Mini-Project (ML for Time Series) - MVA 2021/2022

Mathis Clautrier mclautrier98@gmail.com
Victor Letzelter letzelter.victor@hotmail.fr

1 Introduction and contributions

Recent denoising methods for time series tend to relax as much as possible their assumptions on the noise features, e.g its distribution, its level, the i.i.d behavior, etc. Moussallam et al. in [Mou+14] tend tackle the issue of the unknown noise level in matching pursuit algorithms. The paper proposes a stopping condition for Random Greedy pursuit algorithms, which is independent from the noise level. Their algorithm is called Blind Random Pursuit Denoising (BIRD).

To show this results, Moussallam et al. have performed several experiment to study this influence using Pink and Gaussian Noise. It has been chosen here to study the robustness of this algorithm to broader range of color noises, i.e noises whose spectral power density writes in the form $S : f \mapsto \frac{1}{f^{\alpha}}$.

This project is divided into three parts. At first, a convenient dataset was chosen and preprocessed, and an efficient algorithm for generating efficiently color noise was implemented. Then, a study of the robustness of the blindnoising algorithm has been carried out for $\alpha \in [\![-2, 2]\!]$. The results obtained were compared with other state-of-the-art algorithms in time series denoising, such that Randomized Orthogonal Matching Pursuit [EY09], Stochastic Matching Pursuit [Pee+12], and Wavelet Shrinkage method [Koz05]. Finally, experiments were carried out to study the influence of the noise type with the number of iterations of the BIRD algorithm when compared with the optimal number of iterations using the DTW metric.

For this project, the work was distributed between the two students (M.Clautrier and V.Letzelter). All the work was done in a collaborative manner, although some parts have been the subject of more specific work by one of the students. M.Clautier focuses mainly on the choice, the preprocessing of the dataset, and the implementation of the state-of-the-art matching pursuit algorithms whereas V.Letzelter concentrated on the color noise generation, hypothesis checking, the wavelet theory, the Skrinkage Algorithm and the sensitivity analysis. Maximum effort was made to implement a possible all the methods used ($\simeq$ 80% of the code). The experiments proposed were new and not carried out by the author for all noises types. This study will show us that the method is very sensible to the noise type, as well as the type of dictionary used. This gives room for improving the results of the algorithm.

2 Method

Formalization of the problem Matching pursuit algorithms aim at finding a sparse approximation of a signal $y \triangleq (f(t_i))_{i \in [\![1,N]\!]}$. Given a dictionary $\Phi = (\phi_1, ..., \phi_M) \in \mathbb{R}^{N \times M}$ one wants to select an activation vector $\alpha \in \mathbb{R}^M$ that minimizes the quantity $||y - \Phi\alpha||$.

The *sparse* approximation requires that $||\alpha||_0 \triangleq \sum_{i,\alpha_i \neq 0} 1 \ll k$, where $k \in \mathbb{R}_+^*$ has to be chosen. This decomposition is relevant provided that the dictionary Φ (usually redundant) is composed of functions $(\phi_i(t))$ which are localized in the time and frequency domain. In this case, the dictionary used was Modulated Discrete Cosine Transforms as in [Mou+14]. The aim of the algorithm is to construct a sequence $(\gamma_n) \in [\![1, M]\!]^T$ of wisely chosen T atoms such that the informative part of the signal y is well represented, and the noisy part $\tilde{y}$ is not represented in the decomposition. To do so, the method proposed in [Ste93] consists on defining at each step n the residual $R^n f$ corresponding to the information which remains to extract in the the signal y, with $R^0 f = f$. At each step n, the atom

$$\phi_{\gamma_n} \triangleq \operatorname{argmax}_{\phi \in \Phi} \frac{|\langle y, \phi \rangle|}{\|y\|_2} \tag{1}$$

will be chosen and the residual is updated with the relation $R^{n+1} f \triangleq R^n f - \langle R^n f, \phi_{\gamma_n} \rangle \phi_{\gamma_n}$. We therefore immediately have, $\forall m \geq 1,\ f = \sum_{k=1}^{m} \langle R^k f, \phi_{\gamma_k} \rangle \phi_{\gamma_k} + R^m f$.

The contribution of [Mou+14] relies on two improvements of the initial algorithm. At first, a random component is added the algorithm by substituting, at each iteration, Φ with $\Phi_n \subset \Phi$ in equation 1, with Φ_n about 50 smaller (in cardinal) than Φ. Running J instances of this algorithm, one obtains $\left\{\hat{y}^j\right\}_{j=1..J}$ approximation of the clean signal, and the final result can be obtained by averaging. Secondly, [Mou+14] proposed a stopping criterion for choosing an adequate number of iterations before beginning learning the noise in the signal. More precisely, if we denote p the probability that at a given step k the index γ is added to the sequence of $(\gamma_n)_{n \in [\![0, k-1]\!]}$ when R^k is a pure noise, the quantity $\Lambda_W(\Phi, p) = \frac{\sqrt{2}}{N} \sqrt{1 - \frac{2}{\pi}} \operatorname{erfinv}\left((1-p)^{\frac{1}{M}}\right)$ is defined such that the algorithms stops at step k whenever the condition

$$\lambda_\Phi(R^k f) \triangleq \sup_{\phi \in \Phi} \frac{|\langle R^k f, \phi \rangle|}{\|R^k f\|_2} \leq \Lambda_W(\Phi, p) \tag{2}$$

is satisfied (see [Mou+14]). This condition has been derived supposing notably that (i) the projections of the noise w on the atoms $\forall m \in [1, M], z_m^w = \frac{|\langle w, \phi_m \rangle|}{\|w\|_2}$ are the realizations of a random variable Z^w whose distribution is $\mathcal{N}(0, \cdot)$ and (ii) the distribution of the projections of the informative signal has a heavier tail than the Gaussian Distribution (see these checks in section 4, in Figure 5 and Figure 6).

Moussallam et al. conducted several experiments in [Mou+14]. They notably study the robustness of the algorithm with regards to other state-of-the-art algorithms (Orthogonal MP, Stochastic MP, Shrinkage Algorithm) when applying a Gaussian noise and a Pink noise on a clean signal of interest. On that account, we decided to extend these experiment to a broader range of color noises. Let us precise the method chosen for noise generation.

Generation of color noise For the noise generation, it has been chosen to take profit of the spectral representation of a real valued stationary Gaussian process $(Z(t))_{t \in \mathbb{R}}$,

$$Z(t) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-i\omega t} \sqrt{\hat{c}(\omega)} \hat{n}_\omega \mathrm{d}\omega$$

where $c(t)$ is the covariance function of the process, $\hat{n}_\omega$ and is the Fourier transform of a real values white noise. This representation allows to sample efficiently the realization of a color noise $\boldsymbol{Z} \triangleq (Z(t_1), ..., Z(t_n))$ with $\boldsymbol{c} = (c(t_1), \ldots, c(t_n))^t$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t$ a gaussian random vector using the relation: $\boldsymbol{Z} = \mathrm{IFT}(\sqrt{\mathrm{DFT}(\boldsymbol{c})} \cdot \times \mathrm{DFT}(\boldsymbol{Y}))$. When dealing with color noise, the power spectral density, which is the Fourier transform of the covariance function, is written as $S(f) = \frac{1}{f^{\alpha}}$. Therefore, the term $\mathrm{DFT}(\boldsymbol{c})$ can be replaced with $(S(f_1), ...S(f_n))$, with (f_i) the samples of frequencies considered. To make the experiment more interpretable, it has been otherwise chosen to re scale the noise obtained by a factor in the form $\sigma \triangleq \wp \times \frac{a}{b}$ where a is the maximum value of the signal, and b is the maximum value of the noise (or a quantile of order β) and $\wp \in]0, 1[$. This way, Purple, Blue, White, Pink, and Brown noises, corresponding respectively to $\alpha \in \{-2, .., 2\}$ were generated.

Robustness of matching pursuit bird algorithm when compared with state-of-the-art denoising algorithms For the matching pursuit algorithms we used a dictionary containing all MDCDT atoms (Modified Cosine Transform of window functions) for scale L in $[32, 64, 128]$. As there is no proper stopping criterion for the stochastic matching pursuit (and the orthogonal one) and for the matching pursuit, we used an oracle version of these algorithms (like in the original paper) which stop once the DTW to the original signal (i.e the one without noise) is below a given threshold ($\epsilon = 180$) or once the number of iterations reach a given maximum ($T_{\max} = 800$).

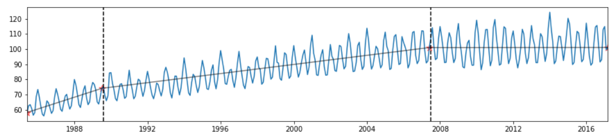
For the SMP [Pee+12], the whole dictionary is used for a first iteration, then a subdictionary is randomly sampled $J - 1$ times for running $J - 1$ pursuit. The J solutions are finally averaged to obtain a final reconstruction. For the SOMP [EY09] the whole dictionary is used for the J runs. However, the new atom a_t is not selected by taking the maximum of $\mathbf{p}_t = \mathbf{D}^T \mathbf{r}_{t-1}$ but by sampling using $\exp(c^2 \mathbf{p}_t^2 / 2\sigma^2)$ as unnormalized distribution. However, as this vector can lead to overflow when taking the exponential, we directly sampled from the log-unnormalized-probability $l_t = c^2 \mathbf{p}_t^2 / 2\sigma^2$ using the Gumbel trick. To do so we chose $a_t = \mathbf{D}_{i^*}^T$ where $i^* = \arg\max l_t + g_t$ and g_t is drawn from the Gumbel standard distribution.

Finally, the Shrinkage algorithm proposed in [Koz05] was applied, using soft and hard shrinkage functions, a debauchies (db4) wavelet dictionary, and using the standard deviation approach for estimating the threshold value (see [Koz05]).

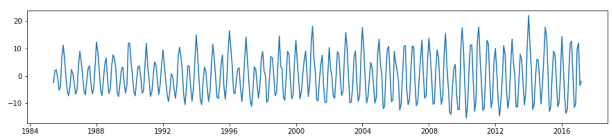
Sensibility of the stopping criterion of the BIRD algorithm to the noise type As a second part, it has been chosen to study the influence of the noise type on the behavior of the [Mou+14] algorithm (namely the BIRD algorithm) specifically when it comes to the number of iterations of the algorithm. More precisely, a statistical study has been carried out to study the influence of the noise type on the relationship between (i) the number of iterations of the algorithm before reaching the condition (2) (for a given value of p), (ii) the optimal number of iteration (i.e such that the DTW between the current signal and the clean signal is minimum), (iii) the number of iterations such that the DTW between the current signal and the noisy signal is minimum (which is close to total number of iteration executed).

3 Data

The dataset used comes from Kaggle and contains electrical production (supposed of Australia) from 1985 to 2017. You can download the processed file here dataset. The unit of these data is not provided and the points are sampled every first of the month. The lowest value is equal to 55.8 while the greatest one is equal to 124.3. As shown in Figure 1, the original signal is clearly not stationary with two detected change points between which the trend seems relatively linear. It is interesting to note that these two points correspond to the end of the cold war and the crisis of 2008 and correspond to a smaller derivative of the trend (i.e the trend is less increasing). A supposed stationary signal has been obtained by removing the linear trend. To evaluate the stationarity of the two trends we performed a Dicker-Fuller test which has been rejected for the initial signal ($p = 0.094$) and accepted for the residual signal ($p = 10^{-7}$).



(a) Initial signal with detected change points



(b) Signal obtained by removing a piecewise-linear trend

Figure 1: Representation of the time series used in this project with and without trend

As the residual signal seems relatively periodic we investigated the autocorrelation of the two signals in Figure 2. First, the initial signal appears again not stationary since its autocorrelation does not decrease with the gap. Second, the residual signal appears to be periodic with a period of 6 month and 12 month. This periods are natural since air conditioning consume a lot during summer the same goes for heating during winter. To further investigate the periods of the residual signal we analyzed its periodogram Figure 3 where the previous frequencies (6 and 12 months) are again revealed. Finally, as the residual signal is an oscillating signal around 0 of periods 6 and 12 we decided to obtain the trend of the original series by average moving of length $T = 12$ so that the oscillations will be removed. The obtained signal Figure 4 is close to the previous assumptions (piece-wise linear with two rupture points). As this signal is very smooth compared to the original one we will be using it for the following experiments on denoising.

4 Results

Let us display results obtained when applying denoising algorithm after applying different noise types. As references values, for a given realization of each noisy the DTW distance between the noisy and the original distance are given in Table 1. Visuals displaying the results on the state-of-the-art denoising algorithms are given in Figure 7. The associated performance of each algorithm is given for each of Purple, Blue, White, Pink, and Brown color noises in Table 2.

Noise Color	Purple	Blue	White	Pink	Brown
DTW w. clean signal	322.23	353.09	325.01	311.07	155.42

Table 1: DTW distances with the original sigal for each algorithms and noise types

Many comments can be made from Table 2. At first, we notice that the Shrinksoft algorithm provides the best results for each noise type. We suppose that this fact is due to the fact that, as proposed in [Mou+14], another dictionary was used when applying this algorithm (namely

Noise color vs. Algorithm	BIRD	SMP	SOMP	MP	Shrinksoft	Shrinkhard
Purple	158.52	206.15	322.23	196.25	**120.96**	255.66
Blue	182.45	273.24	353.09	203.89	**123.32**	260.58
White	190.17	222.04	325.01	226.91	**110.61**	266.79
Pink	144.95	323.83	311.07	198.25	**120.19**	242.15
Brown	159.92	310.33	157.42	197.52	**121.08**	128.79

Table 2: DTW distances with the original clean signal for each algorithms and noise types

the debauchies 4 dictionary). Otherwise, we notice that the Shrinkage algorithm is the only one which denoise the data when dealing with Brownian effectively the data (for the other algorithm, the DTW distance with the clean is higher than the distance between the original noisy signal and the clean signal). Furthermore, we notice that the SOMP does not perform well on the denoising task, as it reconstructs almost perfectly the Noisy signal, and therefore shares the same dtw distances with this one. Likewise, the SMP shows very bad performances when dealing with Pink or Brownian Noise ; we also supposed that this issue is due to the dictionary choice.

Otherwise, the results of the sensitivity analysis is displayed in Table 3. This experiment displays three quantities when applying the BIRD algorithm on different noise types (all averaged on $N = 100$ runs with different noise realizations), namely (i) the iteration index which corresponds to the minimum of the DTW distance with the clean signal (i.e the optimal stopping iteration index), (ii) the index which corresponds to the minimum distance with the noisy signal, and (iii) the index on which the algorithm would have stopped when applying the stopping criterion with $p = 10^{-10}$ (the algorithm was run with $N_{iter\ max} = 1000$).

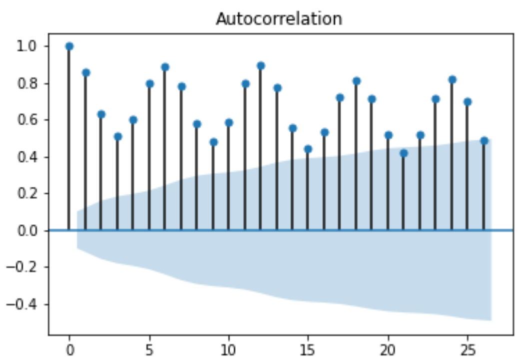
Let us highlight that the column Avg. index min DTW noisy has been given as guide, and has the highest index values, which is coherent as the algorithm tend to overfit with the noisy data when the number of iteration increases. Otherwise several conclusions can be drawn from those results. At first, we notice that the Avg. Index with minimum DTW value is a decreasing function of α ; the BIRD tend to learn faster noises with high α values ($\alpha = 0, 1, 2$) than noises with low α values ($\alpha = -2, -1$). Otherwise, the stopping criterion is almost constant for each α values, and is therefore not dependant on the noise type. More precisely, the absolute difference between the Avg. Index min and the Avg. Index stop tend to be lower for Blue and the White ($\alpha = -1, 0$) noises, which are those on which the stopping index has the most reliable values. This show the limits of this stopping condition proposed by the article.

To interpret this discrepancy of results for different noise types we checked whether the condition of Gaussian Centered distribution of the noise projection on the dictionary chosen is verified in Figure 5. We noticed that, as α increases, this distribution get further from the GD. This results possibly explains the failure of the BIRD stopping condition when dealing with pink or brownian noises. When it comes to the second condition related to the stopping criterion, this one is completely verified (See Figure 6) ; the tail of the distribution of the projection of informative signal on the dictionary (which has significant values until ~ -1000) is heavier than the one of the projection of the color noises.

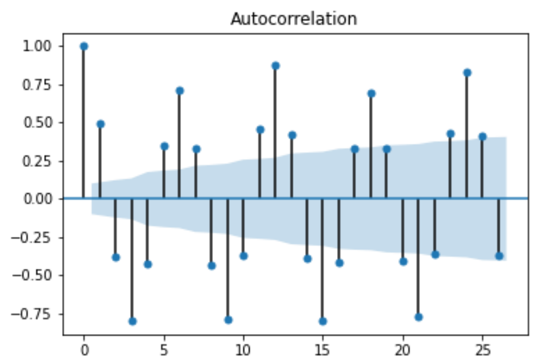
Noise color vs. Algorithm	Avg. Index min DTW	Avg. Index min DTW noisy	Avg. Index stop
Purple ($\alpha = -2$)	877.58 ± 98.79	919.24 ± 76.18	562.44 ± 77.71
Blue ($\alpha = -1$)	673.73 ± 133.42	921.49 ± 76.97	578.97 ± 78.00
White ($\alpha = 0$)	514.08 ± 84.26	950.38 ± 57.56	619.12 ± 88.84
Pink ($\alpha = 1$)	491.95 ± 89.00	948.11 ± 56.52	644.61 ± 92.14
Brown ($\alpha = 2$)	507.41 ± 86.88	943.99 ± 49.43	629.3 ± 84.83

Table 3: Index values for different noise types

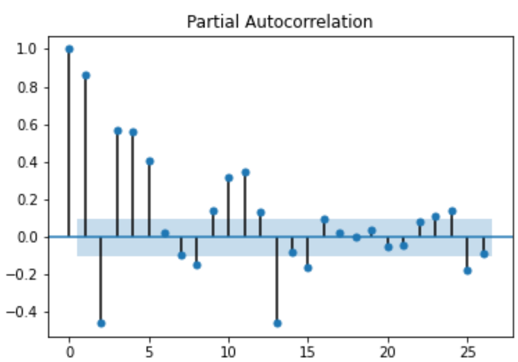
5 Annex



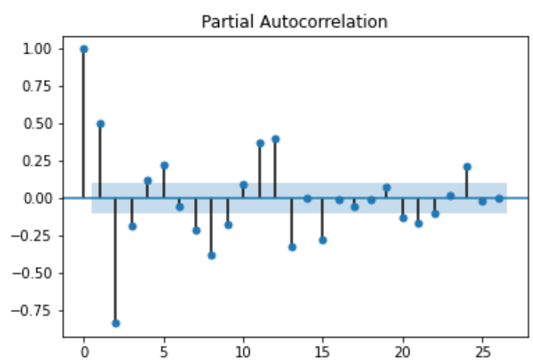
(a) Autocorrelation of the initial signal



(b) Autocorrelation of the residual signal



(c) Partial-autocorrelation of the initial signal



(b) Partial-autocorrelation of the residual signal

Figure 2: Autocorrelation and partial-autocorrelation of the two signals.

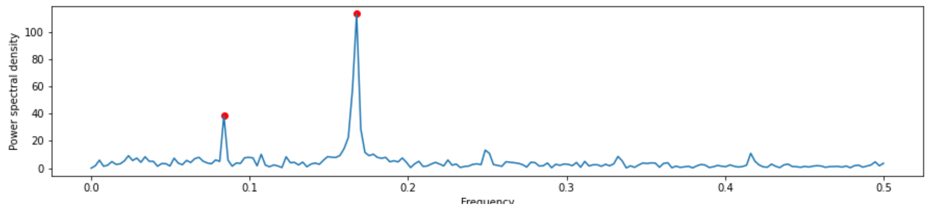


Figure 3: Periodogram of the residual signal

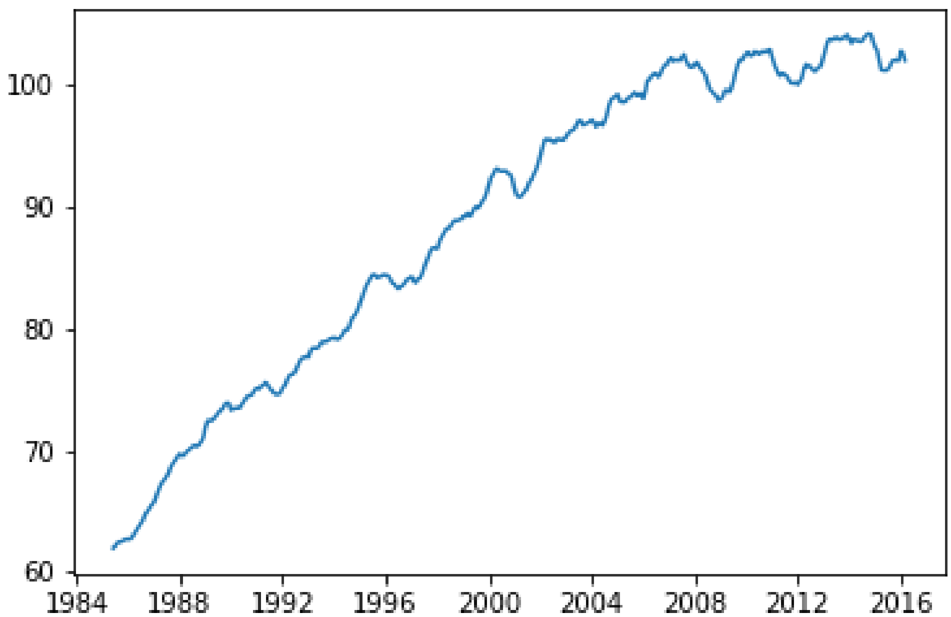


Figure 4: Residual signal obtained by moving average on the original signal with $T = 12$

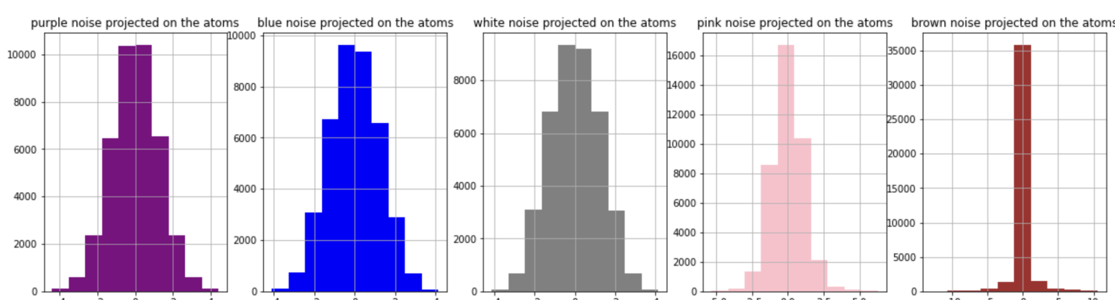


Figure 5: Distribution of the projection of the color noises on the MCDT dictionary

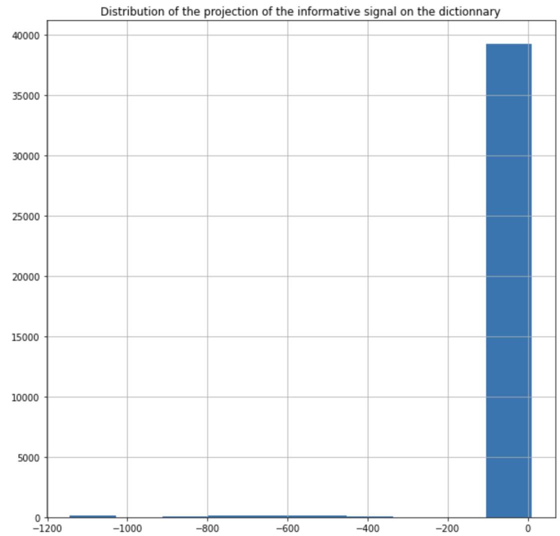


Figure 6: Distribution of the projection of the informative signal on the MCDT dictionary

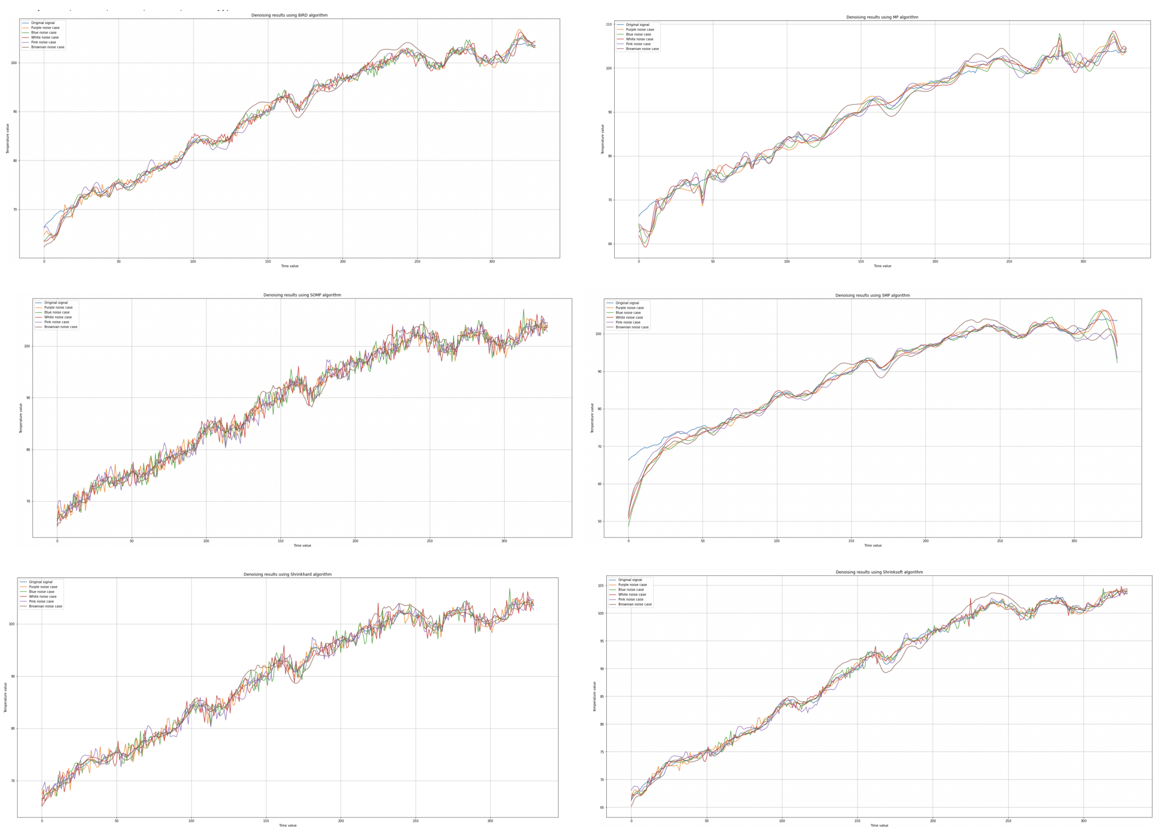


Figure 7: Visual results of the denoising algorithms

6 Bibliography

References

[Ste93] Zhifeng Zhang Stephane Mallat. "Matching pursuit with time-frequency dictionaries". In: *IEEE Transactions on signal processing*. Citeseer. 1993.

[Koz05] Bartosz Kozłowski. "Time series denoising with wavelet transform". In: *Journal of Telecommunications and Information Technology* (2005), pp. 91–95.

[EY09] Michael Elad and Irad Yavneh. "A plurality of sparse representations is better than the sparsest one alone". In: *IEEE Transactions on Information Theory* 55.10 (2009), pp. 4701–4714.

[Pee+12] Thomas Peel et al. "Matching pursuit with stochastic selection". In: *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE. 2012, pp. 879–883.

[Mou+14] Manuel Moussallam et al. "Blind denoising with random greedy pursuits". In: *IEEE Signal Processing Letters* 21.11 (2014), pp. 1341–1345.