Cognitive Inspired Learning based on the Compressive Sensing Postulates

Srdjan Stanković, Irena Orović
University of Montenegro, Faculty of Electrical Engineering, Podgorica
e-mail: srdjan@ac.me, irenao@ac.me

[1]*Abstract* — **This paper introduces a new generalized concept of cognitive inspired learning motivated by the basic principles used in the compressive sensing theory. The aim is to introduce a new perspective on the learning process which uses sparsity as a main premise. Cognitive inspired learning is observed as one of the possible learning modes, where the subject learns about the unknown phenomenon by identifying a sparse set of features belonging to different known basis. Rather than offering an algorithm for gaining the knowledge, we would like to draw attention to the new learning model which could potentially be used in the areas of learning applications.**

Keywords- Cognitive learning, Compressive sensing, Deep learning

I. Introduction

The concept of learning can be defined as a process of describing a certain phenomenon by a set of features bringing a better knowledge and understanding of the particular event, instance or case. Generally, this process is conducted through the stage of observation, foreknowledge-based contemplation, empirical feature selection, optimization toward efficient description, incrementing knowledge with compact information. For simplicity we might observe three levels of learning:

- Observation learning
- Empirical concluding
- Optimization of learning patterns

Observation learning is related to the function of perceiving certain process and its behavior, being able to remember and/or replicate the observed behavior. The observation learning can be also seen as an act of getting aware and storing of another situation, phenomenon, process, behavior or similar. The empirical concluding represents a process of comparing the observed behavior/phenomenon with the previous experience or background knowledge. This process is related to the cognitive ability to select the familiar and unfamiliar features and classify them based on the foreknowledge. This level of learning also includes the comprehension of important and desired range of properties that are identified to describe the observed phenomenon. The third level of learning is devoted to the optimization of phenomenon representation based on the previously identified features and properties called learning patterns. Therefore, starting from the raw observed data in the first learning phase, we obtain compact information about the observed process as a result of the third phase of learning. This compact information brings a certain amount of knowledge that is amenable to further amelioration through a new learning cycle.

The idea of modeling the concept of learning has been widely explored in machine learning applications [1]-[3]. Machine learning rely on computer algorithms for learning which are based on the observation data, instructions for behavior in situations covered by the background knowledge, and training experience from which the system will constantly learn. Certainly, the learning performance of the system will depend on the quality of training data and the type of training experience. Moreover, an important issue in machine learning applications is defining a type of knowledge that will be learnt or in other words the learning task [2]. This issue is usually formulated as a problem of learning certain target function. The target function should be designed or chosen to provide the optimal move in a certain situation. In practice, it is generally not possible to define an ideal optimal target function, but rather an operation description or the so called function approximation. Finally, for the learning algorithm we need also to choose an appropriate representation that will allow the algorithm to make the most of its capability.

In this paper we introduce a new mathematical modeling of learning concept which is based on the principles used in Compressive sensing (CS) theory. Particularly, we explore the concept of sparsity [4]-[6] in order to select the most prominent features from the observed phenomena with the aim to obtain the most compact optimal representation bringing the useful information as an increment of knowledge. The observation data can be viewed as an immense dense forest and as such it is non-optimized and illegible representation. Hence it is just an input set from which it is difficult to gain any knowledge directly. If we are able to classify and count the trees by the type, then we will have a sparse optimized representation that brings the information about the forest. It further means that the learning process progresses through the compact selected information about the observed phenomena. Since the sparsity and compact representation are the essence of the compressive sensing principles, this learning process will be referred as cognitive learning based on the CS postulates. Finally it is

This work is supported by the Montenegrin Ministry of Science, project grant funded by the World Bank loan: CS-ICT "New ICT Compressive sensing based trends applied to: multimedia, biomedicine and communications".

important to note that the purpose of this paper is not in providing a learning algorithm neither the approach for solving the described concept, but rather to bring to light the question of using the CS postulates in certain areas of learning, especially in the part of machine learning.

The paper consists of four sections. The problem formulation is given in Section II. The review of Compressive sensing theory and its prerequisites is given in Section III. The mathematical concept of cognitive inspired learning by sparsifying the representation of phenomena is given in Section IV. The concluding remarks are given in Section V.

II. PROBLEM FORMULATION

Let assume that we have an unknown process, situation or phenomenon of particular interest as the input data. This data are generally called ***observations***. In the physical processes these observation are conveyed in the form of signals with content exhibiting time or space variation captured by sensors. Hence, the observations are something that we need to learn about during the learning process in order to understand and acquire knowledge about the phenomenon we observe. The process of learning is understood as a process of describing the observed phenomenon using a relatively small set of ***features*** that allows an ***optimal sparse representation*** as an approximation function derived from the observations. The concept of sparsity is crucial in Compressive sensing for signal representation and reconstruction. Hence, the sparse representation allows us to learn about signal components that will be revealed using optimal transform domain representation. In other words, this is the way we learn about the signal nature and behavior.

Nevertheless, this problem formulation can be adapted to other learning systems as well. Let us observe an example from the educational system – process of learning in professional education. We can identify the following concepts:

- Observations – a set of actions, works, activities, goals conveying the information about all qualifications required for a certain profession
- Features – a set of a few qualifications that best cover the observed professional behavior.
- Optimal sparse representation - described by the set of subjects/courses matching the identified qualifications.

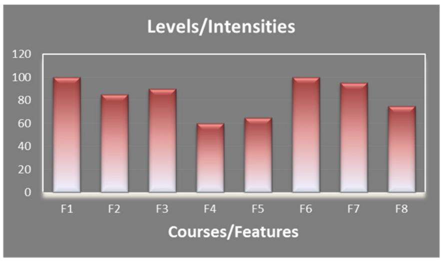


Fig. 1. Sparse representation describing the type and level of features used to describe certain qualifications in a general education system

The optimal sparse representation assumes also different levels for different courses, which indeed corresponds to the intensities of the components. Fig. 1 illustrates a set of 8 selected features (courses) with their own intensities (levels) required to describe certain professional knowledge and qualifications.

III. COMPRESSIVE SENSING THEORY

Compressive sensing is a new sensing theory in signal processing area that benefits from the two important properties, namely the sparsity and the incoherence property [4]-[13]. It has been developed as an alternative to the classical sensing approach where the number of measurements required to gain the complete information about certain process is defined by the sampling theorem: the sampling frequency should be at least twice higher than the maximal signal frequency. In many real cases, it produces a large number of samples/measurements. In CS, this number can be significantly reduced.

Observe a signal in R^n represented by using certain basis vectors $\{\psi_i\}_{i=1}^{N}$ [6]:

$$\mathbf{s}=\mathbf{\Psi x}=\sum\nolimits_{i=1}^{N} x_i\psi_i \,, \tag{1}$$

where $\mathbf{x}$ is $N\times1$ vector of weighting coefficients, while $\mathbf{\Psi}$ is the transform domain matrix (inverse). Thus, a signal $\mathbf{s}$ can be represented as a linear combination of N basis functions ψ_i multiplied by certain weights (coefficients) x_i. As mentioned before, N could have a large value in real applications. In CS, instead of $\mathbf{s}$ we are dealing with an incoherent set of measurements $\mathbf{y}$ having $M<<N$ elements. The incoherent linear measurement process is modelled by the matrix $\mathbf{\Phi}$:

$$\mathbf{y}=\mathbf{\Phi s}=\mathbf{\Phi\Psi x}=\mathbf{Ax} \,, \tag{2}$$

where $\mathbf{A}=\mathbf{\Phi\Psi}$ is usually referred to as CS matrix. Relation (2) represents a system of M linear equations with N unknowns: x=[x_1, …, x_N]. The system seems to be under-determined. However, CS concept relies on a very important sparsity assumption. In most common signal cases, the transform domain vector $\mathbf{x}$ has K out of N non-zero samples, where $K<M<<N$. In that sense, we can say that $\mathbf{s}$ is sparse when represented as a transform domain vector $\mathbf{x}$, in transform basis defined by $\mathbf{\Psi}$. More precisely, only the K elements in $\mathbf{x}$ are significant, while the remaining N-K could be neglected. It can be observed that an important issue in CS is to choose a suitable transform representation, i.e., suitable $\mathbf{\Psi}$, that will provide sparse representation $\mathbf{x}$. The problem of signal reconstruction using a set of M measurements in $\mathbf{y}$ is defined as [8]:

$$\hat{\mathbf{x}}=\min \|\mathbf{x}\|_0 \text{ subject to } \mathbf{y}=\mathbf{Ax} \,, \tag{3}$$

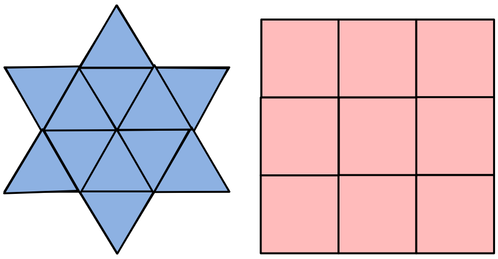
where the ℓ_0-norm represents the measure of sparsity. This means that we are looking for the sparsest $\mathbf{x}$ that corresponds to the set of measurements $\mathbf{y}$. The ℓ_0-norm based optimization problem is NP hard and does not give satisfactory results for signals that not strictly sparse.

Therefore, in practice the ℓ_1- norm minimization is used instead:

$$\hat{\mathbf{x}} = \min \| \mathbf{x} \|_1 \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x} \, . \qquad (4)$$

IV. Cognitive Inspired Learning Based on the Sparse Representation

As discussed in the previous Section, a certain process has sparse representation if it can be completely described as a combination of a few basic elements or features. For instance, if processes are modeled as geometrical shapes as shown in Fig. 2 a and b, then we might say that the star-like shape in Fig.2a can be completely covered using 12 tringles, while the rectangular shape in Fig. 2b can be completely covered by using 9 squares. The triangles and squares are called basis shapes (basis functions in signal analysis) or the features of the learning process. In other words, these features are building elements that allow us to assemble compact information about a form, process, phenomenon or similar.



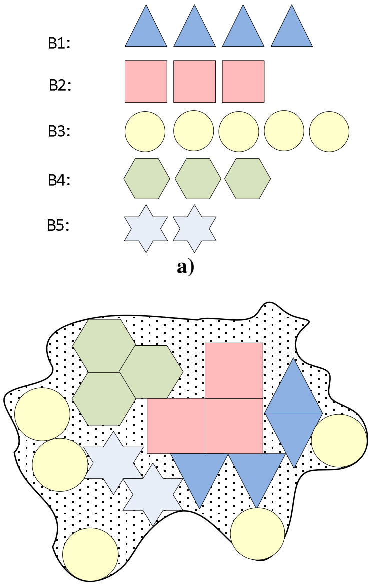
a) b)

Fig. 2 Sparse single basis representations of different spaces (phenomena or processes): a) triangle basis is used for sparse representation of star-like space, b) square basis is used for sparse representation of rectangular space

In real situations, we usually do not deal with processes that are sparse in certain basis (such as triangle basis or square basis in our example). Moreover, we are usually faced with a complex non-sparse process that we want to learn about. It means that the process cannot be represented as a simple combination of elements belonging to one basis. Then, based on the experience or the background knowledge about the observed process we need to choose several basis and a desired number of basis functions i.e., features from every considered basis. An illustration is given in Fig. 3.

We may observe that the unknown shape can be represented as a combination of N=4 elements from basis B_1, M=3 elements from B_2, K=5 elements from B_3, L=3 elements from B_4, and P=2 elements from basis B_5:

$$\begin{aligned}
&B_1: \quad \mathbf{\Psi}_1 = \left\{ \psi_1^i ,\, i=1,\dots,N \right\} \; -N \text{ shapes in } B_1 \\
&B_2: \quad \mathbf{\Psi}_2 = \left\{ \psi_2^j ,\, j=1,\dots,M \right\} \; -M \text{ shapes in } B_2 \\
&B_3: \quad \mathbf{\Psi}_3 = \left\{ \psi_3^k ,\, k=1,\dots,K \right\} \; -K \text{ shapes in } B_3 \qquad (5) \\
&B_4: \quad \mathbf{\Psi}_4 = \left\{ \psi_4^l ,\, l=1,\dots,L \right\} \; -L \text{ shapes in } B_4 \\
&B_5: \quad \mathbf{\Psi}_5 = \left\{ \psi_5^p ,\, p=1,\dots,P \right\} \; -P \text{ shapes in } B_5
\end{aligned}$$



a)

b)

Fig. 3 An illustration of a space that that is not sparse in any of the basis B_i neither as a linear combination of different basis functions

Each of the elements ψ_1^i, ψ_2^j, ψ_3^k, ψ_4^l, ψ_5^p could be also multiplied by the corresponding weights, earlier referred to as levels or intensities. In this case, the weights allow scaling of the element surface to fit best. Hence, according to the notation in the CS theory we might say that unknown process can be approximated as:

$$\begin{aligned}
&\mathit{process} \rightarrow \mathbf{\Psi}_1\mathbf{x}_1 + \mathbf{\Psi}_2\mathbf{x}_2 + \mathbf{\Psi}_3\mathbf{x}_3 + \mathbf{\Psi}_4\mathbf{x}_4 + \mathbf{\Psi}_5\mathbf{x}_5 \\
&\mathbf{x}_1 = \left\{ x_1^i ,\, i=1,\dots,N \right\} \\
&\mathbf{x}_2 = \left\{ x_2^j ,\, j=1,\dots,M \right\} \\
&\mathbf{x}_3 = \left\{ x_3^i ,\, k=1,\dots,K \right\} \qquad (6) \\
&\mathbf{x}_4 = \left\{ x_4^l ,\, l=1,\dots,L \right\} \\
&\mathbf{x}_5 = \left\{ x_5^p ,\, p=1,\dots,P \right\}
\end{aligned}$$

In a more general form, a certain process $\boldsymbol{p}$ that is subject of learning can be modelled using a space $\mathbf{S}$: $\boldsymbol{p}=\wp\{\mathbf{S}\}$, $\wp$ represents the modeling operation (or set of transformations transformations). Describing the space $\mathbf{S}$ means representing $\mathbf{S}$ as a union of subspaces $\mathbf{S}_i$ and remaining area $\mathbf{R}$ (approximation error or remainder).

$$\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_2 \cup \mathbf{S}_3 \cup \dots \cup \mathbf{S}_{n-1} \cup \mathbf{S}_n \cup \mathbf{R} \, , \qquad (7)$$

where:

$$\mathbf{S}_1 \subset \mathbf{S} \, , \mathbf{S}_2 \subset \mathbf{S} \, , \dots , \mathbf{S}_n \subset \mathbf{S} \, , \qquad (8)$$
$$\mathbf{S}_i \cap \mathbf{S}_j = \varnothing \, .$$

The subspaces of the same type do not have to be cohesive but they represent a part of the same basis:

$$\mathbf{S}_4 = \mathbf{S}_{4a} \cup \mathbf{S}_{4b} \cup \mathbf{S}_{4c} \, . \qquad (9)$$

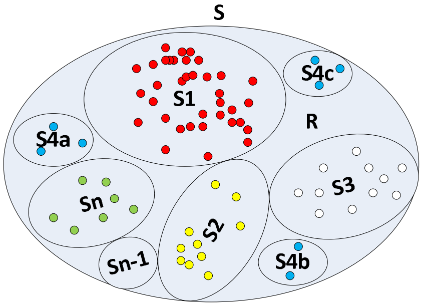


Fig. 4 Space S as a union of subspaces

Hence, the subspaces of the same type such as $\mathbf{S}_{4a}$, $\mathbf{S}_{4b}$ and $\mathbf{S}_{4c}$ could be positioned/located in the remote (distant) instances, but they are treated as a single subspace $\mathbf{S}_4$.

Furthermore, the subspaces S_i may have different number of elements representing a group of features. The total number of features (*NoF*) in **S** is:

$$NoF = card\{\mathbf{S}\} = \sum_{i=1}^{n} card\{\mathbf{S}_i\} \ . \qquad (10)$$

The features within the subspaces S_i may have various intensities which can be described by the non-uniform weighting functions $\mathbf{W}_i$. In that sense the representation of **S** can be modified as follows:

$$\mathbf{S} \to \mathbf{W}_1\mathbf{S}_1 \cup \mathbf{W}_2\mathbf{S}_2 \cup \mathbf{W}_3\mathbf{S}_3 \cup ... \cup \mathbf{W}_n\mathbf{S}_n \cup \mathbf{R}\ , \qquad (11)$$
$$\mathbf{W}_i = \left\{W_1^i, W_2^i, ..., W_p^i\right\},\ and\quad p = card(\mathbf{S}_i)\ .$$

Finally, an optimal approximation of **S**, using the known subspaces $\mathbf{S}_1$, $\mathbf{S}_2$, …, $\mathbf{S}_n$ can be written in the form of minimization problem:

$$\min \|\mathbf{S}\|_0 \ \text{subject to}\ \boldsymbol{p} = \wp\{\mathbf{S}\}\ \text{and}\ \|\mathbf{R}\|_2 < \varepsilon\ , \qquad (12)$$

or,

$$\min \|\mathbf{S}_1 \cup ... \cup \mathbf{S}_n\|_0 \ \text{subject to}$$
$$\mathbf{S} = \mathbf{W}_1\mathbf{S}_1 \cup \mathbf{W}_2\mathbf{S}_2 \cup\ ... \cup \mathbf{W}_n\mathbf{S}_n \cup \mathbf{R}\ \text{and}\ \|\mathbf{R}\|_2 < \varepsilon\ . \ (13)$$

The nonconvex ℓ_0-norm optimization can be written in the form of convex programming using the ℓ_1-norm:

$$\min \bigcup_{i=1}^{n} \|\mathbf{W}_i\mathbf{S}_i\|_1 \ \text{subject to}\ \left\| \mathbf{S} - \bigcup_{i=1}^{n} \mathbf{W}_i\mathbf{S}_i \right\|_2 < \varepsilon\ . \qquad (14)$$

Here it is assumed that the observations and the representation of phenomenon can be interpreted as a set of values (measurable quantities). When the unknown process $\boldsymbol{p}$ is described as an optimal set of features from different subspaces $\mathbf{S}_i$ (basis), then we can say that we have learnt about $\boldsymbol{p}$, such that we are able to make comparison with known processes and to drive conclusions.

Another important and very efficient measure of sparsity that can be applied to the numerical representations is called Gini coefficient. For a given set of representation elements **x=[** **x**(1), **x**(2), …, **x**(*N*)**]**, and its sorted version $\mathbf{x}_s$: $|\mathbf{x}_s(1)| \le |\mathbf{x}_s(2)| \ldots \le |\mathbf{x}_s(N)|$, the Gini coefficient is calculated as follows [14],[15]:

$$G(\mathbf{x}) = 1 - 2\sum_{i=1}^{N} \frac{|\mathbf{x}_s(i)|}{\|\mathbf{x}\|_1}\left(\frac{N-i+1/2}{N}\right). \qquad (15)$$

The Gini coefficient is independent on the size of representation, it is scale invariant, and suitable for comparing sparsity between different representations. Hence it is able to identify the sparsest domain to represent the observations, as will be shown in the numerical example.

Lastly, the remainder **R** can be used as an input of the next level of learning. After learning a set of features from the initial observations, we may continue the learning process using **R** as observations that need to be analyzed further, hence allowing a kind of deep learning. For instance, the remainder can be further represented using sparse set of features corresponding to the subspaces/basis having the highest concentration (Gini coefficient can be used again as an indicator in the case of phenomena with measurable quantities).

V. Numerical Example

This example aims to illustrate the application of the presented concept in the numerical analysis of signals. Hence we are faced with the set of observations which are time domain samples of the unknown signal. In order to simplify the example, we assume only two possible basis: B_1- Discrete Fourier transform (DFT) basis, B_2- Hermite transform (HT) basis [16], [17]. Next we measure the sparsity in both basis using the Gini coefficient such that the G_1(DFT)=0.66 and G_2(HT)=0.42 (first column in Table 1). The DFT and HT of the observation vector is shown in Fig.5. Accordingly, we can firstly identify two components (features) in the DFT domain as shown in Fig. 5. Then we remove identified components from the observation and calculated the Gini coefficient again (second column in Table 1). Now we can for example decide to identify 3 prominent components in the HT domain (Fig.5), and remove them from the observations. We might say that with this step we have finished the first phase of learning/analysis.

The next level of learning starts from the current remainder. According to the values for the Gini coefficient of the DFT and HT calculated for the remainder (third column in Table 1), we may conclude that the dominant features are in the DFT domain. Hence we may select additional three prominent components in the DFT domain (Fig.6). Then we are left with the final remainder which is negligible when compared with initial observations. The analysis has shown that the unknown process can be observed as a mix of 5 DFT components (sinusoids) and 3 Hermite components. This compact representation of features is suitable for machine learning applications.

Table 1. Gini coefficients for DFT representation G_1 and HT representation G_2

	I	II	III
G1	0.66	0.52	0.81
G2	0.42	0.64	0.58

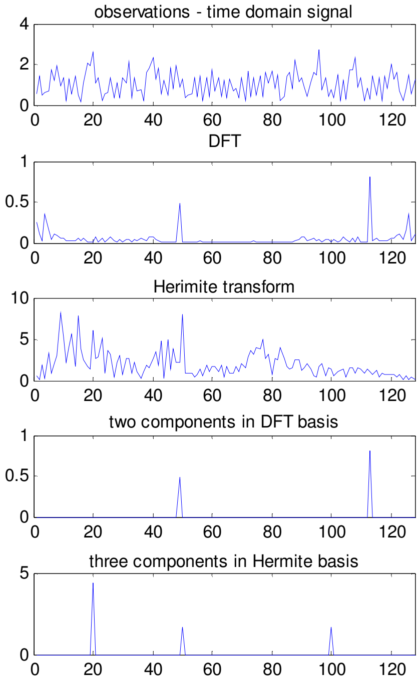


Fig. 5. Observations vector, DFT and HT representation, and identified feature in these two transform domains

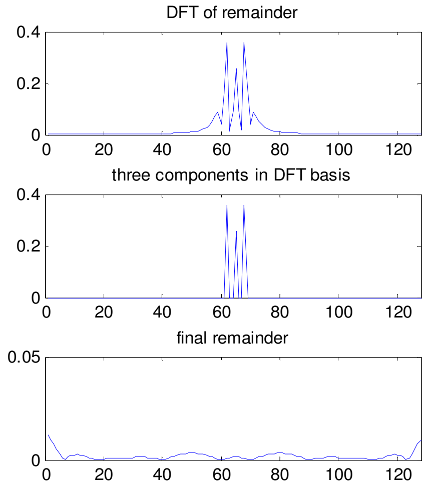


Fig. 6. DFT of the remainder after the first phase, the identified DFT components within the remainder, and the final remainder

VI. Conclusion

A new model of learning based on the sparse representation of unknown observations is presented. The learning is interpreted as process of finding optimal sparse representation of phenomenon consisted of features from different known basis. The results of the learning process will definitely depend on the number of known basis and the number of expected features within the basis. These factors are assumed to be the subject of experience or pre-knowledge, and could be treated separately in different learning areas, which might be interesting for further research. Finally for the phenomena that could be described by the measurable quantities (numerical values), the learning process is modeled using optimization problem and Gini coefficient as a sparsity measure.

References

[1] S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014

[2] T.M. Mitchel, Machine Learning, McGraw Hill, 1997

[3] I. Guyon, A. Elisseeff, “An Introduction to Variable and Feature Selection”, *Journal of Machine Learning Research*, vol. 3, (2003), pp. 1157-1182

[4] E. Candes, J. Romberg, “`l_1-magic: Recovery of Sparse Signals via Convex Programming” Caltech, 2005

[5] S. Stankovic, LJ. Stankovic, I. Orovic, "Relationship between the Robust Statistics Theory and Sparse Compressive Sensed Signals Reconstruction," *IET Signal Processing*, vol. 8, no. 3, pp. 223-229, 2014

[6] S. Stanković, I. Orović, E. Sejdić, Multimedia Signals and Systems: Basic and Advanced Algorithms for Signal Processing, Springer 2015

[7] M. A. Davenport, M. B. Wakin, "Analysis of Orthogonal Matching Pursuit Using the Restricted Isometry Property," *IEEE Transactions on Information Theory*, vol.56, no.9, pp. 4395-4401, Sept. 2010.

[8] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol.50, no.10, pp.2231-2242, Oct. 2004.

[9] S. Stanković, I. Orović, M. Amin, "L-statistics based Modification of Reconstruction Algorithms for Compressive Sensing in the Presence of Impulse Noise," *Signal Processing*, vol.93, no.11, pp. 2927-2931, 2013

[10] I. Orović, S. Stanković, “Improved higher order robust distributions based on compressive sensing reconstruction”, *IET Signal Processing*, 8 (7), 738-748.

[11] R. Monteiro and I. Adler, “Interior path following primal-dual algorithms. Part I: Linear programming”, *Mathematical Programming*, 44 (1989), pp. 27–41

[12] L. Stanković, I. Orović, S. Stanković, M. Amin, “Compressive sensing based separation of nonstationary and stationary signals overlapping in time-frequency”, *IEEE Transactions on Signal Processing*, vol. 61, no. 18, pp. 4562-4572

[13] S. Stanković, I. Orović, LJ. Stanković, "An Automated Signal Reconstruction Method based on Analysis of Compressive Sensed Signals in Noisy Environment," *Sig. Proc.*, vol. 104, pp. 43 - 50, 2014

[14] G. Li, G Chi, “A new Determining objective Weights Method-Gini Coefficient weight” *Proc. of the 2009 First IEEE International Conference on Information Science and Engineering*, pp. 3726-3729

[15] D. Zonoobi, A. A. Kassim, Y. V. Venkatesh “Gini Index as Sparsity Measure for Signal Reconstruction from Compressive Samples”, *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, No. 5, 2011

[16] S. Stanković, L. Stanković, I. Orović, “Compressive sensing approach in the Hermite transform domain”, *Mathematical Problems in Engineering* vol. 2015, Article ID 286590, 9 pages

[17] M. Brajović, I. Orović, M. Daković, S. Stanković, “Gradient-based signal reconstruction algorithm in Hermite transform domain”, *Electronics letters*, vol. 52, no. 1, pp. 41-43