Topological Data Analysis to identify subgroups of type-2 Diabetes Mellitus patients

Juan Vallarta
MSc Health Data Science

David Prieto-Merino
International Chair of Statistical Analysis and Big Data
Introduction

Some evidence suggests that the pathogenesis of the Type-2 Diabetes Mellitus (T2DM) is not only influenced by a deficiency in the pancreatic functions, but from a more complex pathway of the disease.
Some evidence suggests that the pathogenesis of the Type-2 Diabetes Mellitus (T2DM) is not only influenced by a deficiency in the pancreatic functions, but from a more complex pathway of the disease.

The use of electronic medical records and the implementation of new analytical techniques, such as machine learning algorithms, can provide a better understanding of diseases.
Topological Data Analysis (TDA) is an unsupervised algorithm which main characteristic is to study the shape of data. This technique has been previously used to identify subtypes of T2DM in the American population. However, there is not much information regarding subtypes of this disease and the implementation of TDA.

Some evidence suggests that the pathogenesis of the Type-2 Diabetes Mellitus (T2DM) it is not only influenced by a deficiency in the pancreatic functions, but from a more complex pathway of the disease.

The use of electronic medical records and the implementation of new analytical techniques, such as machine learning algorithms, can provide a better understanding of diseases.
To perform a TDA using Electronic Medical Records (CALIBER dataset) to identify unique clusters of T2DM patients.
Methods
Database and Study Population

**CPRD (primary care)**
- Life-styles
- Nutritional Status
- Lab tests
- Diagnostics
- Prescriptions
- Procedures

**HES (secondary care)**
- Sociodemographic
- Administrative
- Lab tests
- Diagnostics
- Prescriptions
- Procedures

**ONS (Dep. statistics)**
- Mortality
- Deprivation

**MINAP**
- Cardiovascular diseases

**CALIBER** England 1998-2010

Filtered

Algorithm needs server limitations

- 101,514 patients with T2DM
- 6,851 patients with T2DM
Data Preprocessing

Numeric features

<table>
<thead>
<tr>
<th>ID</th>
<th>Year1</th>
<th>Year2</th>
<th>Year3</th>
<th>Year4</th>
<th>Year5</th>
<th>Year6</th>
<th>BMI μ</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>26</td>
<td>28</td>
<td>27</td>
<td>30</td>
<td>32</td>
<td>28</td>
<td>1.2</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>30</td>
<td>31</td>
<td>32</td>
<td>30</td>
<td>32</td>
<td>31</td>
<td>1.8</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>22</td>
<td>23</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>24</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HDL μ</th>
<th>HDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.2</td>
</tr>
<tr>
<td>180</td>
<td>2.1</td>
</tr>
<tr>
<td>165</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Mean was obtained for numeric variables (6 years previous to the T2DM diagnosis)
Numeric features were standardised
Data Preprocessing

**Numeric features**

| ID | Year1 | Year2 | Year3 | Year4 | Year5 | Year6 | BMI μ | BMI  
|----|-------|-------|-------|-------|-------|-------|-------|-------
| 1  | 25    | 26    | 28    | 27    | 30    | 32    | 28    | 1.2   |
| 2  | 32    | 30    | 31    | 32    | 30    | 32    | 31    | 1.8   |
| 3  | 24    | 22    | 23    | 25    | 25    | 25    | 24    | 0.6   |

Mean was obtained for numeric variables (6 years previous to the T2DM diagnosis)

**Numeric features were standardised**

| ID | Year1 | Year2 | Year3 | Year4 | Year5 | Year6 | HDL μ | HDL  
|----|-------|-------|-------|-------|-------|-------|-------|-------
|    |       |       |       |       |       |       | 100   | 0.2   |
|    |       |       |       |       |       |       | 180   | 2.1   |
|    |       |       |       |       |       |       | 165   | 1.4   |

**Categorical features**

| ID | Year1 | Year2 | Year3 | Year4 | Year5 | Year6 | MI  
|----|-------|-------|-------|-------|-------|-------|-----
| 1  | Yes   | No    | No    | No    | No    | Yes   | Yes |
| 2  | No    | No    | No    | No    | No    | No    | No  |
| 3  | No    | No    | No    | Yes   | No    | No    | Yes |

Categorical variables were transformed into dummies (6 years previous to the T2DM diagnosis)
Topological Data Analysis
TDA Properties

- **Deformation Invariance**
- **Compressed Representation**
- **Coordinate Freeness**
## Data Analysis (TDA Mapper)

<table>
<thead>
<tr>
<th>ID</th>
<th>IMC</th>
<th>GLU</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>120</td>
<td>150</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>210</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>180</td>
<td>120</td>
</tr>
</tbody>
</table>

Database
Data Analysis (TDA Mapper)

**Database**

<table>
<thead>
<tr>
<th>ID</th>
<th>IMC</th>
<th>GLU</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>120</td>
<td>150</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>210</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>180</td>
<td>120</td>
</tr>
</tbody>
</table>

**Distance Matrix**

```
   1 | 2 | 3
---|---|---
1  | 0 | 4.6| 4.6|
2  | 4.6| 0  | 6.9|
3  | 4.6| 6.9| 0  |
```
Data Analysis (TDA Mapper)

Database

<table>
<thead>
<tr>
<th>ID</th>
<th>IMC</th>
<th>GLU</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>120</td>
<td>150</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>210</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>180</td>
<td>120</td>
</tr>
</tbody>
</table>

Distance Matrix

```
1  2  3
1  0  4.6 4.6
2  4.6 0  6.9
3  4.6 6.9 0
```

Filter functions
Definition of parameters
(Intervals, Overlap)

\[ f(x) = \max d(x, y) \]

\[ A = U \Sigma V^T \]
Data Analysis (TDA Mapper)

<table>
<thead>
<tr>
<th>ID</th>
<th>IMC</th>
<th>GLU</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>120</td>
<td>150</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>210</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>180</td>
<td>120</td>
</tr>
</tbody>
</table>

Database

Distance Matrix

\[
A = U \Sigma V^T
\]

\[
f(x) = \max d(x,y)
\]

Sphere projection

Filter functions
Definition of parameters (Intervals, Overlap)
Data Analysis (TDA Mapper)

Database

<table>
<thead>
<tr>
<th>ID</th>
<th>IMC</th>
<th>GLU</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>120</td>
<td>150</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>210</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>180</td>
<td>120</td>
</tr>
</tbody>
</table>

Distance Matrix

\[
A = U \Sigma V^T
\]

\[
f(x) = \max d(x, y)
\]

Filter functions
Definition of parameters
(Intervals, Overlap)

Sphere projection
Data Analysis (TDA Mapper)

**Database**

<table>
<thead>
<tr>
<th>ID</th>
<th>IMC</th>
<th>GLU</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>120</td>
<td>150</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>210</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>180</td>
<td>120</td>
</tr>
</tbody>
</table>

**Distance Matrix**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>4.6</td>
<td>4.6</td>
</tr>
<tr>
<td>2</td>
<td>4.6</td>
<td>0</td>
<td>6.9</td>
</tr>
<tr>
<td>3</td>
<td>4.6</td>
<td>6.9</td>
<td>0</td>
</tr>
</tbody>
</table>

**Projection at 5 intervals**

(x-coordinates filter)

**Filter functions**

Definition of parameters (Intervals, Overlap)

\[ A = U \Sigma V^T \]

\[ f(x) = \max d(x, y) \]

\[ f(x) \]
Data Analysis (TDA Mapper)

Database

<table>
<thead>
<tr>
<th>ID</th>
<th>IMC</th>
<th>GLU</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>120</td>
<td>150</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>210</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>180</td>
<td>120</td>
</tr>
</tbody>
</table>

Intervals

Distance Matrix

\[
A = U \Sigma V^T
\]

\[
f(x) = \max d(x, y)
\]

Filter functions
Definition of parameters
(Intervals, Overlap)

Projection at 5 intervals
(x-coordinates filter)
Filter functions
Definition of parameters
(Interval, Overlap)

Data Analysis (TDA Mapper)

Database

<table>
<thead>
<tr>
<th>ID</th>
<th>IMC</th>
<th>GLU</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>120</td>
<td>150</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>210</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>180</td>
<td>120</td>
</tr>
</tbody>
</table>

Distance Matrix

\[
A = \Sigma V^T \]

\[
f(x) = \max d(x, y) \]

Projection at 5 intervals
(x-coordinates filter)

Overlap
Data Analysis (TDA Mapper)

Filter functions
Definition of parameters (Intervals, Overlap)

A = UΣVT

f(x) = max d(x, y)

f(x)

Database

Distance Matrix

Projection at 5 intervals (x-coordinates filter)

Clustering
Data Analysis (TDA Mapper)

Filter functions
Definition of parameters (Intervals, Overlap)

\[
A = U \Sigma V^T
\]

\[
f(x) = \max d(x, y)
\]

Projection at 5 intervals (x-coordinates filter)

Distance Matrix

Clustering

TDA

<table>
<thead>
<tr>
<th>ID</th>
<th>IMC</th>
<th>GLU</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>120</td>
<td>150</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>210</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>180</td>
<td>120</td>
</tr>
</tbody>
</table>

Database
Filter Examples

Step 1: Filtration

Step 2: Clustering

Step 3: Topological Representation

It somehow shows that data points are surrounding one center.

Filter 1: L2-Centrality

Filter 2: PCA-PI

Step 1: Filtration

Step 2: Clustering

Step 3: Topological Representation

It somehow shows the linearity of the first principal.

Jing E., 2015
Multinomial Logistic Regression using a backward approach.
Results
Descriptive Results

- 54% women.
- 75% diet to lose weight.
- 12% depression.
- 44% non-smokers and 30% ex-smokers.
- 84% non-alcoholic.
- 7% episode of heart attack.
TDA Output

6-years (5 intervals, 60% overlap, 40 bins when clustering)
TDA alternative outputs

6-years (5 intervals, 50% overlap, 20 bins when clustering)

1-year (5 intervals, 60% overlap, 40 bins when clustering)
Multinomial Logistic Regression

- Asian
- Age
- Deprivation
- PA Moderate

- Smoker
- Depression
- Hypertension
- Heart Attack
Conclusions
TDA is an useful algorithm to visualize and understand high dimensional datasets, and to find clusters in data.

The results suggest the existence of subgroups of T2DM patients with unique clinical, sociodemographic, and behavioural characteristics. This can be useful to target different type of treatments.
Many thanks Sponsors!!
References


