Biomedical data processing
Assignment 5 - ICA/Wavelet

Marine Chaput

December 21, 2018

1 Introduction

For this final project, we will explore in the first part, algorithm in data analysis with a focus on the fastICA to remove noise and artefacts. In addition, this algorithm will be applied to a mixture problem between mother and fetus ECG.
In a second part, we will work on the wavelet analysis to decompose and recompose signals. To finish by using the decomposition signals energy as features to create a better model used to detect epilepsy as in the last assignment.

2 Independent Component Analysis (ICA) and Electroencephalogram (EEG)

2.1 FastICA applied on 4 sources with different mixture matrix

In this first part, the fastICA algorithm will be applied on different artificial mixtures composed of 3 sources (resp and ECG, surface EMG and PPG signal) plus a white Gaussian noise signal.
The input data must be, first, prewhitened which means centered (mean=0) and whitened (variance=1). To respect this condition, all sources are normalized before creating any mixtures.
The output of the algorithm is correlated with original sources. The result of the correlation is used as a similarity metric to match each original source with its estimated signal. To finish, the Root Mean Squared Error is computed between matching signals as a performance metric.

Three different mixtures are tested below :

- First mixture : 4 signals composed from 3 sources + 1 Gaussian noise signal
- Seconde mixture : 6 signals composed from 3 sources + 1 Gaussian noise signal
- Third mixture : 4 signals composed from 3 sources + 1 Gaussian noise signal + 1 random noise

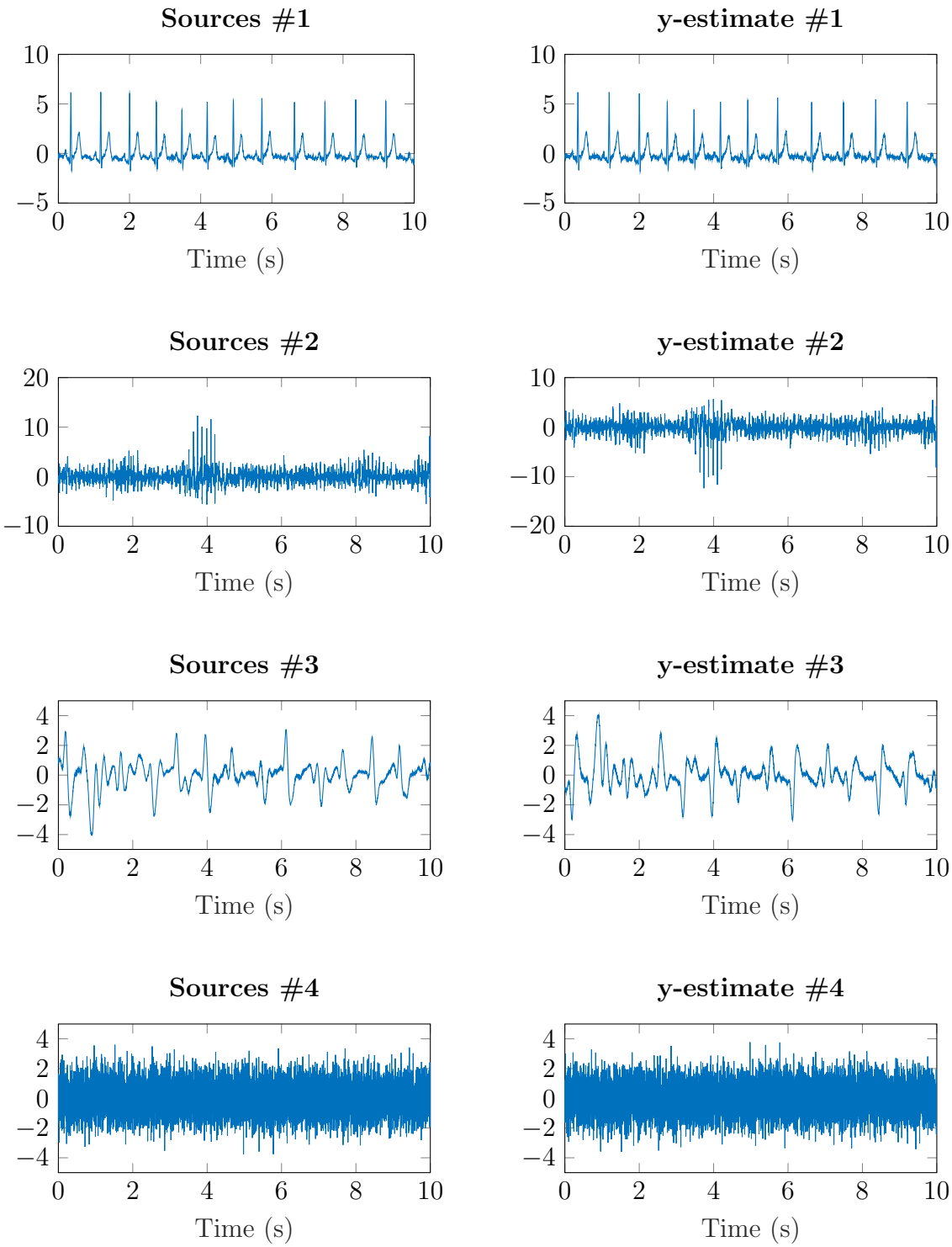


Figure 1: Result of the fastICA after the first mixture

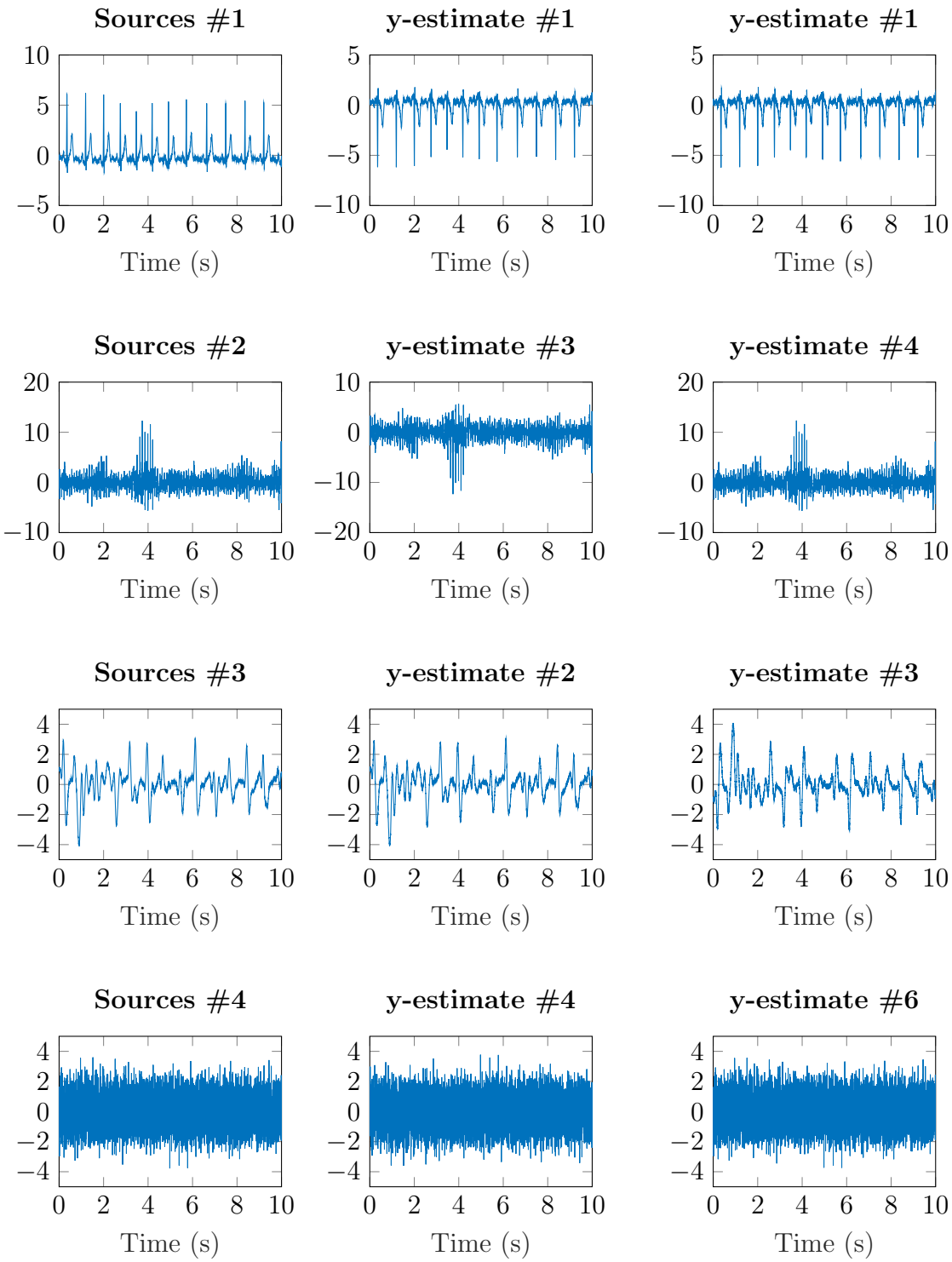


Figure 2: Result of the fastICA after the second and third mixtures

In the first mixture, all signals are correctly identified by the algorithm.
In the second mixture, the algorithm identified the right number of sources by reducing the number of signals from 6 (input) to 4 (output). fastICA is based the minimization of mutual information. Each independent component is computed one by one until optimal result is found. It is not linked to the amount of mixture signals given in input.
In the third mixture, a random noise is added. The algorithm hasn't been able to identify the optimal number of sources. Separations appeared by maximizing non-Gaussianity in the estimated components. So, if an additional Gaussian noise is added, the fastICA wouldn't be able to separate it from the Gaussian noise already present in the mixture. Both noises are statistically dependent.
We notice that estimated signals have a different sign than expected. Since for the algorithm and in real life, sources and mixing matrix are unknown, any multiplicative factor including a change of sign could be absorbed by coefficients in the mixing matrix.
The RMSE gives us an interesting performance metric for each test.

	Mixture_1	Mixture_2	Mixture_3
ECG	0.045032	0.045036	0.053324
EMG	0.03072	0.038197	0.057788
PPG	0.02244	0.031847	0.053009
White noise	0.012833	0.013032	0.043297

Figure 3: RMSE

In every case, RMSE result is very low. Best case is, of course, the first one tests because it is the easiest.

To finish, with this case study, fastICA is used for artefact removal. In the first mixture, we want to remove the noise and the EMG signal.
Coefficients related to these both signals are replaced by zeros in the demixing matrix.
The result is computed by multiplying the new matrix with the estimated sources.
As presented above, signals are free from EMG artefacts and noise.

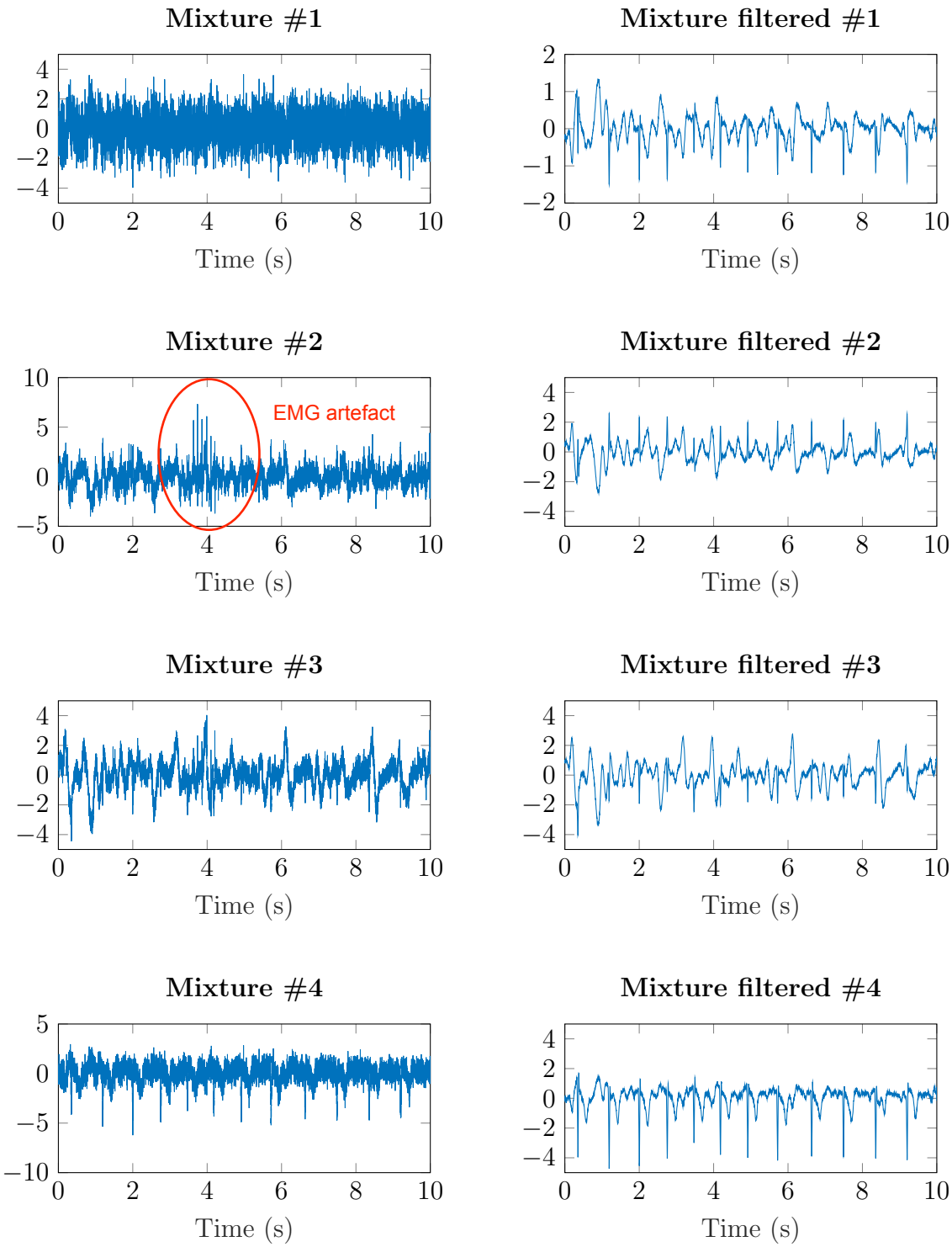


Figure 4: Mixture with and without filtering by ICA

2.2 FastICA applied on a mixed ECG between a fetus and his mother

The data, used in this part, is composed of 4 ECG signals from the mother and a direct measure of the fetus ECG.

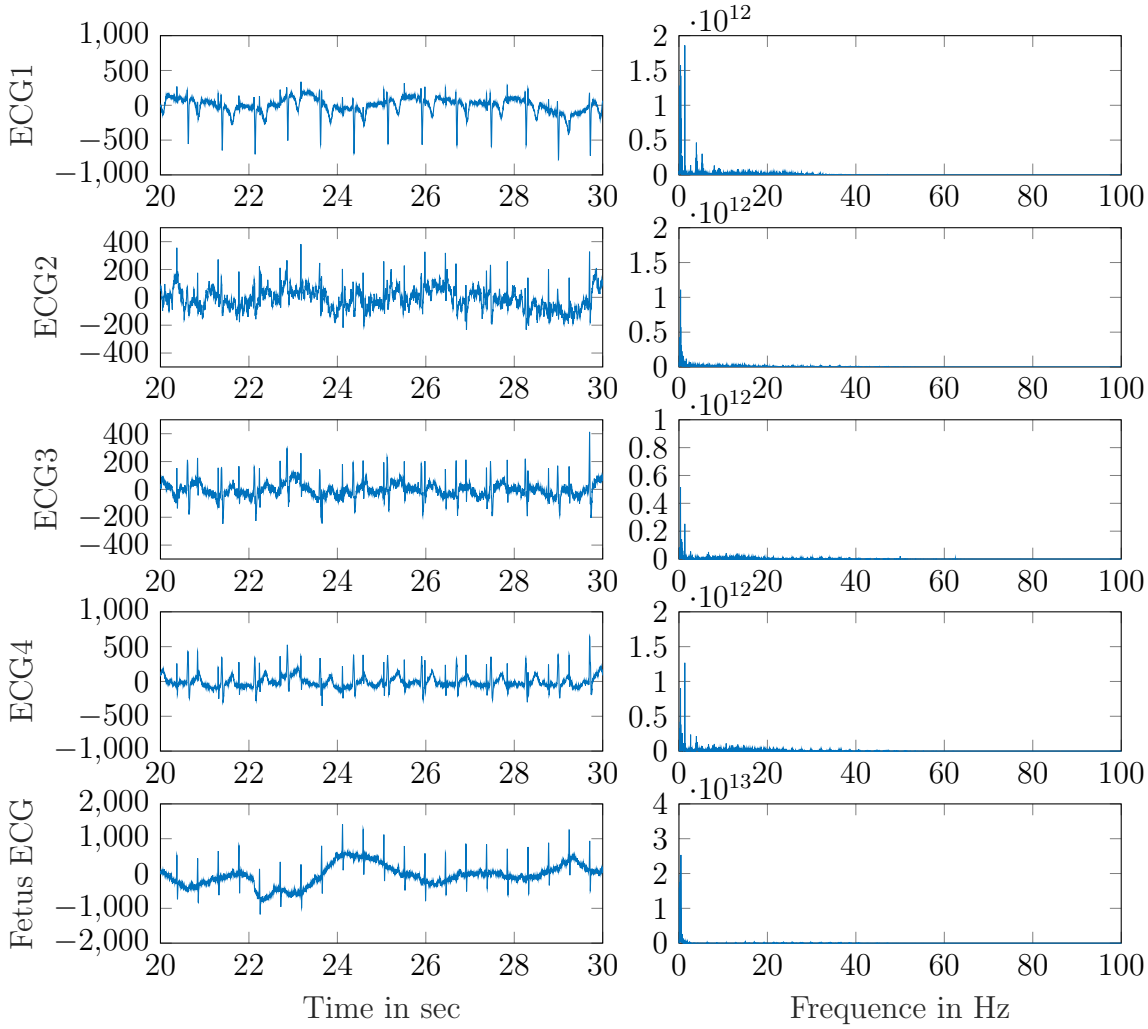


Figure 5: ECG in time and frequency domain

Theses signals have a baseline drift problem, resolved by applying a high-pass filter on it. We notice the direct fetus ECG has twice the amplitude of the ECG signals from the mother.
An ECG signal is identified as the most similar to the direct measure by correlation in 3 cases .

- After filtering
- After PCA
- After ICA

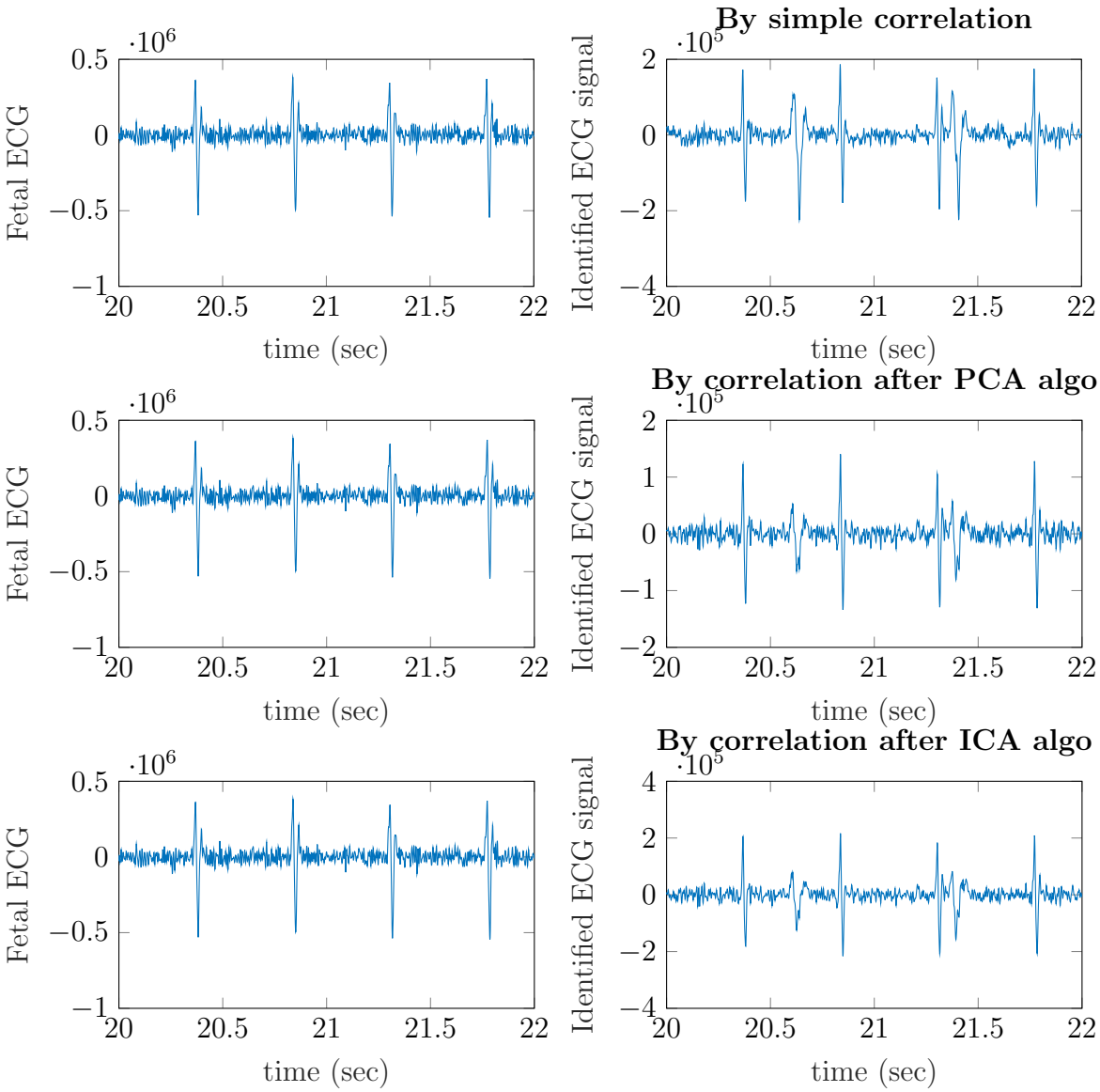


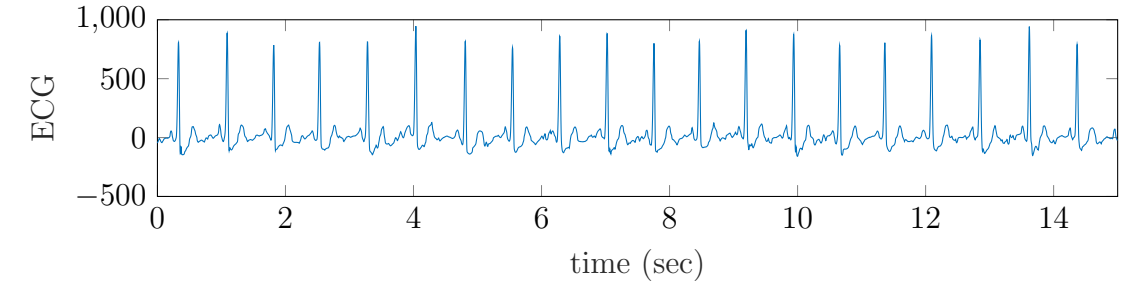
Figure 6: Identification of the fetus ECG in the ECG signal from the mother

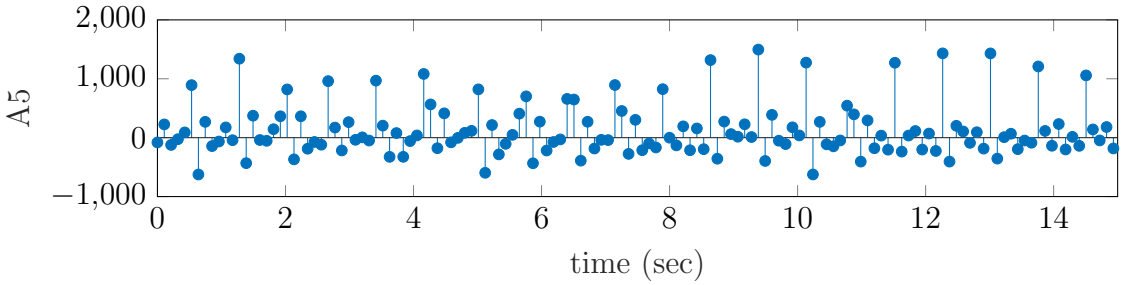
After filtering, it is hard to select a particular signal because three ECGs to four have a correlation coefficient almost equal. It means that all ECG signals are strongly correlated with the fetal ECG which is coherent with the fact that mother ECGs contains part of the fetus one.
After using PCA, signals are separated in linearly uncorrelated variables called principal components ranked by importance order. Only eigenvalues and scores related to the first component are used to compute the new signals. We don't need to correlate this time as all output signals are almost identical.
After using ICA, one of ECG signals has a strong similarity but it is hard to say if it is better as the precedent results just by look at it. It is why a RMSE measure is computed in all cases. All results are high because of the difference of amplitude between mother ECG and fetus ECG but we can see an improvement in the result by using ICA algorithm. However, this signal still has some artefacts from the mother ECG.
Both algorithms, PCA and ICA, try to get a new set of variables, more adapted to the situation, related by linear combination with the original.
The main difference between these 2 algorithms is when PCA attempts to find uncorrelated sources, ICA attempts to find independent sources.
Being independent is a stronger property than being uncorrelated. As PCA give us uncorrelated sources classified by order of importance, we often use this to reduce the dimension of a dataset. ICA doesn't give us a ranking but a set of "demixed" signals which need to be "recognize" or eventually matched by correlation with a signal seeing as ground true before analyze. Therefore, the result obtained is satisfying if signals are prewhitened and statistically independent.

3 Wavelets on ECG and EEG

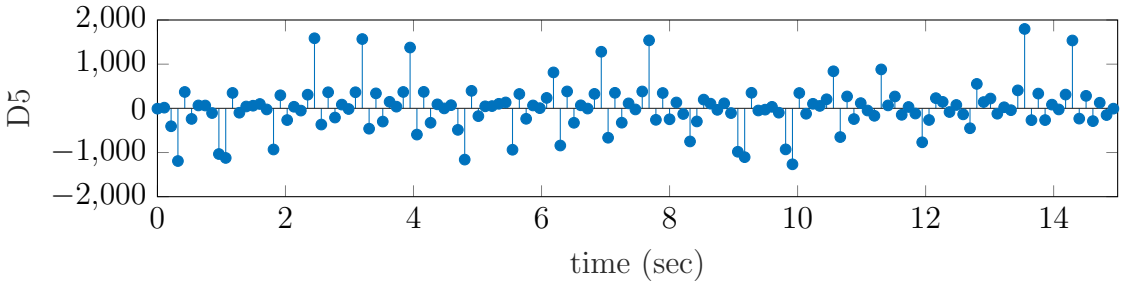
3.1 ECG decomposition—recomposition

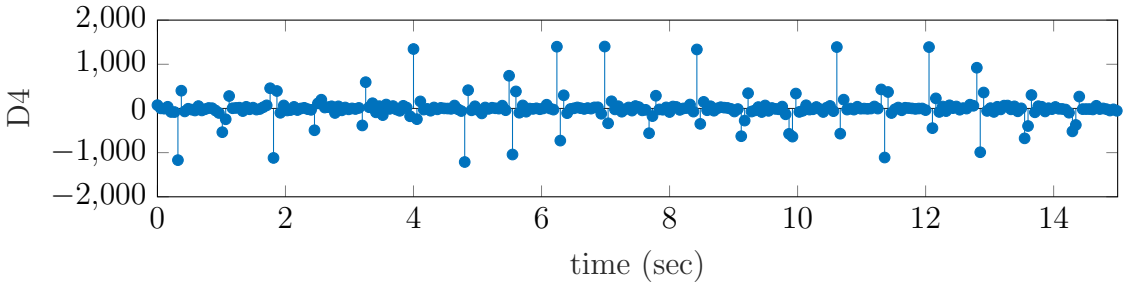
The wavelet analysis is used to decompose a signal, here an ECG, in another domain than usual methods mixing a good frequency and time resolution. For each level, a "detail coefficient" signal is computed with the particularity that the signal is down sampled by 2 each time. Each coefficient is the reconstruction of the signal in a specific frequency band all along the time. In this part, the ECG is decomposed in 5 levels by the Daubechies wavelet 4.





detail coeff





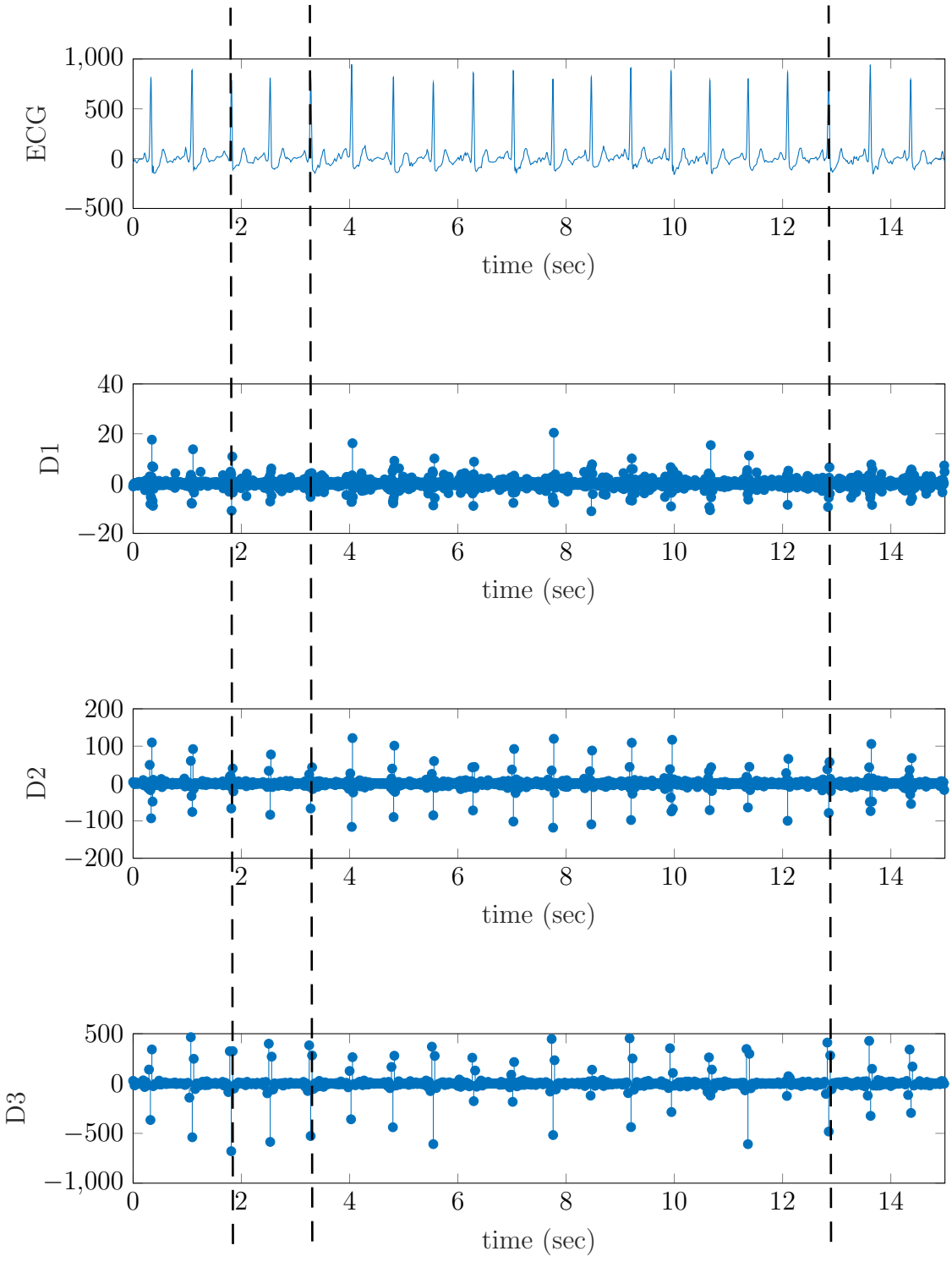


Figure 7: ICA Decomposition - 5 Levels

To analyze the QRS complex, the coefficient D3 seems to be the most appropriate when we look at it. The R-peak is in the range between 37.5 Hz and 75 Hz.
The Daubechies wavelet is the most often used to WT. It has 2 important proprieties : orthogonality and no redundancy.
The orthogonality guarantee the conservation of energy and make the reconstruction easiest. Therefore, in the next application, this conservation of energy will be useful to compute energy of each coefficient using as model features. And as there are no redundancies, a perfect reconstruction without aliasing is possible, useful propriety for the next step of this application. In effect, now, we want to keep only the N largest coefficient in each level of the decomposition to reconstruct the signal. N is tested with 4 different values : 10 ,25 ,50 ,100.

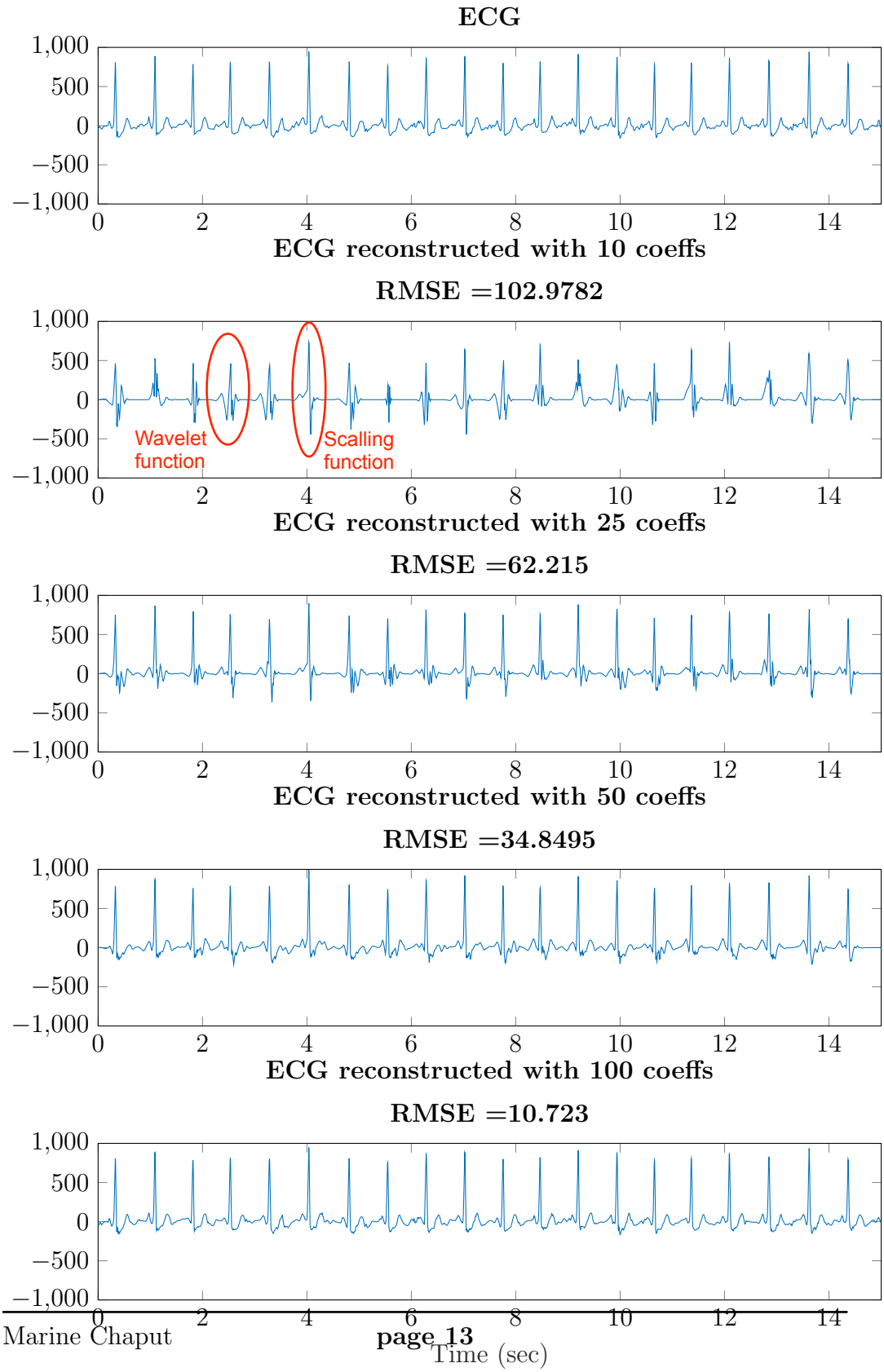
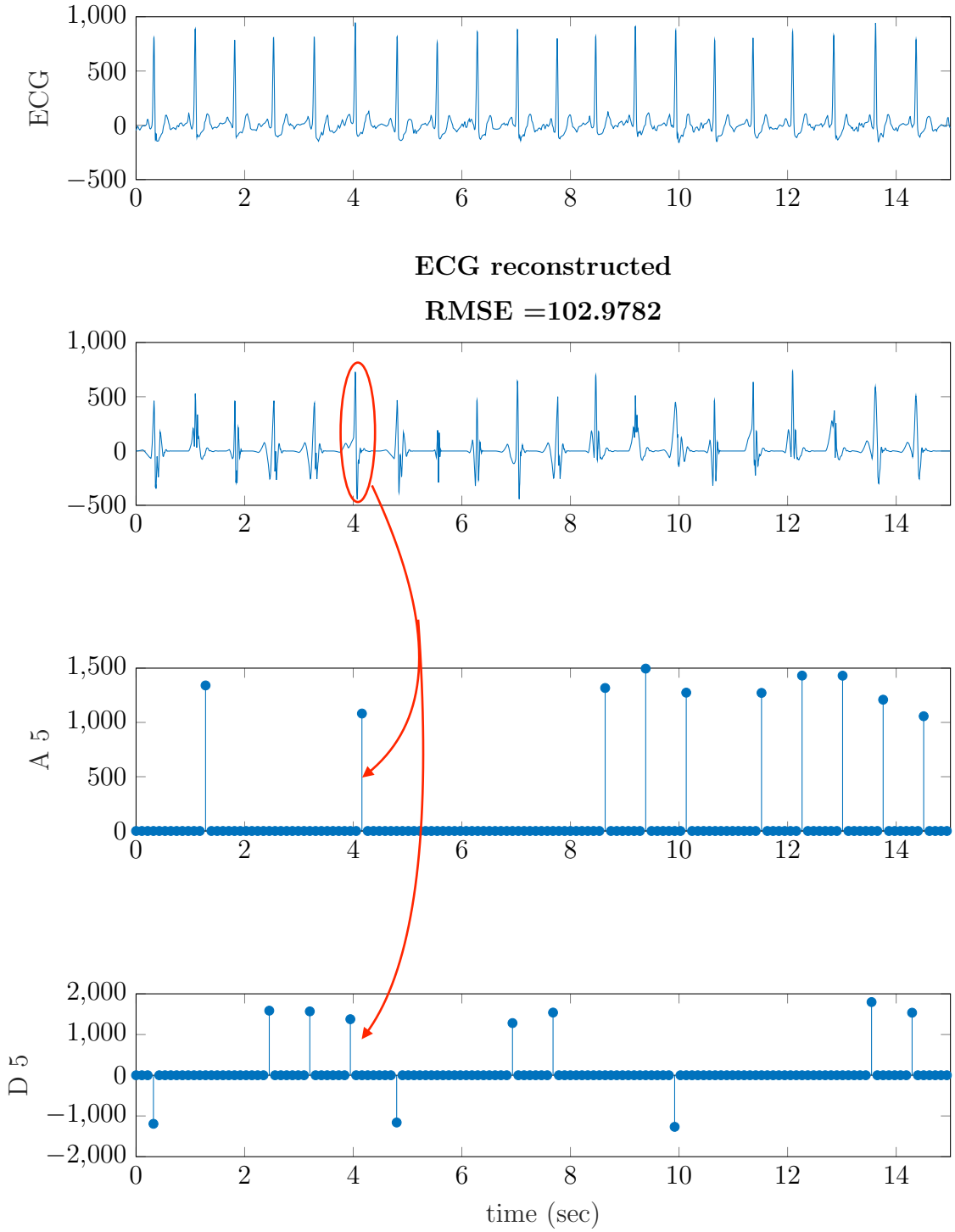


Figure 8: ECG Reconstructed

By adding coefficients, the RMSE is improved. It makes sense as the Daubechies wavelet allows a perfect reconstruction if we are using all computed coefficients. Part of theses reconstruction signals can be identified as being the wavelet and the scaling functions. It is particularly visible in reconstructions with a small amount of coefficients.


ECG
1,000
500
0
−500
0
2
4
6
8
10
12
14
ECG reconstructed
RMSE =102.9782
A 5
1,500
1,000
500
0
D 5
2,000
1,000
0
−1,000
−2,000
time (sec)

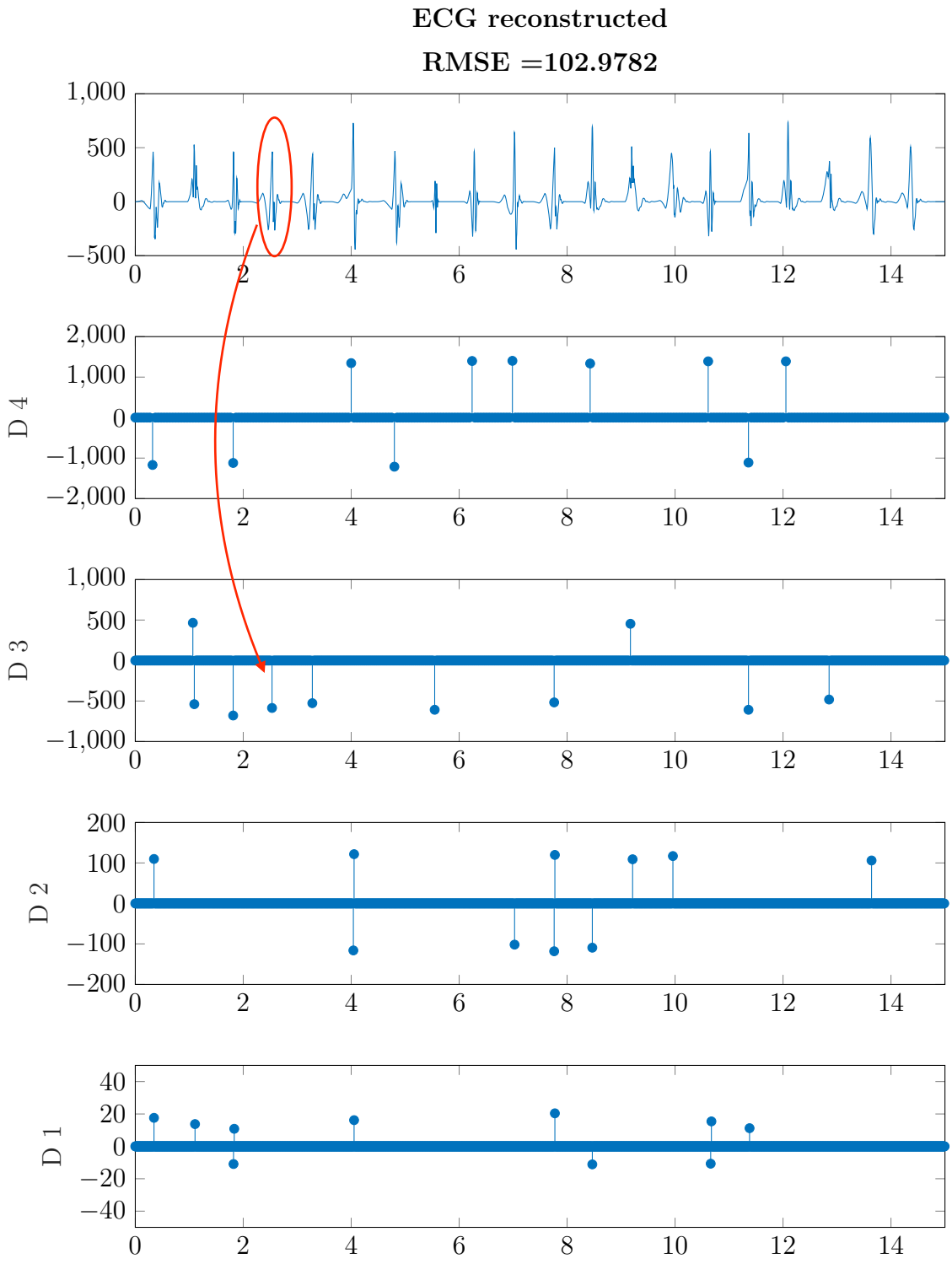


Figure 9: ICA recomposition— 10 coef

The signal reconstructed and 10 coefficients used to it, shows us how the signal is reconstructed:

- Coefficients in low-frequency range are using scaling function
- Coefficients in high-frequency range are using wavelet function

The combination of both functions give us the signal. Similar figures for other cases can be found in annex to keep a clear report.

3.2 EEG classification by subband frequency energy

In the previous assignment, we tried to classify EEG segments to find a detection model of seizure. The result wasn't a clear success. In this assignment, new features are tested to create a better model. First, as in the previous assignment, EEG is separated in overlapping segment of 2 sec and labeled with or without a seizure.

By looking at the figure, the amplitude of both classes is very different. Segments without seizures seems to be more condensed than segment with seizures. It is confirmed by computing the mean standard deviation, twice greater in the case of the seizure than without. By normalizing data, this discriminant property will disappear. Furthermore, data are from the same signal and are in the same range of value, meaning there is no need to normalize it.
The second step is to extract features from data.
Six features are interesting here :

- Energy in the full segment
- Energy in each of 5-level of decomposition
  - A5 : 0 — 8 Hz (δ and θ rhythms)
  - D5 : 8 — 16 Hz (approximately α rhythm)
  - D4 : 16 – 32 Hz (approximately β rhythm)
  - D3 : 32 – 64 Hz
  - D2 : 64 — 128 Hz
  - D1 : 128 — 256 Hz

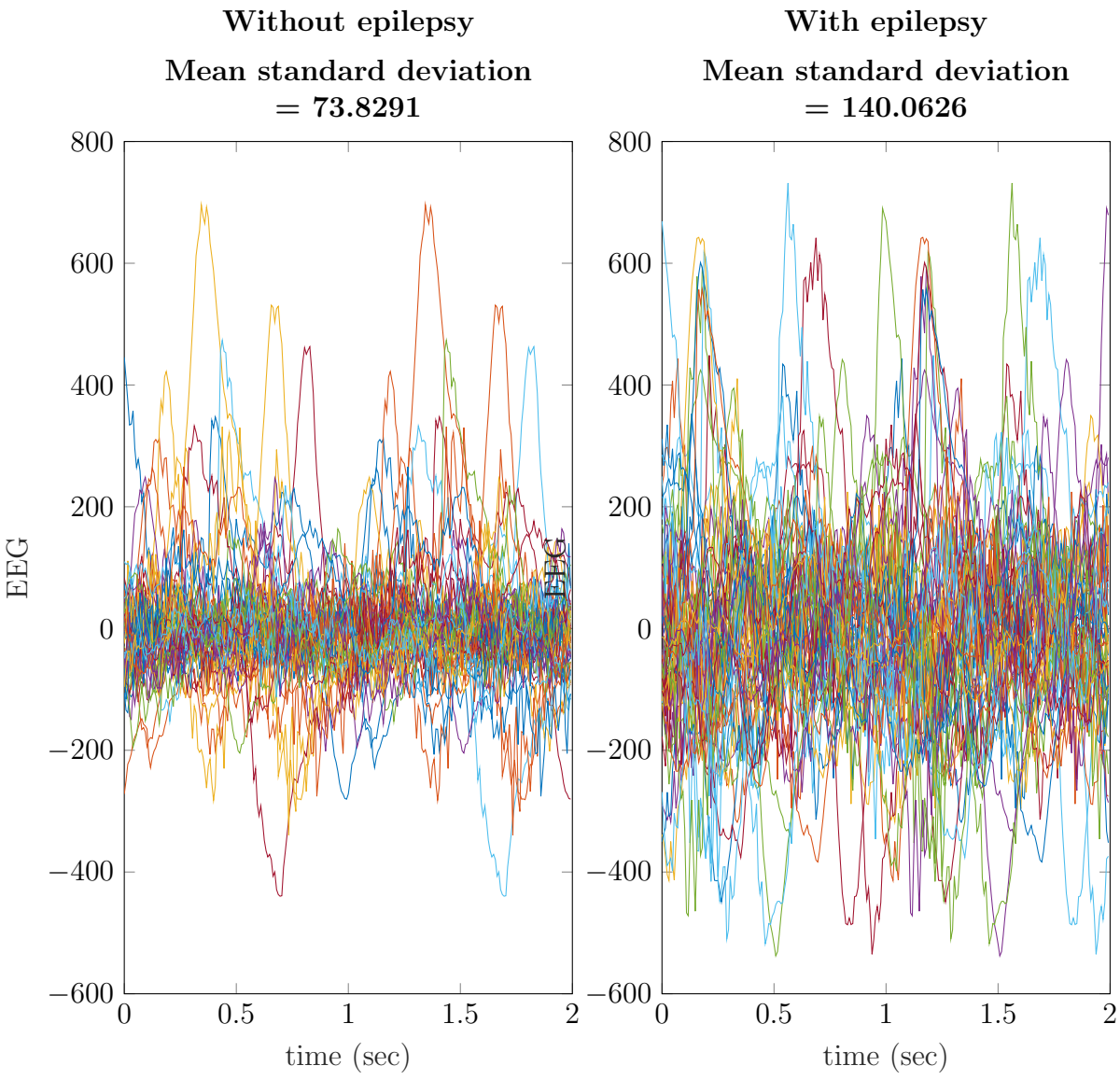


Figure 10: EEG segment represented by classes

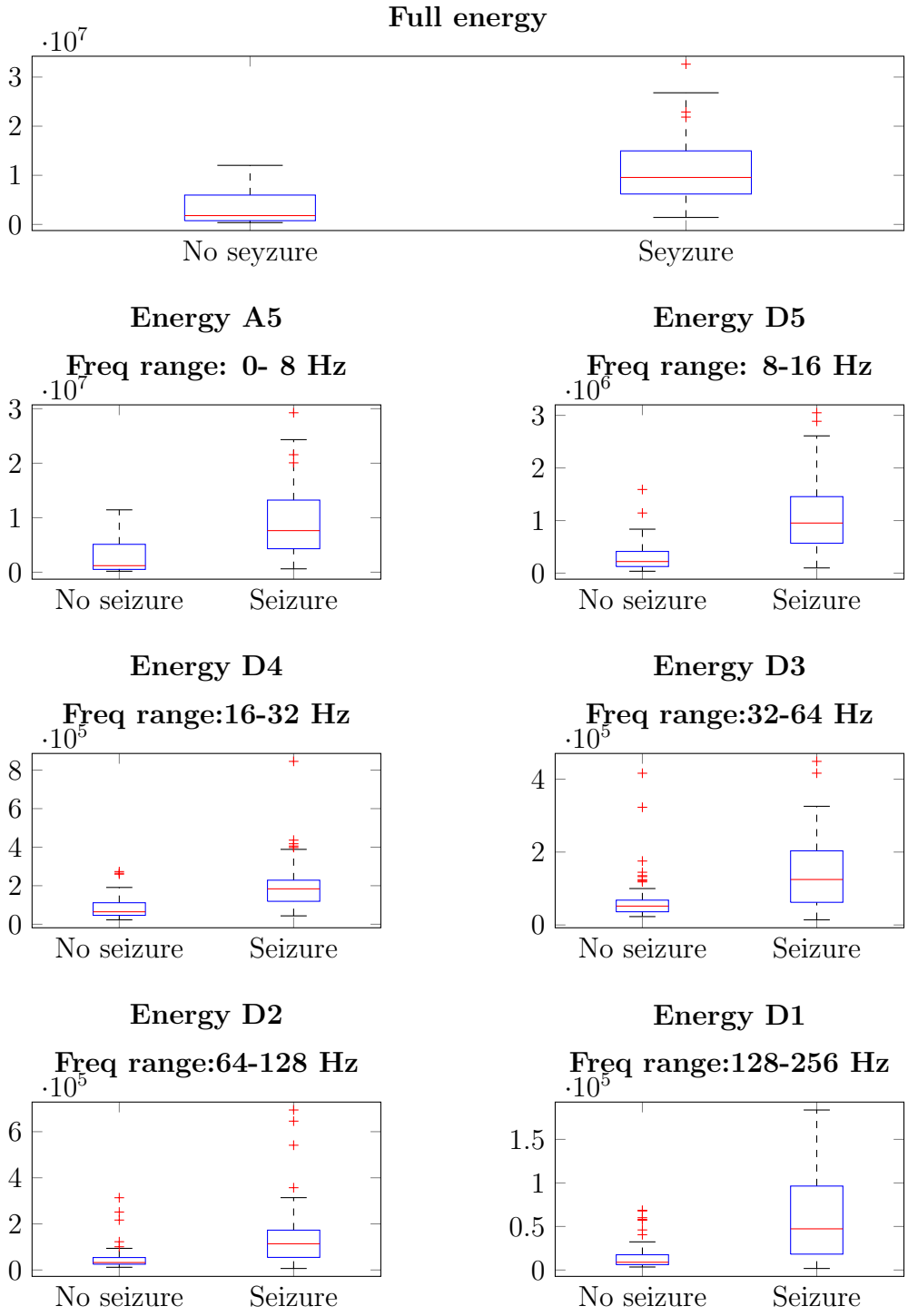


Figure 11: Feature characteristics for each class

High frequencies seem to be interesting because of theses discriminant characteristics between classes. Median values are different and segments without seizures have a little variation around the median value, allowing it to be classified more precisely. A model is computed with training set once with the full energy as a feature and once with all subband energy as features. Performances are evaluated on the test set.

	full_energy	Subband_energy
Accuracy	0.79577	0.92958
Sensitivity	0.64789	0.87324
Specificity	0.94366	0.98592

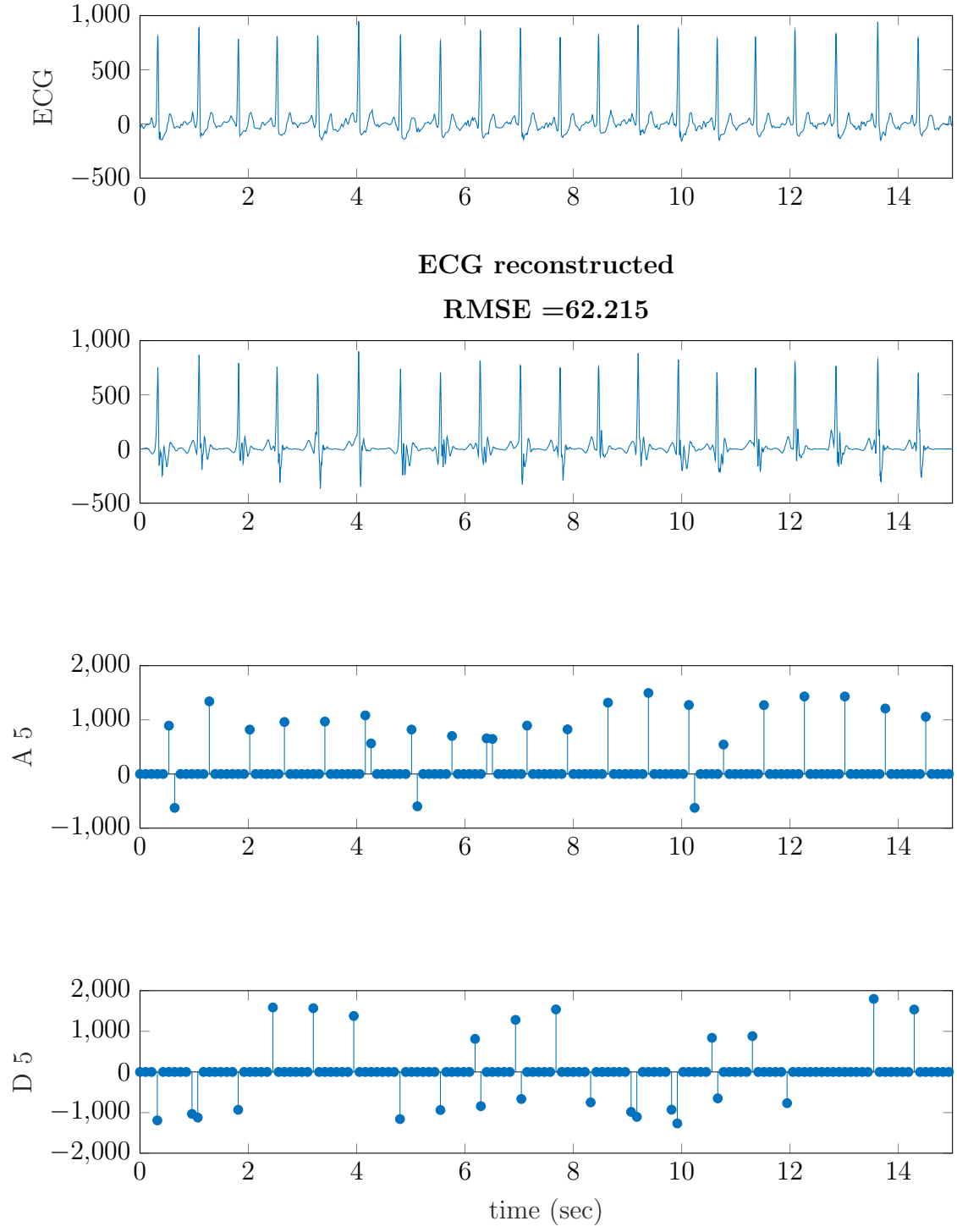
Figure 12: Performance for both models

Performances of the classification is better than in the previous assignment.
Moreover, in the full energy case, we have 94% of cases without epilepsy correctly diagnosticate but only 64% of the case with epilepsy found. It is an important missing. Hopefully, by using subband energy as features, this result increase to 87% with still 98% of success in case of normal EEG.
Subband energy features give us a correct model.
Using both full energy and subband energy to construct the model will conduct to better performance on this dataset but also overfit the model because of the information redundancy between features. Features should stay uncorrelated. At the end, it doesn't guarantee a better result in all situations.

4 Conclusion

In conclusion, in this report, performance of fastICA and PCA has been studied to separate mixing signals between ECG mother and fetus. The result was mitigated. In second part, a new model to detect epilepsy seizures with better results have been created by using energy signals as features. The wavelet analysis has been also tested by decomposing and recomposing an ECG.

5 Annexes


ECG
1,000
500
0
−500
0
2
4
6
8
10
12
14
ECG reconstructed
RMSE =62.215
A 5
2,000
1,000
0
−1,000
D 5
−2,000
time (sec)

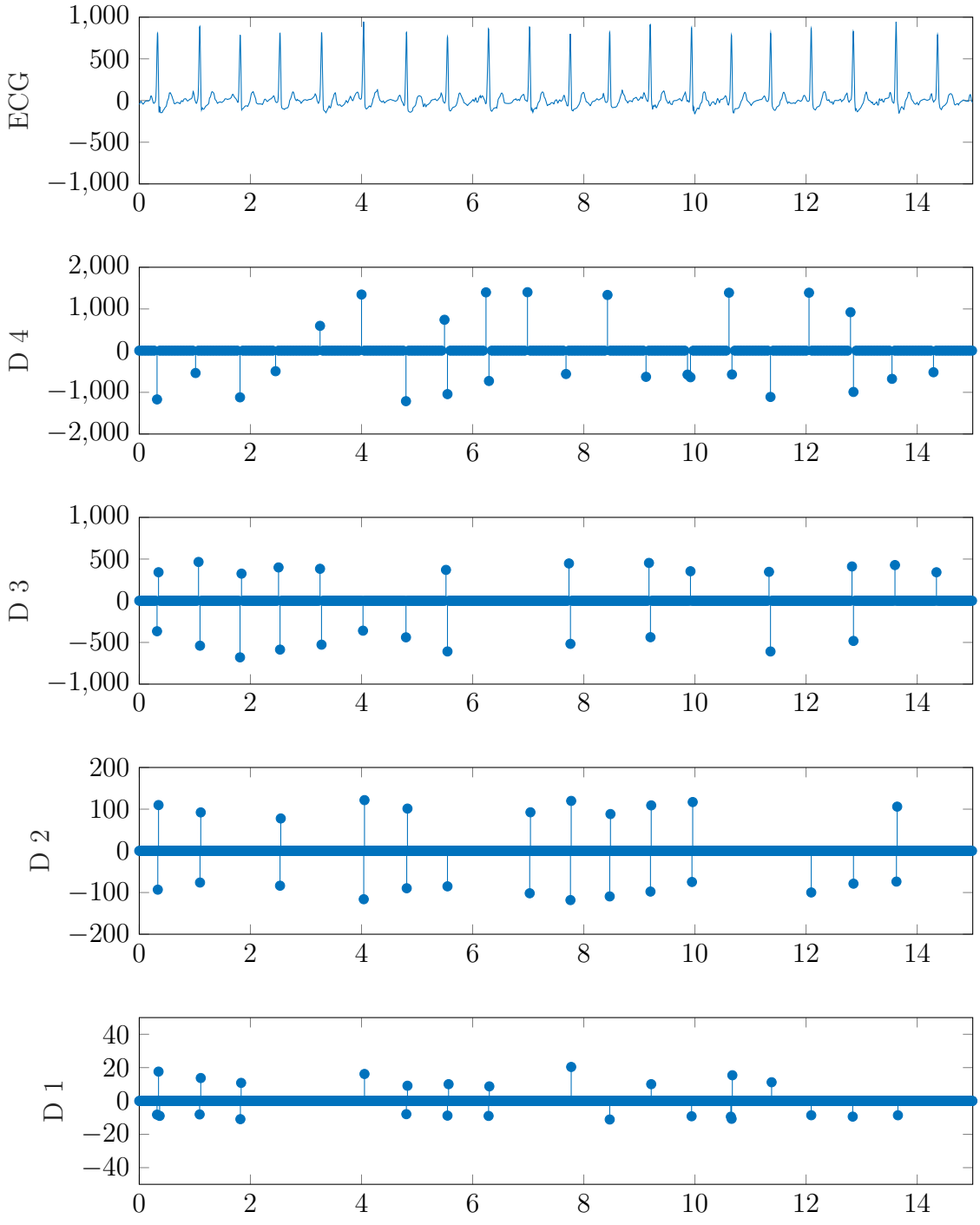
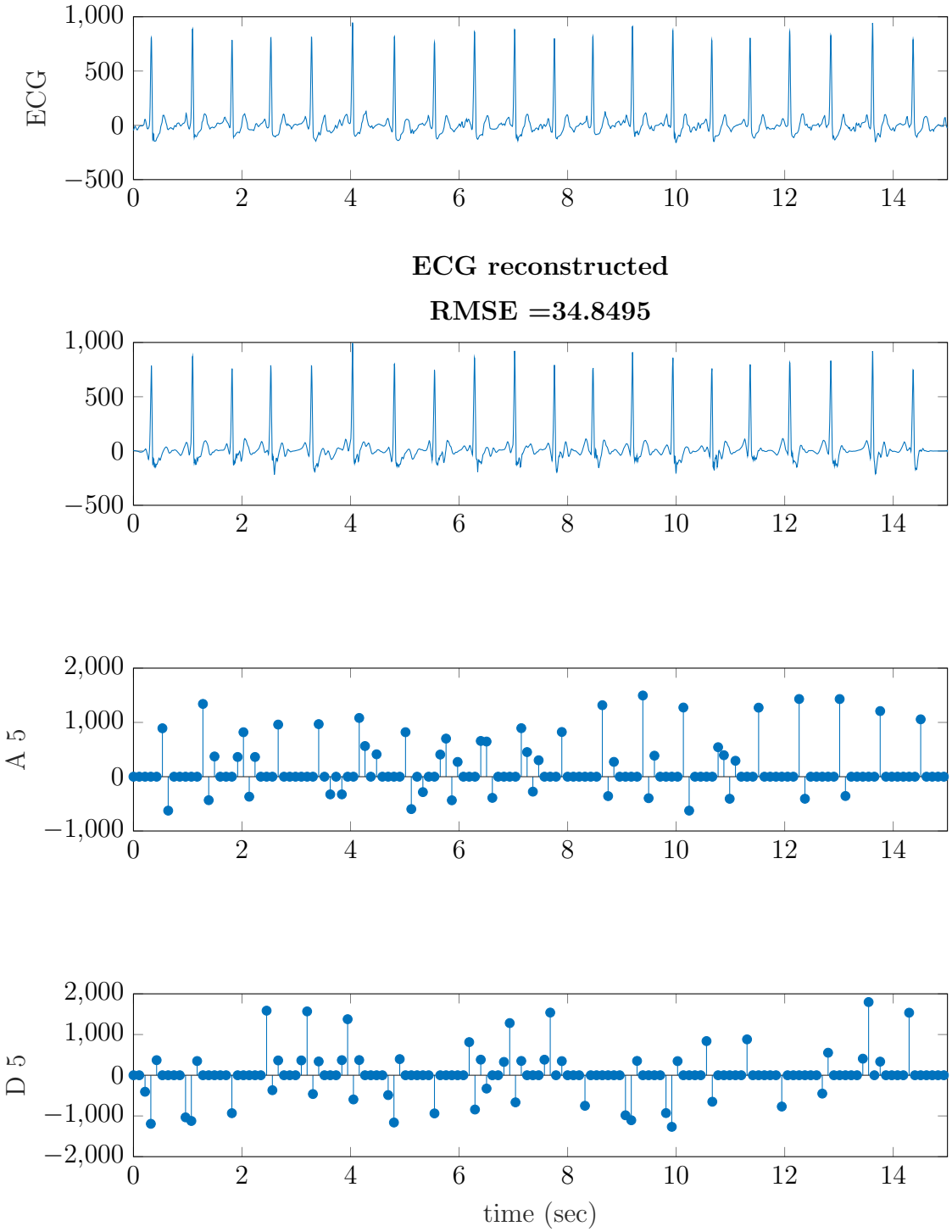


Figure 13: ICA recomposition - 25 coef


ECG
1,000
500
0
−500
0
2
4
6
8
10
12
14
ECG reconstructed
RMSE =34.8495
1,000
500
0
−500
0
2
4
6
8
10
12
14
A 5
2,000
1,000
0
−1,000
0
2
4
6
8
10
12
14
D 5
2,000
1,000
0
−1,000
−2,000
0
2
4
6
8
10
12
14
time (sec)

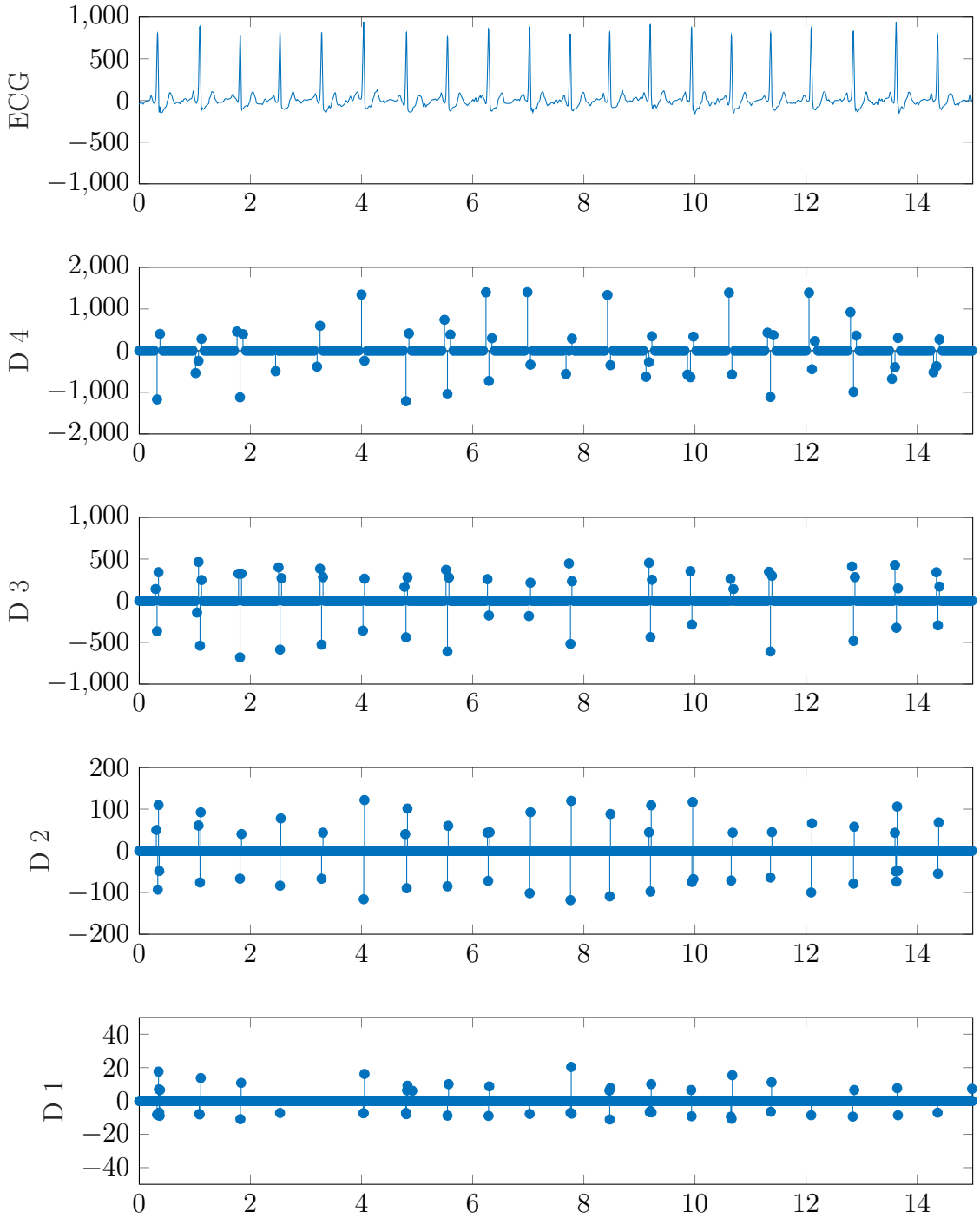
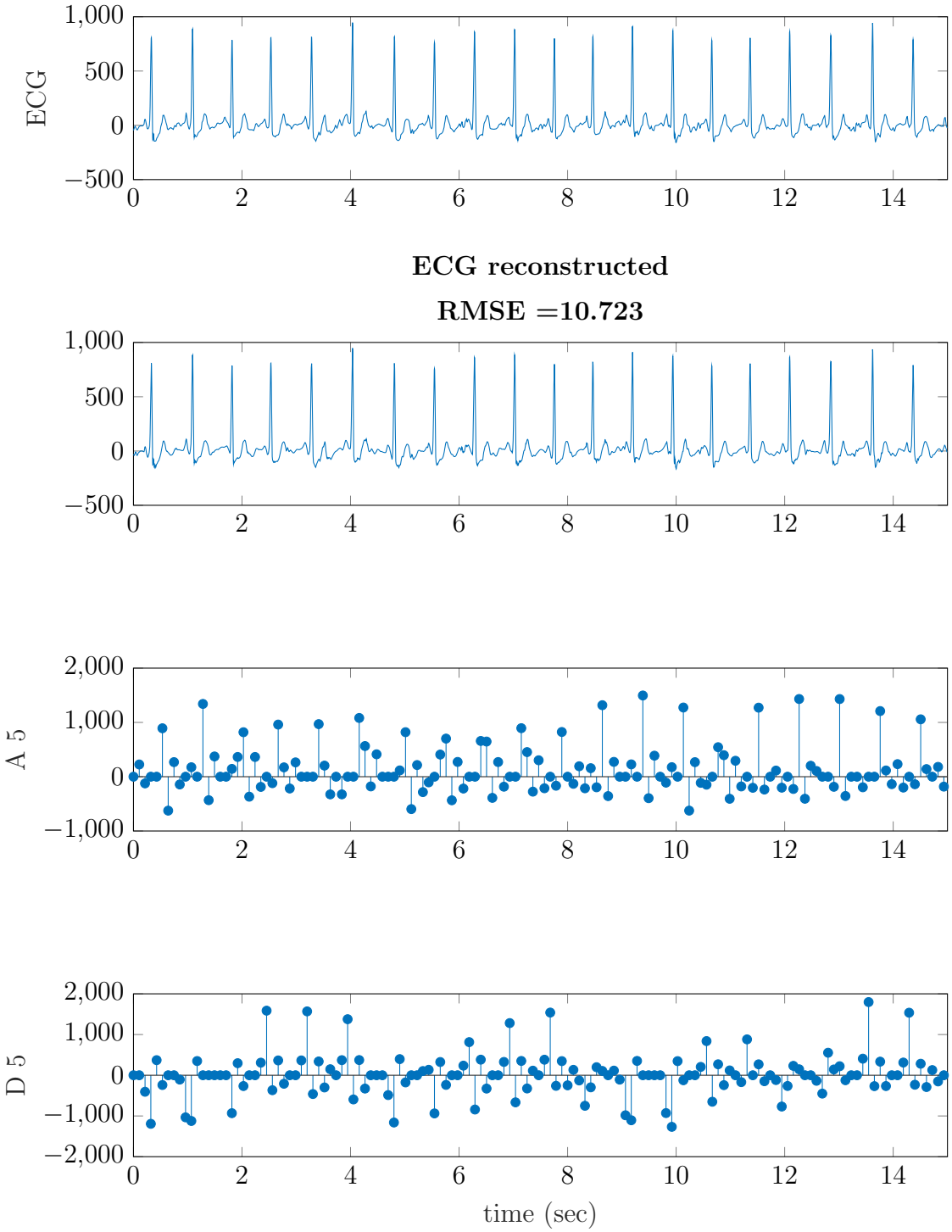


Figure 14: ICA recomposition - 50 coef


ECG
1,000
500
0
−500
0
2
4
6
8
10
12
14
ECG reconstructed
RMSE =10.723
1,000
500
0
−500
0
2
4
6
8
10
12
14
A 5
2,000
1,000
0
−1,000
0
2
4
6
8
10
12
14
D 5
2,000
1,000
0
−1,000
−2,000
0
2
4
6
8
10
12
14
time (sec)

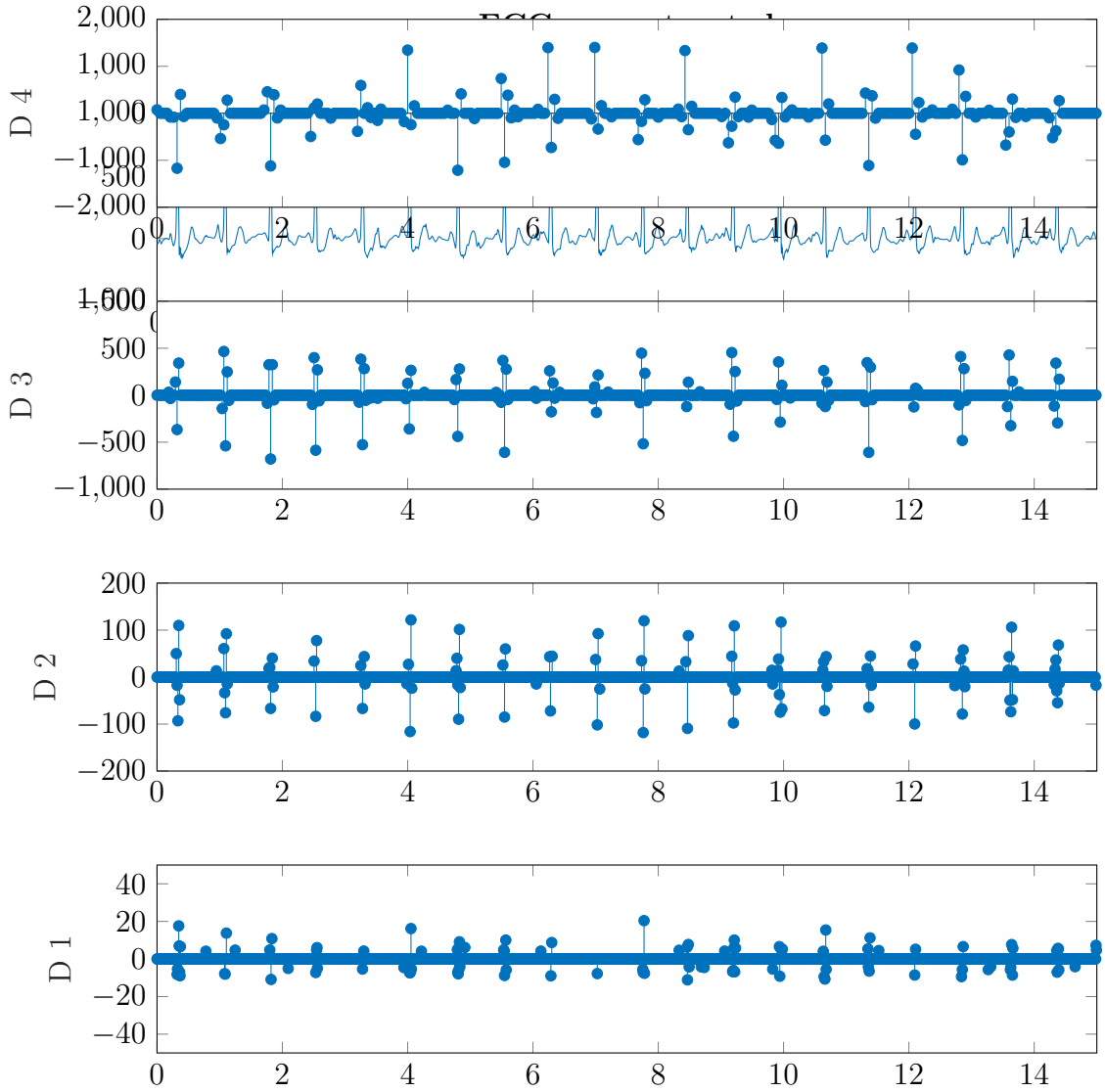


Figure 15: ICA recomposition - 100 coef