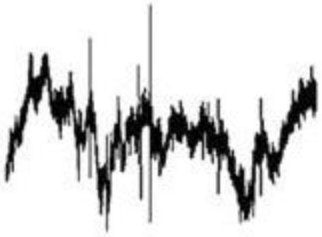
Final Project: Sparse and Redundant Representations And their applications in signal and image processing

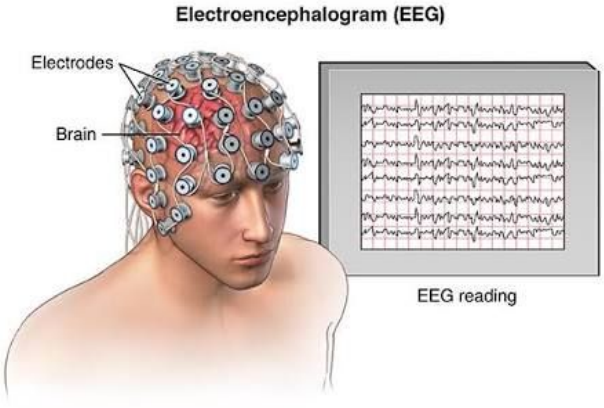
Paper: Learning the Morphology of Brain Signals Using Alpha-Stable Convolutional Sparse Coding
Reut farkash 302629118, Ori Nizan 200955474

Introduction
This paper discuss a new way of analysing neural time series data, using a variant of convolutional sparse coding named alpha stable CSC.



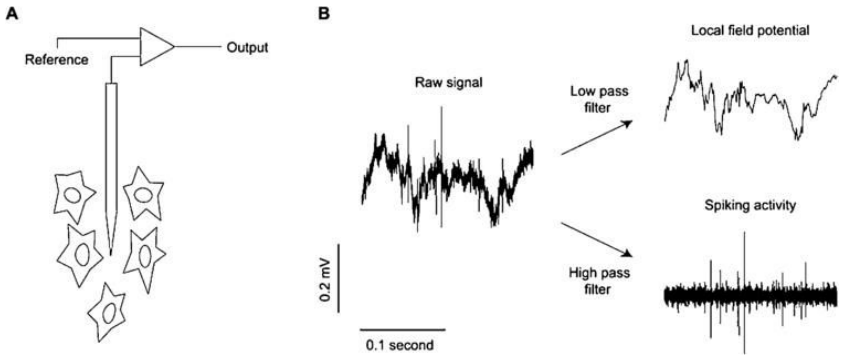
Neural time series data is characterized as having many artifacts and impulsive noise.
Existing algorithms for analysis of such signals generally have a heuristic approach which limits their applicability.
The main elements of the presented model are: a family of heavy-tailed distributions called alpha stable distributions, and a Monte Carlo expectation maximization algorithm for inference.
The connection to CSC is in the maximization step which is a weighted CSC problem for which the paper presents a computationally efficient optimization algorithm.
The main achievements of this paper are: a CSC model that is more robust to artifacts commonly seen in neural signals, and state of the art convergence times.

Examples of neural time series data considered in the paper:
EEG
EEG or Electroencephalography is an electrophysiological monitoring method to record electrical activity of the brain. Electrodes are typically placed along the scalp so it is non invasive. The method measures voltage fluctuations resulting from ionic current within the neurons. EEG is used to diagnose epilepsy, sleep disorders, brain death and other medical conditions. EEG has very high temporal resolution, on the order of milliseconds, sampling rates between 250 and 2000 Hz but on the other hand have relatively low spatial resolution.



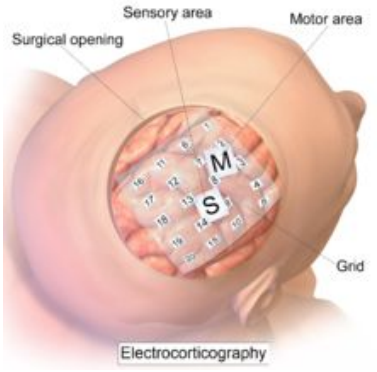
LFP

LFP - local field potential signal is generated by electric current flowing from multiple nearby neurons within a small volume of nervous tissue. voltage is recorded with a microelectrode embedded **within** neuronal tissue.



ECoG

ECoG - Electrocorticography uses electrodes placed **directly** on the exposed surface of the brain to record electrical activity from the cerebral cortex



Older algorithms

Some of the algorithms in the field which the author regardes.

MoFIT

MoTIF - Matching of time invariant filters. The paper presents an algorithm for iterative learning of generating functions that can be translated at all positions in the signal to generate a highly redundant dictionary.

The motivation is solving the following problem: Given a signal s of support of size S in a space of infinite size discrete signals: compute a good approximation $\tilde{s}_N$ as a linear superposition of N basic elements picked up in a huge collection of signals D = {φk}, referred to as a dictionary :

$$\tilde{s}_N = \Sigma_{k=0}^{N-1} c_k \phi_k, \phi_k \in D, \left\| s - \tilde{s}_N \right\|_2 \leq \epsilon$$

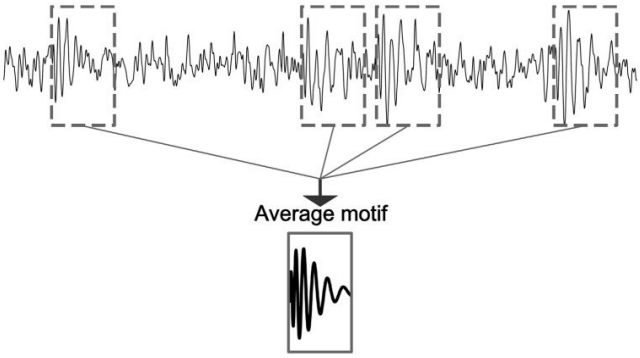
The proposed algorithm solves this problem by iteratively finding generating functions (translation invariant atoms). It does this by finding uncorrelated elements and fitting them to the given signal by way of generalized eigenvector decomposition. This process is done in a greedy manner. Each atom is found by iterating between two steps: first (after initialization) compute its best translation that fit the data and then fixing the translation and finding the best shape of the atom to fit that translation. This process repeats for each atom when the next atom is constrain to be uncorrelated to the previous ones.

The shortcoming of this algorithm is does not handle correlated atoms which are common in practice.

SWM -sliding window matching

SWM works by sliding windows across the input signal in a monte carlo way. The algorithm tries to make the content in these windows to be as similar as possible to each other then averaging their content. The result of this is a signal the paper call a motif which represent a recurring element of the signal.

The main disadvantage of this method is of its very slow speed to find these motifs.



Where older algorithms fail

Other algoritem that are commonly used for analysis of brain signals. Like the wavelet or fourier transform often are efficient to compute. But they fail to capture the morphology of the signals. Good example of the failure can be seen in the disambiguation between alpha rhythm and mu rhythm. They are both have a 10 Hz component but with different morphologies that is not visible by the fourier or wavelet transforms.

Background for the proposed algorithm:

Convolutional Sparse Coding

The basic CSC problem formulation:

$$\min_{d,z} \sum_{n=1}^{N} \left(\frac{1}{2} \|x_n - \sum_{k=1}^{K} d^k * z_n^k\|_2^2 + \lambda \sum_{k=1}^{K} \|z_n^k\|_1 \right), \quad \text{s.t.} \quad \|d^k\|_2^2 \leq 1 \text{ and } z_n^k \geq 0, \forall n, k \ ,$$

Where d^k are the atoms and z^k are their activations, K is the number of atoms and T is the length of the trail and N is the number of observed segments.

Alpha stable distributions

Lévy alpha-stable distributions is a family of distributions for which the generalized limit law exists when the number of random variables become infinite (for a sum of iid random variables). Most distributions in the alpha stable family don't have a closed form pdf, even so, it is simple to draw random samples from them (this is what makes them useful to us).

One of multiple equivalent definitions for alpha stable distributions is: F is a stable distribution if and only if, given any two positive numbers a_1 and a_2, we can find a positive number a and a real number b such that F satisfies

$$F(x/a_1) * F(x/a_2) = F\left((x-b)/a\right),$$

If this relation is satisfied with b=0 in all cases, F is strictly stable.

The family is characterized by 4 parameters:
$\alpha \in (0, 2]$ - characteristic exponent - determines the tail thickness of the distribution, (smaller $\alpha \implies$ heavier tailed distribution).
$\beta \in [-1, 1]$ - skewness parameter ($\beta = 0 \implies$ symmetric distribution).
$\sigma \in (0, \infty)$ - scale parameter - measures the spread of the random variable around its mode.
$\mu \in (-\infty, \infty)$ - location parameter.

The characteristic function of alpha stable distribution is:

$$x \sim \mathcal{S}(\alpha, \beta, \sigma, \mu) \iff \mathbb{E}[\exp(i\omega x)] = \exp(-|\sigma\omega|^{\alpha}\,[1 + i\,\mathrm{sign}(\omega)\beta\psi_{\alpha}(\omega)] + i\mu\omega)$$

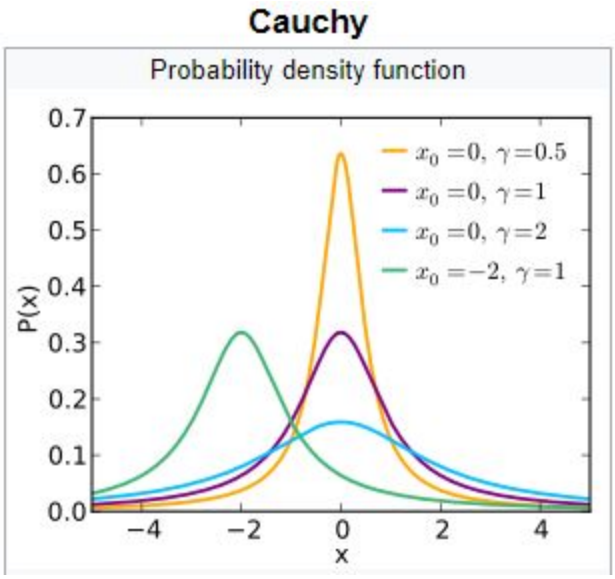
The pdf of most alpha stable distributions can not be written in closed form, except for in the following examples:

- The known Gaussian distributions for $\alpha = 2 \beta = 0$:

$$p(x) = (2\pi\sigma^2)^{-1/2}\, e^{-(x-\mu)^2/2\sigma^2} \quad \text{and} \quad \widetilde{p}(k) = e^{-i\mu\,k - \sigma^2\,k^2/2},$$

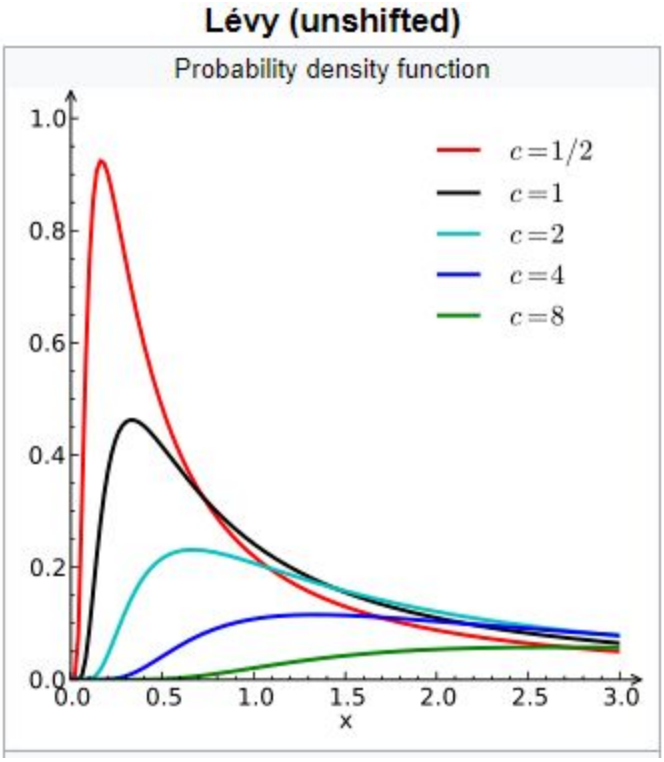
- Cauchy distributions for $\alpha = 1$:

$$p(x) = (\lambda/\pi)[(x-\mu)^2 + \lambda^2]^{-1} \quad \text{and} \quad \widetilde{p}(k) = e^{-i\mu\,k - \lambda\,|k|},$$

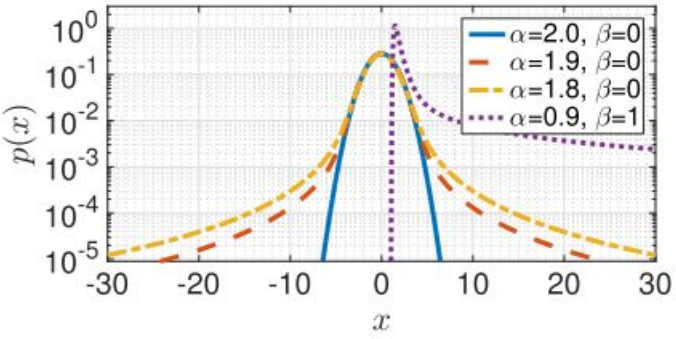


- Levy distribution for $\alpha = 0.5$:

$$p(x) = [c/(2\pi x^3)]^{1/2}\, e^{-c/(2x)} \quad \text{and} \quad \widetilde{p}(k) = e^{-c|k|^{1/2}\,(1 + i\,\mathrm{sgn}\,k)},$$



Some properties: The moments of alpha stable distributions can only be defined up to the order α. $\alpha < 2 \implies$ the distribution is heavy tailed (making Gaussians the only distributions in the alpha stable family to not be heavy tailed and to have a finite variance). $\alpha < 1 \implies$ the first moment does not exist.



PDFs of α-stable distributions

The connection of alpha stable distributions to neural signals

Gaussian models often fail at handling high amplitude noise in observations and at handling outliers (because of their light tail). Neural signals are characterized by many such artifacts (generated by the nature of the signals and the methods of data collection).

This make regular CSC models, which assume a form of Gaussian noise(l_2 error), perform badly on such tasks.

The general alpha stable family, characterized by a heavier tail, are better fitted for modeling signals with such a high variability.

The extension of the CSC model proposed in the paper, makes use of this fact by changing the Gaussian noise model assumption to a more general alpha stable noise assumption.

Expectation maximization

Expectation–maximization - EM algorithm is an iterative method to find maximum likelihood or maximum a posteriori - MAP estimates of parameters in statistical models. The algorithm alternates between performing an expectation - E step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization - M step, which computes parameters maximizing the expected log-likelihood found on the *E* step.

Alpha CSC presentation

The paper formulates the following probabilistic generative model:

$$z_{n,t}^k \sim \mathcal{E}(\lambda), \quad x_{n,t}|z,d \sim \mathcal{N}(\hat{x}_{n,t},1), \quad \text{where,} \quad \hat{x}_n \triangleq \sum_{k=1}^{K} d^k * z_n^k \ .$$

Where $z_{n,t}^k$ are the atoms coefficients, x are the signal to be learned and $\hat{x}$ is the estimated x using the dictionary and the atoms coefficients. The model assumes gaussian noise added to the estimated x.

The paper then asserts that the MAP estimate on this model: $\max_{d,z} \log p(d,z|x)$ is identical to the originally defined optimization problem:

$$\min_{d,z} \sum_{n=1}^{N} \Big(\frac{1}{2} \|x_n - \sum_{k=1}^{K} d^k * z_n^k\|_2^2 + \lambda \sum_{k=1}^{K} \|z_n^k\|_1 \Big), \quad \text{s.t.} \quad \|d^k\|_2^2 \leq 1 \text{ and } z_n^k \geq 0, \forall n,k \ ,$$

The paper suggests that regular CSC uses the assumption of gaussian noise which is why the resulting model is sensitive to outliers and the assumption of heavy tail noise help to alleviate that problem.

The paper proposes the following generalization of the probabilistic generative model:

$$z_{n,t}^k \sim \mathcal{E}(\lambda), \quad x_{n,t}|z,d \sim \mathcal{S}(\alpha, 0, 1/\sqrt{2}, \hat{x}_{n,t}) \ ,$$

where S denotes the α-stable distribution .

The new model replaces the Gaussian likelihood assumption with a more general alpha stable distribution assumption with a alpha parameter, to be learned. Note that for $\alpha = 2$ we get the original Gaussian distribution.

Take note that the paper assumes $\beta = 0$ which means symmetric distribution and $\sigma = 1/\sqrt{2}$. The paper does not present an explanation for why these parameters were fixed to these values. In particular, the choice of setting β to a fixed value of 0 is strange since it limits the possible distributions to be learned to only symmetric ones. Limiting the choice to symmetric distributions is not a reasonable assumption given our knowledge of the real world signals we are aiming to model which. These signals, as we have stated before, are characterized by abrupt and strong changes that are not symmetric in time. Examples of such changes are muscle spasms, electrode detaching at the time of the exam and such.

Now that the model is defined, the paper seek a solution, to the defined optimization problem (presented a MAP estimation):

$$(d^\star, z^\star) = \arg\max_{d,z} \sum_{n,t} \Big(\log p(x_{n,t}|d,z) + \sum_k \log p(z_{n,t}^k) \Big).$$

The solution presented is more involved then the solution to the corresponding Gaussian model due to the fact that the general alpha stable family does not admit an analytical expression. This is also true for the subset of the family the paper presented.
To overcome the lack of an analytical expression,the paper suggest reformulating the formally proposed probabilistic generative model with a conditional gaussian of the form:

$$z_{n,t}^k \sim \mathcal{E}(\lambda), \quad \phi_{n,t} \sim \mathcal{S}\Big(\frac{\alpha}{2}, 1, 2(\cos\frac{\pi\alpha}{4})^{2/\alpha}, 0\Big), \quad x_{n,t}|z,d,\phi \sim \mathcal{N}\Big(\hat{x}_{n,t}, \frac{1}{2}\phi_{n,t}\Big)$$

(here we can see why the symmetric assmetion of the noise distribution is needed, with it we can use the gaussian approximation, as Gaussians are necessarily symmetric.)
φ is called the impulse variable that is drawn from a positive α-stable distribution
(i.e. β = 1)
These two problems are identical according to the paper.
These new formulation lets as separate the problem into two subproblems that are easier to solve in an iterative manner using the EM algorithm.

The algorithm - Expectation maximization

The E step of the algorithm will be defined as follows:

$$\text{E-Step:} \qquad \mathcal{B}^{(i)}(d,z) = \mathbb{E}\left[\log p(x,\phi,z|d)\right]_{p(\phi|x,z^{(i)},d^{(i)})}$$

And the M step:

$$\text{M-Step:} \qquad (d^{(i+1)}, z^{(i+1)}) = \arg\max_{d,z} \mathcal{B}^{(i)}(d,z)$$

B(i) represent a lower bound on log p(d, z|x) and this is a tight bound at the current iterates z(i), d(i).

The E-Step

The B for our problem take the following form :

$$\mathcal{B}^{(i)}(d,z) =^{+} -\sum_{n=1}^{N}\left(\|\sqrt{w_n^{(i)}} \odot (x_n - \sum_{k=1}^{K} d^k * z_n^k)\|_2^2 + \lambda \sum_{k=1}^{K} \|z_n^k\|_1\right)$$

the weights that are defined as follows: $w_{n,t}^{(i)} \triangleq \mathbb{E}\left[1/\phi_{n,t}\right]_{p(\phi|x,z^{(i)},d^{(i)})}$.

ϕ represents the variance of the gaussian noise in our model, so we can interpret these weights as being small when the variance of x is big and vise versa. This is what makes this approach more robust than the regular one.

Evaluating the weights

The wights can only be approximated, this paper uses Markov chain Monte Carlo to find there approximation. The approximation is done as with an average sample as follows:

$$w_{n,t}^{(i)} \approx (1/J)\sum_{j=1}^{J} 1/\phi_{n,t}^{(i,j)}$$

We would like that $\phi_{n,t}^{(i,j)}$ will be drawn from the posterior distribution $p(\phi|x, z^{(i)}, d^{(i)})$ but we don't know this probability either. To draw asymptotically from the desired distribution the paper uses a Metropolis-Hasting algorithm. The algorithm works in two steps first draw a random sample from the distribution $\phi'_{n,t} \sim p(\phi_{n,t})$ for every t,n. The second step is to compute the acceptance probability:

$$\text{acc}(\phi_{n,t}^{(i,j)} \to \phi'_{n,t}) \triangleq \min\left\{1, p(x_{n,t}|d^{(i)}, z^{(i)}, \phi'_{n,t})/p(x_{n,t}|d^{(i)}, z^{(i)}, \phi_{n,t}^{(i,j)})\right\}$$

Then draw from a uniform distribution $u_{n,t} \sim \mathcal{U}([0,1])$ and for all t,n :

If $u_{n,t} < \text{acc}(\phi_{n,t}^{(i)} \to \phi'_{n,t})$ the sample is accepted and we update $\phi_{n,t}^{(i+1)} = \phi'_{n,t}$. Otherwise $\phi_{n,t}^{(i+1)} = \phi_{n,t}^{(i)}$.

M-step

The M-steps basically solves a weighted CSC with the wightes found in the previous stage. This problem is solve by iterating between fixing z and solve for d and vise versa. Fixing d gives the following expression:

$$\min_z \sum_{n=1}^{N} \left(\|\sqrt{w_n} \odot (x_n - \sum_{k=1}^{K} D^k \bar{z}_n^k)\|_2^2 + \lambda \sum_k \|z_n^k\|_1 \right) \quad \text{s.t. } z_n^k \geq 0, \forall n, k$$

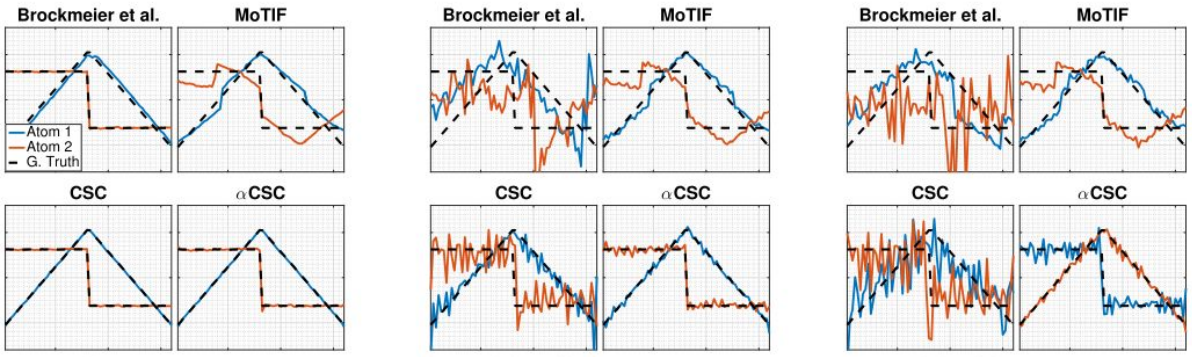
The paper solves this problem using L-BFGS-B algorithm.
For the d the expression to be minimize is:

$$\min_d \sum_{n=1}^{N} \|\sqrt{w_n} \odot (x_n - \sum_{k=1}^{K} Z_n^k d^k)\|_2^2, \quad \text{s.t. } \|d^k\|_2^2 \leq 1$$

Which can be solved using FISTA, or solve the dual problem using L-BFGS-B which the paper suggest is the more efficient way.x

Results

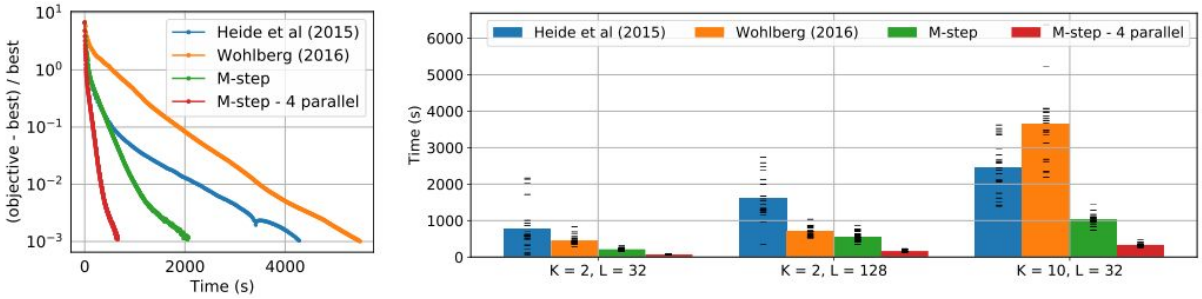
The paper claim the αCSC is more robust to corrupted data then the stat of the art CSC method. It shows that by using artificial signals made by 2 artificial atoms as can bee seen:



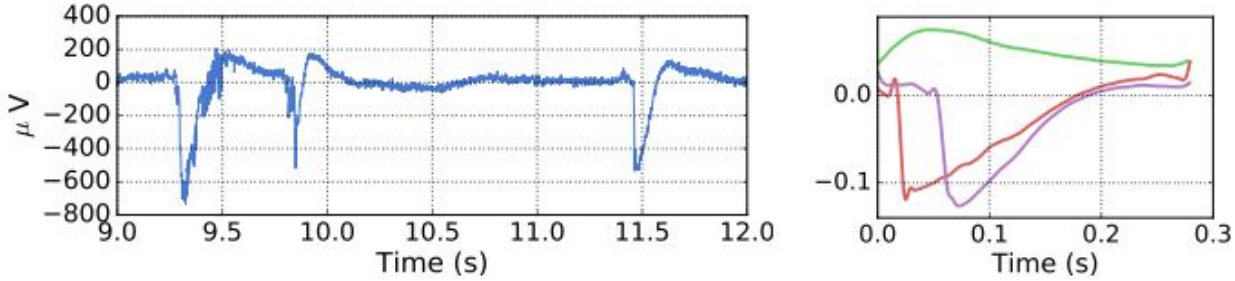
(a) No corruption. (b) 10% corruption. (c) 20% corruption

These atoms can be superimposed on the manufactured signal, in order to test the robustness, the paper added high gaussian noise to some percentage of the signals (σ = 0.1) and checks what atoms were found by different algorithms. From the results we can see that indeed the atoms that were found by αCSC looks better in comparison as the noise increase for this

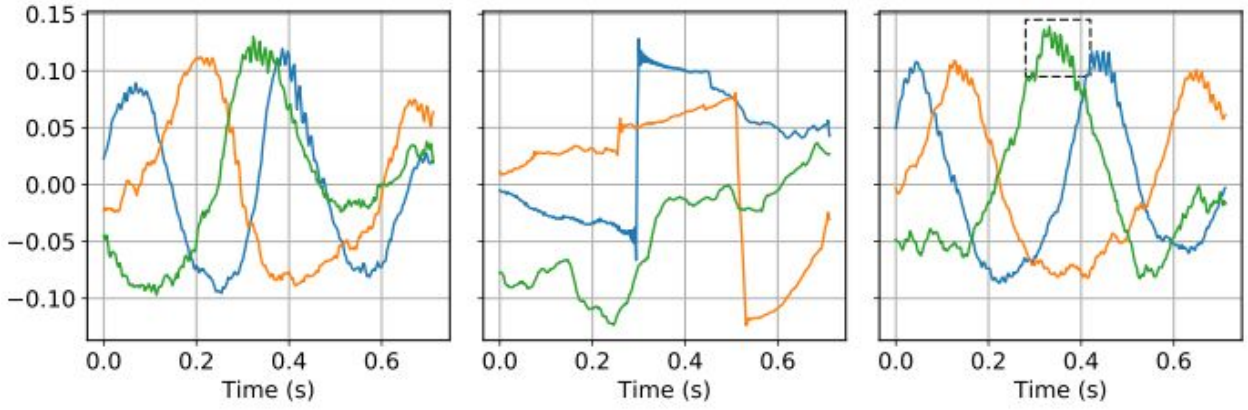
particular signal type at least. The paper does not show more types of singles and how αCSC perform on them or what are the results when more atoms are involved.
The paper also show that it M - step is more efficient and converge faster than the standard approaches on specific manufactured setup:



On real data the paper shows the algorithm achieves atoms reported in other works but without using any heuristic approach:



Farther more, on real data that was processed for noise removal the paper shows that original CSC finds good looking atoms but when the noise is included CSC preform very badly.
However αCSC perform very similar on the noisy signal to what CSC archives on clean signal.



(a) Atoms learnt by: CSC (clean data), CSC (full data), αCSC (full data)

Critic

As stated previously, the paper assumes symmetric signals in its search of the alpha stable family. It makes this assumption in order to create a model that is easier to solve but gives no explanation for why this is a reasonable assumption to make of the model. This point is even

more questionable when we consider the fact that the target signals the paper is attempting to model (neural time series data) are very much not symmetric in nature.
For example: common noise artifacts in such signals are patient muscle spasms or electrode leads being moved or detaching, both being very abrupt events (and thus are good examples for noises the model is said to be immune to) but both have a very non symmetric presentation in time.
Another point that should be noted is the lack of a quantitative analysis of the paper's results that would allow for a comparison with other papers in the field. The paper only gives a visual example of the results of different methods, but there is no metric by which to judge them.
We should not that there paper does give a quantitative metric and analysis of the running time of the different algorithms, and the suggested algorithm outperforms its competitors in certain criteria.
The paper doesn't talk about the method drawbacks and weakness.

Conclusion

We saw in this paper a relative new way of performing CSC coding, taking different assumption of the noise from the regular approaches that are better suited for gaussian type noises. The noise assumption this paper assumes is of an alpha stable distribution which is heavy tailed family of distribution and is more characteristic of noise seen in neural signals. The paper talked about their main idea behind the noise assumption and there way of using it to create a new CSC optimization algorithm. Finally the perper shows improvement on some signals types common in brain research studies.