# Identification of patterns within electrocardiogram of big data

William Mateus Avila

School of Statistic
National University of Colombia
Medellín

2015/10

# The problem

The problem consists of the following things:

- ¿What is a Electrocardiogram (ECG)?, and the importance of study these.
- Work with ECG's of big data.
- The difficult identification of the cycles.
- ¿Where and when the ECG has irregularities?.
- ¿Is possible determine a some pattern for the ECG of big data?.

# Electrocardiogram and its cycles

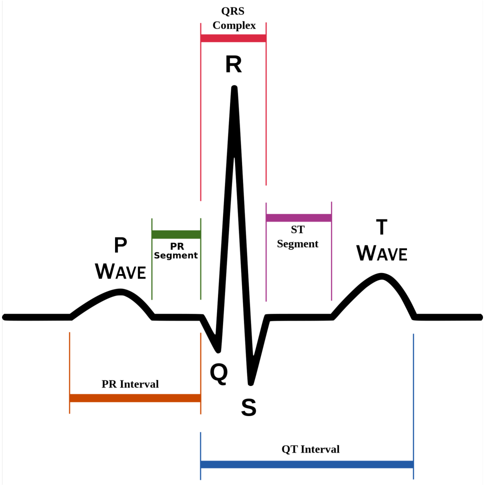



Figura: One cycle of the ECG

## Components of the cycle

One cycle is formed of the following Components:

- P wave
- PR interval
- PR segment
- QRS complex
- QT interval
- ST segment
- T wave
- U wave

# Importance of some components of the cycle

The following components are very important in our work and in the medical meaning:

- P wave
- PR interval
- QRS complex
- QT interval

The data that we studied have the following characteristics:

- Has many irregularities.
- The length of the cycles in one ECG are very different.
- The quantity of cycles is huge; In one ECG we got around 277000 cycles.

# Some characteristics of one ECG

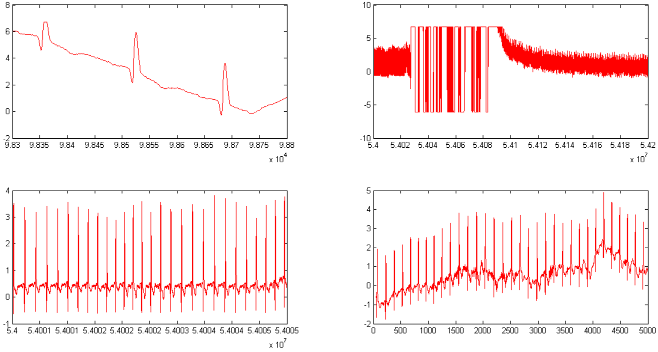



Figura: Muller-ECG

# The method of least square

One way to represent a cycle of the ECG is with the use of a polynomial of degree n, therefore if took m pairs of data $(x_i, y_i)$ with $n \leq m$; it defines the following polynomial:

$$P(x) = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n$$

And it calculates:

$$S = \sum_{i=0}^{m} (P(x_i) - y_i)^2 = \sum_{i=0}^{m} (a_0 + a_1 x_i + a_2 x_i^2 + ... + a_n x_i^n - y_i)^2$$

Therefore if it wants the coefficients of the polynomial, we must to determine the coefficients $a_0, a_1, a_2, ..., a_n$ such that S is minimal.

# The method of least square

We do the partial derivatives of S with respect to $a_0, a_1, a_2, ..., a_n$, finally the outcome be equal to zero, it means

$$\frac{\delta S}{\delta a_0} = 0; 2\sum_{i=0}^{m}(a_0 + a_1 x_i + a_2 x_i^2 + ... + a_n x_i^n - y_i) * 1 = 0$$

$$\frac{\delta S}{\delta a_1} = 0; 2\sum_{i=0}^{m}(a_0 + a_1 x_i + a_2 x_i^2 + ... + a_n x_i^n - y_i) * x_i = 0$$

$$\vdots$$

$$\frac{\delta S}{\delta a_n} = 0; 2\sum_{i=0}^{m}(a_0 + a_1 x_i + a_2 x_i^2 + ... + a_n x_i^n - y_i) * x_i^n = 0$$

Thus we obtain a system of $n+1$ equations with $n+1$ unknown quantity $a_0, a_1, a_2, ..., a_n$, due to that matrix has rank $n+1$ the system has solution.

# Problem

We want that a cycle be approximate using least square with a polynomial of degree n, and we ask if the previous approximation is good?.

We have to remember that the length of the cycles in one ECG are very different and the quantity of cycles is huge.

## Example

With the use of MATLAB, it selects one cycle with length of 149 (The cycle is located between the data 1179 and 1327 of the Muller ECG), the cycle is the following:

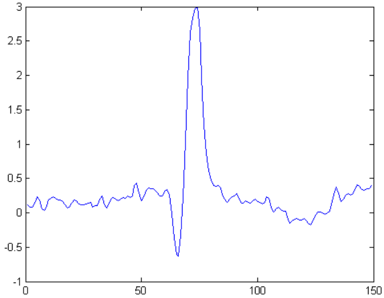



Figura: Ejemplo

# Approximation with the use of least square

We use the polyfit function (this function uses least square) with domain [1179,1327] and the polyval function of MATLAB, the outcome is:

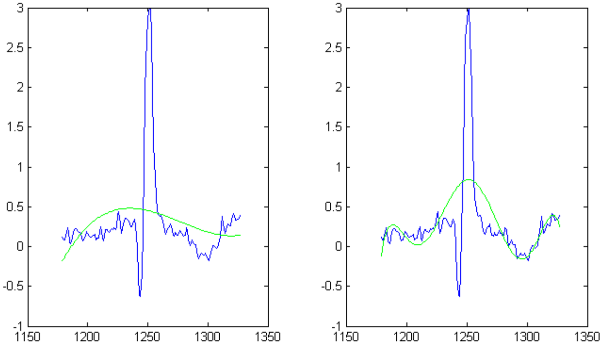



Figura: Polyfit of 3 and 6 degree respectively

# Conclusions

- If it uses the method of least square for approximate one cycle of the ECG, this is not a good way.

  The reason is that the coeficients of the polynomial are extremes.

- The use of other method is necesary.

## ¿What is a B-spline and How to it defines?

- A B-spline is a spline function $(P(t) : \mathbb{R} \Rightarrow \mathbb{R}^l)$, this function is defined with respect to one degree, softness and domain partition.

It defines the B-spline curve associated to:

- The degree n of the spline.
- The control points $P_0, P_1, ..., P_m$, with $P_i \in \mathbb{R}^l$
- The set knots $a_0, a_1, ..., a_{n+m+1}$, with $a_i \in \mathbb{R}$, $a_i \leq a_j$, $i \neq j$.

Like

$$P(t) = \sum_{i=0}^{m} P_i N_i^n(t)$$

Where $t \in [a_n, a_{m+1})$ y $N_i^n(t)$ are the basic functions.

# Basic functions

We take the following sequence $\{a_i\}$ such that $a_i < a_{i+1}\ \forall i$, the previous sequence will help us to define the B-spline

Our basic functions $N_i^n(t)$ are defined through the following recurrence relation:

$$N_i^0(t) = \begin{cases} 1 & t \in [a_i, a_{i+1}) \\ 0 & \text{in other case} \end{cases}$$

For $n \neq 0$

$$N_i^n(t) = \alpha_i^{n-1} N_i^{n-1}(t) + (1 - \alpha_{i+1}^{n-1}) N_{i+1}^{n-1}(t)$$

Where

$$\alpha_i^{n-1} = \frac{(t - a_i)}{(a_{i+n} - a_i)}$$

## Some observations about the B-spline

- The sequence $\{a_i\}$ have the restriction that $a_i \in \mathbb{R}$, $a_i \leq a_j$, $i \neq j$.
- If we want to calculate $N_m^n$, with $n \in \mathbb{N}$ and $m \in \mathbb{N}^*$; $\mathbb{N}^* = \mathbb{N} \cup \{0\}$ it's necessary $(n+m+1)+1$ knots.
- $N_i^n$ It's piecewise polynomial.
- $\sum_{i=0}^{m} N_i^n(t) = 1$ for all $t \in [a_n, a_{m+1})$ and $n \leq m$, this property is called "the partition of unity".
- If it wants that the B-spline is defined in $[a_n, a_{m+1})$ and in the same time meets with the property of the partition of unity, we must reject $2n$ knots.
- The control points has not any restrictions.

# The problem

If it wants to approximate one cycle of the ECG with the use of B-spline of degree 2 or 3. We ask the following questions: ¿How to make this approximation? and ¿which one is the best approximation?

For this problem it must keep in mind the following things:

- The length of the cycle
- $P(t) = \sum_{i=0}^{m} P_i N_i^3(t) \in \mathbb{R}\ \forall t \in [a_3, a_{m+1})$, for degree 3
- ¿What is the importance of the t's?
- ¿Which one is the game of the control points?

## How to make the approximation

If we want to use a B-spline for approximated one cycle of the ECG, is necessary:

- The basic functions of degree 3, i.e $N_0^3, N_1^3, ..., N_m^3$ on the sequence of knots $a_0, a_1, ..., a_{m+n+1}$.
- The points (or data) $c_0, c_1, ..., c_k$ with $m < k$ these must approximated in the X axis $t_0 < t_1 < ... < t_k$ such that $t_j \in [a_n, a_{m+1})$.

With the previous, it has to meet the following equation:

$$\sum_{i=0}^{m} P_i N_i^3(t_j) = c_j$$

Note:

- When $k = m$, we talk of interpolation.
- ¿What kind of condition has the t's?

# Continuation

With the previous equation, it satisfies the following linear system:

$$\begin{pmatrix} N_0^3(t_0) & \cdots & N_m^3(t_0) \\ \vdots & & \vdots \\ \vdots & & \vdots \\ N_0^3(t_k) & \cdots & N_m^3(t_k) \end{pmatrix} \begin{pmatrix} P_0 \\ \vdots \\ P_m \end{pmatrix} = \begin{pmatrix} c_0 \\ \vdots \\ \vdots \\ c_k \end{pmatrix}$$

If it selects a good set of t's, the system has a solution

# Example

With the use of MATLAB, it selects one cycle with length of 149 (The cycle is located between the data 1179 and 1327 of the Muller ECG)

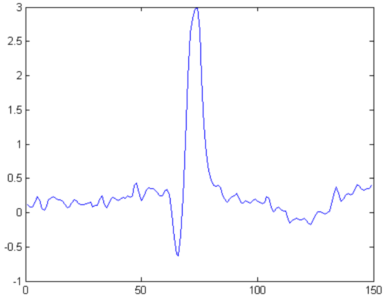
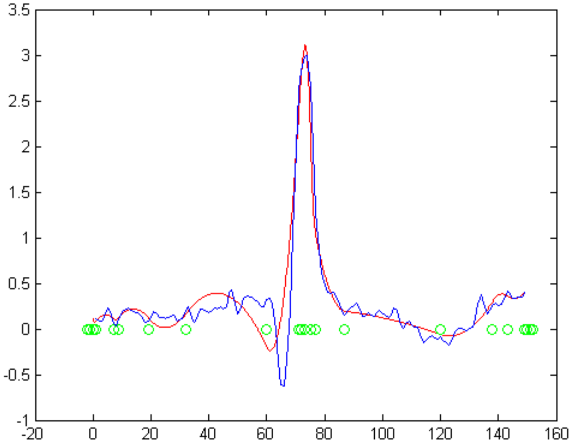


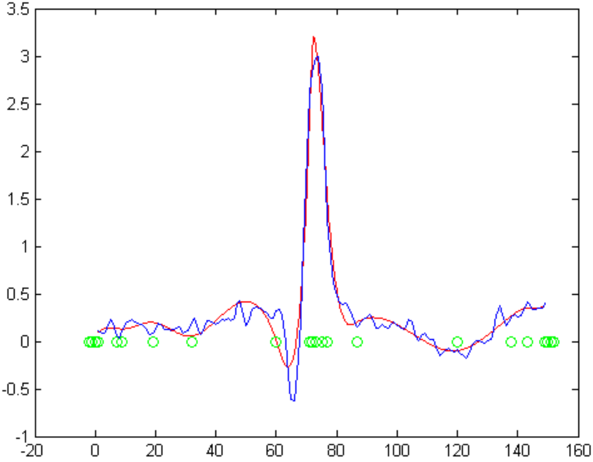

Figura: Example

# Approximation with the method of B-spline's of degree 2 and 3

A set of knot points are chosen (green circles), the example was approximate to B-spline of degree 2 and 3, as following



# Approximation with the method of B-spline's of degree 2 and 3



# Conclusions

- The approximation by the method of B-spline of degree 3 is a good strategy.
- The choise of $t_j's$ and knot points are appropriate and handmade.
- If it knows the ubication of the complex QRS, the aproximation can be better.
- The approximation generated a set of control points, these points represent that cycle, in the example the B-pline of degree 3 has 17 contol points.
- The method of B-splines is useful to characterize the cycles.

## Sequence of cycles

If it wants to identify patterns, keep in mind the following:

- We filter the ECG with the use of convolution

The convolution of two vectors, x and y, represents the area of overlap under the points as y slides across x. Algebraically, convolution is the same operation as multiplying polynomials whose coefficients are the elements of x and y.

$$w(k) = \sum x(j)y(k-j+1)$$

# Sequence of cycles

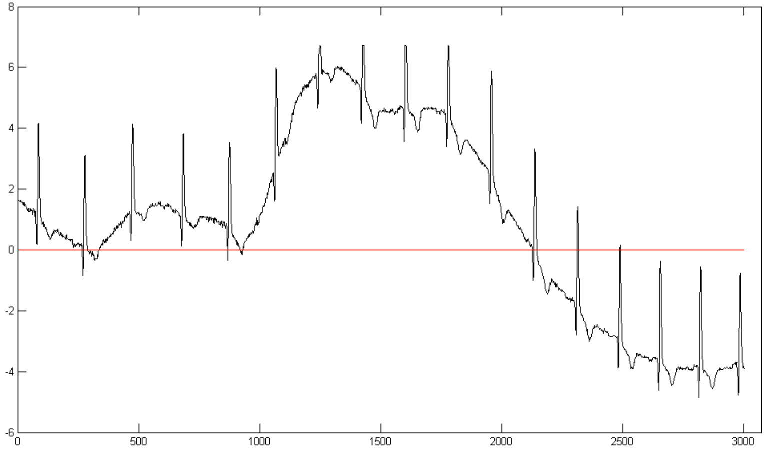



Figura: Sequence of cycles without filter

# Sequence of cycles

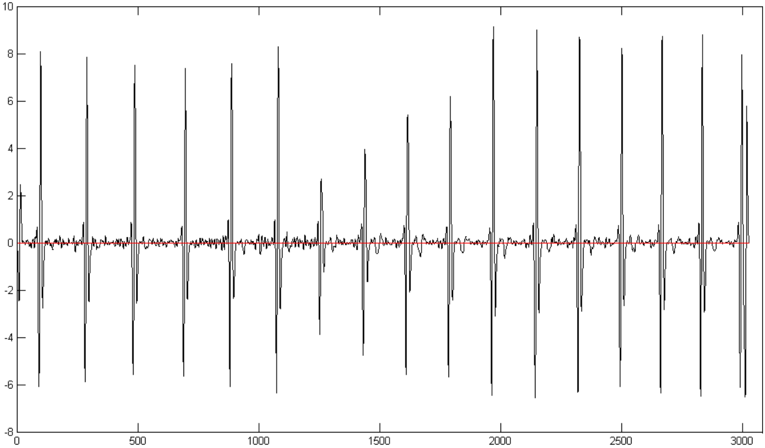



Figura: Sequence of cycles with filter

# Sequence of cycles

- With help of the convolution, we can detect problems like as follow:

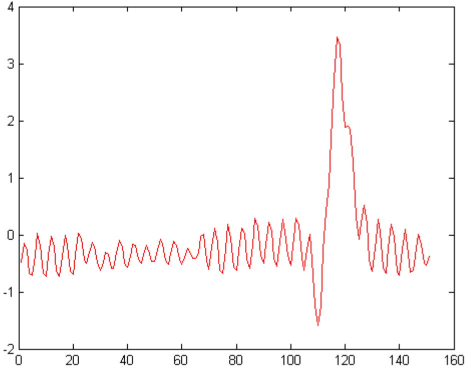



Figura: High frequency

## Sequence of cycles

- Thus we can select a large block of cycles without presence of high frequency and with the signal stabilized around zero, for example:

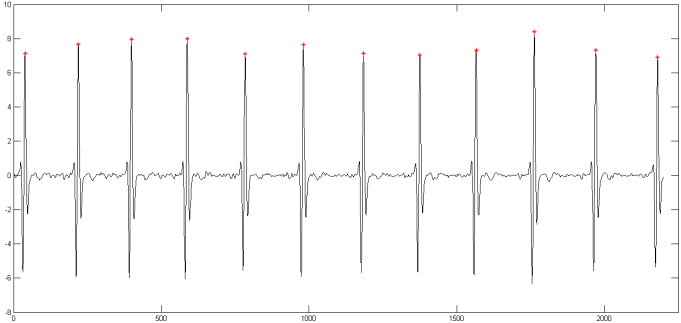



Figura: A block without problems

## Sequence of cycles

- We reduce the length of the cycles within one sequence cycles, for example:

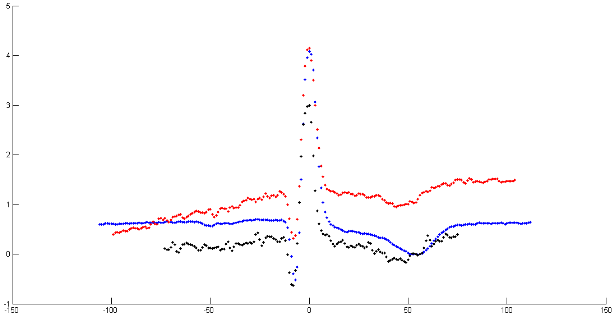



Figura: Set of cycles

## Analysis of a block with 500 cycles

For the previous explanation,we select a large block of cycles (500 cycles) without problems, the following figure shows a segment of that block

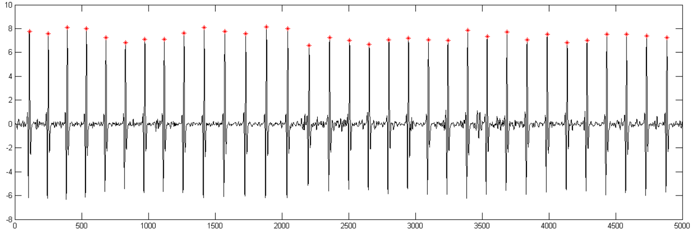
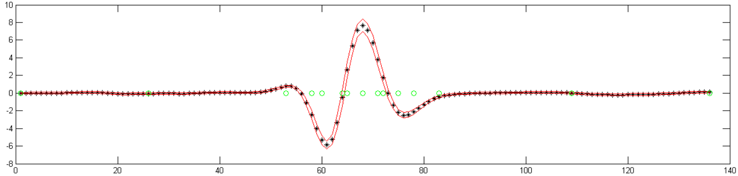



Figura: A short segment of a block of 500 cycles

# Analysis of a block with 500 cycles

With the block of 500 cycles, it uses the approximation by B-splines of degree 3 for each cycle and it calculates the mean and variability both data and control points




| | Punto1 | Punto2 | Punto3 | Punto4 | Punto5 | Punto6 | Punto7 | Punto8 | Punto9 | Punto10 | Punto11 | Punto12 | Punto13 | Punto14 | Punto15 | Punto16 | Punto17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.0028 | 0.0082 | 0.0922 | -0.3294 | 1.2783 | -7.9920 | -4.2859 | 5.1586 | 9.5667 | 5.1940 | 0.3766 | -3.4243 | -0.9158 | 0.3924 | -0.2045 | -0.1938 | 0.1605 |
| 2 | 0.0991 | 0.2204 | 0.3606 | -0.1090 | 1.4954 | -7.1873 | -3.1662 | 6.1291 | 10.4274 | 6.0317 | 0.9864 | -2.9205 | -0.6907 | 0.5650 | 0.0178 | 0.0228 | 0.2887 |
| 3 | -0.1047 | -0.2039 | -0.1762 | -0.5498 | 1.0613 | -8.7967 | -5.4057 | 4.1881 | 8.7059 | 4.3562 | -0.2331 | -3.9281 | -1.1409 | 0.2197 | -0.4268 | -0.4104 | 0.0323 |

# Conclusions

- With help of the convolution, certain problems have a good solution.
- We can characterize the blocks of cycles with the use of the control points.
- We have to use a good statistical technique for the interpretation of some class of pattern.

**THANKS FOR YOUR ATTENTION**