Detecting Premature Heartbeats in ECG Data

Carnegie Mellon University
Statistics & Data Science

By: Prakruthi Pradeep, Sabrina Rodriguez, Xinyi Ke

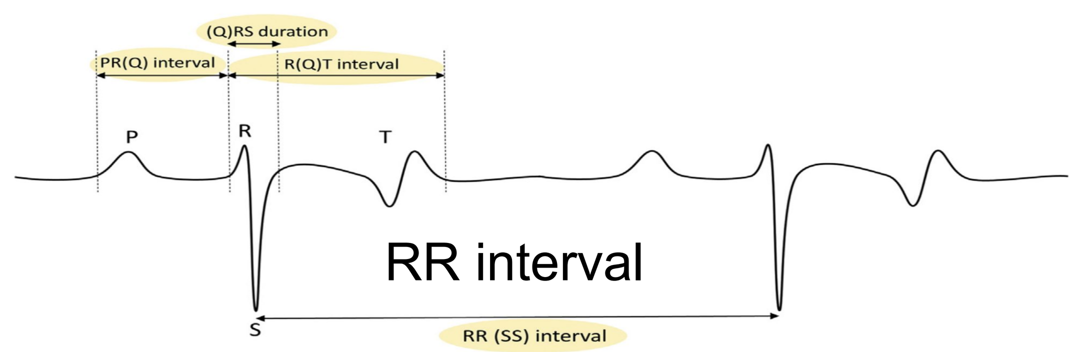
Project Advisor: Ron Yurko External Client: Dr. Katharyn Mitchell

Introduction

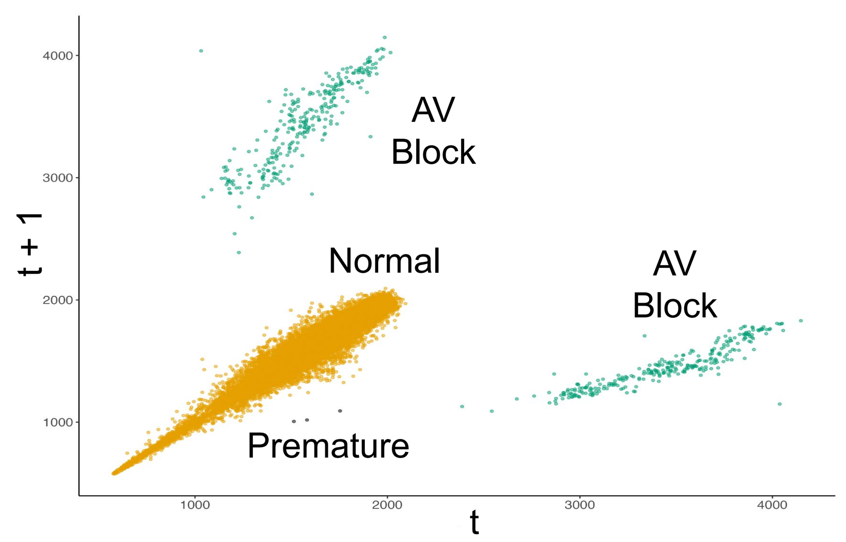
- Cardiac Arrhythmias are irregular rhythm in hearts that can potentially be life-threatening for race horses. An electrocardiogram (ECG) enables clinicians to detect these abnormalities in a non-invasive manner.
- **Research goal**: to **identify premature points in ECG data with statistical and machine learning methods**, to make detection easier and provide a basis of comparison for experts.

Data

- **Univariate** data of RR values
- 7 datasets each containing over **20k** values
- All 7 horses are **healthy**
- **0-5 premature points** in each dataset according to experts
- Preprocessed the data by adding a second column containing RR values from time t+1.
- Speed of heart rates can be examined by time differences between two heartbeats, the RR interval.



- Premature points located near the main block, with the pattern of a normal heartbeat followed by a shorter heartbeat.
- AV blocks are further away from the main block, with the pattern of a larger value between two normal values.



Methods

- Based on the fact that AV blocks are not the points of interest, we conducted the research in **two stages**:
  1. Removed AV blocks with **classification** methods,
  2. Identified premature points with **outlier detection** methods.
- We investigated **Gaussian Mixture Modelling (GMM)** for classification, and **Isolation Forests** and **Mahalanobis Distance** for outlier detection.
- **GMM:** probabilistic model assuming all data points are generated from mixture of finite number of Gaussian distributions with unknown parameters.

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{i=1}^{K} \tilde{\phi}_i \mathcal{N}(\tilde{\boldsymbol{\mu}_i}, \tilde{\boldsymbol{\Sigma}_i})$$

- The EM algorithm is used for unsupervised classification, alternating between steps 1 and 2 until convergence:
  1. Infer the posterior distribution of the latent variables given the model parameters.
  2. Tune parameters to maximize the data likelihood given the latent variable distribution
- **Cluster combination:** Used to determine optimal classification.
- **Isolation forests**: Machine learning algorithm for anomaly detection, that uses decision trees to split data randomly. This causes anomalies to be isolated more often than regular points.
- **Mahalanobis Distance**: After applying GMM on filtered data and selecting number of components according to BIC, used mean and covariance of its component assignment to measure distance. Using Bonferroni correction, outliers were detected.

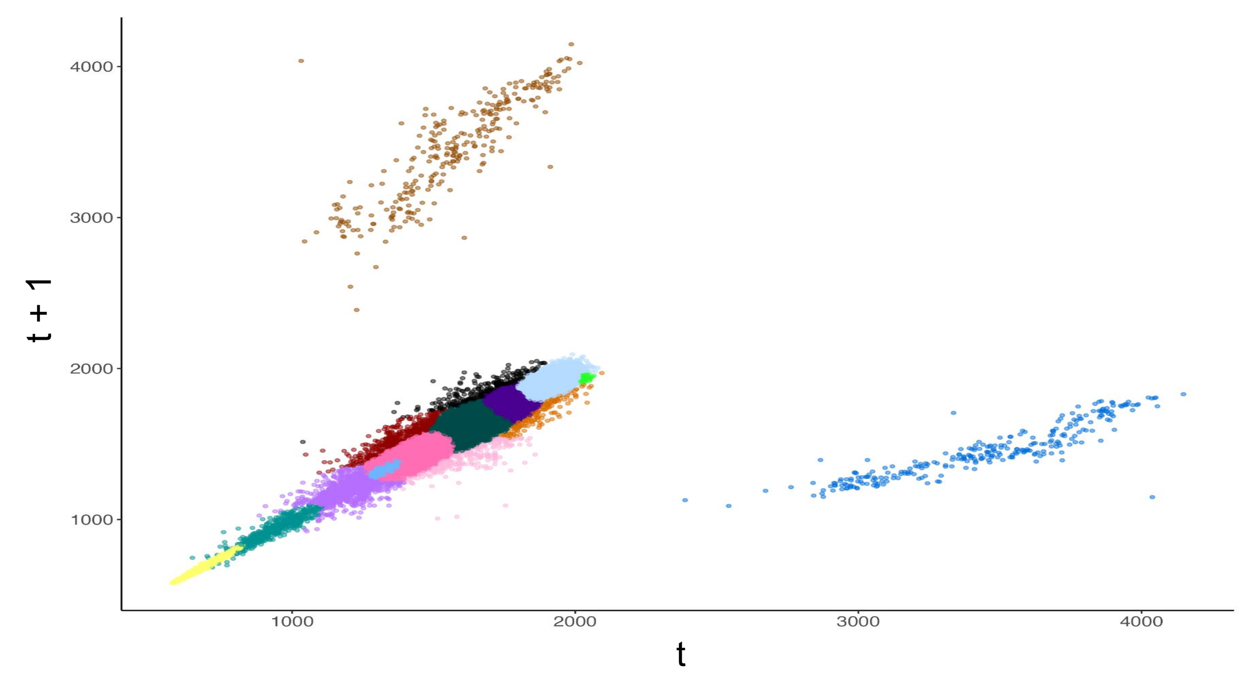
$$\Pr[\vec{x}]\,d\vec{x} = \frac{1}{\sqrt{\det(2\pi\mathbf{S})}}\exp\left(-\frac{d^2}{2}\right)d\vec{x}.$$

Results

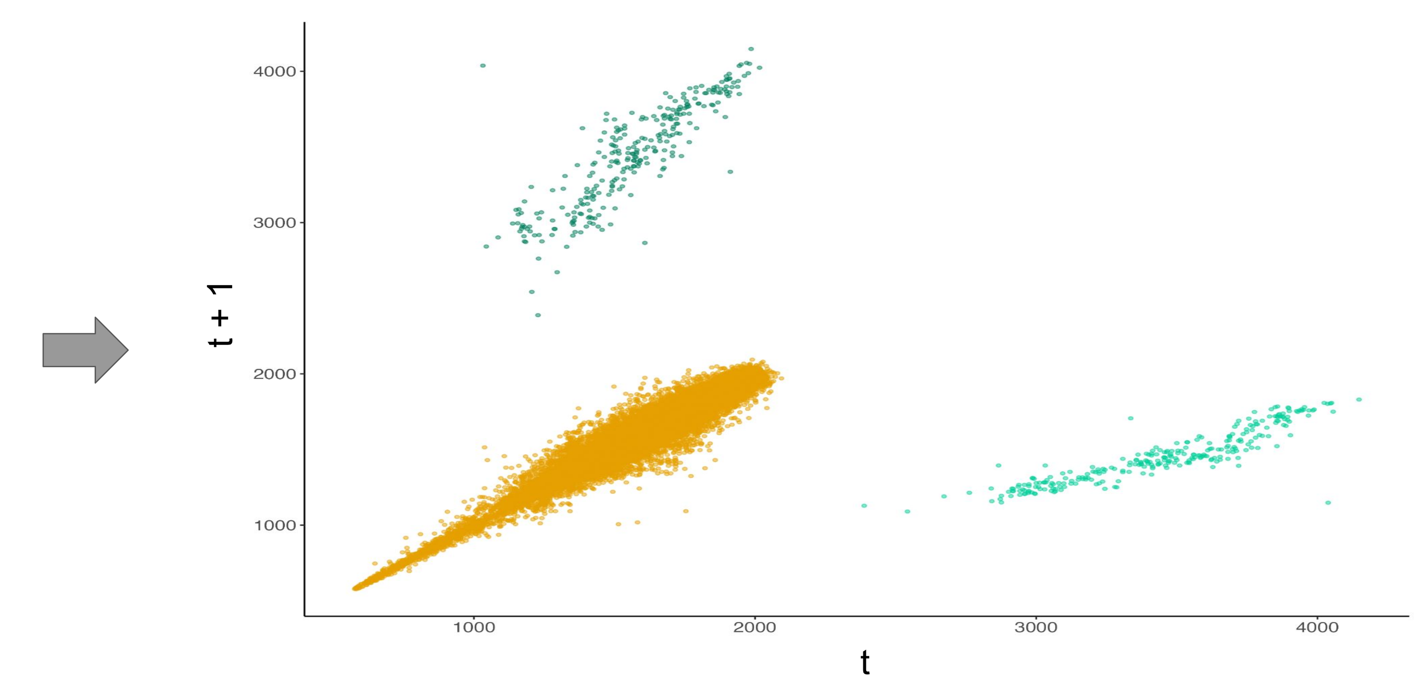
** All of the results displayed in this section are from one dataset/one horse. The datasets all varied in overall results.*

Unsupervised Classification with Gaussian Mixture Modeling

Initial Clustering



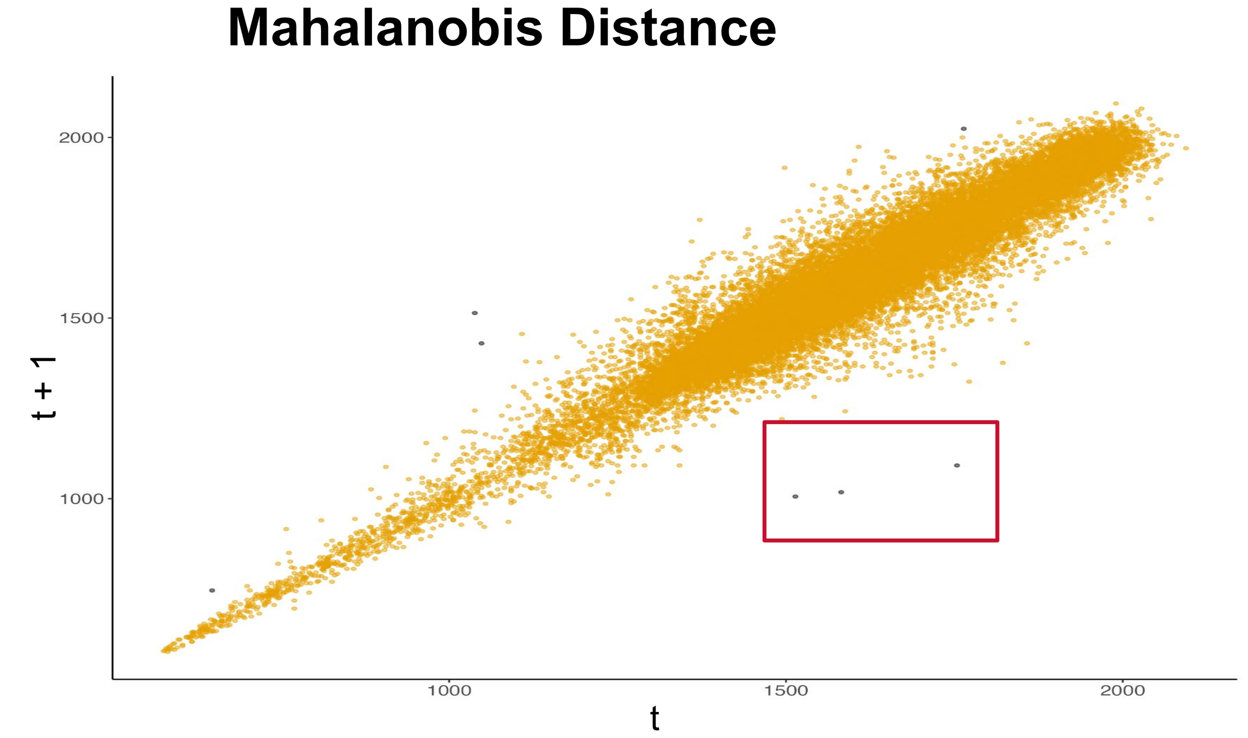
Combined Clustering



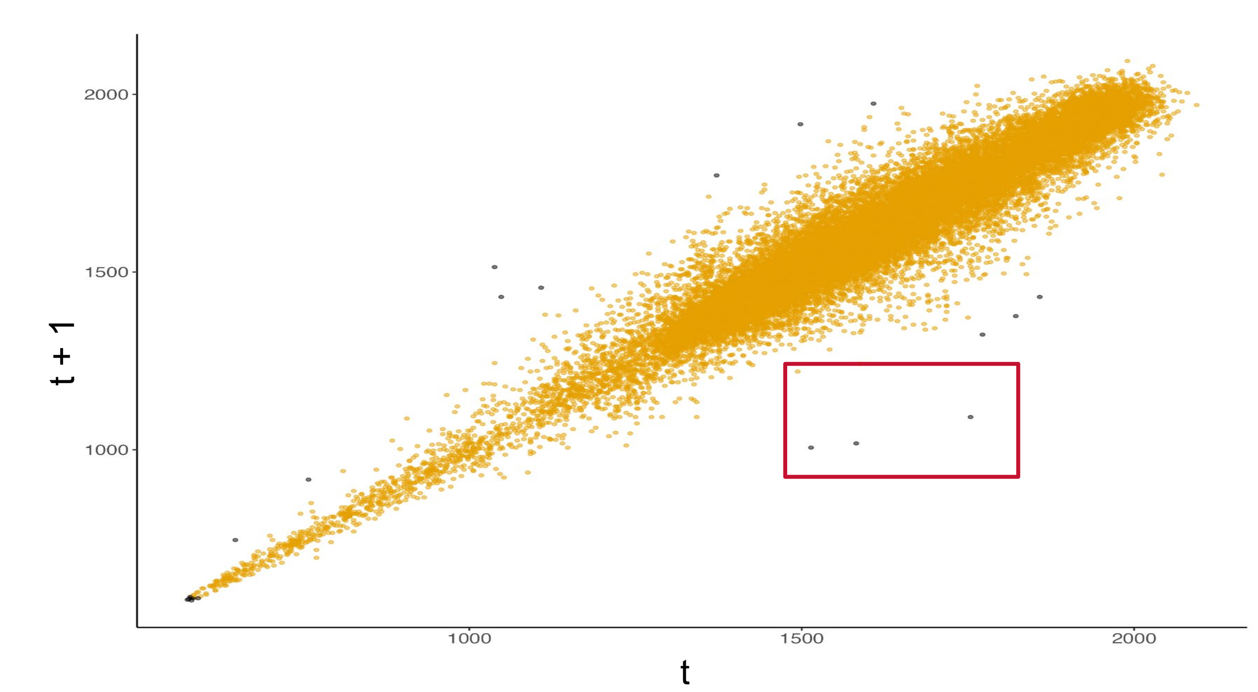
- Model estimated by EM algorithm initialized by hierarchical model-based clustering.
- Initialized at 15 components with most flexible parametrization: ellipsoidal, varying volume, shape, and orientation.
- Clusters are combined according to entropy criterion, and final classification is chosen manually in comparison to idealized Poincare plot.
- Final Classification: three clusters accurately separate AV blocks from main points for this horse.

Anomaly Detection

Mahalanobis Distance



Isolation Forests



- Number of components selected by BIC is 15. Compared against a chi-square with Bonferroni correction, all premature beats are captured, with a narrow range of outliers.
- The isolation forests method was able to capture a much narrower range of outliers and effectively isolate the premature beats (highlighted by the red box).
- Isolation forests method marked 16 points with a threshold level of 0.70, while Mahalanobis distance method marked 5 points as potential premature points.

Conclusions and Future Steps

- Successfully clustered and filtered out AV blocks by GMMs, and highlighted a small subset of potential premature points via outlier detection methods using isolation forest and Mahalanobis distance, reducing the need of manual work in identifying premature points.
- Future investigations include constructing generalized methods for selecting threshold levels for isolation forest and Mahalanobis distance, and further analysis on data from unhealthy horses.

References

Mitchell KJ. Equine Electrocardiography. Vet Clin North Am Equine Pract. 2019 Apr;35(1):65-83.
Mclust package. https://CRAN.R-project.org/package=mclust