MCMC Modeling For the Characterization of ECG Signals

1 Introduction

Electrocardiography is a very common, non-invasive medical procedure used to diagnose and understand heart conditions related to chest pain, palpitations, dizziness, or other cardiac symptoms. An electrocardiogram (also referred to as an ECG or EKG) shows a graph in real time of a patient' s heart voltage vs. time. This graph conveys valuable information to clinicians, who may use the result of an ECG to determine electrical or structural abnormalities of the heart.

Interpreting ECG data is an incredibly important task in the medical world. An ECG signal is relatively complicated and conveys information about all aspects of heart function, so it is important that ECG signals are only evaluated by a trained technician. For the same reasons, there is interest in generating algorithms or models to characterize an ECG signal to potentially aid in the diagnosis of heart conditions and classify patients as healthy or unhealthy on the basis of their ECG data.

Each segment of an ECG signal consists of the following: the P wave, the Q-R-S complex, and the T wave (See Figure 1). The amplitudes, interval length, and shape of these individual components are clinically important, as these characteristics can help a clinician find abnormalities in the heartbeat of a patient, allowing for crucial diagnoses. Previous studies within statistics and machine learning that seek to classify ECG signals in healthy and unhealthy patients have applied models such as wavelet transforms, neural networks, and partially collapsed Gibbs samplers [9, 6].

Parametric modeling is effective when we want to compactly represent wave features as model parameters [3]. Unlike certain "black box" machine learning techniques, it offers a higher level of interpretability, which is especially useful for the clinical setting. Rather than simply outputting a classification result, a parametric model allows us to investigate in more detail which aspects of an ECG signal differ between healthy and unhealthy patients.

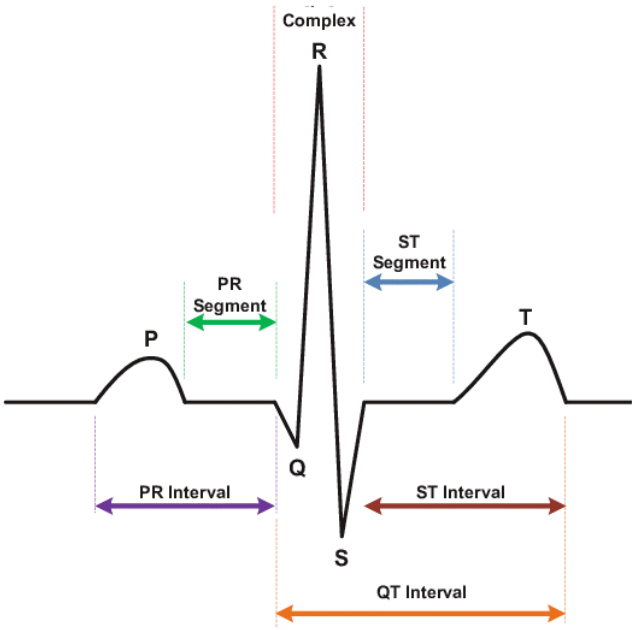

Figure 1: Representation of an ECG signal with major wave components. [4]

In our project, we borrow a parametric model proposed by Bodisco et.al [8] and apply it to interpret the difference in model parameters between a healthy adult (with normal sinus rhythm) and an adult with arrhythmia. To do so, we accomplished the following:

1. Code from scratch the Metropolis-Hasting algorithm proposed by Bosdico (Elda) et.al. [8] (Refer to full code in Appendix I).
2. Determine appropriate proposal distributions which the authors did not specify.
3. Replicate the Pan-Tompkins algorithm (a well-established QRS detector) which allows us to segment a 10 second ECG signal data each from a healthy and unhealthy adult. [7] (Refer to full code in Appendix II).
4. Run the MH for each interval of these segments and calculate the average model parameter values.

2 Data

We decided to use data from the MIT PhysioNet website. The database consists of ECG time series data taken from 47 patients at Boston' s Beth Israel Hospital in the 1970s [1]. Each series consists of 1800 evenly-spaced voltage measurements from a single patient, taken in intervals of 0.003 ms and measured in mV. From PhysioNet, we selected one sample ECG of a patient from the MIT-BIH Arrhythmia data and one from the MIT-BIH Normal Sinus Rhythm Database for comparison.

Since our data comes in the form of 10-second long samples, each of which includes multiple heartbeats, it is necessary that we split each sample into a number of sub-samples representing a single

patient heartbeat. This is because the model is set up to characterise the ECG signal of an individual heartbeat, one interval centered at the QRS complex and bookended by the peaks of the adjacent heartbeats' QRS complexes. Because the model requires splitting at the peaks of the QRS complex, we decided to employ the Pan-Tomkins algorithm to segment the data. This algorithm is a popular technique employed to identify the QRS complex peaks in an ECG signal represented as time series data. Rather than coding the algorithm from scratch, we installed a Python implementation of the Pan-Tomkins algorithm and ran it on our chosen time-series data to split each series into five smaller series representing individual heartbeats. This Python implementation was written by Micha Sznajder and Marta ukowska of Jagiellonian University in Poland [7].

3 Model and Parameters

The data is modeled as a normal distribution. The model generates samples of the ECG signal characterization at time t as $s(t) = y$ which has the distribution $N(\mu(t)+DC, \tau)$. Table 1 is a replication of Bodisco, et al.'s model formulation for the mean $\mu(t)$.

mean $\mu(t)$	time interval
$\alpha_1 cos(\frac{\pi}{\delta_1}t) + \beta$	$[0, \delta_1)$
$\alpha_2[1 - cos(\frac{\pi}{\delta_2 - \delta_1}t(t - \delta_1)] + \beta - \alpha_1$	$[\delta_1, \delta_2)$
$\alpha_3[1 - cos(\frac{\pi}{\delta_3 - \delta_2}(t - \delta_2)] + \beta - \alpha_1 + 2\alpha_2$	$[\delta_2, \delta_3)$
$\frac{\beta+\alpha_1+2\alpha_2+2\alpha_3}{2}[1 + cos(\frac{\pi}{\delta_4 - \delta_3}(t - \delta_3)]$	$[\delta_3, \delta_4)$
0	$[\delta_4, \delta_5)$
$\alpha_4[1 - cos(\frac{\pi}{\delta_6 - \delta_5}(t - \delta_5)]$	$[\delta_5, \delta_6)$
$\alpha_4[1 - cos(\frac{\pi}{\delta_7 - \delta_6}(t - \delta_6)]$	$[\delta_6, \delta_7)$
$\alpha_5[cos(\frac{\pi}{\delta_8 - \delta_7}(t - \delta_7) - 1]$	$[\delta_7, \delta_8)$
$\alpha_6[1 - cos(\frac{\pi}{\delta_9 - \delta_8}(t - \delta_8)] - 2\alpha_5$	$[\delta_8, \delta_9)$
$\alpha_7[cos(\frac{\pi}{\delta_{10} - \delta_9}(t - \delta_9) - 1] + 2\alpha_6 - 2\alpha_5$	$[\delta_9, \delta_{10})$
$\alpha_8[1 - cos(\frac{\pi}{\delta_{11} - \delta_{10}}(t - \delta_{10})] - 2\alpha_7 + 2\alpha_6 - 2\alpha_5$	$[\delta_{10}, \delta_{11})$
$\alpha_9[1 - cos(\frac{\pi}{\delta_{12} - \delta_{11}}(t - \delta_{11})] - +2\alpha_8 - \alpha_7 + 2\alpha_6 - 2\alpha_5$	$[\delta_{11}, \delta_{12})$
$(\alpha_9 + \alpha_8 - \alpha_7 + \alpha_6 - \alpha_5)[1 + cos(\frac{\pi}{\delta_{13} - \delta_{12}}(t - \delta_{12})]$	$[\delta_{12}, \delta_{13})$
0	$[\delta_{13}, \delta_{14})$
$\alpha_{10}[1 - cos(\frac{\pi}{\delta_{15} - \delta_{14}}(t - \delta_{14})]$	$[\delta_{14}, \delta_{15})$
$\alpha_{10}[1 + cos(\frac{\pi}{\delta_{16} - \delta_{15}}(t - \delta_{15})]$	$[\delta_{15}, \delta_{16})$
$\alpha_{11}[cos(\frac{\pi}{\delta_{17} - \delta_{16}}(t - \delta_{16}) - 1]$	$[\delta_{16}, \delta_{17})$
$\alpha_{12}[1 - cos(\frac{\pi}{T - \delta_{17}}(t - \delta_{17})]$	$[\delta_{17}, T)$

Table 1: Model specification as per Bodisco, et al.

The model parameters describes the following: β is an offset to compensate for starting at the top of a qrs complex; $\alpha_1, \ldots, \alpha_{12}$ are related to the amplitudes of the cosine waves; $\delta_1, \ldots, \delta_{17}$ are the turning points in an ECG heart beat; τ represents the variability of the signal output at time t; 't' is time, and T is the time related to the top of the QRS complex in the beat following that of interest [8]. The function, represented as s in the paper, is the ECG signal and y is the model representation of the ECG signal.

As for the choices of our Prior and Proposal Distribution, while the prior distributions are specified by the authors, little explanation is given for their choices. We provide justifications for them below. The prior distribution of τ is given as $Gamma(0.01, 0.01)$. This is reasonable because we have very little information about the variability of the signal output at each time point, which is highly dependent on the patient' s condition and the noise in the electrical system. A flat prior which gives more weight towards the sample variance is thus appropriate.

The prior distribution of δ_i is $Unif(\delta_{i-1}, \delta_{i+1})$. This ensures that the time intervals do not overlap and allow for the shape of the P, Q, R, S, T waves to be unambiguously modeled. The important implication of this prior is that the acceptance ratio for delta will only include the likelihood given all deltas since the conditional distribution of δ_i given the others and the joint prior over all the other deltas (last two terms of the equation below) are the same in the numerator and denominator.

$$\begin{aligned} p(\delta_1, \ldots, \delta_{17}) &\propto p(y|\delta_1, \ldots, \delta_{17}) p(\delta_1, \ldots, \delta_{17}) \\ &= p(y|\delta_1, \ldots, \delta_{17}) p(\delta_i|\delta_1, \ldots, \delta_{i-1}, \delta_{i+1}, \ldots, \delta_{17}) p(\delta_1, \ldots, \delta_{i-1}, \delta_{i+1}, \ldots, \delta_{17}) \end{aligned}$$

The prior for all other model parameters are Gaussian. While the authors opted for a uniform prior between zero and infinity because they are "computationally faster," we opted for normal prior because this allows model parameters to take either positive or negative values and offers more flexibility for modeling signals which themselves can have both positive and negative voltage measurements. Since we are programming for accuracy (and interpretation) rather than efficiency, this additional computational cost is reasonable. Gaussian priors are also widely used in existing literature [6].

Because the proposal distribution is normal and thus symmetric, the proposal ratio is assumed to be unity. Again, this choice was not stated explicitly by the authors. The choice for a unitary proposal distribution is justified because 1) it is computationally more efficient (without the proposal ratio) and 2) there is no need in our case for an asymmetric proposal where we weight towards certain values since we have no a priori information about or constraint for the range of model parameters. Note, the proposal variance (stepsize) was not specified by the authors and we assigned unique values to it depending on the type of ECG signal (healthy or unhealthy) to ensure that the random-walk explores the full parameter space in each case. The mean of the sampling model, $\mu(t) + DC$, is calculated based on the various random model parameters that compose it (deltas, alphas, beta). Unlike the other parameters, we sampled τ from its full conditional distribution. The derivation of its full conditional distribution is shown below.

Assume we are following $s(t) = y$ distributed as $N(\mu(t) + DC, \tau)$. Then if our prior on τ is $Gamma(0.01, 0.01)$ then our prior on $\frac{1}{\tau}$ is $InvGamma(0.01, 0.01)$. This can be parameterized as $InvGamma(\frac{\nu}{2}, \frac{\nu\gamma^2}{2})$ where $\nu = 0.02$ and $\gamma^2 = 1$. This allows us to determine that the full conditional $[\frac{1}{\tau}|\mu(t) + DC, \mathbf{y}]$ is $InvGamma(\frac{\nu^*}{2}, \frac{\nu^*\gamma^{*2}}{2})$, where $\nu^* = \nu + n$ and $\gamma^{*2} = \frac{1}{\nu^*}(\nu\gamma^2 + \Sigma_{i=1}^n(y_i - (\mu(t) + DC)^2)) = \frac{1}{nu^*}(0.02 + \Sigma_{i=1}^n(y_i - (\mu(t) + DC)^2))$. From this distribution we can sample from the full conditional of τ which is instead $Gamma(\frac{\nu^*}{2}, \frac{\nu^*\gamma^{*2}}{2})$.

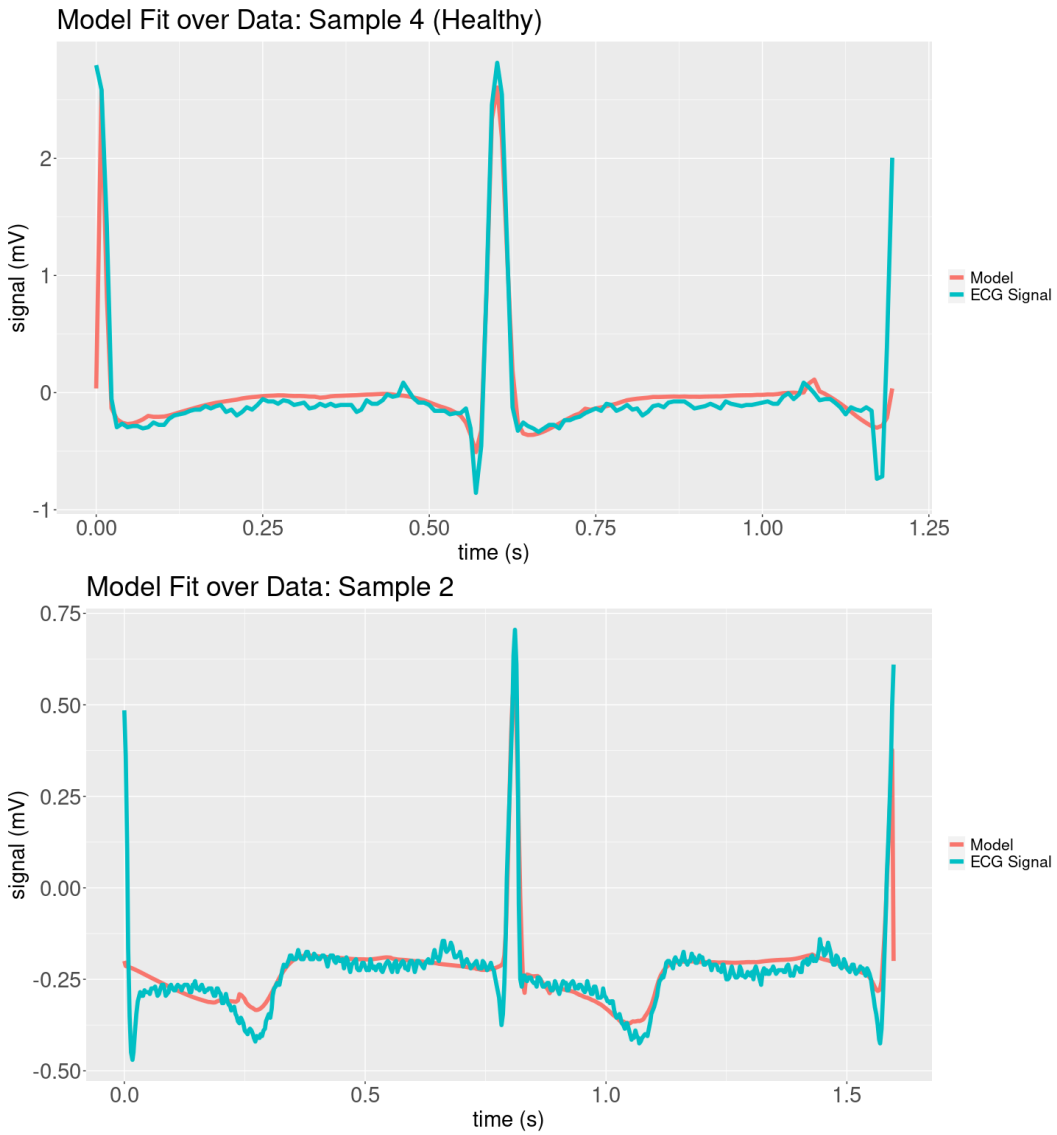

Figure 2: Model posterior fit to healthy patient heartbeat (top) and arrhythmia heartbeat (bottom).

We ran the MCMC algorithm for 3000 samples. Model fit can be assessed via plots of the ECG data with the predictive mean values superimposed to show how well the data characterizes the signal. These plots are shown in Figure 2. Note here that the mean of the predictive distribution of $\mu(t) + DC$ is the same as the mean of the posterior distribution of $\mu(t) + DC$.

$$\begin{aligned}
E_{p(y^*|y_1,...,y_n)}(y^*) &= \int y^* p(y^*|y_1, ..., y_n) dy^* \\
&= \int y^* \int p(y^*|\mu) p(\mu|y_1, ..., y_n) d\mu dy^* \\
&= \int \left[\int y^* p(y^*|\mu) dy^* \right] p(\mu|y_1, ..., y_n) d\mu \\
&= \int \mu p(\mu|y_1, ..., y_n) d\mu \\
&= E_{p(\mu|y_1,...,y_n)}(\mu)
\end{aligned}$$

The MCMC diagnostics (in the Appendix III) suggest that most model parameters only reach some hint of stationarity after the 2500th sample, which is not surprising since highly parametrized models like this tend to have high autocorrelation in the chain. Mixing was also far from ideal. Firstly, the effective sample size for all parameters are less than 100 and those for parameters of the arrhythmia patient do not even exceed 30, because samples are highly correlated. Moreover, most parameters have an acceptance ratio between 0.2 and 0.5 [5] which is satisfactory but not ideal at or close to 0.234 [2].

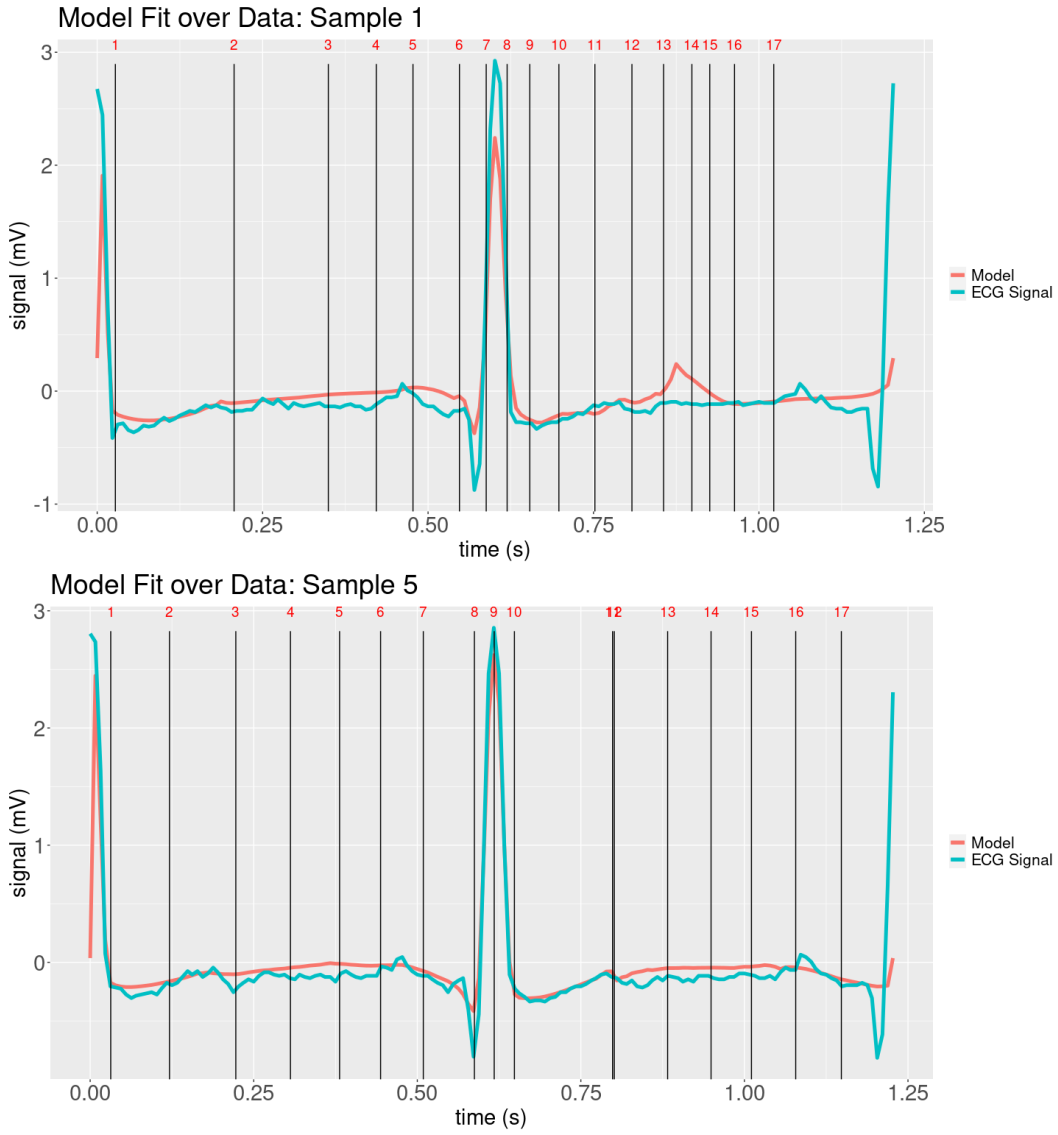

Figure 3: Examples of two heartbeats from the same healthy patient, overlayed with posterior mean values of each δ_i parameter. Shows how these cannot reliably identify portions of the ECG signal.

4 Results

To analyze the results of our posterior distributions, we hope to show that:

1. The model accurately characterises the ECG signal from the MIT-BIH samples as expected, both for healthy patients and those with arrhythmia.

2. Meaningful conclusions can be drawn from this modeling strategy about differences between healthy and unhealthy patients based on our posterior samples of model parameters.

For the first task, we were able to show that our model is able to characterise the ECG signals of healthy patients fairly accurately, although this can only be assessed with somewhat of a judgement call. Plotting the mean of our predictive distribution on $\mu(t)+DC$ over the true, somewhat noisy signal of the healthy ECG patient ' s heartbeat shows a reasonably good, smooth fit to the characteristic waves of the ECG signal (Figure 2). The model had more difficulty characterising the signal of patients with arrhythmias, as can be seen in Figure 1, as it either inserts unnecessary bends or fails to characterise parts of the signal.

The second goal was to determine if, and how, a comparison of the posterior distributions of parameters between the healthy and unhealthy patients can be used to distinguish between different types of heartbeats. We discovered that this question is not readily answered by the model for a few critical reasons.

Firstly, the model is segmented into seventeen partitions based off of the turning points of a healthy ECG signal, so we assumed that the model would adequately and consistently characterise those seventeen partitions for each heartbeat. From our posterior results, it would seem that this is not the case. The model parameters δ_1 through δ_{17} are intended to segment the ECG signal, as described in the model and parameters section. Our hope was that by viewing our posterior fit, we can determine which delta ranges correspond to different parts of an ECG signal, such as the P-wave, T-wave, and QRS complex. However, after analyzing our posterior fit on the deltas for various heartbeats, we saw that the same delta values were not commonly fitting the same portions of the ECG signal, even across heartbeats from the same patient. Figure 3 shows two ECG signals from heartbeats of a healthy patient, as well as vertical lines showing the mean posterior value of each delta parameter. It can be clearly seen from the figure that different delta values are used to characterise the same portions of the wave across the two heartbeats. For example, in the left chart, the QRS complex is contained between δ_8 and δ_9, whereas in the right chart, it is contained between δ_6 and δ_7. For adequate analysis of parameter distributions across patients, we would need to see consistency in the way that these delta values characterise portions of the ECG signal, which we unfortunately do not see. This severely limits the practical utility of this model as a parametric alternative to black box methods for ECG diagnostics.

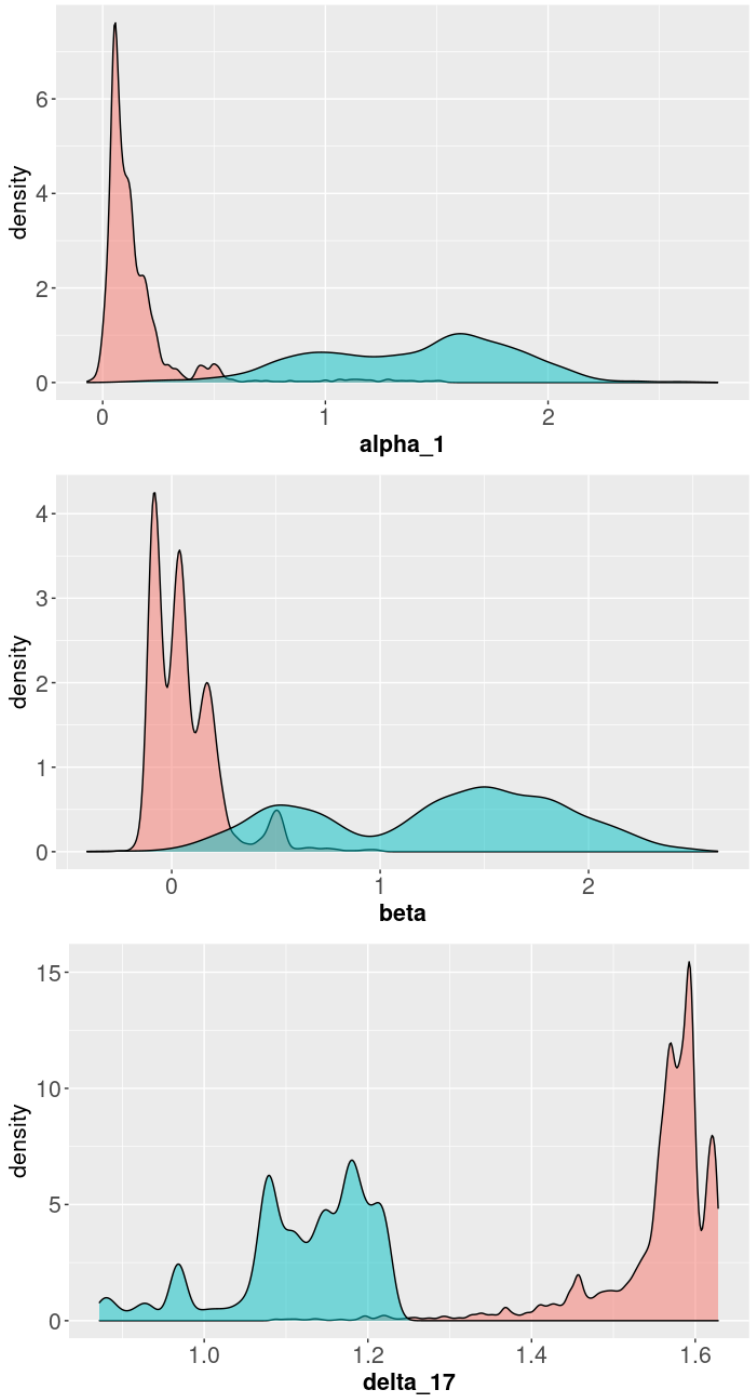

Figure 4: Posterior distributions of selected parameters for a healthy patient (blue) and a patient with an arrhythmia (red). These showed some of the more clearly-defined parameter differences distinguishing the two types of patients.

At the same time, we tested to see if the model parameters' posterior samples varied significantly between the healthy and unhealthy patients. Using a Welch' s two-sample t-test for comparison, given roughly normal distributions for each parameter, a statistically significant difference was found between each parameter between the healthy and unhealthy patients at the 0.001 level of significance. This does give some hope that the parameters can be used to differentiate between patients, but it is nearly impossible to make a definitive conclusion that values of one parameter tend to represent a specific heart function, given the previous statements about the inconsistency of the delta values. For instance, even though α_5 may differ significantly between healthy and unhealthy patients, and this fact may be used to help make diagnoses, it is not an interpretable result, since α_5 encodes information about the QRS complex in some of our samples and information about the T-wave in others, for instance.

That being said, plotting the posterior distributions of model parameters across healthy and unhealthy patients does show some clear differences, some of which are shown in Figure 3.

5 Conclusion

Our analysis shows that while the method presented by Bodisco, et al. does an effective job at tracing ECG signal with noise into a smoother plot, its parameters in practice are not interpretable in the way intended by the authors. The posterior distributions of parameters in the model can be effective in showing differences between patients, but offer little in the way of interpretable medical conclusions. All this being said, given more time to run, it is possible that there could be better MCMC convergence and mixing performance with more samples and thinning. Alternatively, we suggest exploring more advanced statistical techniques like an adaptive metropolis hasting algorithm or simply parallel computing for improved efficiency. Whether this changes our posterior analysis is unclear, and while this work hints at difficulties with the original model, with the previous suggestions, future work can be done to reinforce or reject the conclusions of this study on a larger set of data.

References

[1] Mit-bih arrhythmia database v1.0.0.

[2] A. GELMAN G. O. ROBERTS and W. R. GILKS. Weak convergence and optimal scaling of random walk metropolis algorithms. *Annals of Applied Probability*, 7(1(1997)):110–120, 2002.

[3] Johnson M.A. Grimble M.J. *Principles of Adaptive Filters and Self-learning Systems*. Advanced Textbooks in Control and Signal Processing. Springer, London.

[4] Medina Hadjem and Farid Naït-Abdesselam. An ecg t-wave anomalies detection using a lightweight classification model for wireless body sensors. *2015 IEEE International Conference on Communication Workshop (ICCW)*, pages 278–283, 2015.

[5] Peter D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer, New York, NY.

[6] Chao Lin, Corinne Mailhes, and JeanYves Tourneret. P- and t-wave delineation in ecg signals using a bayesian approach and a partially collapsed gibbs sampler. *IEEE Transactions on Biomedical Engineering*, (57):2840–2849, 2010.

[7] Micha Sznajder and Marta ukowska. Python Online and Offline ECG QRS Detector based on the Pan-Tomkins algorithm, April 2018.

[8] N. Kelson J. Banks R. Hayward T. Bodisco, J. D' Netto and T. Parker. Characterising an ecg signal using statistical modelling: a feasibility study. *ANZIAM Journal*, (55):C1–C15, 2014.

[9] Pengwei Xie. Bidirectional recurrent neural network and convolutional neural network (bircnn) for ecg beat classification. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, page 2555, 2018.

6 Appendix

6.1 I. Full code for Metropolis-Hastings sampling

This code has been submitted in Sakai alongside this document and can also be found online here.

6.2 II. Full code for Pan-Tompkins Algorithm

Python code for the Pan-Tomkins algorithm was taken from the repository Python Online and Offline ECG QRS Detector based on the Pan-Tomkins Algorithm, written by Michal Sznajder and Marta Lukowska. This repository can be downloaded here.

6.3 III. MCMC Mixing

Figures below represent the MCMC convergence and mixing performance of parameters of our model.

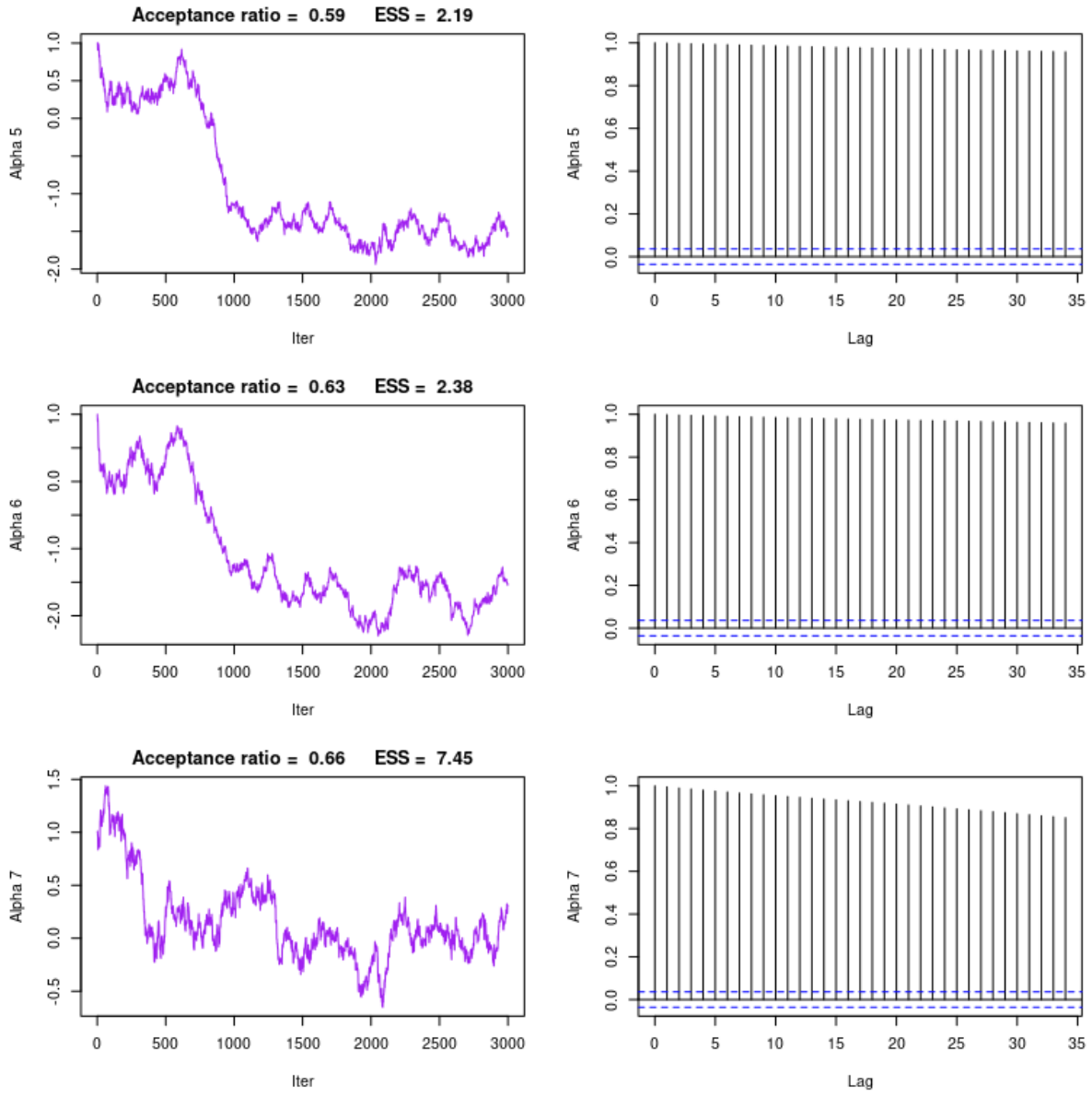

Figure 5: Traceplots and Autocorrelation plots for alpha parameters for the healthy patient.

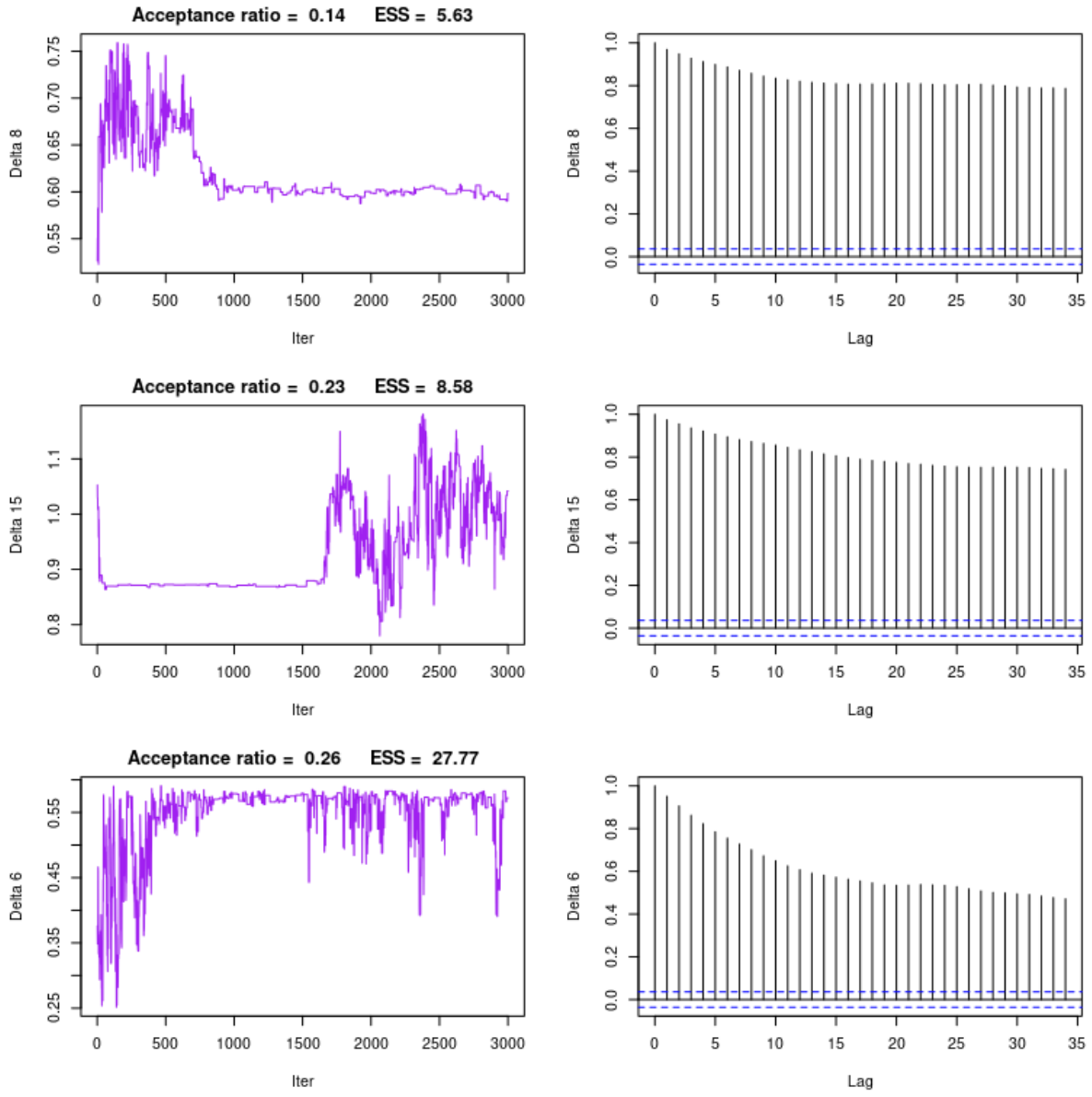

Figure 6: Traceplots and Autocorrelation plots for delta parameters for the healthy patient.

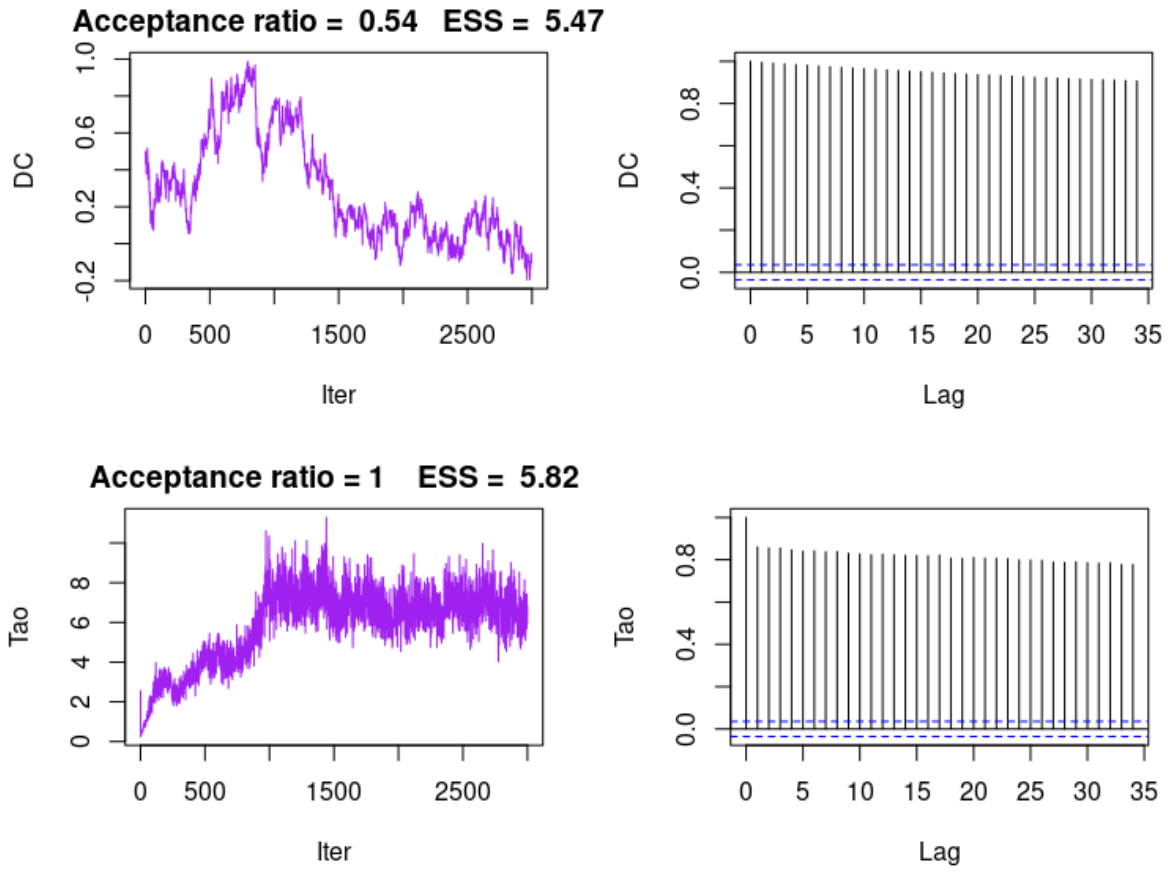

Figure 7: Traceplots and Autocorrelation plots for dc and tao parameters for the healthy patient.

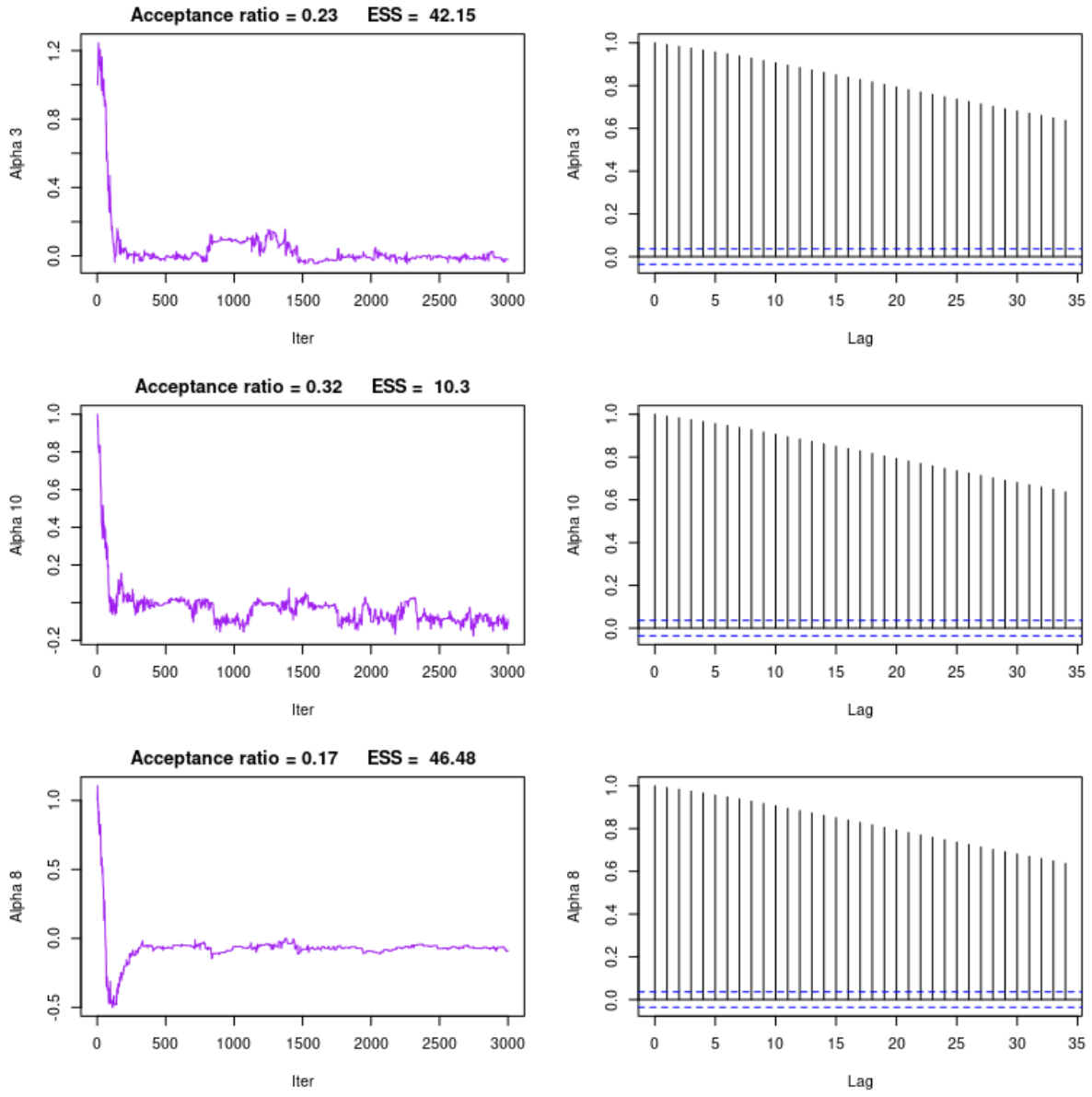

Figure 8: Traceplots and Autocorrelation plots for alpha parameters for the patient with an arrhythmia.

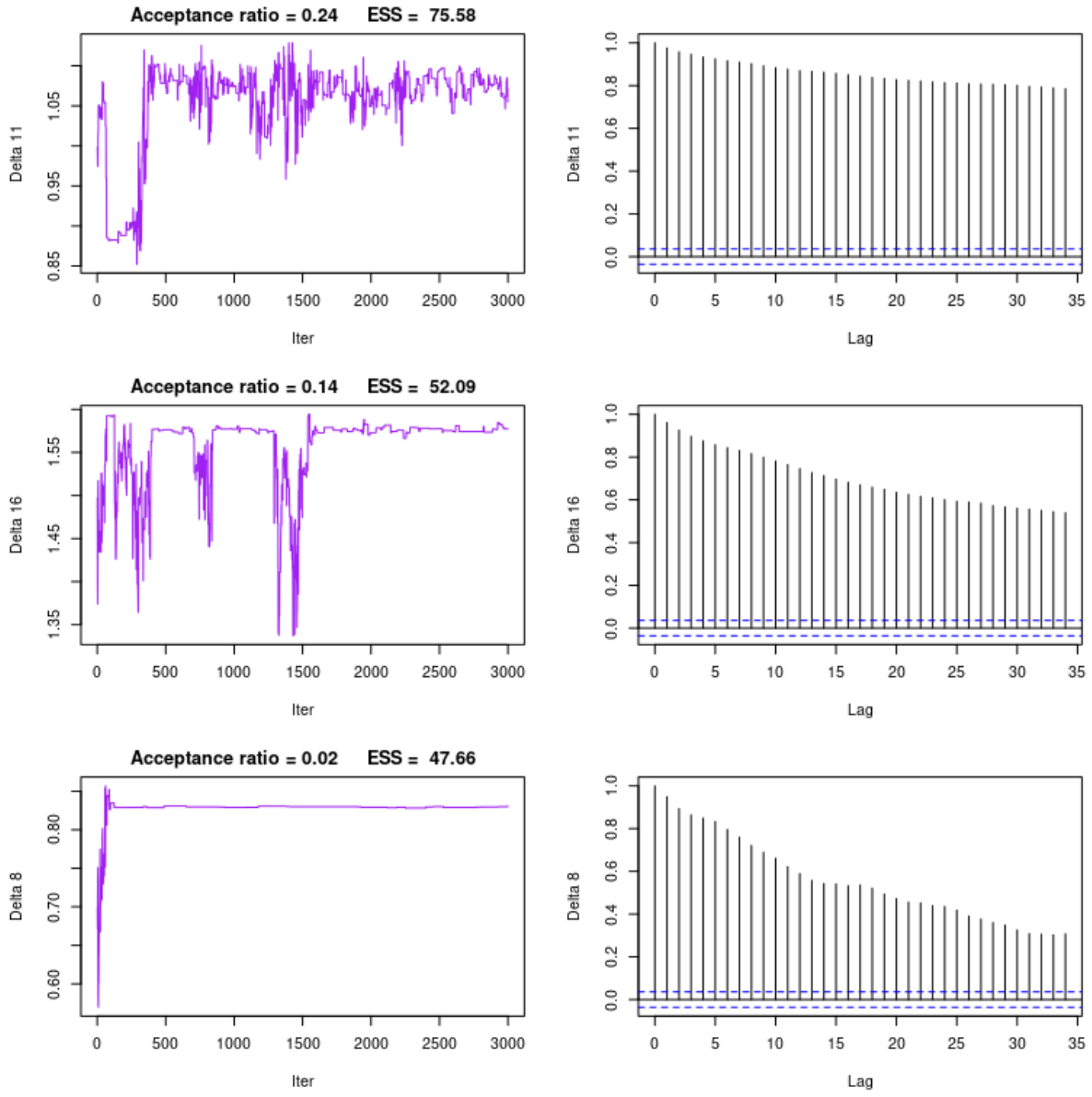

Figure 9: Traceplots and Autocorrelation plots for delta parameters for the patient with an arrhythmia.

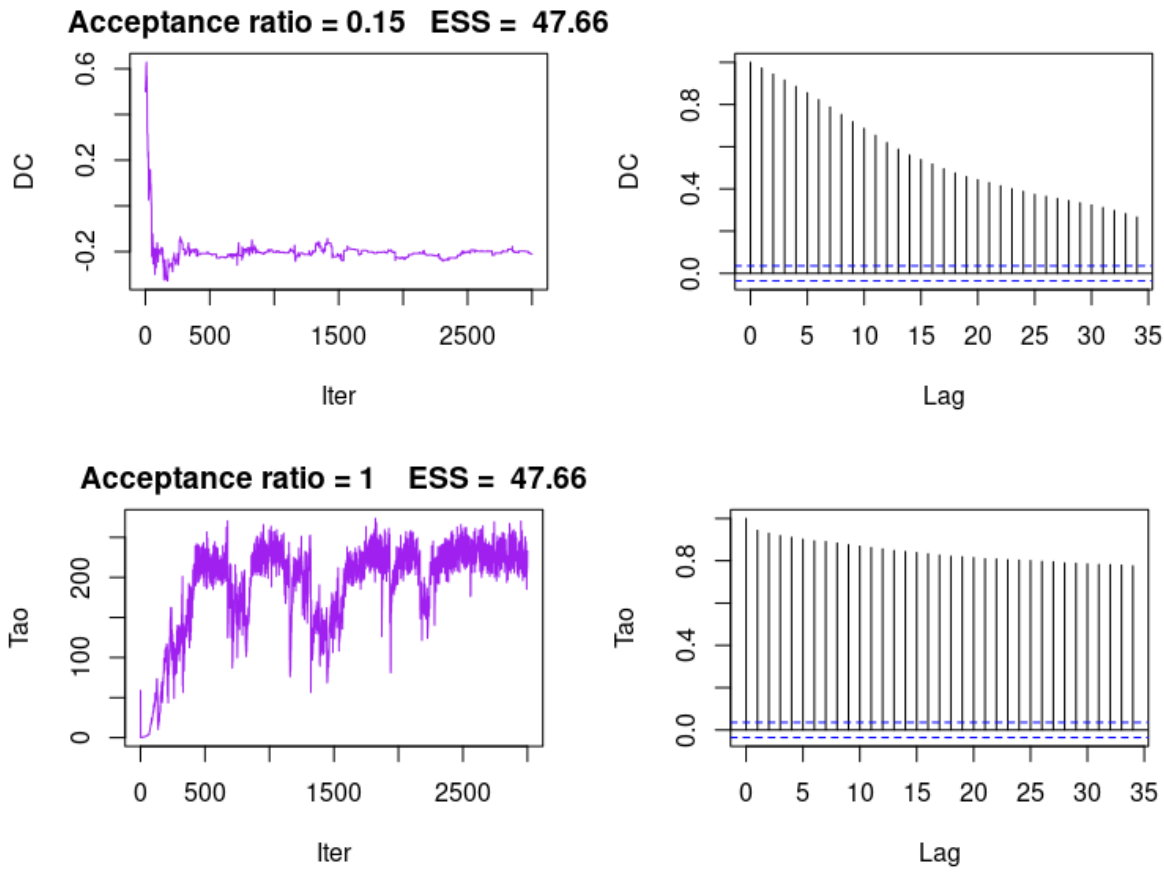

Figure 10: Traceplots and Autocorrelation plots for dc and tao parameters for the patient with an arrhythmia.