Modelling Cardiovascular Condition Evolution in Hypertensive Population Using Graph Signal Processing

Antonio G. Marques[1], Cristina Soguero-Ruiz[1], Javier Ramos[1], Inmaculada Mora-Jiménez[1],
Rebeca Goya-Esteban[1], Rafael García-Carretero[1,2], Óscar Barquero-Pérez[1]

[1] Rey Juan Carlos University, Madrid, Spain
[2] Móstoles University Hospital, Madrid, Spain

Abstract

Graph signal processing (SP) is a new discipline that interprets data as a collection of signals defined on top of a graph. The nodes of the graph correspond to variables (features), with the links between nodes describing pairwise relationships between the different variables. Graph signals are useful in several interesting fields, including medicine and health care. Our aim in this paper is to use graph SP to model and analyze clinical records of a chronic disease such as essential hypertension in a population. The ultimate goal is to identify prognostic factors and to assess the predictive value of features among the participants. Electronic clinical records of 1664 hypertensive patients were collected. The initial cohort was split into two groups: one group with patients with an incident cardiovascular (CV) event, and another group with patients without CV event. Clinical and analytic features were assessed, such as body mass index, blood pressure, cholesterol, albuminuria, and kidney function. By performing graph SP techniques, we provided a better understanding of pairwise interactions, correlation between features and conditional independence among them, which may help caregivers in designing an appropriate medical management in patients with chronic diseases such as essential hypertension, obesity and diabetes.

1. Introduction

Motivated by the desire desire to analyze and process heterogeneous data supported on irregular domains, there has been a growing interest in broadening the scope of traditional signal processing (SP) techniques to signals defined on graphs [1, 2]. Noteworthy representatives include sampling and reconstruction of graph signals, linear graph filtering, the graph Fourier transform (GFT) and graph topology inference [2–4].

This paper is a first attempt to use those graph SP tools to model and analyze clinical records of chronic hypertensive patients and their risk for a cardiovascular (CV) event. Our objective was to consider each patient's record as a graph signal that varies across time. The nodes of the graph that support the signal correspond to the registered variables (record) of a patient. The links (edges) describe pairwise interactions –such as correlation or conditional independence [5, Ch. 7]– among those variables. Our ultimate goal was to provide a better understanding of our clinical dataset, which can be subsequently leveraged to design enhanced preprocessing schemes such as sampling, denoising, or feature extraction, among others.

The key is to interpret the patients' records as signals defined on graphs learned from the own database. Using this interpretation, we can quantify properties associated with those graph signals for different types of graphs. The purpose is twofold: a) to assess how useful the graphs are to describe the data and b) to unveil hidden structures that can help to understand and pre-process our dataset. More specifically, the preliminary results using our dataset show that the graphs built using data from individuals that suffered a CV event are indeed different from those who had not. Moreover, we also observed that the graphs representing a particular set of individuals change with time. These findings support the idea that graph-based representations can be exploited for prognostic purposes as well as to track the evolution of the patients over time.

The fundamental concepts of graph SP are reviewed in Section 2, while the database and pre-processing are described in Section 3. Section 4 explains how to apply the tools in Section 2 to our dataset and discusses results. Conclusions are stated in Section 5.

2. Mathematical foundations

Consider an undirected graph $\mathcal{G}$ with a set of N nodes or vertices $\mathcal{N}$ and a set of links $\mathcal{E}$, such that if node n is connected to m, then both (n, m) and (m, n) belong to $\mathcal{E}$. The neighborhood of n is defined as the set of nodes $\mathcal{N}_n = \{m \,|\, (m, n) \in \mathcal{E}\}$ connected to n. For any given graph we define the adjacency matrix $\mathbf{A}$ as a sparse $N \times N$ matrix

ISSN: 2325-887X DOI:10.22489/CinC.2017.238-354

with non-zero elements $A_{m,n}$ if and only if $(m, n) \in \mathcal{E}$. The value of A_{mn} captures the strength of the connection between m and n.[1]

2.1. Graph Signal Processing

The interest in graph SP is on analyzing not only $\mathcal{G}$, but also graph signals defined on the nodes of $\mathcal{G}$. Formally, each of these signals can be represented as a vector $\mathbf{x} = [x_1, ..., x_N]^\top \in \mathbb{R}^{\mathrm{N}}$ where the n-th element represents the value of the signal at node n. Since this vectorial representation does not convey explicitly the topology of $\mathcal{G}$, the graph is endowed with a sparse *graph-shift operator* (GSO) that captures the local structure of $\mathcal{G}$. Typical choices for this GSO are the adjacency matrix [2], and the graph Laplacian [1].

Since the graph is undirected, the GSO $\mathbf{S}$ is symmetric and therefore admits the eigendecomposition $\mathbf{S} = \mathbf{V}\,\mathbf{\Lambda}\,\mathbf{V}^\top$, where $\mathbf{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$ is a diagonal collecting the N eigenvalues, and $\mathbf{V}$ is a unitary matrix which columns correspond to the different eigenvectors. Matrix $\mathbf{V}$ is critical to generalize the notion of Fourier transform to the graph domain [1,2] . Moreover, when the graph $\mathcal{G}$ is built using either correlation or conditional independence metrics [5, Ch. 7], $\mathbf{V}$ can be shown equivalent to the discrete Karhunen–Loève transform [4].

2.2. Smooth graph signals

If we have *a priori* information indicating that our signal exhibits certain properties or belongs to a particular subclass, that structural information must be exploited to preprocess the signal to, e.g., reduce noise or enhance the most informative parts, leading to a better postprocessing performance. One of the benefits of graph SP is that it provides new graph-based models to describe the signals of interest. More specifically, a number of metrics to quantify how well a given signal matches its supporting graph have been defined. The existing definitions range from spectral-based metrics that exploit the notion of frequency in $\mathbf{V}$ [2], to node-based metrics which define the smoothness of a graph signal using either the Laplacian or the adjacency matrix of the graph.

To elaborate a bit more on smoothness-based metrics, suppose first that $\mathbf{x}$ is a *time varying* signal. Then, we say that signal $\mathbf{x}$ is smooth if the distance between the signal and its shifted version is small; that is, if the total variation $\mathrm{tv}(\mathbf{x}) = \sum_{\mathrm{t}=1}^{\mathrm{N}}(\mathrm{x_t} - \mathrm{x_{t-1}})^2$ is small. The previous definition uses the norm-2 as a distance metric; see, e.g., [6] for alternative definitions and related discussions. Motivated by this and assuming that $\mathbf{S} = \mathbf{A}$, several graph SP works proposed smoothness (total variation) metrics for graph signals that tried to generalize the one in time domain by taking into account the local structure encoded in $\mathbf{A}$ [1,2]. Two of the most popular are $\mathrm{tv}_{\mathbf{A},1}(\mathbf{x}) = \mathbf{x}^\top(\mathbf{I} + \mathbf{A})^{-1}\mathbf{x}$ and $\mathrm{tv}_{\mathbf{A},2}(\mathbf{x}) = \mathbf{x}^\top\,(\mathbf{I} - \alpha\,\mathbf{A})^2\,\mathbf{x} = \|\mathbf{x} - \alpha\mathbf{A}\mathbf{x}\|_2^2 = \sum_{\mathrm{n}}\left(\mathrm{x_n} - \alpha\sum_{\mathrm{m}\in\mathcal{N}_{\mathrm{n}}}\mathrm{A_{nm}x_m}\right)^2$, where α is a normalizing constant set to the spectral radius of $\mathbf{A}$ [2].

2.3. Building the graph

In applications such as smart grids or transportation networks, the graph $\mathcal{G}$ is given by an actual network where data is observed. In most applications, however, the graph $\mathcal{G}$ must be learned from the data itself. Although the literature on graph topology inference is extensive [5, Ch. 7], we describe next the two simplest and most widespread methods: correlation networks and conditional independence networks. In both cases, the graph $\mathcal{G}$ is considered not to have self loops, so that $A_{n,n} = 0\ \forall\ n$.

- In correlation networks, the edge (n, m) exists if the pairwise correlation $C_{nm} = \mathbb{E}[x_n x_m]$ is above a given threshold η. Mathematically, this implies that $A_{nm} = C_{nm}\,\mathbb{I}_{\{|C_{n,m}|\geq\eta\}}$.
- In conditional independence networks, we first compute the precision matrix $\mathbf{P} = \mathbf{C}^\dagger$ where † denotes pseudoinverse and $\mathbf{C} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{N\times N}$ is the covariance matrix. Then, we decide that the edge (n, m) exists if the entry P_{nm} is above a threshold η. This is equivalent to setting the adjacency matrix of the graph as $A_{nm} = P_{nm}\mathbb{I}_{\{|P_{nm}|\geq\eta\}}$. Intuitively, P_{nm} represents the importance of x_m to predict x_n assuming that all other variables are known [5, Ch. 7]. As a result, if $P_{nm} = 0$ we say that x_n and x_m are "conditionally independent"; see [5, Ch. 7].

While due to the space limitations the focus in this paper will be on correlation networks (which are the most popular ones), our future research will compare both methods.

3. Data set description

We collected data from 3,473 patients from Móstoles University Hospital's Hypertension Unit between 2006 and 2016. Patients with less than three follow-up appointments or with prevalent CV disease were excluded, resulting 1,664 patients. Demographic, clinical and biochemical variables (features) were collected in different appointments (each appointment every six months). Table 1 shows features baseline statistics.

We were interested in knowing the predictive value of the features and determine their relevance when assessing

[1] *Notation:* Generically, the entries of a matrix $\mathbf{X}$ and a (column) vector $\mathbf{x}$ will be denoted as X_{ij} and x_i. The superscripts $^\top$ and † stand for transpose and pseudoinverse, respectively; $\mathbf{0}$ is the all-zero vector and $\mathbf{1}$ is the all-one vector; and the ℓ_0 pseudo norm $\|\mathbf{x}\|_0$ is defined as the number of nonzero entries in $\mathbf{x}$. For a vector $\mathbf{x}$, diag($\mathbf{x}$) is a diagonal matrix with the (i, i)th entry equal to x_i. For a Boolean statement b, the indicator function $\mathbb{I}_{\{b\}}$ yields one if b is true and zero otherwise. The expectation operator is denoted as $\mathbb{E}[\cdot]$.

Table 1. Baseline features of our hypertensive patients.

	Non-CV event patients	**CV event patients**	**Total**
Patients	1507	157	1664
Age (years)	56.0 ± 13.1	65.2 ± 11.4	56.8 ± 13.3
Weight (kg)	83.8 ± 16.9	81.2 ± 17.0	83.6 ± 16.9
Height (cm)	162.8 ± 10.2	159.0 ± 10.0	162.5 ± 10.2
BMI	31.5 ± 5.60	32.0 ± 5.60	31.6 ± 5.60
Creatinine (mg/dl)	0.8 ± 0.20	0.9 ± 0.30	0.8 ± 0.20
Cystatin C (mg/dl)	0.8 ± 0.20	0.9 ± 0.30	0.8 ± 0.20
Blood glucose (mg/dl)	124.8 ± 44.3	151.0 ± 71.4	127.3 ± 48.1
HDL cholesterol (mg/dl)	63.9 ± 17.2	65.5 ± 17.8	64.0 ± 17.3
LDL cholesterol (mg/dl)	135.1 ± 32.3	129.7 ± 35.1	134.6 ± 32.6
Systolic BP (mmHg)	140.5 ± 12.3	143.3 ± 13.1	140.7 ± 12.4
Diastolic BP (mmHg)	79.6 ± 8.30	75.9 ± 8.60	79.3 ± 8.40

Data are reported as percentages or median (± interquartile range). BMI: body index mass. BP: blood pressure.

CV risk. In this sense, we considered a CV event as a composite outcome of CV death, incident coronary disease, incident heart failure and cerebrovascular disease.

The dataset was divided into two subsets: $\mathcal{X}_0$, which collects the records of individuals with no CV-event up to the last appointment registered; and $\mathcal{X}_1$, with records of patients collected up to a CV event. In order to provide a first study on the evolution of the health status, the database consisting of 1,664 patients was subsequently filtered to get just the records associated to the first and last appointment, excluding patients with missing data in the first or last appointment. This way, our final cohort included a total of 767 patients, with 66 patients suffering from CV events, and with follow-up of 12.5 years (median 8 years).

4. Results

For each subset $\mathcal{X}_0$ and $\mathcal{X}_1$, we built two correlation-based graphs. The adjacency GSO was investigated to capture the local structure of each $\mathcal{G}$, so a total of four GSO were considered: two were built using just the first record for non-CV and CV event patients (see Fig. 1(a) and (d), respectively) and the other two were constructed using just the last record (see Fig. 1(b) and (e), respectively). Each matrix is symmetric, with size the number of features following the order shown in Table 1. The difference between the adjacency matrices constructed when using the last and the first record is shown in Fig. 1(c) for non-CV patients, and in Fig. 1(f) for CV patients. This setting would allow us to understand the patient evolution along time.

By inspecting the adjacency matrices, we found a positive relationship among creatinine and cystatin C in non-CV-event patients when considering just the last appointment. On the contrary, this relationship was present in the CV-event group both at the beginning and end of the study. Following with the analysis of the group with no CV event, a positive relationship exists between age and cystatine C, which is biologically plausible, higher serum levels of cystatine C are expected as a consequence of natural ageing.

Our hypothesis is that both creatinine and cystatin C play a relevant role in the development of a CV event from the beginning of the observation period regardless of the age. In the non-CV-event group, the natural ageing makes both creatinine and cystatin C levels rose over time, so we only found this relationship at the end of the observation period. However, in the CV-event group, this relationship is present from the beginning, probably due to the loss of physiological homeostasis of those biomarkers, i.e., cystatin C had a significant weight, independently of the age.

Another finding was the positive relationships between creatinine and glucose, and between cystatin C and glucose in the CV-event group when just the last record was considered, pointing out that any hyperglycemia stage put patients at a high risk of having a CV event. However, a negative relationship between HDL-cholesterol and BMI, and between creatinine and LDL-cholesterol in the CV-event group when the event is close in time could not be explained but open the way to future research regarding the role of these interactions. It is noteworthy, however, some interpretation based on the above findings. Although it is well known that high LDL- and low HDL-cholesterol levels are significant risk factors for CV events, these traditional risk factors were less strong than cystatin C and creatinine serum levels when predicting a CV event.

Simultaneously, we found a negative relationship between systolic blood pressure and height in both groups, but it is remarkable in the CV-event patients. Biologically, this relationship lacks any relevant meaning, and we have posed it as a spurious statistical finding.

The last step was to run a preliminary test to asses the smootheness of the signals on the inferred graphs. To that end, we selected the first record of all non-CV individuals and evaluated the mean smoothness of those signals using $\mathrm{tv}_{\mathbf{A},1}$ for the adjancency matrices (a), (b), (d) and (e) in Fig. 1. The observation is that the smallest value is indeed attained when using the matrix in Fig. 1(a), which is the corresponding to the first record of non-CV patients. The second smallest is for the matrix in Fig. 1(b), which corresponds to the graph build with the last record of non-CV

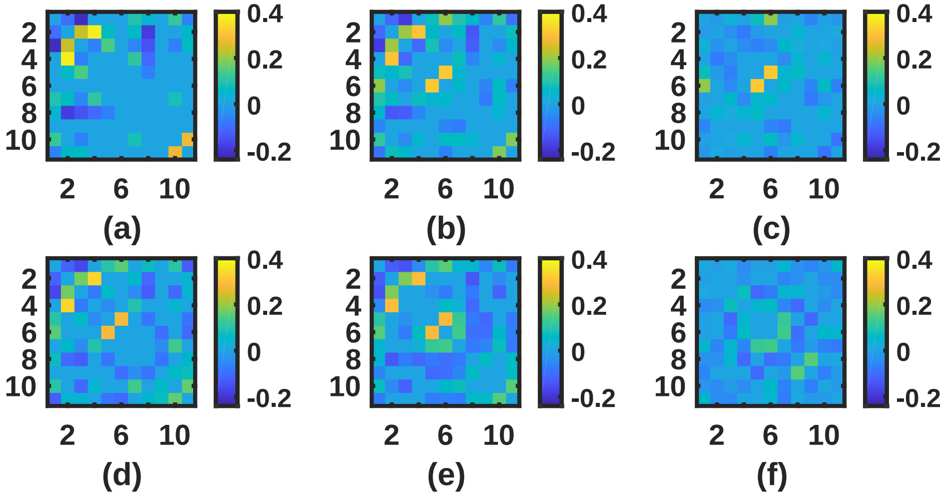

Figure 1. Adjacency matrices built using just the first record for non-CV and CV event patients ((a) and (d)) and using just the last record ((b) and (e)). Panels (c) and (f) show the difference between the adjacency matrices constructed using the last and the first record.

patients. Such a consistent behavior is also observed when evaluating the performance using signals for the last record as well as when selecting those for CV patients.

5. Conclusions

Graph signal processing adds a great advantage in identifying prognostic factors when dealing with fine-scale temporal information. Patients' data were managed taking into account the number of subsequent medical appointments in their follow-up. Each patient had a new appointment every six months, allowing to assess and record every single feature: height, weight, glucose, glycated hemoglobin, and so on. Other techniques are tedious or inefficient when dealing with time series.

By performing graph signal processing techniques, we provided a better understanding of pairwise correlation between features along the evaluation period for both non CV and CV event patients.

Consequently, in our cohort, which included patients with a chronic disease, that is, essential hypertension and obesity, we can claim the strong relationships of both creatinine and cystatin C with a CV event. If those biomarkers arise as relevant prognostic factors in the development of an acute CV condition, they may be used to predict CV events, in agreement with a recent publication[7].

These findings may help physicians to design appropriate diagnostic methods and pharmacological treatment when dealing with patients with certain characteristics.

Acknowledgements

This work has been partly funded by Research Projects TEC2016-75361-R, and TEC2016-75161-C2-1-R from the Spanish Government, and Research Project DTS17/00158 from Carlos III Institute (Spain).

References

[1] Shuman D, Narang S, Frossard P, Ortega A, Vandergheynst P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Process Magazine Mar. 2013; 30(3):83–98.

[2] Sandryhaila A, Moura J. Discrete signal processing on graphs. IEEE Trans Signal Process Apr. 2013;61(7):1644–1656. ISSN 1053-587X.

[3] Segarra S, Marques AG, Leus G, Ribeiro A. Interpolation of graph signals using shift-invariant graph filters. In IEEE European Signal Process. Conf. (EUSIPCO). Nice, France, Aug. 2015; 210–214.

[4] Segarra S, Marques AG, Mateos G, Ribeiro A. Network topology inference from spectral templates. IEEE Transactions on Signal and Information Processing over Networks 2017;467–483.

[5] Kolaczyk ED. Statistical Analysis of Network Data: Methods and Models. Springer, New York, NY, 2009.

[6] Mallat S. A wavelet tour of signal processing. Academic press, 1999.

[7] García-Carretero R, Vigil-Medina L, Barquero-Pérez O, Goya-Esteban R, Mora-Jiménez I, Soguero-Ruiz C, , Ramos-López J. Cystatin c as a predictor of cardiovascular outcomes in a hypertensive population. Journal of Human Hypertension 2017 [Article in press];.

Address for correspondence:

Óscar Barquero Pérez
Department of Signal Theory and Communications
Rey Juan Carlos University
Camino del Molino s/n 28943 - Fuenlabrada (Madrid)
oscar.barquero@urjc.es