a=1}^N x_a^4}{\|x\|^4}\right\}$ is always positive and less than 1 (indeed, $\frac{\sum_{a=1}^N x_a^4}{\|x\|^4} = \frac{\sum_{a=1}^N x_a^4}{\sum_{a=1}^N x_a^4 + \sum_{a=1}^N \sum_{b=1,b\neq a}^N x_a^2 x_b^2}$), so we can define a looser bound which no longer includes dependence on the signal by setting this term to 0, i.e

$$P\left\{\left|\frac{\|\Phi x\|^2}{\|x\|^2} - 1\right| \leq \delta\right\} \geq 1 - \frac{2}{m\delta^2}. \tag{8}$$

Simulations show that for normally distributed random vectors, the difference between (7) and (8) becomes smaller and smaller as N grows. Using (5) in (7) and (8) leads to

$$P\left\{(1-\varepsilon) \leq \frac{\|\Phi x\|_2}{\|x\|_2} \leq (1+\varepsilon)\right\}$$

$$\geq 1 - \frac{2}{m\varepsilon^2}\left(1 - \mathbb{E}\left\{\frac{\sum_{a=1}^N x_a^4}{\|x\|^4}\right\}\right) \geq 1 - \frac{2}{m\varepsilon^2}. \tag{9}$$

3.2. Comparison with Johnson-Lindenstrauss lemma

Similarly to figure 1, figure 2 shows the variation of the minimal value of m for eq. (2) and that induced by (8). If one requires that the distortion remains low with high probability ($P > 0.95$), both approaches lead to similar requirements for the number of projections m. However, if the required P is lower (0.5), the simple Bienaymé-Chebyshev (BC) approach shows that a much smaller m may be satisfactory.

The following simulations were also performed to compare the proposed result to lemma 1: x signals were drawn from a normally distributed random process with mean 0 and variance 1. Matrices were drawn from a Bernoulli process, with $p = 0.5$ and normalized column-wise. The simulations is run over 1000 signals and 1000 matrices. The probability estimation is done by computing the proportion of (the million) realizations when $\left|\frac{\|\Phi x\|}{\|x\|} - 1\right| \leq \varepsilon$ is true.

As it can be seen on figure 3, in the case of small m (and N), our proposed bound is positive only for large value of ϵ. However, it provides information that was not available the Johnson-Lindenstrauss lemma, as the minimal value for m is in the order of the hundreds (fig. 1). The bound is loose (but this is expected as it uses the Bienaymé-Chebyshev inequality) and the simulated probability is much higher and it is possible to have reasonnable distortion with high probability. Note that (8) was used, and that with small N value, the bound (7) would be slightly tighter.

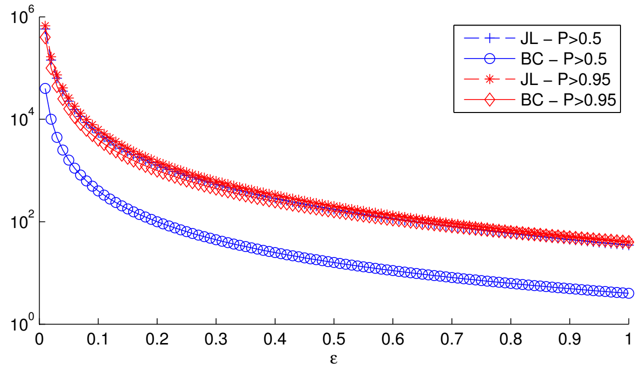

Fig. 2. Comparison of minimal value of m to achieve probabilities higher than 0.5 and 0.95, as a function of ε. The proposed bound (BC) is derived from eq. (8) as $m_{min} = \frac{2}{(1-P)\varepsilon^2}$, and compared to m defined by lemma 1 (JL), as in figure 1 with $\#Q = 1000$.

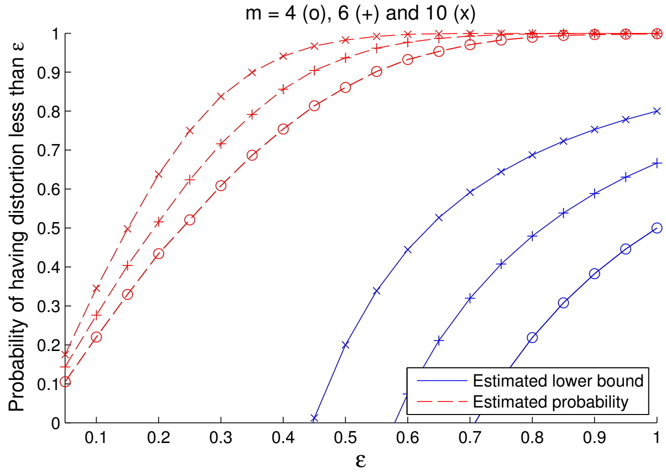

Fig. 3. Computed probability bound according to (8) (full lines), and estimated probability that $\left|\frac{\|\Phi x\|}{\|x\|} - 1\right| \leq \varepsilon$ (dashed lines), for several values (4, 6 and 10) of m. N was set to 32 but since equation (8) was used, it does not influence the bound value.

4. EXPERIMENTAL RESULTS WITH SIMULATED NEURAL SPIKES

We performed experiments on simulated neural spikes data, using the data[1] fully described in [24]. Figure 4 shows the 507 spikes as available in the data, with a color code corresponding to the original class of the spike. In this example, the dimensions m and N are small, but this is not a limitation

[1] http://www2.le.ac.uk/departments/engineering/research/bioengineering/neuroengineering-lab/spike-sorting

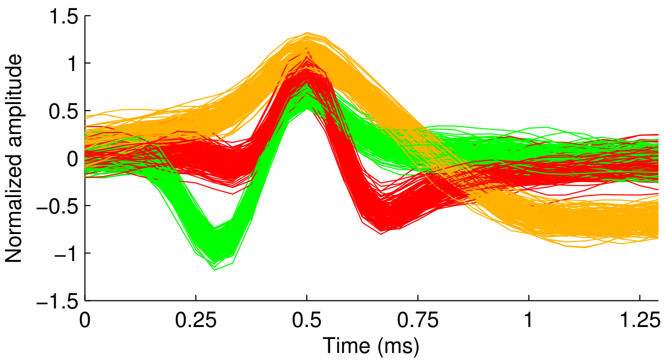

Fig. 4. Synchronized and superposed spikes, with color scheme corresponding to cluster class ($N = 32$ samples per spike).

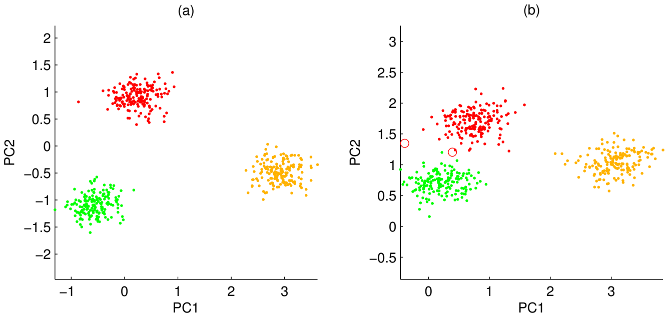

Fig. 5. (a): Projection on the first two principal components of the spikes. (b): Projection on the first two principal components of the compressed ($m = 4$, $N = 32$) spikes. The circles indicate misclassified spikes and the color is then that of the original cluster class.

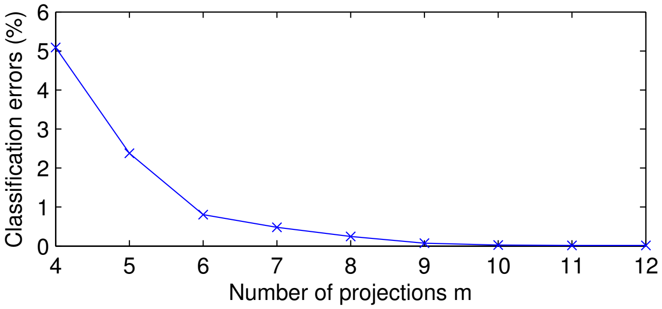

Fig. 6. Average percentage of misclassified spikes as a function of the number of projections m, with $N = 32$.

of the method.

Performing PCA on the spikes allows to see three well-separated clusters, as seen in figure 5 (a). Please note that PCA is only performed for visualization convenience. The experiment consists in projecting the spike traces on a $m \times N$ binary matrix and performing a k-*means* clustering on the projected data. The classification results are then compared to the true classes. The experiment was done for various values of m, and repeated for 1000 different matrices.

To find the cluster centers automatically, we used a method inspired by [25]: the principle is to build a minimum spanning tree with euclidean distance, and forming clusters by thresholding. The threshold limit was set to the mean plus one standard deviation of the distances in the tree, and minimal cluster size was set to $\#S/6$, where $\#S$ is the total number of spikes.

The results in table 1 show that it is possible to achieve simple compression while keeping the clustering ability in most cases. The first line gives raw proportion of misclassified spikes over the 1000 realizations, as shown in figure 6. In our example, there are 507 spikes, so 0.02% corresponds approximately to a single error. The second line shows the proportion of realizations where the number of misclassified spikes was lower than 0.5%, that is, at most 2 errors (fig. 5 (b)). Starting from $m = 6$, which corresponds to a compression ratio larger than 5, it happens with probability higher than 90%, and the average error rate is less than 1%. The two last lines deal with the number of clusters: it occurs often with high compression rates that the number of clusters is less than 3 and usually, this is because two of them appear mixed in a PCA and thus, they are not separable using distance between points. However, when $m = 10$ (that is, a compression ratio higher than 3), it never occurred.

The cost of the compression is only $m \times N$ additions, thus the higher the compression rate, the lower the cost, but it comes at the expense of clustering quality. Choosing $m = 6$ (i.e. a compression ratio of $32/6 \approx 5.33$) seems a good trade-off between clustering performance and compression cost.

5. CONCLUSIONS AND FUTURE WORK

The aim of this communication was to provide an insight in the trade-off between the minimum number of compressive measurements and the probability of successful classification when working directly in the compressed data domain. We provided a simple probability bound for random embeddings distortion that gives results for low dimensional problems, which was not possible with the Johnson-Lindenstrauss lemma. In larger dimensional cases, there is no improvement but it is expected since the bound relies on the Bienaymé-Chebyshev inequality, which is known to be loose. Experiments showed that it is actually possible to compress low dimensional data into lower dimensions with random binary embeddings, at a low computational cost, and maintain clustering results with the k-means algorithm. This gives an ex-

Number of projections m	4	5	6	7	8	9	10	11	12
Average proportion of misclassified spikes	5.09%	2.38%	0.81%	0.48%	0.25%	0.07%	0.03%	0.02%	0.02%
Less than 0.5% misclassified	69.4%	83.1%	91.8%	95.2%	97.6%	98%	99%	99.6%	99.6%
Average number of clusters	2.855	2.934	2.979	2.988	2.994	2.999	3	3	3
Probability of having less than 3 clusters	14.3%	6.6%	2.1%	1.2%	0.6%	0.1%	0%	0%	0%

Table 1. Classification performance in function of the number of projections m, $N = 32$.

ample where it is possible to exploit CS-based ideas whilst sparing the expensive reconstruction process. Future work will attempt to extend the proposed approach to other clustering algorithms such as e.g. graph-Laplacian methods.

References

[1] DL Donoho. "Compressed sensing". In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1289–1306.

[2] EJ Candes, J. Romberg, et al. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on Information Theory* 52.2 (2006), pp. 489–509.

[3] E.J. Candès, J.K. Romberg, et al. "Stable Signal Recovery from Incomplete and Inaccurate Measurements". In: *Communications on Pure and Applied Mathematics* 59 (2006), pp. 1207–1223.

[4] EJ Candes and T. Tao. "Decoding by linear programming". In: *IEEE Transactions on Information Theory* 51.12 (2005), pp. 4203–4215.

[5] MF Duarte, MA Davenport, et al. "Sparse Signal Detection from Incoherent Projections". In: *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*. Vol. 3. 2006.

[6] EJ Candes and T. Tao. "Near-optimal signal recovery from random projections: Universal encoding strategies?" In: *IEEE Transactions on Information Theory* 52.12 (2006), pp. 5406–5425.

[7] J.A. Tropp and A.C. Gilbert. "Signal recovery from random measurements via orthogonal matching pursuit". In: *IEEE Transactions on Information Theory* 53.12 (2007), p. 4655.

[8] E. Candes and J. Romberg. "l1-magic: Recovery of sparse signals via convex programming". In: *California Institute of Technology, Tech. Rep* (2005).

[9] R. Chartrand and W. Yin. "Iteratively reweighted algorithms for compressive sensing". In: *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*. 2008, pp. 3869–3872.

[10] H. Mohimani, M. Babaie-Zadeh, et al. "Sparse Recovery using Smoothed l 0 (SL0): Convergence Analysis". In: *Arxiv preprint cs.IT/1001.5073* ().

[11] MS Lewicki. "A review of methods for spike sorting: the detection and classification of neural action potentials". In: *Network: Computation in Neural Systems* 9.4 (1998), R53–R78.

[12] S. Dasgupta. "Learning mixtures of Gaussians". In: *Foundations of Computer Science, 1999. 40th Annual Symposium on*. 1999, pp. 634 –644. DOI: 10.1109/SFFCS.1999.814639.

[13] S. Dasgupta. "Experiments with random projection". In: *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2000)*. 2000, pp. 143–151.

[14] X.Z. Fern and C.E. Brodley. "Random projection for high dimensional data clustering: A cluster ensemble approach". In: *Proceedings of 20th International Conference on Machine learning*. 2003.

[15] A. Bertoni and G. Valentini. "Ensembles Based on Random Projections to Improve the Accuracy of Clustering Algorithms". In: *Neural nets 2005* 3931 (2006), p. 31.

[16] Pearson K. "On lines and planes of closest fit to systems of points in space." In: *Psychol, J.E.* 6.2 (1901), pp. 559–572.

[17] M. Belkin and P. Niyogi. "Laplacian eigenmaps for dimensionality reduction and data representation". In: *Neural computation* 15.6 (2003), pp. 1373–1396.

[18] F.R. Bach and M.I. Jordan. "Learning spectral clustering, with application to speech separation". In: *The Journal of Machine Learning Research* 7 (2006), pp. 1963–2001.

[19] E.J. Candès. "The restricted isometry property and its implications for compressed sensing". In: *Comptes rendus-Mathématique* 346.9-10 (2008), pp. 589–592.

[20] W.B. Johnson and J. Lindenstrauss. "Extensions of Lipschitz mappings into a Hilbert space". In: *Contemporary mathematics* 26.189-206 (1984), pp. 1–1.

[21] R. Baraniuk, M. Davenport, et al. "A simple proof of the restricted isometry property for random matrices". In: *Constructive Approximation* 28.3 (2008), pp. 253–263.

[22] R.G. Baraniuk and M.B. Wakin. "Random projections of smooth manifolds". In: *Foundations of Computational Mathematics* 9.1 (2009), pp. 51–77.

[23] M. Mishali and Y.C. Eldar. "Expected RIP: Conditioning of The modulated wideband converter". In: *Information Theory Workshop, 2009. ITW 2009. IEEE*. IEEE. 2009, pp. 343–347.

[24] R. Quian Quiroga, Z. Nadasdy, et al. "Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering". In: *Neural Computation* 16.8 (2004), pp. 1661–1667.

[25] L. Galluccio, O.J.J. Michel, et al. "Graph Based k-Means Clustering". In: *Elsevier Signal Processing* (2011). DOI: 10.1016/j.sigpro.2011.12.009.