Offline detection of change-points in the mean for stationary graph signals

Alejandro de la Concha   **Nicolas Vayatis**   **Argyris Kalogeratos**

Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, France

Abstract

This paper addresses the problem of *segmenting a stream of graph signals*: we aim to detect changes in the mean of a multivariate signal defined over the nodes of a known graph. We propose an offline method that relies on the concept of *graph signal stationarity* and allows the convenient translation of the problem from the original vertex domain to the spectral domain (Graph Fourier Transform), where it is much easier to solve. Although the obtained spectral representation is sparse in real applications, to the best of our knowledge this property has not been sufficiently exploited in the existing related literature. Our change-point detection method adopts a model selection approach that takes into account the sparsity of the spectral representation and determines automatically the number of change-points. Our detector comes with a proof of a non-asymptotic oracle inequality. Numerical experiments demonstrate the performance of the proposed method.

1   Introduction

One of the most common tasks in Signal Processing is segmentation. Identifying time intervals where a signal is 'homogeneous' is a way to uncover latent features of its source. The signal segmentation problem can be restated as a change-point detection task: delimiting a segment can be achieved by spotting the timestamps where it starts and ends. This subject has being extensively investigated leading to a vast literature and applications in many domains including computer science, finance, medicine, geology, meteorology, etc. The

majority of the related work so far focuses on temporal signals (Basseville and Nikiforov, 1993; Balzano et al., 2010; Chen and Gupta, 2012; Tartakovsky et al., 2014; Aminikhanghahi and Cook, 2016; Truong et al., 2020).

In this work we study a different type of object: *graph signals appearing as a stream*. In general terms, a graph signal is a function defined over the nodes of a given graph. Intuitively, the graph partially encodes the variability of the function, i.e. nodes that are connected will take similar values. This applies to real situations, e.g. contacts in social networks would share similar tastes, two neighboring sensors of a sensor network would provide similar measurements, etc. Note that this property can be exploited even when the graph is not given. In some applications, the graph itself has to be inferred and most algorithms are built over this property of local similarity of nodes' behavior that corresponds to signal *smoothness* (Kalofolias, 2016; Le Bars et al., 2019). This can be seen also in graphical models or graphs used to approximate manifolds (Perraudin and Vandergheynst, 2017; Friedman et al., 2007; Le Bars et al., 2020; Tenenbaum, 2000).

Despite the plethora of change-point detection methods in the literature, the development of detectors specifically for graph signals is still limited (Balzano et al., 2010; Angelosante and Giannakis, 2011; Chen et al., 2018; Le Bars et al., 2020). Most existing methods do not take into account the interplay between the signal and the graph structure. The main contribution of this article is in exactly this direction: an offline change-point detector aiming to spot jumps in the mean of a Stream of Graph Signals (SGS). Our method leverages many of the techniques developed in Graph Signal Processing, a relatively new field aiming to generalize the tools commonly used in classical Signal Processing (Shuman et al., 2013; Ortega et al., 2018). More specifically, our approach depends on the concept of Graph Fourier Transform (GFT) that, similarly to the usual Fourier Transform, induces a spectral domain and a sparse representation of the signal. The main idea behind our approach is to translate the problem from the vertex domain to the spectral domain, and design a change-point detector operating in

this space that accounts for the sparsity of the data and automatically infers the number of change-points. This is achieved by adding two penalization terms: one aiming to recover the sparsity, and another one penalizing models with a large number of change-points. The performance of the algorithm and the design of this penalization terms are based on the framework introduced in Birgé and Massart (2001) and the innovative perspective of the ℓ_1 norm analyzed in Massart and Meynet (2011).

The paper is organized as follows. Sec. 2 presents the basic definitions and tools that are used later on. Sec. 3 formulates the change-point detection problem in the context of graph signals, and then presents a Lasso-based algorithm and a Variable Selection-based algorithm. Sec. 4 provides theoretical guarantees for these algorithms and, finally, in Sec. 5 our overall approach is tested through experiments on synthetic signals generated over real and random graphs.

2 Basic concepts and notations

In this section we introduce notations and key concepts. Let A_i and $A^{(j)}$ denote the i-th row and j-th column of matrix A, respectively, and $A_{i,j}$ be a specific element in A. Moreover, $A^\top$ and A^* stand for the transpose and the conjugate transpose (i.e. transpose with negative imaginary part) of matrix A. $x^{(i)}$ denotes the i-th entry of vector x, x_t represents the observed vector x at time t, and $\tilde{x}$ stands for the GFT of x, which is introduced in Definition 3. A graph is defined by an ordered tuple $G = (V, E)$, where V and E stand for the set of vertices and edges respectively, and $p := |V|$ is the number of graph nodes.

Definition 1. *A **graph signal** is a tuple (G, y), where $G = (V, E)$ and y is a function $y : V \to \mathbb{R}$.*

Definition 2. *A **graph shift operator** (GSO) S associated with a graph $G = (V, E)$, is a $p \times p$ matrix whose entry $S_{i,j} \neq 0$ iff $i = j$ or $(i, j) \in E$, and it admits an eigenvector decomposition $S = U\Theta U^*$.*

Definition 3. *For a given GSO $S = U\Theta U^*$ associated with a graph G, the **Graph Fourier Transform** (GFT) of a graph signal $y : V \to \mathbb{R}$ is defined by $\tilde{y} = U^*y$.*

The frequencies of the GFT correspond to the elements of the diagonal matrix Θ, that is $\{\Theta_{i,i}\}_{i=1}^p$. Moreover, the eigenvectors $\{u_i\}_{i=1}^p$ provide an orthogonal basis for the graph signals defined over the graph G. Finally, the GFT is the basic tool that allows us to translate operations from the vertex domain to the spectral domain (Sandryhaila and Moura, 2013).

The *graph signal stationarity* over the vertex domain is a property aiming to formalize the notion that the

graph structure can explain to a large degree the observed inter-dependencies between the dimensions of a graph signal. Henceforth, we refer to stationary with respect to a GSO S. The definitions and the properties that are listed bellow can be found in Marques et al. (2017) and in Perraudin and Vandergheynst (2017).

Definition 4. ***Stationarity w.r.t. the vertex domain:*** *Given a normal GSO S, a zero-mean graph signal $y : V \to R$ with covariance matrix Σ_y is stationary with respect to the vertex domain encoded by S, iff Σ_y and S are simultaneously diagonalizable, i.e. $\Sigma_y = U \operatorname{diag}(P_y)U^\top$. The vector $P_y \in \mathbb{R}^p$ is known as the **graph power spectral density** (PSD).*

The next two properties are used to derive our change-point detection algorithm, to perform the estimation of P_y, and to generate synthetic experimental scenarios.

Property 1. *Let y be a stationary graph signal w.r.t. to S, then $\tilde{y} = U^*y$, which means that the GFT of y will have a covariance matrix $\Sigma_{\tilde{y}} = \operatorname{diag}(P_y)$.*

Property 2. *Let y be a stationary graph signal with covariance matrix Σ_y and PSD P_y. The output of a graph filter $H = U \operatorname{diag}(h)U^*$, with a frequency response h, applied to the graph signal is $z = Hy$ and has the following properties:*
1. It's stationary on S with covariance $\Sigma_z = H\Sigma_y H^*$.
2. $P_z^{(i)} = |h^{(i)}|^2 P_y^{(i)}$.

3 Change-point detection for a stream of graph signals

3.1 Problem formulation

Suppose a stream of graph signals (SGS) Y observed over the same graph $G = (E, V, S)$. Let $Y = \{y_t\}_{t=0}^T$, where $\forall t : y_t \in \mathbb{R}^p$, and also $\mu_t = \mathbb{E}[y_t]$ is its unknown mean value. These expected values are the rows of matrix $\mu \in \mathbb{R}^{T \times p}$. We suppose that there is an unknown ordered set $\tau = \{\tau_0 = 0, ..., \tau_d = T\} \subset \{0, ..., T\}$ indicating $d + 1$ *change-points*, which segments the time-series. Our hypothesis is that the expected values in each of the segments induced by τ remains constant. Our goal is to infer the set of change-points τ and the μ that will be an element of the space:

$$F_\tau = \{\mu \in \mathbb{R}^{T \times p} \mid \mu_{\tau_{l-1}+1} = ... = \mu_{\tau_l}, \forall \tau_l \in \tau \backslash \{0\}\}. \ (1)$$

The problem is illustrated in Fig. 1 through an example where we can identify 4 different segments, i.e. $d + 1 = |\tau| = 5$ change-points.

The general idea of our approach is to take advantage of the advances in Graph Signal Processing and model selection fields in order to recover automatically the
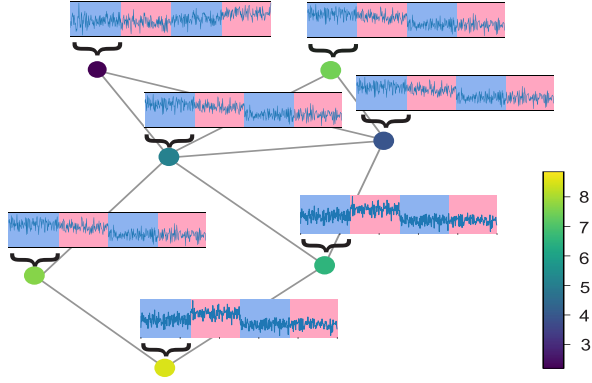
Figure 1: Example stream of graph signals (SGS) with four change-points in the mean (according our problem formulation, the end of the sequence is always a change-point). Successive segments are shown with different colors. The color of the graph nodes represents the mean of the signal during the first observed segment. The signal observed at each node evolves through time, as shown in the line plots next to them. At some timestamps, the mean of the graph signals exhibits a change in a subset of nodes, which also signifies changes in the spectral representation of the signals.

number of change-points as well as sparse representations of the $\{\mu_{\tau_l}\}_{l\in\{1,...,d\}}$ in terms of the GFT. This allows not only to segment the signal, but also to characterize the nature of the change. Such an example is later illustrated in Fig. 2.

We make the following hypotheses over the SGS:

1. The graph signals are independent with respect to the temporal domain.

2. The graph signals follow a multivariate Gaussian distribution sharing the same expectation parameter if they belong to the same segment.

3. If $t \in \{\tau_{l-1}+1, \tau_{l-1}+2, ..., \tau_l\}$, then the residual $y_t - \mu_{\tau_l}$ is stationary with respect to the GSO S.

4. The mean of the graph signals admits a sparse representation with respect to the basis defined by the eigenvectors of U. More specifically, there exists $I \subsetneq \{1,...,p\}$ such that for $l \in \{1,...,d\}$ $\tilde{\mu}_{\tau_l}^{(i)} \neq 0$ for all $i \in I$, and $\tilde{\mu}_{\tau_l}^{(i)} = 0$ otherwise.

5. S is a normal matrix with all its eigenvalues different, and S remains constant throughout the observation and up to the time-horizon T.

The sparsity hypothesis is commonly used in the literature and can have interesting applications in our context since it provides insights about the nature of a transition. For example, if we suppose that S is the Laplacian matrix of the graph G, then its first eigenvector (the one corresponding to the smallest eigenvalue) is a graph signal that is constant over the entire graph; then the next k eigenvectors encode informa-

tion about the cluster structure of the data, and k depends on the complexity of this structure; the higher the eigenvalue, the higher the variance of the associated eigenvector (Ortega et al., 2018). The connection with our algorithm can be made as follows: suppose, we have computed the Graph Fourier coefficients of the means $\tilde{\mu}_{\tau_l}, \tilde{\mu}_{\tau_{l+1}} \in \mathbb{R}^p$, for the segments τ_l and τ_{l+1}. Then, if $|\tilde{\mu}_{\tau_l}^{(i)} - \tilde{\mu}_{\tau_{l+1}}^{(i)}|$ is higher around the i-s (i.e. dimensions in the spectral representation) corresponding to low frequencies, the difference implies a change occurring more uniformly across the graph, e.g. a shift of the same magnitude at most of the nodes or affecting the nodes belonging to the same cluster. On the other hand, if $|\tilde{\mu}_{\tau_l}^{(i)} - \tilde{\mu}_{\tau_{l+1}}^{(i)}|$ is concentrated around i-s lying at high frequencies, the shift should show many sign changes across graph edges, indicating -for example- that a set of random nodes (i.e. the formation of the set cannot be explained by the graph, e.g. it is not a cluster of nodes) has modified their mean. The eigenvectors of the adjacency matrix also provide information about the connectivity of the graph as discussed in Ortega et al. (2018) and verified in a theoretical analysis in Athreya et al. (2017).

As we suppose that S does not change over time, the stationarity of the graph signals with respect to the graph implies that the covariance matrix $\Sigma_y = U \operatorname{diag}(P_y)U^*$ remains unchanged too. Then the average log-likelihood of the SGS can be written as:

$$L(\mu, \tau) = -\sum_{l=1}^{d} \sum_{t=\tau_{l-1}+1}^{\tau_l} \frac{(y_t-\mu_{\tau_l})^T \Sigma^{-1}(y_t-\mu_{\tau_l})}{2T} + \frac{\log \det(\Sigma)}{2T}$$

$$= -\sum_{l=1}^{d} \sum_{t=\tau_{l-1}+1}^{\tau_l} \sum_{i=1}^{p} \frac{(\tilde{y}_t^{(i)}-\tilde{\mu}_{\tau_l}^{(i)})^2}{2TP_y^{(i)}} + \frac{\log P_y^{(i)}}{2T}.$$
(2)

This formulation can be seen as a way to *translate the signal from the vertex domain to the spectral domain* where the sample becomes independent to the graph structure (see Property 1). It is true that we could decorrelate the data via other techniques such as PCA or SVD, however this way we would lose the interpretability of the results with respect to the graph.

Based on Eq. 2, we can define the following ℓ_1-penalized least squares cost function:

$$C_T(\tau, \tilde{\mu}, \tilde{Y}) = \sum_{l=1}^{d} \sum_{t=\tau_{l-1}+1}^{\tau_l} \sum_{i=1}^{p} \frac{(\tilde{y}_t^{(i)} - \tilde{\mu}_{\tau_l}^{(i)})^2}{TP_y^{(i)}}$$
$$+ \sum_{l=1}^{d} \lambda_l \frac{\sum_{i=1}^{p} I_l |\tilde{\mu}_{\tau_l}^{(i)}|}{T},$$
(3)

where λ_l is the penalization constant leading to the desired sparsity of the GFT that a priori is segment-specific, $I_l = \tau_l - \tau_{l-1}$ is the length of the l-th segment.

The hypothesis that the data follow a multivariate Gaussian distribution is actually only required to derive our theoretical results. However, looking at Eq. 3 we can see that it just involves a least squared error (LSE) term standardized by the PSD, and therefore there is no requirement for Gaussian data. We also verify empirically in our experiments (Sec. 5) that we can still get good performance in practice even when this data property does not hold.

3.2 Penalized cost function for an SGS with sparse GFT representation.

Lasso-based GS change-point detector (LGS). Many graph signals observed in real applications can be accurately approximated by a subset of Graph Fourier frequencies, it is necessary to further exploit this feature when computing the means of each segment (Perraudin and Vandergheynst, 2017; Marques et al., 2017; Huang et al., 2016). This justifies adding an ℓ_1-penalization term to Eq. 2. Furthermore, to address the issue of the unknown number of change-points we also add the penalization term $pen(d)$.

The overall optimization problem for the change-point detection is written as:

$$
(\hat{d}, \{\hat{\tau}_0, \hat{\tau}_1, ..., \hat{\tau}_{\hat{d}}\}, \{\hat{\tilde{\mu}}_1, ..., \hat{\tilde{\mu}}_{\hat{d}}\})
$$
$$
:= \underset{\substack{d \in \{1,...,T\} \\ \tau \in \mathcal{T}^d \\ \tilde{\mu} \in F_\tau}}{\arg\min} \ C_T(\tau, \tilde{\mu}, \tilde{Y}) + pen(d)
$$
$$
= \underset{\substack{d \in \{1,...,T\} \\ \tau \in \mathcal{T}^d}}{\arg\min} \sum_{l=1}^d \underset{\hat{\mu} \in F_\tau}{\arg\min} \left[\sum_{t=\tau_{l-1}+1}^{\tau_l} \sum_{i=1}^p \frac{(\tilde{y}_t^{(i)} - \tilde{\mu}_{\tau_l}^{(i)})^2}{TP_y^{(i)}} \right.
$$
$$
\left. + \lambda_l \frac{\sum_{i=1}^p I_l |\tilde{\mu}_{\tau_l}^{(i)}|}{T} \right] + \frac{d}{T} \left(c_1 + c_2 \log \frac{T}{d} \right),
\tag{4}
$$

where $\mathcal{T}^d$ is the set of all possible segmentations of an SGS of size d.

Problem 4 requires estimating the GFT of the mean of the graph signals that is assumed segment-wise constant. The separability of the cost function implies that, each time, this parameter depends only on the observations belonging to one segment, which is delimited by its change-points. This leads to the closed-form solution for $\bar{\tilde{\mu}}_{\tau_l}^{(i)}$:

$$
\bar{\tilde{\mu}}_{\tau_l}^{(i)} = \text{sign}\left(\bar{\tilde{y}}_{\tau_l}^{(i)}\right) \ * \ \max\left(\left| \bar{\tilde{y}}_{\tau_l}^{(i)} \right| - \frac{\lambda_l P_y^{(i)}}{2}, 0 \right), \tag{5}
$$

where $\bar{\tilde{y}}_{\tau_l}^{(i)} = \frac{1}{I_l} \sum_{t=\tau_{l-1}+1}^{\tau_l} \tilde{y}_t^{(i)}$. Thanks to this formulation, we can make use of dynamic programming in Alg. 1 to find the precise change-points.

Even though Problem 4 is computationally easy to solve, it still requires to set the λ_l parameter, related with the sparsity of the graph signals, and a penalization term $pen(d)$ that would allow us to infer the number of change-points. This problem is not trivial since the number of possible solutions depends on the time-horizon T and the number of nodes p; this feature hinders an asymptotic analysis. Following the model selection approach, we can obtain an oracle-type inequality for the estimators $\hat{\tau}(\hat{d})$ and $\hat{\tilde{\mu}}_{\hat{\tau}}(\hat{d})$. Nevertheless such analysis allows us to only infer the shape of $pen(d)$ depending on the unknown constants c_1, c_2 and a lower bound for λ_l. In model selection, it is common to infer the penalization parameters by the slope heuristics: an approach that is based on the idea that the minimum amount of penalization before overfitting can be observed, this is called *minimal penalty* Baudry et al. (2012); Arlot (2019). Unfortunately, in our context it is not trivial to define the minimal penalty, hence we propose the VSGS change-point detector.

Variable Selection-based GS change-point detector (VSGS). The idea of the detector (Alg. 2) is to estimate the parameter λ from a grid of candidates Λ, and the equivalents of the penalization constants c_1 and c_2 via the slope heuristic. To this end it defines the auxiliary optimization Problem 6 made of three terms: the first one aims to reduce the LSE, the second encourages sparse solutions based on the elements of Λ, the third one aims to determine the right number of change-points. We denote by $\hat{d}^{\text{LSE}}$, $\hat{\tilde{\mu}}_{\hat{\tau}^{\text{LSE}}}^{\text{LSE}}$, and $\hat{\tau}^{\text{LSE}}$ the solutions to this auxiliary problem:

$$
(\hat{d}^{\text{LSE}}, \hat{\tilde{\mu}}_{\hat{\tau}^{\text{LSE}}}^{\text{LSE}}, \hat{\tau}^{\text{LSE}})
$$
$$
:= \underset{\substack{d \in 1,..,T \\ \tau \in \mathcal{T}^d, \\ \tilde{\mu} \in S_{(D_m, \tau)}}}{\arg\min} \ C_T^{\text{LSE}}(\tau, \hat{\mu}, \hat{Y}) + pen(m, \tau)
$$
$$
= \underset{\substack{d \in 1,..,T \\ \tau \in \mathcal{T}^d, \\ \tilde{\mu} \in S_{(D_m, \tau)}}}{\arg\min} \left\{ \sum_{l=1}^d \left(\sum_{t=\tau_{l-1}+1}^{\tau_l} \sum_{i=1}^p \frac{(\tilde{y}_t^{(i)} - \tilde{\mu}_{\tau_l}^{(i)})^2}{TP_y^{(i)}} \right) \right.
\tag{6}
$$
$$
\left. + K_1 \frac{D_m}{T} + \frac{d}{T} \left(K_2 + K_3 \log \frac{T}{d} \right) \right\}.
$$

Problem 6 should be minimized over sets of the form:

$$
S_{(D_m, \tau)} := \{ \tilde{\mu} \in F_\tau \mid \tilde{\mu}_{\tau_l} \in S_{D_m}, \forall l \in \{1, ..., d\} \}, \tag{7}
$$

where we denote by S_{D_m} the space generated by m specific elements of the standard basis of $\mathbb{R}^p$. In simple terms, we restrict the means defined in each of the segments to be elements of S_{D_m}. The quality of the solutions of Problem 6 is measured by comparing against an oracle inequality in Theorem 2, and is further discussed in Sec. 4.

Algorithm 1: Lasso-based GS change-point detector (LGS)

Input : Y: matrix $\mathbb{R}^{T \times p}$ representing the SGS
w: length of the warming period
$d_{\max}$: maximum number of change-points
U: eigenvectors of the GSO
(c_1, c_2): constants of the penalization $pen(\tau)$
λ: constant of the penalization $pen(\mu_\tau)$

Output: $\hat{\tau}(\hat{d})$: set of change-points
$\hat{\tilde{\mu}}_{\hat{\tau}}(\hat{d})$: matrix $\mathbb{R}^{\hat{d} \times p}$ with rows being the mean of each segment

1 Estimate the GFT of the dataset $\tilde{Y} = YU$
2 Compute an estimation of P_y using w observations
3 **for** $d \in \{1, ..., d_{\max}\}$ **do**
4 $\quad$ Apply dynamic programming to solve:
$$\hat{\tau}(d), \hat{\tilde{\mu}}_{\hat{\tau}}(d) := \underset{\substack{\tau \in \mathcal{T}^d, \\ \tilde{\mu} \in F_\tau}}{\arg\min} \ C_T(\tau, \tilde{\mu}, \tilde{Y})$$
5 **end**
6 Choose:
$$\hat{d} := \underset{d = \{1, ..., d_{\max}\}}{\arg\min} \ C_T(\hat{\tau}(d), \hat{\tilde{\mu}}_{\hat{\tau}}(d), \tilde{Y}) + \frac{d}{T}\left(c_1 + c_2 \log \frac{T}{d}\right) \quad (8)$$

7 Return $\hat{\tau}(\hat{d})$ and $\hat{\tilde{\mu}}_{\hat{\tau}}(\hat{d})$

The full description of this new algorithm can be found in Alg. 2. Note that for each element of Λ a level of sparsity D_{m_λ} is induced (Step 5). Then we use the auxiliary optimization Problem 6 to recover the right number of change-points and the level of sparsity D_{m_λ} (Steps 6-11) . We use this information to recover the right parameter λ. The final solution will be the partition found via the auxiliary optimization problem and the means obtained after applying Eq. 5.

How to estimate P_y. Both algorithms require the knowledge of P_y of the SGS (Step 2 in Alg. 1 and Alg. 2). The empirical covariance estimator of the GFT, $\hat{Y}$, has the disadvantage that its variance scales with the norm of P_y, which is a problem in high dimensions when the distance between change-points is not long enough. There are several algorithms aiming to solve this problem (Marques et al. (2017)). In particular, the estimator proposed by Perraudin and Vandergheynst (2017) requires a smaller number of samples and its computation scales with the number of edges in the graph (sparse in most applications). The idea of the estimator is based on Property 2: once the vertex domain of a stationary graph signal is known, it is possible to use different filters to generate synthetic observations corresponding to different regions of the graph, and then use them to reconstruct the PSD. This enables a good approximation of the PSD even when having a single graph signal as input. This is the method we use in our experiments.

Algorithm 2: Variable Selection-based GS change-point detector (VSGS)

Input : Y: matrix $\mathbb{R}^{T \times p}$ representing the stream of the graph signal
$d_{\max}$: Maximum number of change-points
w: length of the warming period
U: eigenvectors of the GSO
Λ: values of the penalty term $pen(\mu_\tau)$

Output: $\hat{\tau}(\hat{d})$: set of change-points
$\hat{\tilde{\mu}}_{\hat{\tau}}(\hat{d})$: matrix $\mathbb{R}^{\hat{d} \times p}$ with rows being the mean of each segment

1 Estimate the GFT of the dataset $\tilde{Y} = YU$
2 Compute an estimation of P_y using w observations
3 **for** $\lambda \in \Lambda$ **do**
4 $\quad$ Solve the Lasso problem:
$$\tilde{\mu}_{\text{Lasso}} := \underset{\tilde{\mu} \in R^{T \times p}}{\arg\min} \sum_{t=1}^{T} \sum_{i=1}^{p} \frac{\left\| \tilde{y}_t^{(i)} - \tilde{\mu}^i \right\|^2}{T P_y^{(i)}} + \lambda \left\| \tilde{\mu} \right\|_1$$
5 $\quad$ Define $D_{m_\lambda} := \|\tilde{\mu}_{\text{Lasso}}\|_0$
6 $\quad$ **for** $d \in \{1, ..., d_{\max}\}$ **do**
7 $\quad\quad$ Solve the change-point detection problem via dynamic programming
$\hat{\tau}(d, D_{m_\lambda}), \hat{\tilde{\mu}}_{\hat{\tau}}^{`}(d, D_{m_\lambda}) :=$
$$\underset{(\tilde{\mu}, \tau) \in S_{(D_{m_\lambda}, \tau(d))}}{\arg\min} C^{\text{LSE}}(\tilde{\mu}(d, D_{m_\lambda}), \tau(d, D_{m_\lambda}))$$
8 $\quad$ **end**
9 **end**
10 Find K_1, K_2, K_3 using the slope heuristic
11 Solve: $(\hat{\lambda}, \hat{d}) :=$
$$\underset{\substack{\lambda \in \Lambda, \\ d = \{1, ..., d_{\max}\}}}{\arg\min} \ C^{\text{LSE}}(\hat{\tau}(d, D_{m_\lambda}), \hat{\tilde{\mu}}_{\hat{\tau}}^{\text{LSE}}(d, D_{m_\lambda}))$$
$$+ K_1 \frac{D_{m_\lambda}}{T} + \frac{d}{T}\left(K_2 + K_3 \log \frac{T}{d}\right) \quad (9)$$

12 Keeping the segmentation $\hat{\tau}(\hat{d}, \hat{D}_{m_{\hat{\lambda}}})$ and $\hat{\lambda}$ fixed, recover $\hat{\tilde{\mu}}_{\hat{\tau}}(\hat{d})$ via Eq. 5
13 Return $\hat{\tau}(\hat{d}, \hat{D}_{m_{\hat{\lambda}}})$ and $\hat{\tilde{\mu}}_{\hat{\tau}}(\hat{d})$

---

4 Model selection approach

The change-point detection problem for the mean of an SGS can be written as a generalized linear Gaussian model: we will detect the change-points over the GFT of the SGS, that is $\tilde{Y}$ instead of Y; we have standardized $\tilde{Y}$ so that the variance of all the GFT coefficients to be $\epsilon = 1$. Under these conditions, we define an isonormal process $(W(\tilde{\mu}))_{\tilde{\mu} \in \mathbb{R}^{T \times p}}$: $W(\tilde{\mu}) := \frac{\text{tr}(\eta^\top \tilde{\mu})}{T}$, where $\eta \in \mathbb{R}^{T \times p}$ is a matrix whose rows follow a centered multivariate Gaussian distribution with covariance matrix $\mathbb{I}_p$. The generalized Gaussian process related to the SGS can be written as:

$$\tilde{Y}_\epsilon(\tilde{\mu}) = \frac{\text{tr}((\tilde{\mu}^*)^\top \tilde{\mu})}{T} + \epsilon W(\tilde{\mu}). \quad (10)$$

This formulation enables the use techniques from the model selection literature (Massart and Picard, 2003) in order to design the penalized term $pen(d)$ related with the number of change-points, and derive oracle-type inequalities for the performance of the proposed estimators described in Alg. 1 and Alg. 2.

Theorem 1 is an *oracle inequality* that provides insights on how Alg. 1 behaves with respect to the time-horizon T and the graph size p. Furthermore, it gives us a guideline towards choosing λ_l and the number of change-points in order to minimize the penalized mean-squared criteria, which is one of the differences of our work to the change-point detection algorithms analyzed in Lebarbier (2005); Arlot et al. (2019) that are based in model selection too, but they focus on mean-squared criteria.

Theorem 1. *Assume that:*

$$\lambda_l = \lambda \geq \frac{(3\sqrt{2})\epsilon\sqrt{\log p + L}}{T},$$
$$pen(d) = \frac{d}{T}\left(c_1 + c_2 \log \frac{T}{d}\right), \quad (11)$$

where $c_1 \geq 6\sqrt{2}\epsilon^2$, $c_2 \geq 3\sqrt{2}\epsilon^2$, and L is such that $L > \log 2$. Then, there exists an absolute constant $C > 0$ satisfying:

$$\mathbb{E}\left[\frac{\left\|\hat{\tilde{\mu}}_{\hat{\tau}} - \tilde{\mu}^*\right\|_F^2}{T} + \lambda \left\|\hat{\tilde{\mu}}_{\hat{\tau}}\right\|_{[\hat{\tau}]} + pen(\hat{d})\right] \leq$$

$$C(K)\left[\left(\inf_{\tau \in \mathcal{T}}\left(\inf_{\substack{\tilde{\mu} \in F_\tau \\ \|\tilde{\mu}\|_{[\tau]} < +\infty}}\frac{\|\tilde{\mu} - \tilde{\mu}^*\|_F^2}{T} + \lambda \|\tilde{\mu}\|_{[\tau]}\right)\right. \quad (12)$$

$$\left. + pen(d_\tau)\right) + 2\lambda\epsilon + \left(1 + \frac{1}{(e^\gamma - 1)(e - 1)}\right)\epsilon^2\right],$$

where $\|\tilde{\mu}\|_{[\tau]} := \frac{1}{T}\sum_{l=1}^d I_{\tau_l} \|\tilde{\mu}_{\tau_l}\|_1$, $\mathcal{T}$ is the set of all possible segmentations of the SGS, $\gamma = \frac{1}{K}(\sqrt{\log p + L} - \sqrt{\log p + \log 2})$, and $K > 1$ is a given constant.

The proof follows similar arguments to Massart and Meynet (2011) and can be found in the supplementary material. Specifically, it requires first to define the set of models of interest, namely the list of candidate models indexed by the possible segmentations and ℓ_1-balls of length $m\epsilon$, where $m \in \mathbb{N}^*$.

The following lemma is a direct consequence of Corollary 4.3 in Giraud (2015).

Lemma 1. *For any $L > 0$, the estimator obtained by solving Problem 4 with tuning parameter*

$$\lambda = 3\epsilon\sqrt{2(\log p + L)}, \quad (13)$$

fulfills with probability at least $1 - e^{-L}$ the risk bound:

$$\frac{\left\|\hat{\tilde{\mu}}_{\hat{\tau}} - \tilde{\mu}^*\right\|_F^2}{T} \leq \sum_{l=1}^d \sum_{t=\tau_{l-1}+1}^{\tau_l} \inf_{\substack{\tilde{\mu}\neq 0, \\ \tilde{\mu} \in R^p}} \frac{\|\tilde{\mu} - \tilde{\mu}_t^*\|_2^2}{T}$$
$$+ \frac{18\epsilon^2(L + \log p)}{T\Phi(\tilde{\mu})^2} \|\tilde{\mu}\|_0, \quad (14)$$

where $\Phi(\tilde{\mu})$ is broadly known as compatibility constant.

Both results, Theorem 1 and Lemma 1, provide details of the performance of the algorithm, when applied in practice, with respect to λ. The combination of both results suggests that *the value of λ_l should be the same for all the segments* in order to recover the sparsity of the signal. We can see that there is a trade-off between the performance of the estimator and its ability to recover the sparsity of the signal: On one side, we need a low λ value in order to reduce the bias of the estimator (see Ineq. 12), while on the other side we need a high λ value that will allow us to recover the sparsity of the signal with higher probability Ineq. 14.

Theorem 1 provides lower bounds for the values of c_1 and c_2. Nevertheless, in practice, when fixing c_1 and c_2 at these values, the obtained results were not satisfying. Finding the right constants in model selection is a common difficult problem (Arlot, 2019). The slope heuristic recovers the constants using a linear regression of the empirical risk against the elements of a penalization term. However, the curve defined by the cost function including the ℓ_1 term does not tend to remain constant as the number of change-points increases, a feature that is used in the slope heuristics.

In Alg. 2, we replace the ℓ_1-penalization term by a Variable Selection penalization term. For each of the elements of a given set of penalization parameters Λ, we solve a Lasso problem over the whole SGS. This allows us to keep all the relevant frequencies. Then, we solve multiple change-point detection problems for different levels of sparsity. We can deduce the right level of sparsity controlled by K_1, as well as the two constants K_2 and K_3 related with the number of change-points, via the slope heuristic. This approach is validated by Theorem 2 and the experiments in Sec. 5.

Theorem 2. *Let $\hat{\tau}$ and $\hat{\tilde{\mu}}_{\hat{\tau}}$ be solutions to the optimization Problem 6. Then, there exist constants K_1, K_2, K_3 defining the penalty term for all $(m, \tau) \in M$, where $M \subset \{1, ..., p\} \times \mathcal{T}$:*

$$pen(m, \tau) = K_1\frac{D_m}{T} + \frac{d_\tau}{T}\left(K_2 + K_3 \log \frac{T}{d_\tau}\right), \quad (15)$$

there exists a positive constant $C(K)$, and $K > 1$ a

given constant, such that:

$$
\mathbb{E}\left[\frac{\left\|\hat{\tilde{\mu}}_{\hat{\tau}} - \tilde{\mu}^*\right\|_F^2}{T}\right] \le C(K)\left[\inf_{(m,\tau)\in M}\inf_{\tilde{\mu}\in S(D_m,\tau)}\frac{\|\tilde{\mu} - \tilde{\mu}^*\|_F^2}{T}\right.
$$

$$
\left. + pen(m,\tau) + \left(1 + \left(\frac{1}{(e^\gamma - 1)(e - 1)}\right)\right)\epsilon^2\right],
$$

(16)

where $\mathcal{T}$ is the set of all possible segmentations of the SGS, $\gamma = \frac{1}{K}(\sqrt{\log p + L} - \sqrt{\log p + \log 2})$.

The proof of Theorem 2 is a consequence of Theorem 4.18 in Massart and Picard (2003), and is provided in the supplementary material.

5 Numerical experiments

Setup. As mentioned earlier, a key hypothesis in our approach is the stationarity of the graph signals. An alternative definition states that a stationary graph signal is what we get as output after applying a graph filter H, with a frequency response $h(\theta)$, to white noise. This is equivalent to Definition 4 when the GSO is normal and all its eigenvalues are different (Perraudin and Vandergheynst, 2017). In this case we will use the Laplacian of the graph as GSO. We generate our graph signals via this definition. We design three different scenarios to test the capabilities of our method. For each scenario, we generate 100 instances.

Scenario I: We generate Erdős–Rényi (ER) graphs with a fixed link creation probability $p_{\text{ER}} = 0.3$. The spectral profile of the filter is defined by: $h(\theta) \propto \frac{1}{\log(\theta+10)+1}$. The white noise generating the signal is a zero mean uniform distribution with variance 1. We generate change-points via a Poisson distribution with mean value 5 and the distance between them is generated at random via $30 + \epsilon$, where ϵ follows an exponential distribution with mean value equal to 20 (the expected distance is 50). Before the first change-point, the mean of the graph signals is a linear combination of the first 20 eigenvectors of the Laplacian matrix; 20 random coefficients of the GFT are changed right after each change-point, the mean is then this new linear combination of eigenvectors. In all cases the coefficients of the linear combinations were generated uniformly at random in the interval $[-5, 5]$.

Scenario II: The graph structure is generated by a Barabasi-Albert (BA) model in which each incoming node is connected to 4 nodes. The spectral profile of the filter is proportional to the density function of a Gamma distribution, $h(\theta) \propto p_{\Gamma(20,5)}(\theta)$. The white noise generating the signal is a standardized Gaussian distribution. Then, 4 change-points are generated and

the distance between them is generated at random via $30 + \epsilon$, with ϵ following an exponential distribution with mean value equal to 20 (the expected distance is 50). Before the first change-point, the mean of the graph signals is a linear combination of the first 20 eigenvectors of the Laplacian matrix; after the first change, the node with the highest degree and all its neighbors change their mean; after the second change-point the first 5 nodes with the highest degrees modify their mean; after the third change-point, 20 nodes at random of the graph get their mean changed. In all cases, the mean is generated uniformly at random in the interval $[-5, 5]$.

Scenario III: We generate a SGS over the Minnesota Road Network (Rossi and Ahmed, 2015) that contains 2.6k nodes, 3.3k edges, and maximum k-core of 3 (i.e. the spatial network is quasi-planar, see Fig. 2). The spectral profile of the filter is defined by: $h(\theta) \propto \frac{1}{\log(\theta+10)+1}$. The white noise generating the signal is a Student-t distribution with 100 degrees of freedom. We generate 3 change-points and the distance between them is generated at random via $120 + \epsilon$, with ϵ following an exponential distribution with mean value equal to 30 (the expected distance is 150). In the first segment the mean of the graph signals is a linear combination of the first 500 eigenvectors of the Laplacian matrix. Then, the first change is generated in a way that can be explained by the graph structure: a number of random non-overlapping regions are created, each one containing a random node and all nodes in its 5-hop neighborhood that were not selected earlier. In each region, the direction of the signal change is kept same, either positive or negative for all of its nodes, and the intensity of the change is generated for each node uniformly at random in the interval $[1, 5]$. Contrary, the second change does not respect the graph structure: random nodes are chosen and their value of their means is increased uniformly by a value in $[-10, -5] \cup [5, 10]$.

We proceed to the analysis of the performance of our method. Recall that our approach requires the PSD P_y of the signal as a parameter. Alg. 2 is mentioned as VSGS when we use the real value of the P_y and Approx. VSGS when we approximate it. For the later case, we estimate the P_y of the signal with the technique of Perraudin and Vandergheynst (2017) on the first 50 graph signals of the SGS.

We implement the slope-heuristic to recover the parameters K_1, K_2, and K_3: that is we make a linear regression of the cost-functions of the list of models with high complexity against the penalization terms, then we multiply the linear regression's coefficients by -2 as suggested in Arlot et al. (2019).

Table 1: Performance evaluation of the VSGS and Approximate VSGS change-point detectors for the synthetic Scenarios I, II, and III. The mean and (std) of the different evaluation metrics are estimated over 100 generated instances of each scenario (↑/↓ indicate our preference for a higher/lower value for a metric). The Recall and the Precision are evaluated allowing a difference of 10 timestamps between the estimated and the real change-points.

Scenario	#Nodes	Detector	Hausdorff (↓)	Rand (↑)	Recall (↑)	Precision (↑)	F1 (↑)
I	100	VSGS	0.94 (00.24)	0.99 (0.00)	1.00 (0.00)	0.80 (0.15)	0.88 (0.10)
		Approx. VSGS	1.73 (70.88)	0.99 (0.02)	1.00 (0.05)	0.88 (0.13)	0.93 (0.08)
II	100	VSGS	0.84 (00.37)	0.99 (0.01)	1.00 (0.00)	0.98 (0.07)	0.99 (0.04)
		Approx. VSGS	1.57 (07.28)	0.99 (0.01)	1.00 (0.00)	0.98 (0.07)	0.99 (0.04)
I	500	VSGS	0.94 (00.24)	0.99 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		Approx. VSGS	6.29 (17.13)	0.98 (0.04)	0.97 (0.11)	1.00 (0.05)	0.98 (0.09)
II	500	VSGS	10.43 (15.61)	0.96 (0.05)	0.91 (0.15)	1.00 (0.00)	0.94 (0.09)
		Approx. VSGS	12.48 (16.54)	0.96 (0.06)	0.89 (0.16)	1.00 (0.00)	0.93 (0.09)
I	1000	VSGS	0.94 (00.23)	0.99 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		Approx. VSGS	7.36 (23.46)	0.98 (0.07)	0.96 (0.12)	1.00 (0.00)	0.98 (0.09)
II	1000	VSGS	33.02 (17.70)	0.89 (0.06)	0.71 (0.14)	1.00 (0.00)	0.83 (0.09)
		Approx. VSGS	33.81 (17.05)	0.89 (0.05)	0.71 (0.12)	1.00 (0.00)	0.83 (0.08)
III - 5 rand. regions 10 rand. nodes	2642	VSGS	119.48 (64.49)	0.82 (0.09)	0.60 (0.02)	1.00 (0.00)	0.73 (0.13)
		Approx. VSGS	120.81 (63.38)	0.82 (0.09)	0.60 (0.20)	1.00 (0.00)	0.73 (0.13)
III - 10 rand. regions 20 rand. nodes	2642	VSGS	7.54 (29.82)	0.99 (0.04)	0.98 (0.11)	1.00 (0.00)	0.98 (0.07)
		Approx. VSGS	8.85 (32.28)	0.99 (0.04)	0.97 (0.12)	1.00 (0.00)	0.98 (0.08)
III - 20 rand. regions 40 rand. nodes	2642	VSGS	0.72 (00.45)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		Approx. VSGS	0.72 (00.45)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)

A code repository including the implementation of the proposed algorithms and the tested experimental scenarios is publicly available online.[1]

Results. Table 1 summarizes the results of our experiments by reporting several evaluation metrics (Truong et al., 2020). The first thing to realize is that, generally, our method has good performance regardless of the PSD estimation or whether the Gaussian hypothesis holds, that is a the distance with respect to the real change-points (Hausdorff distance) is small given the minimum gap between change-points. Almost all the timestamps get correctly classified as whether they are change-points or not (Rand Index → 1), the majority of the change-points are recovered (Recall → 1) and not too many spurious change-points are wrongly identified (Precision → 1). The effect of the PSD estimation can be seen in the Hausdorff distance, which tends to increase in mean and variance.

Fig. 2 shows one instance of Scenario III, namely when the number of affected regions is 10 and the number of random nodes modifying their mean is 20. After the first change-point, as expected, it can be seen how the shift in the mean related to the structure of the graph makes $|\hat{\mu}_{\tau_l}^{(i)} - \hat{\mu}_{\tau_{l+1}}^{(i)}|$ higher in low frequencies. In this particular example, our algorithm was able to recover the right number of change-points and a good approximation of the means for each segment in both domains.

The detailed evaluation metrics that are reported in Tab. 1 confirm the quality of the obtained solutions. However, the intensity of the change-point is impor-

tant for our algorithms to perform well: when a change is small in terms of the spectral domain, it is essentially harder to recover the respective change-points and the sparse representation. This can be seen in Scenario III where the performance of VSGS improves with the increase of the number nodes that get affected by the change.

6 Conclusion

We presented an offline change-point detection method for shifts in the mean of a stream of graph signals. The proposed approach infers automatically the number of change-points and the level of sparsity of the signal in its Graph Fourier representation. The formulation has the advantage of being easy to resolve via dynamic programming while also allowing interesting theoretical guarantees such as the oracle-type inequality that we provide. The techniques and results of this paper could be generalized to situations where we aim to spot change-points in a stream of multivariate signals that supports a sparse representation in a orthonormal basis. Proving the consistency of the detected change-points and its generalization to types of graph signals representation other than the GFT, such as wavelets, is among our plans for future work.

Acknowledgment

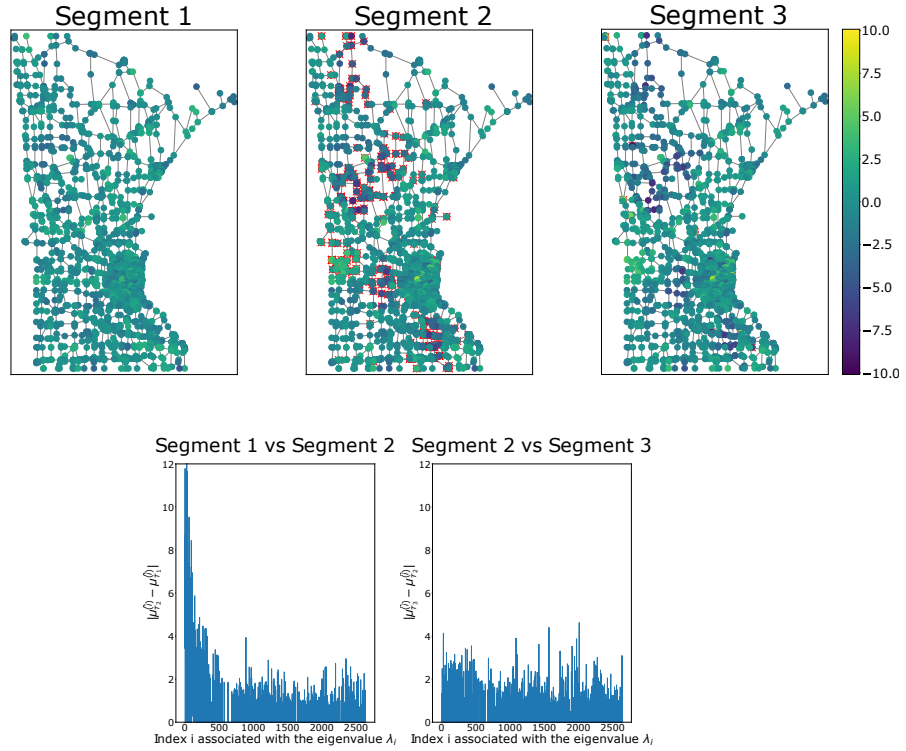[1]https://github.com/AlejandrodelaConcha/VSGS
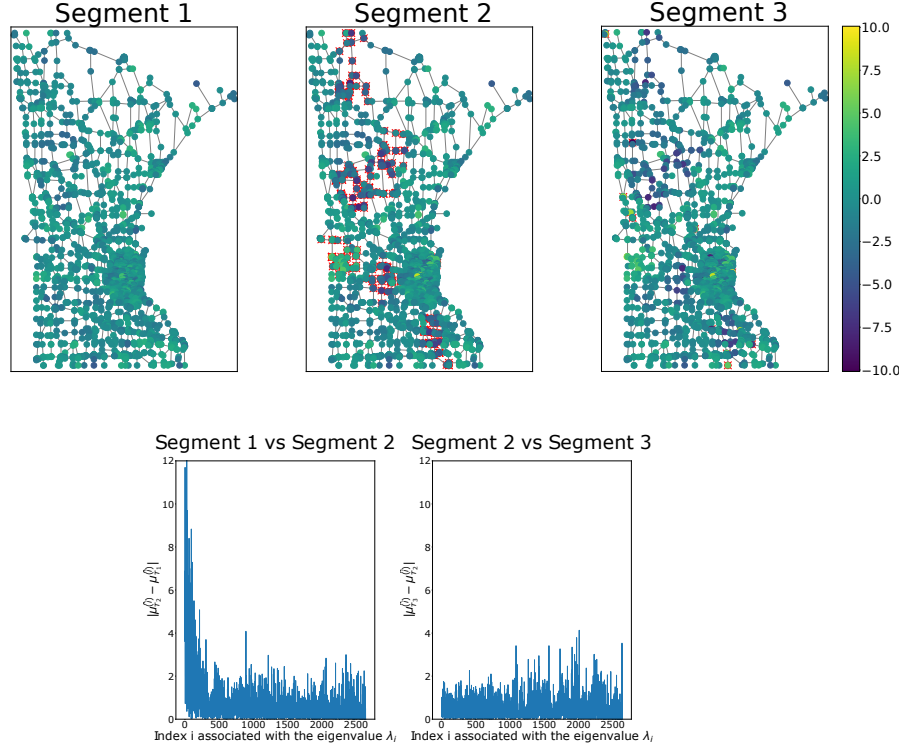
Figure 2: Instance of Scenario III (after the first change-point: 10 regions; after the second change-point: 20 random nodes). The node colors indicate their expected values. **First two rows:** In the first row, the figures show the true mean in each of the 3 segments induced by the 2 change-points. The red contour around a node indicates that its mean has changed compared to the previous segment. In the second row, the two plots show the true difference between the GFT of the mean of two consecutive segments (1st and 2nd, 2nd and 3rd). **Last two rows:** The estimated mean is shown. The red contour around a node indicates that it was detected by our algorithm as having a change in its mean. In the last row, we show the difference between the GFT of the estimated mean of two consecutive segments.

References

Aminikhanghahi, S. and Cook, D. (2016). A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51:339–367.

Angelosante, D. and Giannakis, G. (2011). Sparse graphical modeling of piecewise-stationary time series. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1960–1963.

Arlot, S. (2019). Rejoinder on: Minimal penalties and the slope heuristics: a survey. *J. de la Societe Française de Statistique*, 160(3):158–168.

Arlot, S., Celisse, A., and Harchaoui, Z. (2019). A kernel multiple change-point algorithm via model selection. *J. of Machine Learning Research*, 20(162):1–56.

Athreya, A., Fishkind, D., Levin, K., Lyzinski, V., Park, Y., Qin, Y., Sussman, D., Tang, M., Vogelstein, J., and Priebe, C. (2017). Statistical inference on random dot product graphs: A survey. *J. of Machine Learning Research*, 18.

Balzano, L., Recht, B., and Nowak, R. (2010). High-dimensional matched subspace detection when data are missing. In *IEEE Int. Symp. on Information Theory*, pages 1638–1642.

Basseville, M. and Nikiforov, I. (1993). *Detection of abrupt changes: Theory and application*. Prentice-Hall, Inc.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope Heuristics: Overview and Implementation. *Statistics and Computing*, 22:455—470.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. European Mathematical Society*, 3:203–268.

Chen, J. and Gupta, A. (2012). *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. Birkhäuser Basel, 2nd edition.

Chen, Y., Mao, X., Ling, D., and Gu, Y. (2018). Change-point detection of gaussian graph signals with partial information. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 3934–3938.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441.

Giraud, C. (2015). *Introduction to high-dimensional statistics*. Monographs on statistics and applied probability (Series) ; 139. CRC Press.

Huang, W., Goldsberry, L., Wymbs, N. F., Grafton, S. T., Bassett, D. S., and Ribeiro, A. (2016). Graph frequency analysis of brain signals. *IEEE J. of Selected Topics in Signal Processing*, 10(7):1189–1203.

Kalofolias, V. (2016). How to learn a graph from smooth signals. In *Int. Conf. on Artificial Intelligence and Statistics*, volume 51, pages 920–929.

Le Bars, B., Humbert, P., Kalogeratos, A., and Vayatis, N. (2020). Learning the piece-wise constant graph structure of a varying Ising model. In *Int. Conf. on Machine Learning*.

Le Bars, B., Humbert, P., Oudre, L., and Kalogeratos, A. (2019). Learning laplacian matrix from bandlimited graph signals. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 2937–2941.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing*, 85(4):717–736.

Marques, A. G., Segarra, S., Leus, G., and Ribeiro, A. (2017). Stationary graph processes and spectral estimation. *IEEE Trans. on Signal Processing*, 65(22):5911–5926.

Massart, P. and Meynet, C. (2011). The Lasso as an ℓ1-ball model selection procedure. *Electr. J. of Statistics*, 5.

Massart, P. and Picard, J. (2003). *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Lecture Notes in Mathematics. Springer Berlin Heidelberg.

Ortega, A., Frossard, P., Kovačević, J., Moura, J., and Vandergheynst, P. (2018). Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828.

Perraudin, N. and Vandergheynst, P. (2017). Stationary signal processing on graphs. *IEEE Trans. on Signal Processing*, 65(13):3462–3477.

Rossi, R. A. and Ahmed, N. K. (2015). The network data repository with interactive graph analytics and visualization. In *AAAI Conf. on Artificial Intelligence*.

Sandryhaila, A. and Moura, J. (2013). Discrete signal processing on graphs. *IEEE Trans. on Signal Processing*, 61(7):1644–1656.

Shuman, D., Narang, S., Frossard, P., Ortega, A., and Vandergheynst, P. (2013). 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.

Tartakovsky, A., Nikiforov, I., and Basseville, M. (2014). *Sequential Analysis: Hypothesis Testing and Change-point Detection*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, CRC Press.

Tenenbaum, J. B. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.