CHAPTER

Statistical Graph Signal Processing: Stationarity and Spectral Estimation

1

Santiago Segarra[a,*], Sundeep Prabhakar Chepuri[**], Antonio G. Marques[†] and Geert Leus[**]

[*]Massachusetts Institute of Technology, USA [**]Delft University of Technology, The Netherlands [†]King Juan Carlos University, Spain

[a]Corresponding: segarra@mit.edu

CHAPTER OUTLINE HEAD

ABSTRACT
Stationarity is a cornerstone property that facilitates the analysis and processing of random signals in the time domain. Although time-varying signals are abundant in nature, in many contemporary applications the information of interest resides in more irregular domains which can be conveniently represented using a graph. This chapter reviews recent advances in extending the notion of stationarity to random graph signals. This is a challenging task due to the irregularity of the underlying graph domain. To that end, we start by presenting coexisting stationarity definitions along with explanations of their genesis, advantages, and disadvantages. Secondly, we introduce the concept of power spectral density for graph processes and propose a number of methods for its estimation. These methods include nonparametric approaches such as correlograms and windowed average periodograms, as well as parametric approaches. To account for distributed scenarios where the supporting graph is related to an actual network infrastructure, the last part of the chapter discusses how to estimate the power spectral density of a graph process when having access to only a subset of the nodes. To gain intuition and insights, the concepts and schemes presented throughout the chapter are illustrated with a running example based on a real-world social graph.

Keywords: Random graph processes and signals, weak stationarity, power spectral density, ARMA estimation, periodogram, windowing, sampling, covariance matching.

1.1 RANDOM GRAPH PROCESSES

1.1.1 INTRODUCTION

Most of the tools in graph signal processing are deterministic in nature, e.g., graph signal denoising using diffusion [1], sampling and reconstruction of graph signals [2, 3, 4, 5, 6, 7], graph filter design [8, 9, 10, 11], and so on. Only recently, statistical signal processing methods tailored to graph signals have been introduced. As we know from classical signal processing focusing on spatiotemporal signals, statistical methods allow one to take statistical information into account when designing optimal sampling and reconstruction schemes, e.g., Wiener filtering for denoising, interpolation, prediction, and so on [12]. This generally leads to a better average performance compared to deterministic methods. Key to the majority of statistical methods is the concept of weak stationarity, which means that the first and second-order statistics of the random process do not change over space and/or time. The extension of this concept to graph signals is not trivial due to the fact that these signals have an irregular structure, which is generally characterized by a so-called graph shift (a generalization of the shift in time and/or space). This is what will be discussed in the current chapter.

The first works discussing stationary graph processes observe that in contrast to a shift in time and/or space, a graph shift is not energy preserving [13, 14]. Hence,

these papers base their definition of a weakly stationary graph process on a new isometric graph shift. However, this new shift cannot be carried out by means of local operations and hence the connection between stationarity and locality is lost. Therefore, in this chapter, we present definitions based on the original graph shift, allowing for stationarity tests and estimation schemes based on local information. Stationary graph processes are also characterized by a power spectral density (PSD) and this chapter provides a rigorous treatment of various PSD estimators including nonparametric and parametric methods. Our treatment of stationary graph processes is based on the comprehensive study presented in [15]. Graph stationarity was also studied in [16], where the analysis is carried out using the Laplacian matrix as graph shift operator. In this chapter, the proposed framework is also extended to random processes that are jointly stationary in the time and vertex domain [17]. This paves the way to statistical tools for random processes over two domains, the regular time domain and the irregular graph domain.

The field of compressive sensing has recently been extended to compressive covariance sensing [18], which is based on the idea that the covariance matrix or PSD of a spatiotemporal process can be estimated from compressed measurements without any prior assumptions on sparsity or smoothness. A special case of compressive covariance sensing occurs when the compression is realized by subsampling (below the Nyquist rate), also known as sparse covariance sensing. This allows one to design statistical signal processing tools from only a subset of measurements. The last part of the chapter explains how these ideas can be extended to random graph processes, where the covariance matrix does not have any apparent structure as for spatiotemporal processes [19]. We demonstrate how the covariance matrix – and thus the PSD – of a graph process can be estimated from a subset of the nodes without the use of priors. Again nonparametric as well as parametric methods are considered and we additionally show how to select the nodes in a greedy fashion.

1.1.2 CHAPTER ORGANIZATION

The definition of weakly stationary graph processes is presented in Section 1.2 along with discussions about the relation with the classical definition in time. Section 1.2.1 introduces the notion of power spectral density (PSD) followed by a recollection of relevant examples and useful properties. The characterization of stationarity for graph processes that also vary over time is presented in Section 1.2.2. Since stationary processes are easier to understand in the frequency domain, Section 1.3 is devoted to the study of different methods for *spectral estimation*, which can also be used to improve the estimate of the covariance matrix itself. These include both nonparametric and parametric approaches. Finally, in Section 1.4 we discuss methods to estimate the PSD and the covariance of random graph processes using only observations from a subset of nodes. We will also develop a low-complexity and near-optimal method to select the nodes in greedy manner.

1.1.3 NOTATION

Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ be a directed graph or network with a set of N nodes $\mathcal{N}$ and directed edges $\mathcal{E}$ such that $(i, j) \in \mathcal{E}$ if there exists an edge from node i to node j. We associate with $\mathcal{G}$ the graph shift operator (GSO) $\mathbf{S}$, defined as an $N \times N$ matrix whose entry $S_{ji} \neq 0$ only if $i = j$ or if $(i, j) \in \mathcal{E}$ [20, 9]. The sparsity pattern of $\mathbf{S}$ captures the local structure of $\mathcal{G}$, but we make no specific assumptions on the values of the nonzero entries of $\mathbf{S}$; hence the GSO can represent the adjacency matrix, the Laplacian, or other graph-related matrices. In this chapter we assume that $\mathbf{S}$ is *normal* to guarantee the existence of a unitary matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$ and a diagonal matrix $\mathbf{\Lambda}$ such that $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H$. We use $\mathbf{x} = [x_1, ..., x_N]^T \in \mathbb{R}^N$ to denote a generic graph signal and $\tilde{\mathbf{x}} := \mathbf{V}^H\mathbf{x}$ to denote its frequency representation, with $\mathbf{V}^H$ being the graph Fourier transform (GFT) [9]. Finally, we use $\mathbf{H}: \ \mathbb{R}^N \to \mathbb{R}^N$ to denote a linear graph filter of the form

$$\mathbf{H} = \sum_{l=0}^{L-1} h_l \mathbf{S}^l = \mathbf{V}\text{diag}(\tilde{\mathbf{h}})\mathbf{V}^H = \mathbf{V}\text{diag}(\mathbf{\Psi}_L\mathbf{h})\mathbf{V}^H, \tag{1.1}$$

where $\tilde{\mathbf{h}}$ denotes the frequency response of the filter $\mathbf{H}$ and $\mathbf{\Psi}_L$ is an $N \times L$ Vandermonde matrix with entries $\Psi_{kl} = \lambda_k^{l-1}$. The notation $\circ$, $\otimes$, and $\odot$ denote the elementwise, Kronecker, and Khatri-Rao matrix products, respectively. The notation $\oplus$ stands for the Kronecker sum.

1.2 WEAKLY STATIONARY GRAPH PROCESSES

We extend three equivalent definitions of weak stationarity in time to the graph domain, the most common being the invariance of the first and second moments to time shifts. We will see that under certain conditions those definitions can be rendered equivalent for the graph domain as well. Intuitively, stating that a *graph* process is stationary is an inherently incomplete assertion, since we need to declare which graph we are referring to. Hence, the proposed definitions depend on the GSO $\mathbf{S}$, so that a process $\mathbf{x}$ can be stationary in $\mathbf{S}$ but not in $\mathbf{S}' \neq \mathbf{S}$.

Defining a standard zero-mean white random process $\mathbf{n}$ as one with mean $\mathbb{E}[\mathbf{n}] = \mathbf{0}$ and covariance $\mathbb{E}\left[\mathbf{n}\mathbf{n}^H\right] = \mathbf{I}$, we state our first definition of graph stationarity.

Definition 1. Given a normal shift operator $\mathbf{S}$, a zero-mean random process $\mathbf{x}$ is weakly stationary with respect to $\mathbf{S}$ if it can be written as the response of a linear shift-invariant graph filter $\mathbf{H} = \sum_{l=0}^{N-1} h_l \mathbf{S}^l$ to a zero-mean white input $\mathbf{n}$.

The definition states that stationary graph processes can be written as the output of graph filters when excited with a white input. This generalizes the well-known fact that stationary processes in time can be expressed as the output of linear time invariant systems with white noise as input. If we write $\mathbf{x} = \mathbf{H}\mathbf{n}$, the covariance

matrix $\mathbf{C}_\mathbf{x} := \mathbb{E}\left[\mathbf{x}\mathbf{x}^H\right]$ of the process $\mathbf{x}$ is given by

$$\mathbf{C}_\mathbf{x} = \mathbb{E}\left[(\mathbf{H}\mathbf{n})(\mathbf{H}\mathbf{n})^H\right] = \mathbf{H}\mathbb{E}\left[\mathbf{n}\mathbf{n}^H\right]\mathbf{H} = \mathbf{H}\mathbf{H}^H, \tag{1.2}$$

which shows that the color of $\mathbf{x}$ is determined by the filter $\mathbf{H}$. We can think of Def. 1 as a constructive definition of stationarity since it describes how a stationary process can be generated. Alternatively, one can define stationarity from a descriptive perspective, by imposing requirements on the moments of the random graph process in either the vertex or the frequency domain.

Definition 2. Given a normal shift operator $\mathbf{S}$, a zero-mean random process $\mathbf{x}$ is weakly stationary with respect to $\mathbf{S}$ if the following two equivalent properties hold

(a) For any set of nonnegative integers a, b, and $c \leq b$ it holds that

$$\mathbb{E}\left[(\mathbf{S}^a\mathbf{x})\left((\mathbf{S}^H)^b\mathbf{x}\right)^H\right] = \mathbb{E}\left[(\mathbf{S}^{a+c}\mathbf{x})\left((\mathbf{S}^H)^{b-c}\mathbf{x}\right)^H\right]. \tag{1.3}$$

(b) Matrices $\mathbf{C}_\mathbf{x}$ and $\mathbf{S}$ are simultaneously diagonalizable.

The statements in Defs. 2.a and 2.b can indeed be shown to be equivalent [15]. These two statements generalize known definitions of stationarity in time. Def. 2.a generalizes the requirement that the second moment of a stationary process must be invariant to time shifts whereas Def. 2.b extends the requisite that the covariance of time stationary processes must be circulant.

In Def. 2.a we require the correlation to be invariant to *how* we shift our signal – namely, forward $\mathbf{S}\mathbf{x}$ or backward $\mathbf{S}^H\mathbf{x}$ –, as long as the total number of shifts remains constant. Indeed, in both the left and right hand sides of (1.3) the signal is shifted a total of $a + b$ times. This generalizes what happens for stationary signals in time, where correlation depends on the total number of shifts, but not on the particular time instants. More specifically, when $\mathbf{S}$ is a directed cycle we have that $\mathbf{S}^H = \mathbf{S}^{-1}$. Also notice that for the directed cycle $\mathbf{S}^N = \mathbf{I}$. Then, if we set $a = 0$, $b = N$ and $c = l$, (1.3) boils down to $\mathbb{E}\left[\mathbf{x}\mathbf{x}^H\right] = \mathbb{E}\left[\mathbf{S}^l\mathbf{x}(\mathbf{S}^l\mathbf{x})^H\right]$, which is the definition of a stationary signal in time. Intuitively, accumulating the same number of shifts in both sides of (1.3) is necessary because the operator $\mathbf{S}$ in general does not preserve the energy. Thus, requiring $\mathbb{E}\left[\mathbf{x}\mathbf{x}^H\right] = \mathbb{E}\left[\mathbf{S}^l\mathbf{x}(\mathbf{S}^l\mathbf{x})^H\right]$ for stationarity with respect to a general GSO would be infeasible. Def. 2.a strikes the right balance of being valid for general normal GSOs while particularizing to the accepted classical definition when $\mathbf{S}$ represents the domain of time signals.

Def. 2.b characterizes stationarity from a graph frequency perspective by requiring the covariance $\mathbf{C}_\mathbf{x}$ to be diagonalized by the GFT matrix $\mathbf{V}$. When particularized to time, Def. 2.b requires $\mathbf{C}_\mathbf{x}$ to be diagonalized by the Fourier matrix and, therefore, must be circulant. This fact is exploited in classical signal processing to define the power spectral density of a stationary process as the eigenvalues of the circulant covariance matrix, motivating the PSD definition in Section 1.2.1.

Thus far, we have presented three extensions of the concept of stationarity into the realm of graph processes, two of which are equivalent and, hence, grouped in

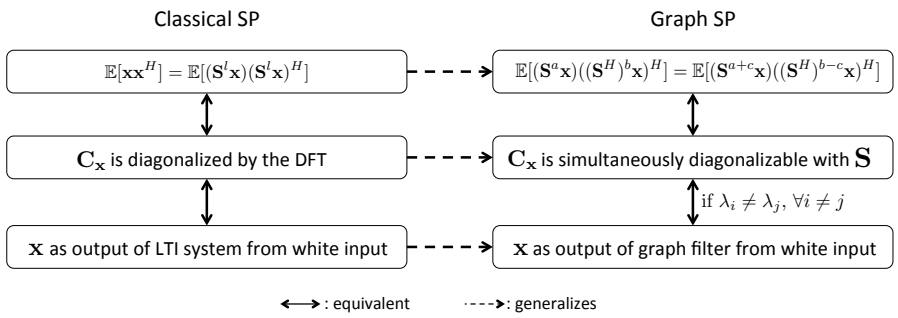

FIGURE 1.1 Equivalent definitions of a weakly stationary graph process

Three equivalent definitions for weak stationarity in time and their corresponding extensions to the graph domain. In graphs, two of the definitions are always equivalent and the third one is equivalent for shifts with distinct eigenvalues.

Def. 2. At this point, the attentive reader might have a natural inquiry. Are Defs. 1 and 2 equivalent for general graphs, as they are for stationarity in time? In fact, it can be shown that Defs. 1 and 2 are equivalent for any graph $\mathbf{S}$ that is normal and whose eigenvalues are all distinct [15]. Figure 1.1 presents a concise summary of the definitions discussed in this section.

Coexisting approaches. Stationary graph processes were first defined and analyzed in [13]. The fundamental problem identified there is that GSOs do not preserve energy in general and therefore cannot be isometric [21]. This problem is addressed in [14] with the definition of an isometric graph shift that preserves the eigenvector space of the Laplacian GSO but modifies its eigenvalues. A stationary graph process is then defined as one whose probability distributions are invariant with respect to multiplications with the isometric shift. One drawback of this approach is that the isometric shift is a complex-valued operator and has a sparsity structure (if any) different from $\mathbf{S}$. By contrast, the vertex-based definition in (1.3) is based on the original GSO $\mathbf{S}$, which is local and real-valued. As a result, (1.3) provides intuition on the relations between stationarity and locality, which can be leveraged to develop stationarity tests or estimation schemes that work with local information. Graph stationarity was also studied in [16] where the requirement of having a covariance matrix diagonalizable by the eigenvectors of the Laplacian GSO is adopted as a definition. This condition is shown to be equivalent to statistical invariance with respect to the translation operator introduced in [22]. When the shift $\mathbf{S}$ coincides with the Laplacian of the graph and the eigenvalues of $\mathbf{S}$ are all distinct, Defs. 1 and 2 are equivalent to those in [13] and [16]. Hence, the definitions presented here differ from [16] in that we consider general normal shifts instead of Laplacians and that we see Def. 1 as a definition, not a property. These are mathematically minor differences that are important in practice though; see [15, 23] for more details.

1.2.1 POWER SPECTRAL DENSITY

Stationarity reduces the degrees of freedom of a random graph process, thus facilitating its description and understanding. It follows from Def. 2.b that one can express the remaining degrees of freedom in the frequency domain via the notion of power spectral density, as defined next.

Definition 3. The power spectral density (PSD) of a random process $\mathbf{x}$ that is stationary with respect to $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H$ is the nonnegative $N \times 1$ vector $\mathbf{p}$

$$\mathbf{p} := \operatorname{diag}\left(\mathbf{V}^H \mathbf{C}_\mathbf{x} \mathbf{V}\right). \tag{1.4}$$

Observe that since $\mathbf{C}_\mathbf{x}$ is diagonalized by $\mathbf{V}$ (see Def. 2.b) the matrix $\mathbf{V}^H\mathbf{C}_\mathbf{x}\mathbf{V}$ is diagonal and it follows that the PSD in (1.4) corresponds to the eigenvalues of the positive semidefinite covariance matrix $\mathbf{C}_\mathbf{x}$. Thus, (1.4) is equivalent to

$$\mathbf{C}_\mathbf{x} = \mathbf{V}\operatorname{diag}(\mathbf{p})\mathbf{V}^H. \tag{1.5}$$

Zero-mean white noise is an example of a random process that is stationary with respect to *any* graph shift $\mathbf{S}$. The PSD of white noise with covariance $\mathbb{E}[\mathbf{n}\mathbf{n}^H] = \sigma^2\mathbf{I}$ is $\mathbf{p} = \sigma^2\mathbf{1}$. Also notice that, by definition, any random process $\mathbf{x}$ is stationary with respect to the shift $\mathbf{S} = \mathbf{C}_\mathbf{x}$ defined by its covariance matrix, with corresponding PSD $\mathbf{p} = \operatorname{diag}(\mathbf{\Lambda})$. This can be exploited in the context of network topology inference. Given a set of graph signals $\{\mathbf{x}_r\}_{r=1}^R$ it is common to infer the underlying topology by building a graph $\mathcal{G}_{corr}$ whose edge weights correspond to cross-correlations among the entries of the signals. In that case, the process generating those signals is stationary in the shift given by the adjacency of $\mathcal{G}_{corr}$; see [23] for details. A random process $\mathbf{x}$ is also stationary with respect to the shift given by its precision matrix, which is defined as the (pseudo-)inverse $\mathbf{\Theta} = \mathbf{C}_\mathbf{x}^\dagger$. The PSD in this case is $\mathbf{p} = \operatorname{diag}(\mathbf{\Lambda})^\dagger$. This is particularly important when $\mathbf{x}$ is a Gaussian Markov Random Field (GMRF) whose Markovian dependence is captured by the unweighted graph $\mathcal{G}_{MF}$. It is well known [24, Ch. 19] that in these cases Θ_{ij} can be non-zero only if (i, j) is either a link of $\mathcal{G}_{MF}$, or an element in the diagonal. Thus, *any* GMRF is stationary with respect to the *sparse* shift $\mathbf{S} = \mathbf{\Theta}$, which captures the conditional dependence between the elements of $\mathbf{x}$.

Two important properties that hold for random processes in time can be shown to be true as well for the PSD of graph processes.

Property 1. Let $\mathbf{x}$ be stationary in $\mathbf{S}$ with covariance $\mathbf{C}_\mathbf{x}$ and PSD $\mathbf{p}_\mathbf{x}$. Consider a filter $\mathbf{H}$ with frequency response $\tilde{\mathbf{h}}$ and define $\mathbf{y} := \mathbf{H}\mathbf{x}$. Then, the process $\mathbf{y}$:

(a) Is stationary in $\mathbf{S}$ with covariance $\mathbf{C}_\mathbf{y} = \mathbf{H}\mathbf{C}_\mathbf{x}\mathbf{H}^H$.

(b) Has a PSD given by $\mathbf{p}_\mathbf{y} = |\tilde{\mathbf{h}}|^2 \circ \mathbf{p}_\mathbf{x}$, where $|\cdot|^2$ is applied elementwise.

Property 2. Given a process $\mathbf{x}$ stationary in $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H$ with PSD $\mathbf{p}$, define the GFT process as $\tilde{\mathbf{x}} = \mathbf{V}^H\mathbf{x}$. Then, it holds that $\tilde{\mathbf{x}}$ is uncorrelated and its covariance matrix is

$$\mathbf{C}_{\tilde{\mathbf{x}}} := \mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^H\right] = \mathbb{E}\left[(\mathbf{V}^H\mathbf{x})(\mathbf{V}^H\mathbf{x})^H\right] = \operatorname{diag}(\mathbf{p}). \tag{1.6}$$

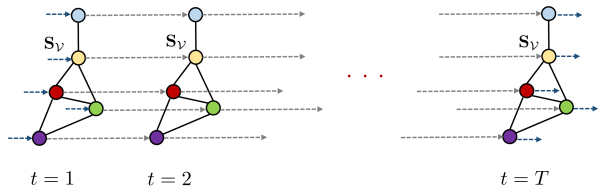

FIGURE 1.2 Support of a vertex-time process

The shift $\mathbf{S}_{\mathcal{V}}$ captures the dependence across the nodes of the underlying network. Solid lines represent the edges in $\mathbf{S}_{\mathcal{V}}$. Dashed lines represent connections between the same node at two consecutive time instants.

Property 1 is a statement of the spectral convolution theorem for graph signals. Property 2 is fundamental to motivate the analysis and modeling of stationary graph processes in the frequency domain, which we undertake in the remainder of this chapter. It also shows that if a process $\mathbf{x}$ is stationary in the shift $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H$, then the GFT $\mathbf{V}^H$ provides the Karhunen-Loève expansion of the process.

The concept of stationarity and, consequently, that of PSD can be extended to processes defined jointly in a graph and over time. Before we review this extension in the ensuing section, we discuss requirements on the first moment of stationary graph processes.

The mean of stationary graph processes. While Defs. 1 and 2 assume that the random process $\mathbf{x}$ has mean $\bar{\mathbf{x}} := \mathbb{E}[\mathbf{x}] = \mathbf{0}$, traditional stationary time processes are allowed to have a (non-zero) constant mean $\bar{\mathbf{x}} = \alpha\mathbf{1}$, with α being an arbitrary scalar. Stationary graph processes, by contrast, are required to have a first-order moment of the form $\bar{\mathbf{x}} = \alpha\mathbf{v}_k$, i.e., a scaled version of an eigenvector of $\mathbf{S}$. This choice: i) takes into account the structure of the underlying graph; ii) maintains the validity of Property 1; and iii) encompasses the case $\mathbf{v}_k = \mathbf{1}$ when $\mathbf{S}$ is either the adjacency matrix of a directed cycle or the Laplacian of any graph, recovering the classical first-order requirement for weak stationarity.

1.2.2 JOINT TIME AND GRAPH STATIONARITY

In many real-world network applications, observations are taken periodically, giving rise to a *sequence* $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T] \in \mathbb{R}^{N \times T}$ of graph signals. Each signal has size N – the number of nodes in the network – and there are T of those signals. Up to this point we have been focusing on the statistical variation across the vertices of the network graph. That is, we took one particular column of $\mathbf{X}$ and analyzed the statistical relations between the signal values at different vertices. The purpose of this section is to carry out this analysis jointly across rows and columns of $\mathbf{X}$. The ultimate goal is to present the conditions under which a random process is considered to be *jointly* stationary in both the vertex and the time domain [17].

The first step to analyze the statistical properties of the vertex-time process $\mathbf{X}$,

which whenever convenient will be represented as $\mathbf{x} = \text{vec}(\mathbf{X})$, is to identify its graph support. As shown in Figure 1.2, for every time instant one can plot a graph that accounts for the graph support of the corresponding column of $\mathbf{X}$. With this representation, a horizontal path in the picture represents a particular node at different time instants. To account for the time variation, node n at time t is the origin of a link toward its successor (node n at time $t + 1$), as well as the destination of a link from its predecessor (node n at time $t - 1$). Suppose that the spatial graph $\mathbf{S}_\mathcal{V} = \mathbf{V}_\mathcal{V}\mathbf{\Lambda}_\mathcal{V}\mathbf{V}_\mathcal{V}^H$ is the same for all columns, as is the case in Figure 1.2, and let us use $\mathbf{S}_\mathcal{T} = \mathbf{V}_\mathcal{T}\mathbf{\Lambda}_\mathcal{T}\mathbf{V}_\mathcal{T}^H$, the adjacency of the directed cycle, to denote the support of the time domain. Then, it holds that the graph support of $\mathbf{X}$, which will be denoted as $\mathbf{S}_\mathcal{J}$, is given by the Cartesian product [25] of $\mathbf{S}_\mathcal{V}$ and $\mathbf{S}_\mathcal{T}$. Mathematically, this implies that the joint shift $\mathbf{S}_\mathcal{J} \in \mathbb{R}^{NT\times NT}$ can be written as

$$\mathbf{S}_\mathcal{J} = \mathbf{S}_\mathcal{T} \oplus \mathbf{S}_\mathcal{V} = \mathbf{I}_\mathcal{T} \otimes \mathbf{S}_\mathcal{V} + \mathbf{S}_\mathcal{T} \otimes \mathbf{I}_\mathcal{V}, \tag{1.7}$$

where $\mathbf{I}_\mathcal{T}$ and $\mathbf{I}_\mathcal{V}$ are identity matrices of appropriate size. Using basic properties of the Kronecker product, it follows from (1.7) that the eigendecomposition of the joint shift is given by $\mathbf{S}_\mathcal{J} = (\mathbf{V}_\mathcal{T} \otimes \mathbf{V}_\mathcal{V})(\mathbf{\Lambda}_\mathcal{T} \oplus \mathbf{\Lambda}_\mathcal{V})(\mathbf{V}_\mathcal{T} \otimes \mathbf{V}_\mathcal{V})^H$, revealing that the GFT associated with $\mathbf{S}_\mathcal{J}$ is $\mathbf{V}_\mathcal{T} \otimes \mathbf{V}_\mathcal{V}$, the Kronecker product ot the GFTs associated[1] with $\mathbf{S}_\mathcal{T}$ and $\mathbf{S}_\mathcal{V}$ [26].

Once the graph support of the joint process and its corresponding GFT have been identified, for $\mathbf{X}$ to be jointly stationary in $\mathbf{S}_\mathcal{V}$ and $\mathbf{S}_\mathcal{T}$ it suffices to particularize the definitions presented in the previous section for the shift $\mathbf{S}_\mathcal{J}$, giving rise to the following result.

Definition 4. A process $\mathbf{X}$ is *jointly stationary* in $\mathbf{S}_\mathcal{V}$ and $\mathbf{S}_\mathcal{T}$ if the covariance matrix $\mathbf{C}_\mathbf{x} = \mathbb{E}[\text{vec}(\mathbf{X})\text{vec}(\mathbf{X})^T]$ can be written as $\mathbf{C}_\mathbf{x} = (\mathbf{V}_\mathcal{T} \otimes \mathbf{V}_\mathcal{V})\text{diag}(\mathbf{p}_x)(\mathbf{V}_\mathcal{T} \otimes \mathbf{V}_\mathcal{V})^H$.

Clearly, the nonnegative vector $\mathbf{p}_\mathbf{x}$ of length NT stands for the PSD of $\mathbf{X}$. If the eigenvalues of $\mathbf{S}_\mathcal{J}$ are non-repeated, the definition is equivalent to requiring $\mathbf{C}_\mathbf{x}$ to be written as a (positive semidefinite) graph filter on the shift operator $\mathbf{S}_\mathcal{J}$.

While Definition 4 describes the *spectral* properties of the covariance of a jointly stationary process, it is also of interest to understand its implications in the vertex and time domains. To that end, recall that $\mathbf{e}_i$ represents the i-th canonical vector, the signal $\mathbf{x}_t = \mathbf{X}\mathbf{e}_t \in \mathbb{R}^N$ collects the values of the process at time instant t, and the signal $\boldsymbol{\chi}_n = \mathbf{X}^T\mathbf{e}_n \in \mathbb{R}^T$ collects the values of the process at node n for the different time instants. Noting that *submatrices* of $\mathbf{C}_\mathbf{x}$ will describe how *subsets* of the elements of $\mathbf{X}$ are correlated, the result stated next follows from Definition 4.

Property 3. If $\mathbf{X}$ is jointly stationary in $\mathbf{S}_\mathcal{V}$ and $\mathbf{S}_\mathcal{T}$, then it holds that:

1. Any submatrix of $\mathbf{C}_\mathbf{x}$ of the form $\mathbf{C}_{t,t'}^\mathcal{V} = \mathbb{E}[\mathbf{x}_t\mathbf{x}_{t'}^T] = \mathbb{E}[\mathbf{X}\mathbf{e}_t\mathbf{e}_{t'}^T\mathbf{X}^T] \in \mathbb{R}^{N\times N}$ is a polynomial on $\mathbf{S}_\mathcal{V}$.

[1] Recall that the fact of $\mathbf{S}_\mathcal{T}$ being the directed cycle implies that the GFT $\mathbf{V}_\mathcal{T}^H$ is the $T \times T$ DFT matrix, so that $[\mathbf{V}_\mathcal{T}]_{k,k'} = \frac{1}{\sqrt{T}}\exp(\mathrm{j}\frac{2\pi}{T}kk')$.

2. Any submatrix of $\mathbf{C}_\mathbf{x}$ of the form $\mathbf{C}^{\mathcal{T}}_{n,n'} = \mathbb{E}[\boldsymbol{\chi}_n \boldsymbol{\chi}_{n'}^T] = \mathbb{E}[\mathbf{X}^T \mathbf{e}_n \mathbf{e}_{n'}^T \mathbf{X}] \in \mathbb{R}^{T\times T}$ is a polynomial on $\mathbf{S}_\mathcal{T}$ and, hence, it is circulant.

The statement in Property 3.2 is equivalent to saying that $\mathbb{E}[X_{n,t} X_{n',t'}] = \mathbb{E}[X_{n,t+a} X_{n',t'+a}]$, which is the classical requirement for a *multivariate* time series being considered stationary [27, Sec. 2.1.3]. Particularizing the results in Property 3 for $t = t'$ and $n = n'$ yield the subsequent property.

Property 4. If $\mathbf{X}$ is jointly stationary in $\mathbf{S}_\mathcal{V}$ and $\mathbf{S}_\mathcal{T}$, then it holds that:

1. All the graph signals $\mathbf{x}_t = \mathbf{X}\mathbf{e}_t$ are stationary in $\mathbf{S}_\mathcal{V}$
2. All the time-varying signals $\boldsymbol{\chi}_n = \mathbf{X}^T \mathbf{e}_n$ are stationary in $\mathbf{S}_\mathcal{T}$.

The result above is not an equivalence. That is, there may be processes that satisfy the two conditions stated in Property 4 but do not posses the structure in Definition 4. Even if those processes cannot be considered jointly stationary, they are likely to arise in practice, so that the design of signal processing schemes that leverage their structure is of interest.

Remark: Definition 4 is also valid if the joint shift $\mathbf{S}_\mathcal{J}$ is defined as either the Kronecker product or the strong product [25] between graphs $\mathbf{S}_\mathcal{V}$ and $\mathbf{S}_\mathcal{T}$. The reason is that for any of these three graph products, the eigenbasis of the joint shift is $\mathbf{V}_\mathcal{T} \otimes \mathbf{V}_\mathcal{V}$ [25, 26].

Jointly stationary and separable processes

We close this section by elaborating on a *subclass* of jointly stationary processes of particular relevance. To that end, let matrix $\mathbf{H}_\mathcal{V}$ be a generic graph filter in the shift $\mathbf{S}_\mathcal{V}$ and, similarly, $\mathbf{H}_\mathcal{T}$ a generic linear time-invariant filter. Those filters are used in the following definition.

Definition 5. Let $\mathbf{X}$ be a process jointly stationary in $\mathbf{S}_\mathcal{V}$ and $\mathbf{S}_\mathcal{T}$. Then, the process $\mathbf{X}$ is called *separable* if it can be written as $\mathbf{X} = \mathbf{H}_\mathcal{V} \mathbf{W} \mathbf{H}_\mathcal{T}^T$, where $\mathbf{W} \in \mathbb{R}^{N\times T}$ is a zero-mean white process with $\mathbb{E}[W_{ij} W_{ij}] = 1$ and $\mathbb{E}[W_{ij} W_{i'j'}] = 0$ for all $(ij) \neq (i'j')$.

From the previous definition one can view the jointly stationary and separable process $\mathbf{X}$ as one generated by processing each of the columns of $\mathbf{W}$ with the same graph filter and, then, each of the resultant rows with the same linear time-invariant filter. Note that one can also apply first the time-invariant filter $\mathbf{H}_\mathcal{T}$ and then the graph filter $\mathbf{H}_\mathcal{V}$. Upon defining $\mathbf{C}_{x,\mathcal{V}} = \mathbf{H}_\mathcal{V} \mathbf{H}_\mathcal{V}^T$, $\mathbf{C}_{x,\mathcal{T}} = \mathbf{H}_\mathcal{T} \mathbf{H}_\mathcal{T}^T$, $\mathbf{p}_{x,\mathcal{V}} = \text{diag}(\mathbf{V}_\mathcal{V}^H \mathbf{C}_{x,\mathcal{V}} \mathbf{V}_\mathcal{V})$ and $\mathbf{p}_{x,\mathcal{V}} = \text{diag}(\mathbf{V}_\mathcal{V}^H \mathbf{C}_{x,\mathcal{T}} \mathbf{V}_\mathcal{T})$, it is easy to show that the following properties hold.

Property 5. Let $\mathbf{X} = \mathbf{H}_\mathcal{V} \mathbf{W} \mathbf{H}_\mathcal{T}^T$ be a jointly stationary and separable process in $\mathbf{S}_\mathcal{V}$ and $\mathbf{S}_\mathcal{T}$. Then, it holds that:

1. The correlation of $\mathbf{X}$ can be factorized as $\mathbf{C}_x = \mathbf{C}_{x,\mathcal{T}} \otimes \mathbf{C}_{x,\mathcal{V}}$.
2. The PSD of $\mathbf{x}$ can be written as $\mathbf{p}_x = \mathbf{p}_{x,\mathcal{T}} \otimes \mathbf{p}_{x,\mathcal{V}}$.

The factorable structure of the correlation implies that, for any given (t, t'), the covariance $\mathbf{C}^{\mathcal{V}}_{t,t'} = \mathbb{E}[\mathbf{x}_t \mathbf{x}_{t'}^T]$ is a *scaled* version of $\mathbf{C}_{x,\mathcal{V}}$. In other words, after a trivial

scaling, the covariance of any of the columns of the separable process $\mathbf{X}$ is the same. Similarly, it holds that $\mathbf{C}^{\mathcal{V}}_{n,n'} = \mathbb{E}[\boldsymbol{\chi}_n \boldsymbol{\chi}^T_{n'}]$ is a scaled version of $\mathbf{C}_{x,\mathcal{T}}$ for all (n, n'). The fact of the PSD being factorable reveals that the number of degrees of freedom of the PSD of a jointly stationary and separable process is $N + T$, which contrasts with the NT degrees of freedom of a generic jointly stationary process. This more parsimonious description of the PSD vector –equivalently, of the covariance matrix– can be exploited when designing spectral (covariance) estimation schemes for processes obeying Definition 5.

1.3 POWER SPECTRAL DENSITY ESTIMATORS

We can exploit the fact that $\mathbf{x}$ is a stationary graph process in $\mathbf{S} = \mathbf{V}\text{diag}(\boldsymbol{\Lambda})\mathbf{V}^H$ to design efficient estimators of the covariance $\mathbf{C}_\mathbf{x}$. In particular, instead of estimating $\mathbf{C}_\mathbf{x}$ directly, which has $N(N + 1)/2$ degrees of freedom, one can estimate $\mathbf{p}$ first, which only has N degrees of freedom, and then leverage that $\mathbf{C}_\mathbf{x} = \mathbf{V}\text{diag}(\mathbf{p})\mathbf{V}^H$.

Motivated by this, the focus of this section is on estimating $\mathbf{p}$, the PSD of a stationary random graph process $\mathbf{x}$, using as input either *one* or a few realizations $\{\mathbf{x}_r\}_{r=1}^R$ of $\mathbf{x}$. To illustrate the developments in Sections 1.3 and 1.4, we will use as a running example a random process defined on the well-known Zachary's Karate club network [28] (Figs. 1.3 and 1.4). As shown in Fig. 1.4, this graph consists of 34 nodes or members of the club and 78 undirected edges symbolizing friendships among members.[2]

1.3.1 NONPARAMETRIC PSD ESTIMATORS

Nonparametric estimators – as opposed to their parametric counterparts – do not assume any specific generating model on the process $\mathbf{x}$. This more agnostic view of $\mathbf{x}$ comes with the price of needing in general to observe more graph signals to achieve satisfactory performance. In this section we extend to the graph setting the periodogram, the correlogram, and the least squares (LS) estimator, which are classical unbiased nonparametric estimators. Moreover, for the special case where the observations are Gaussian, we derive the Cramér-Rao lower bound. We also discuss the windowed average periodogram, which attain a better performance when a few observations are available by introducing bias in a controlled manner while drastically reducing the variance.

[2] The process to assess the performance of the different PSD estimators was created using the generating filter $\mathbf{H} = \sum_{l=0}^{3} h_l \mathbf{S}^l$ where $\mathbf{S}$ was set as the Laplacian matrix and the filter coefficients as $\mathbf{h} = [1, -0.15, 0.075, -10^{-4}]^T$ (cf. Definition 1). The coefficients were chosen for the filter to be of low-order and to have a low-pass behavior, as can be appreciated from the *"True PSD"* curves in Fig. 1.3, where most of the energy is concentrated in the low-frequencies.

Periodogram, correlogram, and LS estimator

From (1.6) it follows that one may express the PSD as $\mathbf{p} = \mathbb{E}\left[|\mathbf{V}^H\mathbf{x}|^2\right]$. That is, the PSD is given by the expected value of the squared frequency components of the random process. This leads to a natural approach for the estimation of $\mathbf{p}$ from a finite set of R realizations of the process $\mathbf{x}$. Indeed, we compute the GFT $\tilde{\mathbf{x}}_r = \mathbf{V}^H\mathbf{x}_r$ of each observed signal $\mathbf{x}_r$ and estimate $\mathbf{p}$ as

$$\hat{\mathbf{p}}_{\text{pg}} \;:=\; \frac{1}{R}\sum_{r=1}^{R} |\tilde{\mathbf{x}}_r|^2 \;=\; \frac{1}{R}\sum_{r=1}^{R} \left|\mathbf{V}^H\mathbf{x}_r\right|^2. \tag{1.8}$$

The estimator in (1.8) is termed *periodogram*, due to its evident similarity with its homonym in classical estimation. It is simple to show that $\hat{\mathbf{p}}_{\text{pg}}$ is an unbiased estimator, that is, $\mathbb{E}\left[\hat{\mathbf{p}}_{\text{pg}}\right] = \mathbf{p}$. A more detailed analysis of the performance of $\hat{\mathbf{p}}_{\text{pg}}$ for the case where the observations are Gaussian is given in Proposition 1.

An alternative nonparametric estimation scheme, denominated *correlogram*, can be devised by starting from the definition of $\mathbf{p}$ in (1.4). Namely, one may substitute $\mathbf{C}_{\mathbf{x}}$ in (1.4) by the *sample covariance* $\hat{\mathbf{C}}_{\mathbf{x}} = (1/R)\sum_{r=1}^{R}\mathbf{x}_r\mathbf{x}_r^H$ computed based on the available observations to obtain

$$\hat{\mathbf{p}}_{\text{cg}} := \text{diag}\left(\mathbf{V}^H\hat{\mathbf{C}}_{\mathbf{x}}\mathbf{V}\right) := \text{diag}\left[\mathbf{V}^H\left[\tfrac{1}{R}\textstyle\sum_{r=1}^{R}\mathbf{x}_r\mathbf{x}_r^H\right]\mathbf{V}\right]. \tag{1.9}$$

Notice that the matrix $\mathbf{V}^H\hat{\mathbf{C}}_{\mathbf{x}}\mathbf{V}$ is in general not diagonal, since the eigenbasis of $\hat{\mathbf{C}}_{\mathbf{x}}$ differs from $\mathbf{V}$, the eigenbasis of $\mathbf{C}_{\mathbf{x}}$. Nonetheless, we keep only the diagonal elements $\mathbf{v}_i^H\hat{\mathbf{C}}_{\mathbf{x}}\mathbf{v}_i$ for $i = 1, \ldots, N$ as our PSD estimator. It can be shown that the correlogram $\hat{\mathbf{p}}_{\text{cg}}$ in (1.9) and the periodogram $\hat{\mathbf{p}}_{\text{pg}}$ in (1.8) lead to identical estimators, as is the case in classical signal processing.

The correlogram can also be interpreted as a LS estimator. The decomposition in (1.5) allows a linear parametrization of the covariance matrix $\mathbf{C}_{\mathbf{x}}$ as

$$\mathbf{C}_{\mathbf{x}}(\mathbf{p}) = \sum_{i=1}^{N} p_i\mathbf{v}_i\mathbf{v}_i^H. \tag{1.10}$$

This linear parametrization will also be useful for the sampling schemes developed in Section 1.4. Vectorizing $\mathbf{C}_{\mathbf{x}}$ in (1.10) results in a set of N^2 equations in $\mathbf{p}$

$$\mathbf{c}_{\mathbf{x}} = \text{vec}(\mathbf{C}_{\mathbf{x}}) = \textstyle\sum_{i=1}^{N} p_i\text{vec}(\mathbf{v}_i\mathbf{v}_i^H) = \mathbf{G}_{\text{np}}\mathbf{p}, \tag{1.11}$$

where $\text{vec}(\mathbf{v}_i\mathbf{v}_i^H) = \mathbf{v}_i^* \otimes \mathbf{v}_i$. Relying on the Khatri-Rao product, we then form the $N^2 \times N$ matrix $\mathbf{G}_{\text{np}}$ as

$$\mathbf{G}_{\text{np}} := [\mathbf{v}_1^* \otimes \mathbf{v}_1, \cdots, \mathbf{v}_N^* \otimes \mathbf{v}_N] = \mathbf{V}^* \odot \mathbf{V}.$$

Using the sample covariance matrix $\hat{\mathbf{C}}_{\mathbf{x}}$ as an estimate of $\mathbf{C}_{\mathbf{x}}$, we can *match* the estimated covariance vector $\hat{\mathbf{c}}_{\mathbf{x}} = \text{vec}(\hat{\mathbf{C}}_{\mathbf{x}})$ to the true covariance vector $\mathbf{c}_{\mathbf{x}}$ in the LS

sense as

$$\hat{\mathbf{p}}_{\text{ls}} = \underset{\mathbf{p}}{\operatorname{argmin}} \|\hat{\mathbf{c}}_{\mathbf{x}} - \mathbf{G}_{\text{np}}\mathbf{p}\|_2^2 = (\mathbf{G}_{\text{np}}^H \mathbf{G}_{\text{np}})^{-1}\mathbf{G}_{\text{np}}^H \hat{\mathbf{c}}_{\mathbf{x}}. \tag{1.12}$$

In other words, the LS estimator minimizes the squared error $\text{tr}[(\hat{\mathbf{C}}_{\mathbf{x}} - \mathbf{C}_{\mathbf{x}}(\mathbf{p}))^T(\hat{\mathbf{C}}_{\mathbf{x}} - \mathbf{C}_{\mathbf{x}}(\mathbf{p}))]$. From expression (1.12) it can be shown that the ith element of $\hat{\mathbf{p}}_{\text{ls}}$ is $\mathbf{v}_i^H \hat{\mathbf{C}}_{\mathbf{x}} \mathbf{v}_i$. Combining this with (1.9), we get that the LS estimator $\hat{\mathbf{p}}_{\text{ls}}$ and the correlogram $\hat{\mathbf{p}}_{\text{cg}}$ – and hence the periodogram as well – are all identical estimators.

The estimators derived in this subsection do not assume any data distribution and are well suited for cases where the data probability density function is not available. In what follows, we provide performance bounds for these estimators under the condition that the observed signals are Gaussian.

Mean squared error and the Cramér-Rao bound

Suppose that the data consists of realizations from a sequence of independent and identically distributed (i.i.d.) Gaussian random vectors $\{\mathbf{x}_r\}_{r=1}^R$, where for each r, the vector $\mathbf{x}_r \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{x}}(\mathbf{p}))$. Under this setting, we can characterize the variance, hence the mean squared error (MSE), of the periodogram estimator (as well as the equivalent correlogram and LS estimators). In the following proposition we present expressions for its bias and variance.

Proposition 1. Let $\{\mathbf{x}_r\}_{r=1}^R$ be independent samples of the process $\mathbf{x}$ stationary in $\mathbf{S}$ with PSD $\mathbf{p}$. Then, the bias $\mathbf{b}_{\text{pg}}$ of the periodogram estimator in (1.8) is zero,

$$\mathbf{b}_{\text{pg}} \;:=\; \mathbb{E}\left[\hat{\mathbf{p}}_{\text{pg}}\right] - \mathbf{p} \;=\; \mathbf{0}. \tag{1.13}$$

Further define the covariance of the periodogram as $\mathbf{\Sigma}_{\text{pg}} := \mathbb{E}\left[(\hat{\mathbf{p}}_{\text{pg}} - \mathbf{p})(\hat{\mathbf{p}}_{\text{pg}} - \mathbf{p})^H\right]$. If the process $\mathbf{x}$ is Gaussian and $\mathbf{S}$ is symmetric, then $\mathbf{\Sigma}_{\text{pg}}$ can be written as

$$\mathbf{\Sigma}_{\text{pg}} := \mathbb{E}\left[(\hat{\mathbf{p}}_{\text{pg}} - \mathbf{p})(\hat{\mathbf{p}}_{\text{pg}} - \mathbf{p})^H\right] = (2/R)\text{diag}^2(\mathbf{p}). \tag{1.14}$$

As was mentioned before, Proposition 1 states that the periodogram is an unbiased estimator, i.e. $\mathbb{E}\left[\hat{\mathbf{p}}_{\text{pg}}\right] = \mathbf{p}$, as expected given its classical counterpart. While (1.13) is valid for any distribution, observe that the covariance expression in (1.14) requires the process $\mathbf{x}$ to be Gaussian. This requirement stems from the fact that the derivation of $\mathbf{\Sigma}_{\text{pg}}$ involves fourth order moments of $\mathbf{x}$. This is natural since an analogous limitation arises for time signals [29, Sec. 8.2]. Notice also that the PSD estimates of different frequencies are uncorrelated, since (1.14) reveals that $\mathbf{\Sigma}_{\text{pg}}$ is a diagonal matrix. A proof of the above result along with generalizations for the cases in which $\mathbf{S}$ is not necessarily symmetric (but normal) can be found in [15].

The MSE of the periodogram, defined as $\text{MSE}(\hat{\mathbf{p}}_{\text{pg}}) := \mathbb{E}\left[\|(\hat{\mathbf{p}}_{\text{pg}} - \mathbf{p})\|_2^2\right]$, can be readily computed using the result in Proposition 1

$$\text{MSE}(\hat{\mathbf{p}}_{\text{pg}}) = \|\mathbf{b}_{\text{pg}}\|_2^2 + \text{tr}[\mathbf{\Sigma}_{\text{pg}}] = (2/R)\|\mathbf{p}\|_2^2. \tag{1.15}$$

As becomes apparent from (1.15), the periodogram is expected to yield large relative

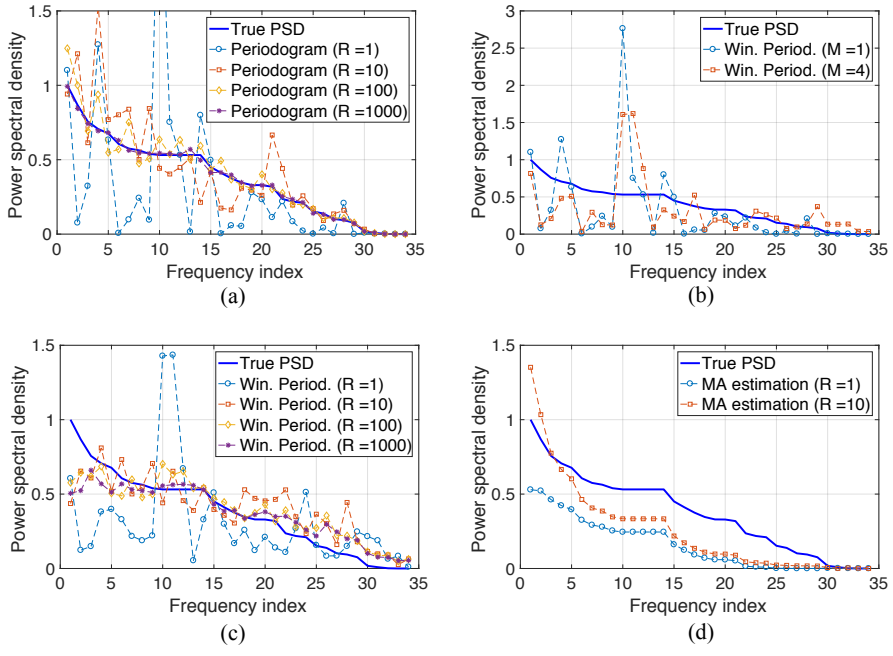

FIGURE 1.3 Power spectral density estimation

All estimators are based on the same random process defined on the Karate club network [28]. (a) Periodogram estimation with different numbers of observations. (b) Windowed average periodogram from a single realization and different number of windows. (c) Windowed average periodogram for 4 windows and varying number of realizations. (d) Parametric MA estimation for 1 and 10 realizations.

errors when only a few observations R are available. In Fig. 1.3(a) we show the periodogram estimation for different numbers of observations R. Notice that, indeed, when $R = 1$ the estimation is very poor. Nonetheless, when increasing R the estimation tends to the true PSD. A method that can achieve better performance for lower values of R – windowed average periodogram – will be introduced after showing that the periodogram is an efficient estimator.

The Cramér-Rao bound provides a lower bound on the covariance of unbiased estimators, when the available data records are finite. The Cramér-Rao bound matrix is equal to the inverse of the Fisher information matrix, $\mathbf{F}$, and it is given by $\mathbf{F} = (R/2)\,\mathrm{diag}^{-2}(\mathbf{p})$; see, e.g., [30, Ch. 6.13]. The efficiency of the periodogram follows readily by comparing $\mathbf{F}^{-1}$ with (1.14).

Windowed average periodogram

When only one or just a few observations of the process $\mathbf{x}$ are available, the periodogram and correlogram yield large errors [cf. (1.15)]. A way to overcome this

roadblock is to artificially generate multiple signals from the few available ones. Bartlett and Welch methods are classical examples of this procedure since they utilize windows to generate multiple samples of the process even if only a single realization is given [31, Sec. 2.7]. Intuitively, a long signal is partitioned into pieces, where each piece can be considered as a different signal. This operation introduces bias in the estimator but reduces variance to the point that the overall MSE can be improved. The frequency counterpart of such classical methods are filter banks, where the signal is partitioned in the Fourier domain. Both the windowed average periodogram – including Bartlett and Welch methods – and the filter banks can be extended for the estimation of graph processes. In this section, we only focus on the former, but extensions of this analysis as well as a full derivation of filter-bank estimators can be found in [15].

The application of a window $\mathbf{w}$ to a signal[3] $\mathbf{x}$ entails a componentwise multiplication to produce the signal $\mathbf{x}_\mathbf{w} = \text{diag}(\mathbf{w})\mathbf{x}$, where we assume that windows are normalized to have energy $\|\mathbf{w}\|_2^2 = N$. We may leverage the definition of the GFT to write

$$\tilde{\mathbf{x}}_\mathbf{w} = \mathbf{V}^H \mathbf{x}_\mathbf{w} = \mathbf{V}^H \text{diag}(\mathbf{w})\mathbf{x} = \mathbf{V}^H \text{diag}(\mathbf{w})\mathbf{V}\tilde{\mathbf{x}} =: \tilde{\mathbf{W}}\tilde{\mathbf{x}}, \tag{1.16}$$

where we implicitly defined $\tilde{\mathbf{W}} := \mathbf{V}^H \text{diag}(\mathbf{w})\mathbf{V}$ as the dual of the windowing operator in the frequency domain. For time signals the frequency representation of a window is its Fourier transform and the dual operator of windowing is the convolution between the spectra of the window and the signal. This parallelism is lost for graph signals. Nonetheless, (1.16) can be used to design windows with small spectral distortion, i.e., windows for which $\tilde{\mathbf{W}} \approx \mathbf{I}$. Recall that our objective is to generate multiple signals from only one, thus instead of focusing on a single window we consider a bank of M windows $\mathcal{W} = \{\mathbf{w}_m\}_{m=1}^M$ and use it to construct the windowed signals $\mathbf{x}_m := \text{diag}(\mathbf{w}_m)\mathbf{x}$. Based on these windowed signals, we build the *windowed average periodogram* as

$$\hat{\mathbf{p}}_\mathcal{W} := \frac{1}{M}\sum_{m=1}^{M} \left|\mathbf{V}^H \mathbf{x}_m\right|^2 = \frac{1}{M}\sum_{m=1}^{M} \left|\mathbf{V}^H \text{diag}(\mathbf{w}_m)\mathbf{x}\right|^2. \tag{1.17}$$

The name given to $\hat{\mathbf{p}}_\mathcal{W}$ becomes apparent when comparing (1.17) with (1.8). Indeed, the former is almost equivalent to the latter with the caveat that the M signals considered in (1.17) are not independent. As a consequence, the variance decreases slower than $1/M$ with the number of windows, this being the rate found in Proposition 1 for the averaging of R independent signals. Moreover, the dependence between the different $\mathbf{x}_m$ introduces a distortion (bias) in the estimator. To state these effects more formally, we construct the dual operators associated with each window $\tilde{\mathbf{W}}_m := \mathbf{V}^H \text{diag}(\mathbf{w}_m)\mathbf{V}$ [cf. (1.16)], and use them to define the power *spectrum mixing* matrix of windows m and m' as the componentwise product $\tilde{\mathbf{W}}_{mm'} := \tilde{\mathbf{W}}_m \circ \tilde{\mathbf{W}}_{m'}^*$.

[3] To keep notation simple, in this subsection we use $\mathbf{x}$ to denote a realization of process $\mathbf{x}$.

Based on the spectrum mixing matrices, the following proposition presents the bias and covariance of $\hat{\mathbf{p}}_{\mathcal{W}}$.

Proposition 2. Let $\hat{\mathbf{p}}_{\mathcal{W}}$ be the windowed average periodogram computed based on a window bank $\mathcal{W} = \{\mathbf{w}_m\}_{m=1}^M$ and single observation $\mathbf{x}$ of a stationary process in $\mathbf{S}$. Then, the bias of $\hat{\mathbf{p}}_{\mathcal{W}}$ is given by

$$\mathbf{b}_{\mathcal{W}} := \mathbb{E}\left[\hat{\mathbf{p}}_{\mathcal{W}}\right] - \mathbf{p} = \left(\frac{1}{M}\sum_{m=1}^{M} \tilde{\mathbf{W}}_{mm} - \mathbf{I}\right)\mathbf{p}. \tag{1.18}$$

Furthermore, if $\mathbf{x}$ is Gaussian and $\mathbf{S}$ is symmetric, the trace of the covariance $\mathbf{\Sigma}_{\mathcal{W}} := \mathbb{E}\left[(\hat{\mathbf{p}}_{\mathcal{W}} - \mathbb{E}\left[\hat{\mathbf{p}}_{\mathcal{W}}\right])(\hat{\mathbf{p}}_{\mathcal{W}} - \mathbb{E}\left[\hat{\mathbf{p}}_{\mathcal{W}}\right])^H\right]$ is given by

$$\operatorname{tr}[\mathbf{\Sigma}_{\mathcal{W}}] = \frac{2}{M^2}\sum_{m=1,\, m'=1}^{M} \operatorname{tr}\left[(\tilde{\mathbf{W}}_{mm'}\mathbf{p})(\tilde{\mathbf{W}}_{mm'}\mathbf{p})^H\right]. \tag{1.19}$$

Expression (1.18) reveals that the bias of $\hat{\mathbf{p}}_{\mathcal{W}}$ is given by the discrepancy between the average spectrum mixing of the windows – depending on both the window silhouette $\mathbf{w}_m$ and the underlying graph through $\mathbf{V}$ – and the identity matrix. Notice that even if the individual spectrum mixing matrices $\tilde{\mathbf{W}}_{mm}$ are far from the identity, a small bias can still be achieved by controlling their average. The covariance expression in (1.19) can be further decomposed into a term akin to (1.14) plus another one that quantifies the added effect of dependency between the windowed signals; see [15] for more details. Furthermore, as done in (1.15), we can use Proposition 2 to obtain a closed-form expression for the MSE of $\hat{\mathbf{p}}_{\mathcal{W}}$ that can then guide the design criteria for optimal window banks. However, the associated optimization problems are non-convex. Although some basic developments in this area are presented in [15], the efficient design of optimal windows is still an open problem.

In Fig. 1.3(b) we illustrate the windowed average periodogram estimation for $M = 4$ random windows for a single observation of the random process on the Karate club network. Notice that the estimation is better than the one obtained by the (regular) periodogram, i.e., $M = 1$. In Fig. 1.3(c) we present the windowed average periodogram estimation for $M = 4$ but with an increasing number of observations. Notice that for a low number of observations ($R = 1$ and $R = 10$) the estimation improves that of the periodogram [cf. Fig. 1.3(a)]. Nonetheless, it can be seen that this estimator is biased since there is still a residual error even for large values of R.

1.3.2 PARAMETRIC PSD ESTIMATORS

A stationary graph process $\mathbf{x}$ can always be represented as the response of a graph filter $\mathbf{H}$ when applied to a white input [cf. Definition 1]. The cases where $\mathbf{H}$ depends on just a few parameters – much less than N – ultimately result in a further reduction of the degrees of freedom of the process $\mathbf{x}$. In particular, we may obtain a parametric description of the PSD of $\mathbf{x}$ as a function of the few coefficients of $\mathbf{H}$. In this section, we leverage this reduction in degrees of freedom to design PSD estimators. We

discuss in detail the case where $\mathbf{H}$ corresponds to a moving average (MA) model, and then briefly review the constructions for an autoregressive (AR) model. For the combined ARMA model, the developments for MA and AR processes can be mimicked; see [15] for more details on the ARMA model.

Moving average graph processes

Consider a vector of coefficients $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_{L-1}]^T$, for $L \ll N$, and assume that the stationary process $\mathbf{x}$ is generated as $\mathbf{x} = \mathbf{H}(\boldsymbol{\beta})\mathbf{n}$ where $\mathbf{n}$ is white and $\mathbf{H}(\boldsymbol{\beta}) = \sum_{l=0}^{L-1} \beta_l \mathbf{S}^l$. From this generative model it immediately follows that the covariance of $\mathbf{x}$ can be written as a function of $\boldsymbol{\beta}$, i.e., $\mathbf{C}_\mathbf{x}(\boldsymbol{\beta}) = \mathbf{H}(\boldsymbol{\beta})\mathbf{H}^H(\boldsymbol{\beta})$. Regarding the PSD of $\mathbf{x}$, from the definition in (1.4) we have that $\mathbf{p}(\boldsymbol{\beta}) = \text{diag}\left(\mathbf{V}^H \mathbf{C}_\mathbf{x}(\boldsymbol{\beta})\mathbf{V}\right)$, from where it follows that the PSD of $\mathbf{x}$ is equal to the squared magnitude of the frequency representation of the filter. The dependence of $\mathbf{C}_\mathbf{x}$ and $\mathbf{p}$ on $\boldsymbol{\beta}$ are explicitly stated below

$$\mathbf{C}_\mathbf{x}(\boldsymbol{\beta}) = \sum_{l=0,l'=0}^{L-1} (\beta_l \mathbf{S}^l)(\beta_{l'} \mathbf{S}^H)^{l'}, \qquad \mathbf{p}(\boldsymbol{\beta}) = |\tilde{\mathbf{h}}(\boldsymbol{\beta})|^2 = |\boldsymbol{\Psi}_L \boldsymbol{\beta}|^2. \tag{1.20}$$

The covariance and PSD expressions in (1.20) correspond to the natural graph counterparts of MA time processes generated by FIR filters; see [15] for discussions on the relevance of these processes.

The estimation of $\boldsymbol{\beta}$ can now be pursued in either the graph or frequency domain, through respectively, covariance or PSD fitting. More specifically, in the graph domain we compute the sample covariance $\hat{\mathbf{C}}_\mathbf{x}$ and use a matrix distortion function $D_\mathbf{C}(\hat{\mathbf{C}}_\mathbf{x}, \mathbf{C}_\mathbf{x}(\boldsymbol{\beta}))$ to measure the dissimilarity between $\hat{\mathbf{C}}_\mathbf{x}$ and $\mathbf{C}_\mathbf{x}(\boldsymbol{\beta})$. Alternatively, in the frequency domain, we compute the periodogram $\hat{\mathbf{p}}_{\text{pg}}$ as in (1.8) and use a vector distortion function $D_\mathbf{p}(\hat{\mathbf{p}}_{\text{pg}}, |\boldsymbol{\Psi}_L \boldsymbol{\beta}|^2)$ to compare the periodogram $\hat{\mathbf{p}}_{\text{pg}}$ with the PSD $|\boldsymbol{\Psi}_L \boldsymbol{\beta}|^2$. We then select the coefficients $\boldsymbol{\beta}$ that lead to the minimal distortion, as specified below, for either the graph or the frequency domain

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\text{argmin}}\ D_\mathbf{C}(\hat{\mathbf{C}}_\mathbf{x}, \mathbf{C}_\mathbf{x}(\boldsymbol{\beta})), \qquad \hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\text{argmin}}\ D_\mathbf{p}(\hat{\mathbf{p}}_{\text{pg}}, |\boldsymbol{\Psi}_L \boldsymbol{\beta}|^2). \tag{1.21}$$

Notice that both the functional forms of $\mathbf{C}_\mathbf{x}(\boldsymbol{\beta})$ and $\mathbf{p}(\boldsymbol{\beta})$ in (1.20) are indefinite quadratics in $\boldsymbol{\beta}$. Hence, the optimization problems in (1.21) will not be convex in general. In the particular case where the distortion $D_\mathbf{p}$ is given by the squared ℓ_2 norm of the difference, i.e., $D_\mathbf{p}(\hat{\mathbf{p}}_{\text{pg}}, |\boldsymbol{\Psi}_L \boldsymbol{\beta}|^2) = \|\hat{\mathbf{p}}_{\text{pg}} - |\boldsymbol{\Psi}_L \boldsymbol{\beta}|^2\|_2^2$, efficient (phase-retrieval) solvers with probabilistic guarantees are available [32, 33]. Alternative tractable formulations of (1.21) are discussed in [15], one of which is described next.

When $\mathbf{S}$ is symmetric, the expression (1.20) reduces to

$$\mathbf{C}_\mathbf{x} = \sum_{k=0}^{Q-1} b_k \mathbf{S}^k, \qquad p_n = \sum_{k=0}^{Q-1} b_k \lambda_n^k. \tag{1.22}$$

Here, $Q := \min\{2L-1, N\}$ unknown expansion coefficients $\{b_k\}_{k=0}^{Q-1}$ are collected in the vector $\mathbf{b} = [b_0, b_1, \ldots, b_{Q-1}]^T \in \mathbb{R}^Q$. By ignoring the structure in $\mathbf{b}$, i.e., the relation between $\mathbf{b}$ and $\boldsymbol{\beta}$, we arrive at a linear parametrization of $\mathbf{C_x}$ using the set of Q symmetric matrices $\{\mathbf{S}^0, \mathbf{S}, \ldots, \mathbf{S}^{Q-1}\}$ as a basis. Vectorizing $\mathbf{C_x}$ in (1.22), we obtain

$$\mathbf{c_x} = \text{vec}(\mathbf{C_x}) = \sum_{k=0}^{Q-1} b_k \text{vec}(\mathbf{S}^q) = \mathbf{G}_{\text{ma}}\mathbf{b}, \tag{1.23}$$

where we implicitly defined the matrix $\mathbf{G}_{\text{ma}} := \left[\text{vec}(\mathbf{S}^0), \cdots, \text{vec}(\mathbf{S}^{Q-1})\right]$. Since $\mathbf{C_x}$ depends linearly on $\mathbf{b}$ – as opposed to quadratically on $\boldsymbol{\beta}$ – we may efficiently solve (1.21) for some choices of $D_{\mathbf{C}}$. For example, the LS estimate of $\mathbf{b}$ is given by $\hat{\mathbf{b}} = (\mathbf{G}_{\text{ma}}^H \mathbf{G}_{\text{ma}})^{-1} \mathbf{G}_{\text{ma}}^H \hat{\mathbf{c}}_{\mathbf{x}}$. We illustrate the implementation of this relaxation in Fig. 1.3(d). Notice that the PSD estimation is quite faithful even for $R = 1$, where the shape of the PSD is captured but the scale is missed. The estimation slightly improves for $R = 10$.

Autoregressive graph processes

A stationary process can be better and better approximated as a MA process by increasing the order of the associated FIR filter. However, the merits of the parametric estimators depend on having a small number of parameters describing the generating process. For some stationary processes, an AR model using an infinite impulse response filter leads to a more parsimonious description. For example, consider the diffusion process driven by the graph filter $\mathbf{H} = \sum_{l=0}^{\infty} \alpha^l \mathbf{S}^l$, where α represents the diffusion rate. For small enough α, the filter can be rewritten as $\mathbf{H} = (\mathbf{I} - \alpha\mathbf{S})^{-1}$, with frequency response $\tilde{\mathbf{h}} = \text{diag}(\mathbf{I} - \alpha\boldsymbol{\Lambda}^{-1})$. Thus, $\mathbf{H}$ can be viewed as a single-pole AR filter, leading to a more meager description. More generally, an AR filter of order M can be described as $\mathbf{H} = \alpha_0 \prod_{m=1}^{M} (\mathbf{I} - \alpha_m \mathbf{S})^{-1}$ for some vector of parameters $\boldsymbol{\alpha} = [\alpha_0, \ldots, \alpha_M]^T$. Correspondingly, the frequency response of this filter is given by $\tilde{\mathbf{h}} = \alpha_0 \,\text{diag}(\prod_{m=1}^{M} (\mathbf{I} - \alpha_m \boldsymbol{\Lambda})^{-1})$. If we define the graph process $\mathbf{x} = \mathbf{H}\mathbf{n}$ with $\mathbf{n}$ white, we may leverage the previous expressions to obtain explicit formulas for the covariance and PSD of $\mathbf{x}$ as a function of the parameters $\boldsymbol{\alpha}$,

$$\mathbf{C_x}(\boldsymbol{\alpha}) = \alpha_0^2 \prod_{m=1}^{M} (\mathbf{I} - \alpha_m \mathbf{S})^{-1} (\mathbf{I} - \alpha_m \mathbf{S})^{-H}, \qquad \mathbf{p}(\boldsymbol{\alpha}) = \alpha_0^2 \,\text{diag}\Big(\prod_{m=1}^{M} |\mathbf{I} - \alpha_m \boldsymbol{\Lambda}|^{-2} \Big). \tag{1.24}$$

The mechanism to obtain the corresponding parametric PSD estimator is equivalent to the one explained for MA processes, where $\mathbf{C_x}(\boldsymbol{\beta})$ and $\mathbf{p}(\boldsymbol{\beta})$ in (1.20) are replaced by $\mathbf{C_x}(\boldsymbol{\alpha})$ and $\mathbf{p}(\boldsymbol{\alpha})$ in (1.24). The associated optimization problems [cf. (1.21)] will be non-convex in general and become intractable for large orders M.

Yule-Walker schemes [31, Sec. 3.4] tailored to graph signals may be of help, as discussed next. The all-pole filter $\mathbf{H}^{-1}(\boldsymbol{\alpha}) = \prod_{k=1}^{M} (\mathbf{I} - \alpha_k \mathbf{S})$ can be alternatively expressed as $\mathbf{H}^{-1}(\mathbf{a}) = \mathbf{I} - \sum_{k=1}^{M} a_k \mathbf{S}^k$, where $\mathbf{a} = [a_1, a_2, \ldots, a_M]^T$. Thus, the AR signal

satisfies the equations

$$\mathbf{x} = \sum_{k=1}^{M} a_k \mathbf{S}^k \mathbf{x} + \mathbf{n}. \tag{1.25}$$

In other words, the graph signal $\mathbf{x}$ depends *linearly* on the M shifted graph signals $\{\mathbf{S}^k\mathbf{x}\}_{k=1}^{M}$ according to the above AR model. As a result, the covariance matrix of $\mathbf{x}$ and its vectorized form can be expressed as

$$\mathbf{C}_\mathbf{x} = \sum_{k=1}^{M} a_k \mathbf{S}^k \mathbf{C}_\mathbf{x} + \mathbf{C}_\mathbf{nx}, \quad \mathbf{c}_\mathbf{x} = \mathrm{vec}(\mathbf{C}_\mathbf{x}) = \sum_{k=1}^{M} a_k \mathrm{vec}(\mathbf{S}^k \mathbf{C}_\mathbf{x}) + \mathrm{vec}(\mathbf{C}_\mathbf{nx}) \approx \mathbf{G}_{\mathrm{ar}}\mathbf{a}, \tag{1.26}$$

where we have defined $\mathbf{G}_{\mathrm{ar}} := [\mathrm{vec}(\mathbf{S}\mathbf{C}_\mathbf{x}), \ldots, \mathrm{vec}(\mathbf{S}^M\mathbf{C}_\mathbf{x})]$ and where we have assumed that $\mathbf{C}_\mathbf{nx} = \mathbb{E}[\mathbf{n}\mathbf{x}^H]$ is a small error term. Note that in contrast to the previous linear equations for the nonparametric (1.11) and MA (1.23) models, the system matrix $\mathbf{G}_{\mathrm{ar}}$ now explicitly depends on the unknown covariance $\mathbf{C}_\mathbf{x}$. Still, when the sample covariance matrix $\hat{\mathbf{C}}_\mathbf{x}$ is available, we can solve (1.26) through LS as $\hat{\mathbf{a}} = (\hat{\mathbf{G}}_{\mathrm{ar}}^H \hat{\mathbf{G}}_{\mathrm{ar}})^{-1} \hat{\mathbf{G}}_{\mathrm{ar}}^H \hat{\mathbf{c}}_\mathbf{x}$, where $\hat{\mathbf{G}}_{\mathrm{ar}}$ is defined as $\mathbf{G}_{\mathrm{ar}}$ replacing $\mathbf{C}_\mathbf{x}$ by $\hat{\mathbf{C}}_\mathbf{x}$.

1.4 NODE SUBSAMPLING FOR PSD ESTIMATION

Compression or data reduction is preferred for large-scale graph processes as the size of the datasets often inhibits a direct computation of the second-order statistics. In this section, we focus on recovering the second-order statistics of stationary graph processes from subsampled graph signals. We refer to this problem as *graph covariance sampling* [19].

The fact that we reconstruct the power spectrum, instead of the graph signal itself, enables us to sparsely sample the nodes, even in the absence of any spectral priors such as smoothness, sparsity, or bandlimitedness with known support. The proposed concept basically generalizes the field of compressive covariance sensing [18] to the graph setting, which is not trivial. This is because for weakly stationary signals with a regular support or signals supported on a circulant graph, the covariance matrix has a clear structure (e.g., Toeplitz, circulant) that enables an elegant subsampler design, but for second-order stationary graph signals residing on arbitrary graphs, the covariance matrix in general does not admit *any* clear structure that can be easily exploited.

1.4.1 THE SAMPLING PROBLEM

Consider the problem of estimating the graph power spectrum of the weakly stationary graph signal $\mathbf{x} \in \mathbb{R}^N$ from a set of $K \ll N$ linear observations stacked in $\mathbf{y} \in \mathbb{R}^K$,

given by

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x}, \tag{1.27}$$

where $\mathbf{\Phi}$ is a known $K \times N$ selection matrix with Boolean entries, i.e., $\mathbf{\Phi} \in \{0, 1\}^{K\times N}$ and where several realizations of $\mathbf{y}$ may be available. The matrix $\mathbf{\Phi}$ is referred to as the *subsampling* or *sparse sampling* matrix. Such a sparse sampling scheme generally results in a reduction in the storage and processing costs. Moreover, for applications where nodes correspond to sensing devices – such as weather stations in climatology and electroencephalography probes in brain networks – it also leads to smaller hardware and communications costs.

The covariance matrices $\mathbf{C}_\mathbf{x} = \mathbb{E}[\mathbf{x}\mathbf{x}^H] \in \mathbb{R}^{N\times N}$ and $\mathbf{C}_\mathbf{y} = \mathbb{E}[\mathbf{y}\mathbf{y}^H] \in \mathbb{R}^{K\times K}$ contain the second-order statistics of $\mathbf{x}$ and $\mathbf{y}$, respectively. In practice, a sample covariance matrix is computed based on R signal observations. More precisely, suppose that R observations of the uncompressed and compressed graph signals are available, denoted by the vectors $\{\mathbf{x}_r\}_{r=1}^R$ and $\{\mathbf{y}_r\}_{r=1}^R$, respectively. Then forming the sample covariance matrix, $\hat{\mathbf{C}}_\mathbf{x} = (1/R)\sum_{r=1}^R \mathbf{x}_r\mathbf{x}_r^H$, from R snapshots of $\mathbf{x}$ costs $O(N^2R)$, while forming the sample covariance matrix, $\hat{\mathbf{C}}_\mathbf{y} = (1/R)\sum_{r=1}^R \mathbf{y}_r\mathbf{y}_r^H$, from R snapshots of $\mathbf{y}$ *only* costs $O(K^2R)$. Therefore, when $K \ll N$, there will clearly be a significant reduction in the storage and processing costs due to compression.

1.4.2 COMPRESSED LS ESTIMATOR

In this section, we will extend the previously derived LS estimators (for nonparametric as well as parametric PSD estimation) to the case where only compressed graph signals are available. The reason we only focus on those estimators is not because they lead to the best performance, but because they can be used to design the best subset of nodes to sample.

Let us condense the linearly structured covariance matrix $\mathbf{C}_\mathbf{x}$ for the nonparametric case (see (1.10)), the parametric MA case with symmetric shifts (see (1.22)), and the parametric AR case (see (1.26)), in a single expression as

$$\mathbf{C}_\mathbf{x}(\boldsymbol{\theta}) = \sum_{i=1}^{L} \theta_i \mathbf{Q}_i; \quad \boldsymbol{\theta} = [\theta_1, \ldots, \theta_L]^T, \tag{1.28}$$

where for the nonparametric case, we have $L = N$, $\boldsymbol{\theta} := \mathbf{p}$, and $\mathbf{Q}_i := \mathbf{v}_i\mathbf{v}_i^H$, for the MA case with symmetric shifts, we have $L = Q$, $\boldsymbol{\theta} := \mathbf{b}$, and $\mathbf{Q}_i := \mathbf{S}^{i-1}$, and for the AR case, we have $L = M$, $\boldsymbol{\theta} := \mathbf{a}$, and $\mathbf{Q}_i := \mathbf{S}^{i-1}\mathbf{C}_\mathbf{x}$.

Using the compression scheme described in (1.27), the covariance matrix $\mathbf{C}_\mathbf{y}$ of the subsampled graph signal $\mathbf{y}$ can be related to $\mathbf{C}_\mathbf{x}$ as

$$\mathbf{C}_\mathbf{y}(\boldsymbol{\theta}) = \mathbf{\Phi}\mathbf{C}_\mathbf{x}\mathbf{\Phi}^T = \sum_{i=1}^{L} \theta_i \mathbf{\Phi}\mathbf{Q}_i\mathbf{\Phi}^T. \tag{1.29}$$

This means that the expansion coefficients of $\mathbf{C}_\mathbf{y}$ with respect to the set $\{\mathbf{\Phi}\mathbf{Q}_1\mathbf{\Phi}^T, \cdots,$

$\mathbf{\Phi}\mathbf{Q}_L\mathbf{\Phi}^T\}$ are the *same* as those of $\mathbf{C_x}$ with respect to the set $\{\mathbf{Q}_1, \cdots, \mathbf{Q}_L\}$, and they are preserved under linear compression. It is not clear at this point whether these expansion coefficients, which basically characterize the graph power spectrum, can be uniquely recovered from $\mathbf{C_y}(\boldsymbol{\theta})$.

Vectorizing $\mathbf{C_y}$ as $\mathbf{c_y} = \text{vec}(\mathbf{C_y}) = (\mathbf{\Phi} \otimes \mathbf{\Phi})\text{vec}(\mathbf{C_x}) \in \mathbb{R}^{K^2}$ we obtain

$$\mathbf{c_y} = (\mathbf{\Phi} \otimes \mathbf{\Phi})\mathbf{G}\boldsymbol{\theta}, \tag{1.30}$$

where $\mathbf{G} = [\text{vec}(\mathbf{Q}_1), \ldots, \text{vec}(\mathbf{Q}_L)]$. When only a finite number of observations are available, we use the compressed sample data covariance matrix $\hat{\mathbf{C}}_\mathbf{y}$ instead of $\mathbf{C_y}$, leading to the approximation $\hat{\mathbf{c}}_\mathbf{y} \approx (\mathbf{\Phi} \otimes \mathbf{\Phi})\mathbf{G}\boldsymbol{\theta}$.

The parameter $\boldsymbol{\theta}$ is identifiable from this system of equations if $(\mathbf{\Phi} \otimes \mathbf{\Phi})\mathbf{G}$ has full column rank, which requires $K^2 \geq L$. Assuming that this is the case, the graph power spectrum (thus the second-order statistics of $\mathbf{x}$) can be estimated in closed form via LS as

$$\hat{\boldsymbol{\theta}} = [(\mathbf{\Phi} \otimes \mathbf{\Phi})\mathbf{G}]^\dagger \hat{\mathbf{c}}_\mathbf{y}. \tag{1.31}$$

It can be shown that a full row rank (wide) matrix $\mathbf{\Phi} \in \mathbb{R}^{K \times N}$ yields a full column rank matrix $(\mathbf{\Phi} \otimes \mathbf{\Phi})\mathbf{G}$ if and only if the matrix $(\mathbf{\Phi} \otimes \mathbf{\Phi})\mathbf{G}$ is tall, i.e., $K^2 \geq L$, and $\text{null}(\mathbf{\Phi} \otimes \mathbf{\Phi}) \cap \text{range}(\mathbf{G}) = \{\mathbf{0}\}$. When this is the case, we can recover the graph power spectrum by observing *only* $O(\sqrt{L})$ nodes.

An important remark is required at this point with respect to the parametric AR model. Note from (1.26) that in this case the matrix $\mathbf{G}$ depends itself on the *uncompressed* covariance matrix $\mathbf{C_x}$, which is unknown. Hence, (1.31) cannot be directly applied. One option is to simply assume we roughly know it and although this is not going to lead to a good estimate, it might be good enough for designing a suboptimal sampling scheme (see Sec. 1.4.3). Another option is to restrict ourselves to particular subsampling schemes that preserve the linear structure in (1.26) but for the compressed data instead of the uncompressed data; see [19] for more details.

1.4.3 SPARSE SAMPLER DESIGN

We have seen so far that the design of the subsampling matrix $\mathbf{\Phi}$ is crucial for the reconstruction of the covariance of the random graph process. In this subsection, we design a sparse subsampling matrix $\mathbf{\Phi}$ to ensure that the observation matrix $(\mathbf{\Phi} \otimes \mathbf{\Phi})\mathbf{G}$ has full column rank and the solution for $\boldsymbol{\theta}$ has a small error.

Consider a structured sparse sampling matrix $\mathbf{\Phi}(\mathbf{z}) \in \{0, 1\}^{K \times N}$, such that the entries of this matrix are determined by a binary *component selection* vector $\mathbf{z} = [z_1, \cdots, z_N]^T \in \{0, 1\}^N$, where $z_i = 1$ indicates that the ith node is selected by $\mathbf{\Phi}(\mathbf{z})$.

Uniqueness and sensitivity of the LS solution developed in the previous subsection depends on the spectrum (i.e., the set of eigenvalues) of the matrix

$$\mathbf{T}(\mathbf{z}) = [(\mathbf{\Phi}(\mathbf{z}) \otimes \mathbf{\Phi}(\mathbf{z}))\mathbf{G}]^T [(\mathbf{\Phi}(\mathbf{z}) \otimes \mathbf{\Phi}(\mathbf{z}))\mathbf{G}] = \mathbf{G}^T \left(\text{diag}(\mathbf{z}) \otimes \text{diag}(\mathbf{z})\right) \mathbf{G}.$$

More specifically, the performance of LS is better if the spectrum of the matrix $(\mathbf{\Phi} \otimes$

Algorithm 1 Greedy algorithm

1. **Require** $\mathcal{X} = \emptyset, K$.
2. **for** $k = 1$ to K
3. $\quad s^* = \underset{s \notin \mathcal{X}}{\text{argmax}}\ f(\mathcal{X} \cup \{s\})$
4. $\quad \mathcal{X} \leftarrow \mathcal{X} \cup \{s^*\}$
5. **end**
6. **Return** $\mathcal{X}$

Φ)G is more uniform [34]. Thus, a sparse sampler $\mathbf{z}$ can be obtained by solving

$$\underset{\mathbf{z}\in\{0,1\}^N}{\text{argmax}}\ f(\mathbf{z}) \qquad \text{s.to } \|\mathbf{z}\|_0 = K, \tag{1.32}$$

with either $f(\mathbf{z}) = -\operatorname{tr}[\mathbf{T}^{-1}(\mathbf{z})]$, $f(\mathbf{z}) = \lambda_{\min}(\mathbf{T}(\mathbf{z}))$, or $f(\mathbf{z}) = \log\det[\mathbf{T}(\mathbf{z})]$. These functions balance the spectrum of $\mathbf{T}(\mathbf{z})$.

Although the above problem can be solved using standard convex relaxation techniques [35], due to the involved complexity of solving the relaxed convex problem and keeping in mind the large scale problems that might arise in the graph setting, we will now focus on the optimization problem (1.32) with $f(\mathbf{z}) = \log\det[\mathbf{T}(\mathbf{z})]$ as it can be solved near-optimally using a low-complexity greedy algorithm. To do so, we introduce the concept of submodularity, a notion based on the property of diminishing returns. This is useful for solving discrete combinatorial optimization problems of the form (1.32) (see e.g., [36]). Submodularity can be formally defined as follows.

Definition 6. Given two sets $\mathcal{X}$ and $\mathcal{Y}$ such that for every $\mathcal{X} \subseteq \mathcal{Y} \subseteq \mathcal{N}$ and $s \in \mathcal{N}\backslash\mathcal{Y}$, the set function $f : 2^N \to \mathbb{R}$ defined on the subsets of $\mathcal{N}$ is said to be submodular, if it satisfies $f(\mathcal{X} \cup \{s\}) - f(\mathcal{X}) \geq f(\mathcal{Y} \cup \{s\}) - f(\mathcal{Y})$.

Suppose the submodular function is monotone nondecreasing, i.e., $f(\mathcal{X}) \leq f(\mathcal{Y})$ for all $\mathcal{X} \subseteq \mathcal{Y} \subseteq \mathcal{N}$ and normalized, i.e., $f(\emptyset) = 0$, then a greedy maximization of such a function as summarized in Algorithm 1 is *near optimal* with an approximation factor of $(1 - 1/e)$; see [37].

To use this framework, we have to rewrite $f(\mathbf{z}) = \log\det[\mathbf{T}(\mathbf{z})]$ as a set function:

$$f(\mathcal{X}) = \log\det\left[\sum_{(i,j)\in\mathcal{X}\times\mathcal{X}} \mathbf{g}_{i,j}\mathbf{g}_{i,j}^T\right], \tag{1.33}$$

where the index set $\mathcal{X}$ is related to the component selection vector $\mathbf{z}$ as $\mathcal{X} = \{m \,|\, z_m = 1, m = 1, \ldots, N\}$ and the column vectors $\{\mathbf{g}_{1,1}, \mathbf{g}_{1,2}, \cdots, \mathbf{g}_{N,N}\}$ correspond to the rows of $\mathbf{G}$ as $\mathbf{G} = [\mathbf{g}_{1,1}, \mathbf{g}_{1,2}, \cdots, \mathbf{g}_{N,N}]^T$. We use such an indexing because the sampling matrix $\mathbf{\Phi} \otimes \mathbf{\Phi}$ results in a Kronecker structured (row) subset selection.

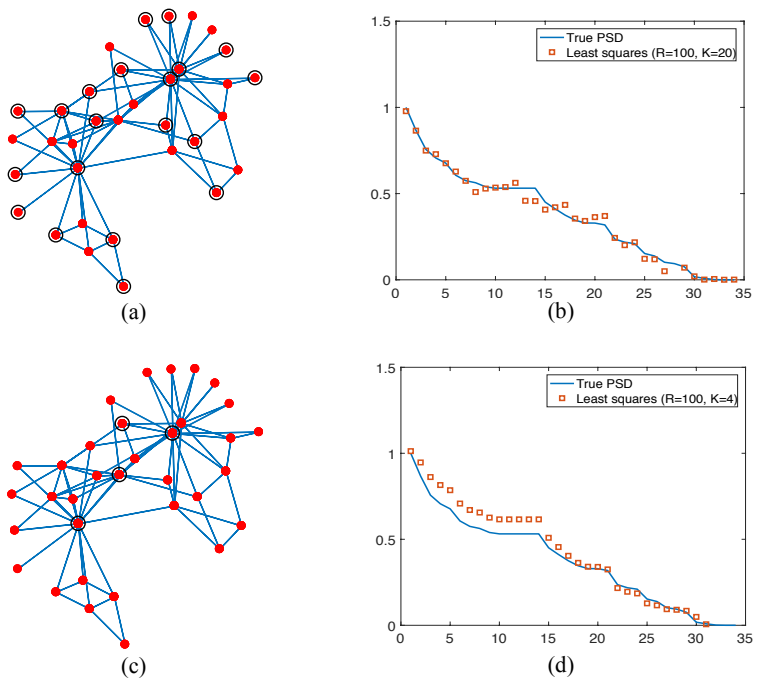

FIGURE 1.4 PSD estimation from a subset of nodes

Estimators are based on a random process defined on the Karate club network [28]. (a) Graph sampling for nonparametric PSD estimation. Here, 20 out of 34 nodes are observed. The sampled nodes are highlighted by the circles around the nodes. (b) Nonparametric PSD estimation based on observations from 20 nodes and 100 data snapshots. (c) Graph sampling for parametric MA PSD estimation. Here, 4 out of 34 nodes are observed. (d) Parametric MA PSD estimation based on observations from 4 nodes and 100 data snapshots.

Modifying this set function slightly to

$$f(\mathcal{X}) = \log\det\left[\sum_{(i,j)\in\mathcal{X}\times\mathcal{X}} \mathbf{g}_{i,j}\mathbf{g}_{i,j}^T + \epsilon\mathbf{I}\right] - N\log\epsilon, \tag{1.34}$$

we obtain a normalized, nondecreasing, submodular function on the set $\mathcal{X} \subset \mathcal{N}$. Here, $\epsilon > 0$ is a small constant. In (1.34), $\epsilon\mathbf{I}$ is needed to carry out the first few iterations of Algorithm 1 and $-N\log\epsilon$ ensures that $f(\emptyset)$ is zero. It is worth mentioning that the greedy algorithm is linear in K, while computing (1.34) dominates the computational complexity. Finally, random subsampling (i.e., $\mathbf{z}$ having random 0 or 1 entries) is not suitable as it might not always result in a full-column rank model matrix.

In Fig. 1.4, we illustrate the PSD estimation based on the observations from a subset of nodes for 100 realizations of the random process on the Karate club network. For the nonparametric model, the selected graph nodes obtained from Algorithm 1

are indicated with a black circle in Fig. 1.4(a). Based on the observations from these 20 selected graph nodes, the PSD estimate obtained using LS is shown in Fig. 1.4(b). It can be seen that the PSD estimate based on the observations from a subset of nodes fits reasonably well to the true PSD.

For the parametric MA model, wherein the PSD is parametrized with $Q = 7$ MA parameters, the selected graph nodes obtained from Algorithm 1 are shown in Fig. 1.4(c) and the reconstructed PSD using LS is shown in Fig. 1.4(d). In this case, we sample only 4 out of 34 graph nodes, and yet obtain a PSD estimate that fits very well to the true PSD.

1.5 DISCUSSION AND THE ROAD AHEAD

In this chapter, we have introduced the concept of weakly stationary graph processes and their related power spectral density. We discussed the links between the different definitions as well as the relations with classical signal processing. Furthermore, we extended this idea to processes that are jointly stationary in the vertex and time domain, where the subclass of separable processes is of particular importance due to their more parsimonious description. The chapter has also focused on estimating the PSD and the covariance using non-parametric as well as parametric methods. Equivalences and differences with classical PSD techniques for spatiotemporal signals have been established. Finally, we presented different techniques to estimate the PSD and the covariance from only a subset of the nodes, without any loss of identifiability. This can be viewed as a particular instance of sparse covariance sampling. In this context, we also proposed a greedy method to select the best nodes to sample in order to guarantee a satisfying estimation of the PSD and the covariance.

While this chapter only covers weakly stationary graph processes, a definition of strict stationarity is still open. One option could be to define a strictly stationary graph process as the output of filtering i.i.d. noise. Ergodicity is also a concept that we did not discuss in this chapter. Ergodicity in a graph signal processing context would mean that the statistics of the graph process could be derived from successive graph shifts of a single realization (observed at one or multiple nodes) [38]. Due to the finite length of graph signals, this will entail certain problems and exact estimates of the statistics (even asymptotically) will rarely be possible. How to model non-stationary graph processes in an intuitively pleasing way is another unexplored area. A way forward in this direction could be the introduction of so-called node-varying graph filters [39], where the variation of the filter taps can be expanded in a particular basis. Filtering white or i.i.d. noise using such filters leads to a non-stationary graph process that is parametrized by a limited number of coefficients. Yet, other parametrized graph filter structures could be employed as a model for non-stationary graph processes, e.g., edge-variant graph filters [40] or median graph filters [41, 42]. Finally, in this chapter, we limited ourselves to normal graph shift operators that are endowed with a unitary matrix of eigenvectors. Stationarity for non-normal graph

shifts (whether diagonalizable or not) is a topic for future research. Some of the concepts discussed in this chapter can be easily extended to non-orthonormal and/or generalized eigenvectors but others require more research.

REFERENCES

1. Zhang F, Hancock ER, Graph spectral image smoothing using the heat kernel. Pattern Recognition 2008; 41(11):3328–3342.
2. Pesenson I, Sampling in Paley-Wiener spaces on combinatorial graphs. Trans of the American Mathematical Society 2008; 360(10):5603–5627.
3. Chen S, Sandryhaila A, Moura J, Kovačević J, Signal recovery on graphs: Variation minimization. IEEE Trans Signal Process 2015; 63(17):4609–4624.
4. Chen S, Varma R, Sandryhaila A, Kovačević J, Discrete signal processing on graphs: Sampling theory. IEEE Trans Signal Process 2015; 63(24):6510–6523.
5. Anis A, Gadde A, Ortega A, Towards a sampling theorem for signals on arbitrary graphs. In: IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP), 2014, pp. 3864–3868.
6. Marques AG, Segarra S, Leus G, Ribeiro A, Sampling of graph signals with successive local aggregations. IEEE Trans Signal Process 2016; 64(7):1832–1843.
7. Segarra S, Marques AG, Leus G, Ribeiro A, Reconstruction of graph signals through percolation from seeding nodes. IEEE Trans Signal Process 2016; 64(16):4363 – 4378.
8. Shuman DI, Vandergheynst P, Frossard P, Distributed signal processing via Chebyshev polynomial approximation. CoRR 2011; abs/1111.5239.
9. Sandryhaila A, Moura JMF, Discrete signal processing on graphs: Frequency analysis. IEEE Trans Signal Process 2014; 62(12):3042–3054.
10. Shi X, Feng H, Zhai M, Yang T, Hu B, Infinite impulse response graph filters in wireless sensor networks. IEEE Signal Processing Letters 2015; 22(8):1113–1117.
11. Isufi E, Loukas A, Simonetto A, Leus G, Autoregressive moving average graph filtering. IEEE Trans Signal Process 2017; 65(2):274–288.
12. Hayes MH, Statistical Digital Signal Processing and Modeling. John Wiley and Sons, 2009.
13. Girault B, Stationary graph signals using an isometric graph translation. In: European Signal Process. Conf. (EUSIPCO), 2015, pp. 1516–1520.
14. Girault B, Gonçalves P, Fleury E, Translation on graphs: An isometric shift operator. IEEE Signal Process Lett 2015; 22(12):2416–2420.
15. Marques AG, Segarra S, Leus G, Ribeiro A, Stationary graph processes and spectral estimation. IEEE Trans Signal Process 2017; 65(22):5911–5926.
16. Perraudin N, Vandergheynst P, Stationary signal processing on graphs. IEEE Trans Signal Process 2017; 65(13):3462–3477.
17. Perraudin N, Loukas A, Grassi F, Vandergheynst P, Towards stationary time-vertex signal processing. In: IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP), 2017, pp. 3914–3918.
18. Romero D, Ariananda DD, Tian Z, Leus G, Compressive covariance sensing: Structure-based compressive sensing beyond sparsity. IEEE Signal Process Mag 2016; 33(1):78–93.
19. Chepuri SP, Leus G, Graph sampling for covariance estimation. IEEE Trans Signal and Info Process over Networks 2017; 3(3):451–466.
20. Sandryhaila A, Moura JMF, Discrete signal processing on graphs. IEEE Trans Signal Process 2013; 61(7):1644–1656.

21. Gavili A, Zhang XP, On the shift operator, graph frequency and optimal filtering in graph signal processing. IEEE Trans Signal Process 2017; 65(23):6303–6318.
22. Shuman DI, Ricaud B, Vandergheynst P, Vertex-frequency analysis on graphs. Applied and Computational Harmonic Analysis 2016; 40(2):260–291.
23. Segarra S, Marques AG, Mateos G, Ribeiro A, Network topology inference from spectral templates. IEEE Trans Signal and Info Process over Networks 2017; 3(3):467–483.
24. Murphy KP, Machine Learning: A Probabilistic Perspective. MIT press, 2012.
25. Imrich W, Klavzar S, Product graphs: Structure and recognition. Wiley, 2000.
26. Sandryhaila A, Moura J, Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure. IEEE Signal Process Mag 2014; 31(5):80–90. doi:10.1109/MSP.2014.2329213.
27. Lütkepohl H, New introduction to multiple time series analysis. Springer, 2007.
28. Zachary WW, An information flow model for conflict and fission in small groups. J Anthropol Res 1977; 33(4):pp. 452–473.
29. Hayes MH, Statistical Digital Signal Processing and Modeling. John Wiley & Sons, 2009.
30. Scharf LL, Statistical signal processing. Reading, MA, USA: Addison-Wesley, 1991.
31. Stoica P, Moses RL, Spectral Analysis of Signals. Pearson/Prentice Hall Upper Saddle River, NJ, 2005.
32. Fienup JR, Phase retrieval algorithms: A comparison. Applied optics 1982; 21(15):2758–2769.
33. Candes EJ, Li X, Soltanolkotabi M, Phase retrieval via Wirtinger flow: Theory and algorithms. IEEE Trans Inf Theory 2015; 61(4):1985–2007.
34. Golub GH, Van Loan CF, Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences, Baltimore, MD, USA: Johns Hopkins Univ. Press, 1996.
35. Chepuri SP, Leus G, Sparse sensing for statistical inference. Foundations and Trends® in Signal Processing 2016; 9(3–4):233–368.
36. Krause A, Optimizing sensing: Theory and applications. Ph.D. dissertation, School of Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, United States, 2008.
37. Nemhauser GL, Wolsey LA, Fisher ML, An analysis of approximations for maximizing submodular set functions— I. Mathematical Programming 1978; 14(1):265–294.
38. Gama F, Ribeiro A, Weak law of large numbers for stationary graph processes. In: IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP), 2017, pp. 4124–4128.
39. Segarra S, Marques AG, Ribeiro A, Optimal graph-filter design and applications to distributed linear network operators. IEEE Trans Signal Process 2017; 65(15):4117–4131.
40. Coutino M, Isufi E, Leus G, Distributed edge-variant graph filters. In: IEEE Intl. Wrksp. Computational Adv. in Multi-Sensor Adaptive Process. (CAMSAP), 2017.
41. Segarra S, Marques AG, Arce GR, Ribeiro A, Center-weighted median graph filters. In: Global Conf. Signal and Info. Process. (GlobalSIP), 2016, pp. 336–340.
42. Segarra S, Marques AG, Arce G, Ribeiro A, Design of weighted median graph filters. In: IEEE Intl. Wrksp. Computational Adv. in Multi-Sensor Adaptive Process. (CAMSAP), 2017.