Online Graph Topology Inference with Kernels

Mircea Moscu, Ricardo Borsoi, Cédric Richard

Lagrange Seminar

June 30th 2020

Structure

Motivation

- data are abundant and diverse, and are often supported by irregular domains that can be naturally modeled as graphs
- most graph signal processing algorithms assume prior knowledge of the graph structure
- examples where topology needs to be inferred from data include brain networks, gene regulation systems [1] or social and economical interactions [2]

Figure 1: Potential brain network. *(Sapien Labs)*

Definitions and notations

Definitions

- graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$
- $\mathcal{N}$ set of $N+1$ nodes
- $\mathcal{E}$ set of edges; if m and n are linked, $(m,n) \in \mathcal{E}$
- *adjacency matrix* $\boldsymbol{A}$ [3, 4]
  - $(N+1) \times (N+1)$ matrix
  - a_{nm} is 0 if $(m,n) \notin \mathcal{E}$, 1 otherwise
  - encodes the underlying graph connectivity
- signal $\boldsymbol{y}(i) \triangleq [y_1(i), \dots, y_{N+1}(i)]^\top$, $i \in \mathbb{N}_+$

Background

Goal

Estimating the graph topology encoded in a (possibly directed) adjacency matrix $\boldsymbol{A}$ from online nodal measurements $\boldsymbol{y}(i) = [y_1(i), \ldots, y_{N+1}(i)]^\top$, $i \in \mathbb{N}_+$ acquired over $\mathcal{G}$

- The dynamic graph signal $\boldsymbol{y}(i)$ can denote, e.g., the electrical activity of different brain-regions [5, 6], or the voltage angle per bus [7]
- The signal at each node $y_n(i)$ influences and is influenced by the signals at the other nodes ($y_m(i)$, $m \in \mathcal{N} \setminus \{n\}$)

Additive signal model

Nonlinear interactions are being reported in many applications (e.g., in brain connectivity [8, 9])

Recent methods have considered models of the form [10]:

$$y_n(i) = \sum_{m \in \mathcal{N} \setminus \{n\}} a_{nm} f_m(y_m(i)) + v_n(i)\,, \tag{1}$$

where

- a_{nm} is the $(n,m)^{\text{th}}$ entry of the graph adjacency matrix $\boldsymbol{A}$
- f_m is a nonlinear function
- $v_n(i)$ represents innovation noise

Optimization problem in the input space

Using model (1), the topology estimation problem can formulated using all available measurements $(y_n(\ell)$ for $\ell \leq i)$ as:

$$\underset{\boldsymbol{a}_n, f_1, \ldots, f_N}{\operatorname{argmin}} \; \frac{1}{2i} \sum_{\ell=1}^{i} \Big\| y_n(\ell) - \sum_{m \in \mathcal{N} \setminus \{n\}} a_{nm} f_m(y_m(\ell)) \Big\|^2 + \vartheta(\boldsymbol{a}_n)$$

$$\text{subject to } a_{nm} \in \{0, 1\}\,, \tag{2}$$

where $\boldsymbol{a}_n$ is the n^{th} row of $\boldsymbol{A}$, and function ϑ is a sparsity promoting regularization (e.g., ℓ_0 or ℓ_1 (semi)-norm).

However, problem (2) is difficult to solve: it is non-convex, and has infinite dimensional decision variables f_m.

Optimization problem in the RKHS

UNIVERSITÉ CÔTE D'AZUR | CNRS | OBSERVATOIRE DE LA CÔTE D'AZUR

To obtain an efficient algorithm without sacrificing representation power, we:

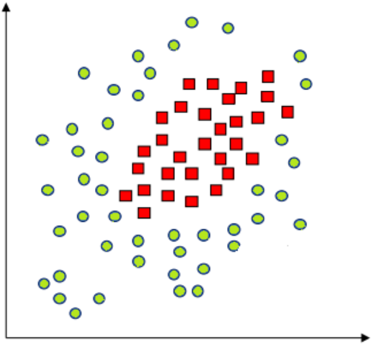
- denote $\phi_{nm} = a_{nm} f_m$, which allows us to incorporate the binary variable a_{nm} and turn (2) into a problem that is quadratic in ϕ_{nm}.
- constrain ϕ_{nm}, for $m \in \mathcal{N} \setminus \{n\}$, to a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_\kappa$ associated with a positive definite reproducing kernel $\kappa(\cdot,\cdot)$.
- Thus, $a_{nm} = 0$ becomes equivalent to $\|\phi_{nm}\|_{\mathcal{H}_\kappa} = 0$.

The optimization problem becomes:

$$\underset{\substack{\phi_{nm} \in \mathcal{H}_\kappa \\ m=1,\ldots,N}}{\operatorname{argmin}} \; \frac{1}{2i} \sum_{\ell=1}^{i} \left\| y_n(\ell) - \sum_{m \in \mathcal{N} \setminus \{n\}} \phi_{nm}(y_m(\ell)) \right\|^2 + \sum_{m \in \mathcal{N} \setminus \{n\}} \psi_{\mathcal{H}_\kappa}(\|\phi_{nm}\|_{\mathcal{H}_\kappa}), \tag{3}$$

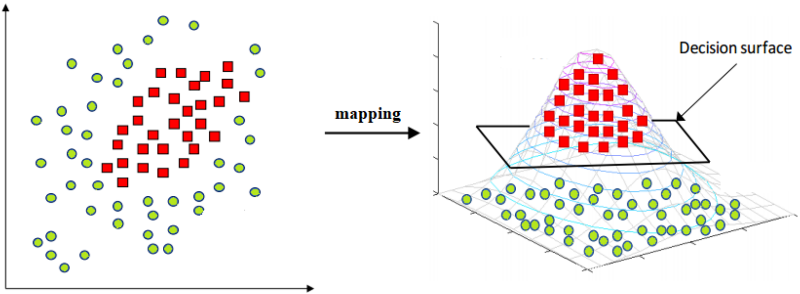
where $\psi_{\mathcal{H}_\kappa} : \mathbb{R} \to [0, \infty[$ is a non-decreasing function.

The *Kernel Trick*



https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-important-98a98db0961d

The *Kernel Trick*



Mapping $\varphi : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}, d_2 \gg d_1, \boldsymbol{x} \in \mathbb{R}^{d_1} \mapsto \varphi(\boldsymbol{x}) \in \mathbb{R}^{d_2}$

$\implies \forall f(\cdot) \in \mathbb{R}^{d_2}, f(\boldsymbol{x}) = \langle f(\cdot), \kappa(\cdot, \boldsymbol{x}) \rangle_{\mathbb{R}^{d_2}}$, with $\kappa(\cdot, \boldsymbol{x}) = \varphi(\boldsymbol{x})$

$\implies \kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \varphi(\boldsymbol{x}_1), \varphi(\boldsymbol{x}_2) \rangle_{\mathbb{R}^{d_2}}$

Examples: $\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-\gamma \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2), \kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \tanh(a\boldsymbol{x}_1^\top \boldsymbol{x}_2 + b)$

---

https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-important-98a98db0961d

Finite dimensional representation

The representer theorem [11] implies that the solution to (3) admits a finite-dimensional representation:

$$\phi^*_{nm}(\cdot) = \sum_{p=1}^{i} \alpha_{nmp}\kappa_m(\cdot, y_m(p)), \quad m = 1, \ldots, N, \quad \alpha_{nmp} \in \mathbb{R} \tag{4}$$

However, the number of coefficients $\{\alpha_{nmp}\}$ increases with i, which is a problem for online processing.

Sparse kernel dictionaries

A solution to this problem is to consider sparse kernel dictionaries $\mathcal{D}_m$:

Kernel dictionary and sparsification rule

- each node m in the network creates, updates, and stores a dictionary of kernel functions, $\mathcal{D}_m = \{\kappa_m(\cdot, y_m(\omega_j)) : \omega_j \in \mathcal{I}_m^i \subset \{1, \ldots, i-1\}\}$
- a candidate kernel function $\kappa_m(\cdot, y_m(i))$ is added in $\mathcal{D}_m$ if the following sparsification condition holds [12]:

$$\max_{\omega_j \in \mathcal{I}_m^i} |\kappa_m(y_m(i), y_m(\omega_j))| \leq \xi_m, \tag{5}$$

  where $\xi_m \in [0, 1[$ determines the level of sparsity and coherence [12]
- the size of the dictionary remains bounded as $i \to \infty$

Introducing sparsity

Since ϕ_{nm} encodes the interaction from node m to node n in model (1), promoting sparsity over $\boldsymbol{A}$ $\Leftrightarrow$ promoting sparsity over the functions ϕ_{nm}.

The coefficient-based representation (4) means that sparsity of groups of variables $\{\alpha_{nm\omega_j}\}_{\omega_j \in \mathcal{I}_m^i}$, $m \in \mathcal{N} \setminus \{n\}$, can be promoted by using a block-sparse regularization [13]:

$$\boldsymbol{\alpha}_n^* = \underset{\boldsymbol{\alpha}_n}{\operatorname{argmin}} \frac{1}{2} \left\| y_n(i) - \boldsymbol{\alpha}_n^\top \tilde{\boldsymbol{k}}(i) \right\|^2 + \eta_n \sum_{m \in \mathcal{N} \setminus \{n\}} \|\tilde{\boldsymbol{\alpha}}_{nm}\|_2, \qquad (6)$$

where we considered the instantaneous MSE estimate (measured only at instant i), with block vectors $\boldsymbol{\alpha}_n$ and $\tilde{\boldsymbol{k}}(i)$ are defined as:

$$\begin{aligned} \tilde{\boldsymbol{k}}(i) &= \left[\boldsymbol{k}_1^\top(i), \ldots, \boldsymbol{k}_N^\top(i)\right]^\top, \quad \boldsymbol{k}_m(i) = \text{col}\{k_m(y_m(i), y_m(\omega_j))\}_{\omega_j \in \mathcal{I}_m^i}, \\ \boldsymbol{\alpha}_n &= \left[\tilde{\boldsymbol{\alpha}}_{n1}^\top, \ldots, \tilde{\boldsymbol{\alpha}}_{nN}^\top\right]^\top, \quad \tilde{\boldsymbol{\alpha}}_{nm} = \text{col}\{\alpha_{nm\omega_j}\}_{\omega_j \in \mathcal{I}_m^i}. \end{aligned} \qquad (7)$$

Algorithm update

Using the subgradient descent algorithm [14] leads to the update:

Update rule

$$\hat{\boldsymbol{\alpha}}_n(i+1) = \hat{\boldsymbol{\alpha}}_n(i) + \mu_n \tilde{\boldsymbol{k}}(i)[y_n(i) - \tilde{\boldsymbol{k}}^\top(i)\hat{\boldsymbol{\alpha}}_n(i)] - \mu_n \eta_n \boldsymbol{\Gamma}_n(i)\,. \tag{8}$$

with $\boldsymbol{\Gamma}_n(i) = [\boldsymbol{\Gamma}_{n1}^\top(i), \ldots, \boldsymbol{\Gamma}_{nN}^\top(i)]^\top$ [14], where each block $\boldsymbol{\Gamma}_{nm}(i)$ is:

$$\boldsymbol{\Gamma}_{nm}(i) = \begin{cases} \dfrac{\tilde{\boldsymbol{\alpha}}_{nm}(i)}{\|\tilde{\boldsymbol{\alpha}}_{mn}(i)\|_2} & \text{if } \|\tilde{\boldsymbol{\alpha}}_{mn}(i)\|_2 \neq 0 \\ \mathbf{0} & \text{if } \|\tilde{\boldsymbol{\alpha}}_{mn}(i)\|_2 = 0 \end{cases}. \tag{9}$$

Edge identification

Set $\hat{a}_{nm}(i)$ to 1 if $\|\hat{\boldsymbol{\alpha}}_{nm}(i)\| \geq \tau_n$, to 0 otherwise

Epilepsy dataset setting

The used data come from a 39-year-old female subject suffering from intractable epilepsy [15]. The data-set contains 8 instances of electrocorticography (ECoG) time series, each instance representing one seizure and contains voltage measurements from 76 different regions on and inside the brain, during:

- the 10 seconds before the epilepsy seizure (*preictal* interval)
- the first 10 seconds during the seizure (*ictal* interval)

Epilepsy dataset results

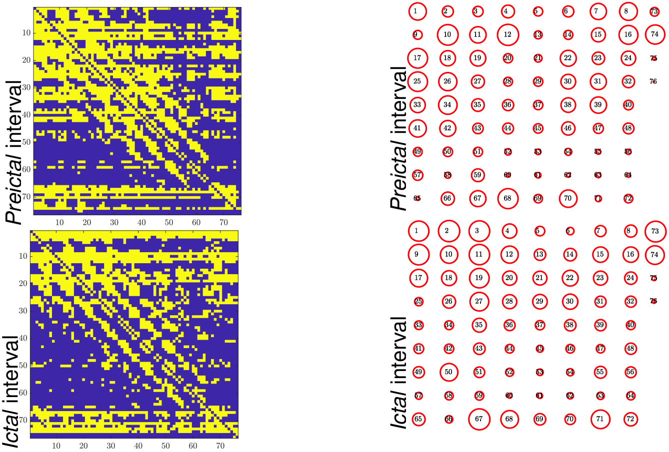


Figure 2: Estimated adjacency matrices (left) and summed in- and out-degrees for the estimated graphs (right). The larger the radius corresponding to node n, the larger the summed degree of node n.

Conclusion and perspectives

Conclusion

- online adaptive graph topology algorithm
- the use of kernels allows for inferring nonlinear relationships
- kernel dictionaries mitigate the increasing number of data points inherently present in an online setting
- consistent results on real data

Perspectives

- the use of multi-kernels
- the use of other sparsity-inducing techniques

References I

[1] The International HapMap Consortium.
A second generation human haplotype map of over 3.1 million SNPs.
Nature, 449:851 EP –, 10 2007.

[2] U.S. Bureau of Economic Analysis.
The use of commodities by industries.
https://apps.bea.gov/iTable/iTable.cfm?reqid=52&step=3&isuri=1.
Accessed: 07/05/2019.

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

[4] N. Biggs.
Algebraic Graph Theory.
Cambridge University Press, 1993.

[5] Mikail Rubinov and Olaf Sporns.
Complex network measures of brain connectivity: uses and interpretations.
Neuroimage, 52(3):1059–1069, 2010.

[6] Yanning Shen, Brian Baingana, and Georgios B Giannakis.
Nonlinear structural vector autoregressive models for inferring effective brain network connectivity.
arXiv preprint arXiv:1610.06551, 2016.

[7] L. Zhang, G. Wang, and G. B. Giannakis.
Going beyond linear dependencies to unveil connectivity of meshed grids.
In *Proc. IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–5, Curaçao, Dutch Antilles, 2017.

[8] W. J. Freeman.
EEG analysis gives model of neuronal template-matching mechanism for sensory search with olfactory bulb.
Biological cybernetics, 35(4):221–234, 1979.

References II

[9] J. A. de Zwart, P. van Gelderen, J. M. Jansma, M. Fukunaga, M. Bianciardi, and J. H. Duyn.
Hemodynamic nonlinearities affect BOLD fMRI response timing and amplitude.
Neuroimage, 47(4):1649–1658, 2009.

[10] Yanning Shen, Brian Baingana, and Georgios B Giannakis.
Kernel-based structural equation models for topology identification of directed networks.
IEEE Transactions on Signal Processing, 65(10):2503–2516, 2017.

[11] B. Schölkopf, R. Herbrich, and A. J. Smola.
A generalized representer theorem.
In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

[12] C. Richard, J.-C. M. Bermudez, and P. Honeine.
Online prediction of time series data with kernels.
IEEE Transactions on Signal Processing, 57(3):1058–1067, 2009.

[13] M. Yuan and Y. Lin.
Model selection and estimation in regression with grouped variables.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67, 2006.

[14] D. Jin, J. Chen, C. Richard, and J. Chen.
Adaptive parameters adjustement for group reweighted zero-attracting LMS.
In *Acoustics, Speech and Signal Processing (ICASSP), Proc. 2018 IEEE International Conference on*, 2017.

[15] M. Kramer, E. D. Kolaczyk, and H. Kirsch.
Emergent network topology at seizure onset in humans.
Epilepsy research, 79:173–86, 2008.

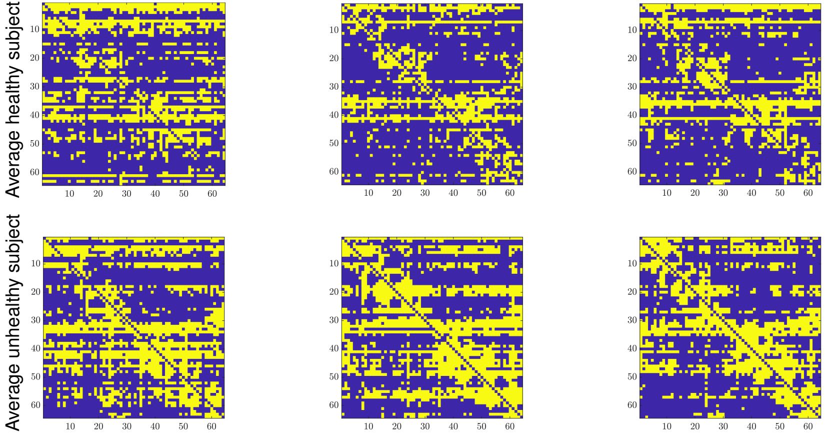
[16] J. M. Ford, V. A. Palzes, B. J. Roach, and D. H. Mathalon.
Did I do that? Abnormal predictive processes in schizophrenia when button pressing to deliver a tone.
Schizophrenia bulletin, 40(4):804–812, 2013.

Schizophrenia dataset setting

We used electroencephalography (EEG) measurements [16] taken from a group of six subjects, half of which are healthy and half suffer from schizophrenia. A simple button-pressing task is set up, in three separate settings where subjects either:

1. Task 1: pressed the button and a tone was immediately played
2. Task 2: listened to the tone without the button press
3. Task 3: pressed the button and the tone was not played

Schizophrenia dataset results



(a) Task 1 (b) Task 2 (c) Task 3

Figure 3: Estimated topologies per task, averaged per group.