



Delft University of Technology

Document Version

Final published version

Licence

Dutch Copyright Act (Article 25fa)

Citation (APA)

Liu, C., Tenorio, V. M., Marques, A. G., & Isufi, E. (2026). Matched Topological Subspace Detector. *IEEE Transactions on Signal Processing*, 74, 1000-1014. <https://doi.org/10.1109/TSP.2026.3656668>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Matched Topological Subspace Detector

Chengen Liu , *Graduate Student Member, IEEE*, Victor M. Tenorio , *Graduate Student Member, IEEE*, Antonio G. Marques , *Senior Member, IEEE*, and Elvin Isufi , *Senior Member, IEEE*

Abstract—Topological spaces capture richer relationships than graphs by modeling interactions not only between nodes but also among higher-order entities, such as edges or triangles. This motivates the representation of information defined in irregular domains as topological signals. We focus on simplicial complexes, which are collections of simplices (nodes, edges, triangles and so on) adhering to the inclusion property: a higher-order element cannot exist unless its lower-order components are also part of the complex (e.g., a triangle requires its edges). Accordingly, simplicial signals are functions defined over this set of simplices. By leveraging the spectral dualities of Hodge and Dirac theory, practical simplicial signals often concentrate in specific spectral subspaces (e.g., gradient or curl). For instance, in a foreign currency exchange network, the exchange flow signals typically satisfy the arbitrage-free condition and hence are curl-free. However, the presence of anomalies can disrupt these conditions, causing the signals to deviate from such subspaces. In this work, we formulate a hypothesis testing framework to detect whether simplicial complex signals lie in specific subspaces in a principled and tractable manner. Concretely, we propose Neyman-Pearson matched topological subspace detectors for signals defined at a single simplicial level (such as edges) or jointly across all levels of a simplicial complex. The (energy-based projection) proposed detectors handle missing values, provide closed-form performance analysis, and effectively capture the unique topological properties of the data. We demonstrate the effectiveness of the proposed topological detectors on various real-world data, including foreign currency exchange networks.

Index Terms—Simplicial signal processing, detection theory, topological signal processing, matched subspace detection.

Received 30 March 2025; revised 2 December 2025; accepted 9 January 2026. Date of publication 21 January 2026; date of current version 13 March 2026. This work was supported in part by the NWO under Grant Veni (222.032) financed by the Netherlands Organization for Scientific Research; in part by the Spanish AEI (AEI/10.13039/501100011033) under Grant PID2022-136887NB-I00, Grant PID2023-149457OB-I00, and Grant FPU20/05554; in part by the Community of Madrid via IDEA-CM under Grant TEC-2024/COM-89; in part by the Ellis Madrid Unit; and in part by the EU H2020 Grant Tailor (No 952215, agreements 76 and 82). The work of Chengen Liu was supported by China Scholarship Council. An earlier version of this paper was presented at the 32nd European Signal Processing Conference (EUSIPCO), 2024 [DOI: 10.23919/EUSIPCO63174.2024.10714960]. The associate editor coordinating the review of this article and approving it for publication was Qing Ling. (*Corresponding author: Chengen Liu.*)

Chengen Liu and Elvin Isufi are with the Faculty of Electrical Engineering, Mathematics and Computer Science, Department of Intelligent Systems, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: c.liu-15@tudelft.nl; e.isufi-1@tudelft.nl).

Victor M. Tenorio and Antonio G. Marques are with the Department of Signal Theory and Communications, King Juan Carlos University, 28933 Madrid, Spain (e-mail: victor.tenorio@urjc.es; antonio.garcia.marques@urjc.es).

Digital Object Identifier 10.1109/TSP.2026.3656668

I. INTRODUCTION

TOPOLOGICAL signals, such as those arising in simplicial complexes [2], [3], encode a more nuanced structure compared to graph signals by supporting multiway relationships among higher-order elements. A simplicial complex is a mathematical structure composed of nodes, edges, triangles, and their higher-order counterparts. Each higher-order simplex includes all its lower-order faces. Simplicial signals are functions defined on these simplices, assigning values to nodes (0-simplices), edges (1-simplices), triangles (2-simplices), or higher-order simplices, enabling the representation of complex interactions in irregular domains. Graph signals primarily focus on pairwise interactions between nodes, whereas topological signals can represent interactions among multiple entities simultaneously. In financial markets, for example, transactions may involve more than two companies at a time, and in protein molecules, the functional relationships may extend beyond simple binary interactions. Recent advances in signal processing and machine learning have introduced a variety of tools to handle topological signals [4], [5], [6], including specialized convolutional and trend filtering techniques [7], [8], neural networks [9], [10], Fourier analysis [5], autoregressive models [11], signal recovery methods [12], and simplicial random walks [13].

The Hodge Laplacian provides an algebraic representation of topological structures and enables a spectral decomposition of simplicial signals [5], [6], [14]. Specifically, any simplicial signal of a given order can be written as the sum of three orthogonal components, each lying on a subspace given by the decomposition of the Hodge Laplacian of that order. Focusing on edge signals, for instance, one can decompose them into three mutually orthogonal components: gradient, curl, and harmonic. Each component lives in a corresponding Hodge subspace and offers distinct insights into the nature of the signal [7], [15]. To jointly incorporate signals defined at multiple orders (e.g., node, edge, and triangle signals), the Dirac operator [16], [17] extends this idea, decomposing the entire space of simplicial signals into Dirac gradient, Dirac curl, and Dirac harmonic subspaces.

These subspaces provide a better characterization of practical topological signals, as they often exhibit special properties such as being divergence-free or curl-free [18]. A divergence-free signal implies that the inflow equals the outflow at each node, meaning there is no gradient component. For example, in traffic networks, where nodes represent intersections and edges correspond to roads, the traffic flow edge signal is nearly divergence-free, as vehicles entering a node will eventually exit it, assuming no congestion [7], [19]. Similarly, a curl-free signal implies that

circulation within each triangle is zero. A notable example is found in the foreign exchange market, where nodes represent currencies and edges denote exchange possibilities. Under the arbitrage-free condition, the exchange rate edge flow is curl-free, ensuring that no profit can be obtained through a closed loop of transactions involving three currencies [20]. However, when abnormalities occur, such as noisy or incomplete measurements [19], [20] or adversarial attacks, these conditions no longer hold. In the case of traffic networks, congestion disrupts the divergence-free condition, introducing gradient components into the edge signal. Similarly, inaccuracies in exchange rate values violate the arbitrage-free condition, causing the edge signal to deviate from being curl-free.

To detect topological anomalies in a principled and mathematically tractable manner, we develop a topological matched subspace detection (MSD) framework inspired by the standard MSD approach [21]. MSD has a long and successful history across various applications, including communications [22], radar [23], and anomaly detection [24]. MSD formulates the detection of a signal residing in a specific subspace as a hypothesis testing problem, leveraging the energy of the projected signal in the orthogonal complement of the target subspace. More recently, MSD has been applied to subgraph detection on graphs [25]. However, existing graph-based MSD methods cannot effectively handle topological signals, which exhibit more intricate and intrinsic relationships among them. Additionally, prior works typically assume full availability of all signals, whereas in real-world applications, this assumption often does not hold. To address this limitation, we further extend the MSD framework to accommodate incomplete topological signals.

More specifically, we make the following contributions:

- 1) We develop a topological MSD framework based on Hodge theory, generalizing MSD on graphs (node signals) without imposing any assumptions on the order of the underlying simplicial signal. More precisely, we formulate a hypothesis testing problem to determine whether a simplicial signal resides in a specific Hodge subspace. The test statistic for this detection task is derived using the generalized likelihood ratio test (GLRT), and its performance is characterized in closed form.
- 2) We extend topological MSD to jointly detect simplicial complex signals across all orders via Dirac theory. Additionally, we establish connections between the Hodge and Dirac MSD tasks, analyze their asymptotic performance, and demonstrate how joint signals can enhance Hodge-based detection tasks.
- 3) We address topological MSD in the presence of missing values. Specifically, we derive the optimal detector based on GLRT by projecting onto the subspace of interest. Furthermore, we analyze how the relationship between the dimension of the target subspace and the number of missing values leads to overdetermined and underdetermined cases.

The effectiveness of these detectors is validated through experiments on real-world datasets, including currency exchange markets, user-item interactions, water networks, and football games.

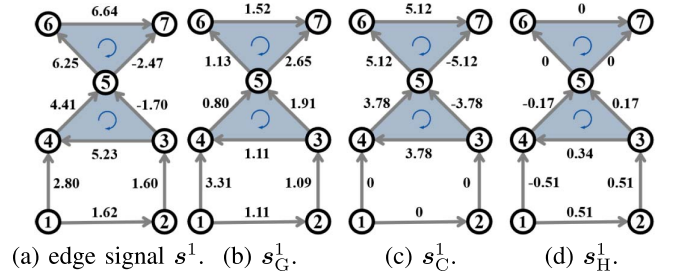


Fig. 1. Hodge decomposition of a 1-signal (edge signal) on a simplicial complex of order two. This edge signal can be decomposed into three different components: the gradient s_G^1 , the curl s_C^1 , and the harmonic component s_H^1 .

The remainder of the paper is organized as follows. Sec. II introduces preliminary concepts, while Sec. III motivates and formulates the problem of interest. Sec. IV presents the MSD framework for both simplicial and simplicial complex signals based on Hodge and Dirac theory. Sec. V discusses the optimal detector for scenarios with missing values. Sec. VI reports numerical experiments, and Sec. VII concludes the paper.

II. PRELIMINARIES

A. Simplicial Complexes

Let $\mathcal{V}$ be a set containing N_0 vertices. Our goal is to define $\mathcal{P}^K$, which is a simplicial complex of order $K \leq N_0$ defined over $\mathcal{V}$. To that end, we first introduce the k -simplex $\mathcal{W}^k$, which is a set containing $k+1 \leq N_0$ vertices of $\mathcal{V}$. Then, a simplicial complex $\mathcal{P}^K$ of order K is a collection of k -simplices (all defined over $\mathcal{V}$ with $k=0, 1, \dots, K$) that satisfy the so-called “inclusion property”. To be specific, let N_k denote the number of k -simplices in $\mathcal{P}^K$. Then, the simplicial complex $\mathcal{P}^K$ is formed by $\{\mathcal{W}_n^0\}_{n=1}^{N_0}$, $\{\mathcal{W}_n^1\}_{n=1}^{N_1}$, $\dots$, $\{\mathcal{W}_n^K\}_{n=1}^{N_K}$, with $N = \sum_{k=0}^K N_k$ being the total number of simplices in $\mathcal{P}^K$. Additionally, to satisfy the inclusion property, it must hold that for any $\mathcal{W}_n^k \in \mathcal{P}^K$, all the subsets of $\mathcal{W}_n^k$ are also part of the simplicial complex $\mathcal{P}^K$. To gain intuition, when embedding the simplicial complex in the Euclidean space, a 0-simplex corresponds to a node, a 1-simplex to an edge, and a 2-simplex to a triangle; see Fig. 1. The inclusion property implies that for a triangle to exist, all its edges and nodes must also be part of the simplicial complex. Additionally, it follows that a graph can be regarded as a simplicial complex of order $K=1$, as it contains only nodes and edges.

We consider the reference orientation of a simplex as the lexicographical ordering of the vertices, and represent the connections between different simplices by the incidence matrices $\mathbf{B}_k \in \mathbb{R}^{N_{k-1} \times N_k}$ describing the relationship between $(k-1)$ -simplices and k -simplices [5]. Based on these incidence matrices, the structure of a simplicial complex can be represented by the *Hodge Laplacian* matrices defined as

$$\begin{cases} \mathbf{L}_0 = \mathbf{B}_1 \mathbf{B}_1^\top, \\ \mathbf{L}_k = \underbrace{\mathbf{B}_k^\top \mathbf{B}_k}_{\mathbf{L}_{k,\ell}} + \underbrace{\mathbf{B}_{k+1} \mathbf{B}_{k+1}^\top}_{\mathbf{L}_{k,u}}, k = 1, \dots, K-1, \\ \mathbf{L}_K = \mathbf{B}_K^\top \mathbf{B}_K. \end{cases} \quad (1)$$

Any *intermediate* Laplacian matrix of order $k = 1, \dots, K - 1$ contains two terms, which are the *lower Laplacian* $\mathbf{L}_{k,\ell} = \mathbf{B}_k^\top \mathbf{B}_k$ and the *upper Laplacian* $\mathbf{L}_{k,u} = \mathbf{B}_{k+1} \mathbf{B}_{k+1}^\top$. They encode respectively the lower adjacencies (e.g., two edges are adjacent via a common node) and upper adjacencies (e.g., two edges are adjacent by being the faces of the same triangle). For example, in Fig. 1, the edges (1, 2) and (2, 3) are lower adjacent, while the edges (3, 4) and (4, 5) are upper adjacent.

B. Simplicial Signals

Simplicial *signals* are mappings from simplices to the set of real numbers. A k -simplicial signal, for short k -signal, $\mathbf{s}^k = [s_1^k, \dots, s_{N_k}^k]^\top \in \mathbb{R}^{N_k}$ is a vector supported on k -simplices where each entry s_n^k corresponds to the n -th k -simplex [5]. If the element s_n^k is positive, the orientation of the signal is the same as the reference, and opposite otherwise. For example, in Fig. 1, the reference orientations of the 1-simplices (edges) are denoted by the arrows. A simplicial complex signal is defined as the concatenation of all k -signals

$$\mathbf{s} = \begin{bmatrix} \mathbf{s}^0 \\ \vdots \\ \mathbf{s}^K \end{bmatrix} \in \mathbb{R}^N, \quad (2)$$

where we recall that $N = \sum_{k=0}^K N_k$.

C. Hodge Decomposition

Hodge Laplacians admit a *Hodge decomposition* stating that the space of k -signals can be decomposed into three orthogonal subspaces [14]

$$\mathbb{R}^{N_k} \equiv \text{span}(\mathbf{B}_k^\top) \oplus \text{kernel}(\mathbf{L}_k) \oplus \text{span}(\mathbf{B}_{k+1}) \quad (3)$$

where $\oplus$ denotes the direct sum, and span and kernel denotes the column space and kernel (nullspace) of a matrix. It implies that any simplicial signal $\mathbf{s}^k$ of order k can be expressed as a sum of three signals of order $k - 1$, k and $k + 1$ fulfilling that, when multiplied by the respective incidence matrices as¹

$$\mathbf{s}^k = \mathbf{B}_k^\top \underline{\mathbf{s}}^{k-1} + \mathbf{s}_H^k + \mathbf{B}_{k+1} \bar{\mathbf{s}}^{k+1}, \quad (4)$$

are orthogonal to each other. Here, the harmonic component $\mathbf{s}_H^k \in \text{kernel}(\mathbf{L}_k)$ is a solution of $\mathbf{L}_k \mathbf{s}_H^k = \mathbf{0}$.

Without loss of generality, consider the edge space (1-signal) to illustrate the Hodge decomposition. The $\text{span}(\mathbf{B}_1^\top)$, $\text{span}(\mathbf{B}_2)$ and $\text{kernel}(\mathbf{L}_1)$ are the gradient, the curl, and the harmonic subspace with dimension $N_{1,G}$, $N_{1,C}$ and $N_{1,H}$, respectively. These subspaces have a direct connection with the eigenvectors of the corresponding Hodge Laplacian. More specifically, let us denote eigendecomposition of the Hodge Laplacian as

$$\mathbf{L}_1 = \mathbf{U}_1 \mathbf{\Lambda}_1 \mathbf{U}_1^\top \quad (5)$$

where the column vectors of $\mathbf{U}_1 \in \mathbb{R}^{N_1 \times N_1}$ form an orthonormal basis, and $\mathbf{\Lambda}_1 = \text{diag}(\lambda_1, \dots, \lambda_{N_1}) \in \mathbb{R}^{N_1 \times N_1}$ is a

diagonal matrix containing the eigenvalues λ_i . The columns of the matrix $\mathbf{U}_1$ can be rearranged as $[\mathbf{U}_{1,G} \mathbf{U}_{1,C} \mathbf{U}_{1,H}]$ where $\mathbf{U}_{1,G}$, $\mathbf{U}_{1,C}$ and $\mathbf{U}_{1,H}$ collect the eigenvectors that span the gradient, curl and harmonic orthogonal subspaces [6]. Then, the Hodge decomposition implies that

$$\mathbf{s}^1 = \mathbf{s}_G^1 + \mathbf{s}_H^1 + \mathbf{s}_C^1, \text{ with } \mathbf{s}_G^1 = \mathbf{B}_1^\top \underline{\mathbf{s}}^0 \text{ and } \mathbf{s}_C^1 = \mathbf{B}_2 \bar{\mathbf{s}}^2 \quad (6)$$

where $\mathbf{s}_G^1$, $\mathbf{s}_C^1$ and $\mathbf{s}_H^1$ are defined as the gradient, curl and harmonic component, respectively. The explanation of the subspace eigenvectors and the corresponding component (see also Fig. 1) are as follows:

- *Gradient eigenvectors and gradient component*: the columns of $\mathbf{U}_{1,G} \in \mathbb{R}^{N_1 \times N_{1,G}}$ are the eigenvectors of $\mathbf{L}_{1,\ell}$ corresponding to the eigenvalues $\lambda_{G,i} > 0$. The gradient component $\mathbf{s}_G^1 = \mathbf{B}_1^\top \underline{\mathbf{s}}^0 \in \text{span}(\mathbf{B}_1^\top)$ is a 1-signal (edge signal) induced by a 0-signal (node signal) and lives in the gradient space. It is computed by taking the difference between the node signal in the nodes connected by an edge. The projection of $\mathbf{s}^1$ onto the gradient subspace $\hat{\mathbf{s}}_G^1 = \mathbf{U}_{1,G}^\top \mathbf{s}^1 = \mathbf{U}_{1,G}^\top \mathbf{s}_G^1 \in \mathbb{R}^{N_{1,G}}$ is the gradient embedding [7].
- *Curl eigenvectors and curl component*: the columns of $\mathbf{U}_{1,C} \in \mathbb{R}^{N_1 \times N_{1,C}}$ are the eigenvectors of $\mathbf{L}_{1,u}$ corresponding to the eigenvalues $\lambda_{C,i} > 0$. The curl component $\mathbf{s}_C^1 = \mathbf{B}_2 \bar{\mathbf{s}}^2 \in \text{span}(\mathbf{B}_2)$ is an 1-signal induced by a 2-signal (triangle signal) and lives in the curl space. It is a local flow circulating along each triangle. The projection of $\mathbf{s}^1$ onto the curl subspace $\hat{\mathbf{s}}_C^1 = \mathbf{U}_{1,C}^\top \mathbf{s}^1 = \mathbf{U}_{1,C}^\top \mathbf{s}_C^1 \in \mathbb{R}^{N_{1,C}}$ is the curl embedding [7].
- *Harmonic eigenvectors and harmonic component*: the columns of $\mathbf{U}_H \in \mathbb{R}^{N_1 \times N_{1,H}}$ are the eigenvectors of $\mathbf{L}_1$ corresponding to the zero eigenvalues $\lambda_{H,i} = 0$. The harmonic component $\mathbf{s}_H^1 \in \text{kernel}(\mathbf{L}_1)$ is an 1-signal in the harmonic space $\text{kernel}(\mathbf{L}_1)$ satisfying $\mathbf{L}_1 \mathbf{s}_H^1 = \mathbf{0}$. The projection of $\mathbf{s}^1$ onto the harmonic subspace $\hat{\mathbf{s}}_H^1 = \mathbf{U}_{1,H}^\top \mathbf{s}^1 = \mathbf{U}_{1,H}^\top \mathbf{s}_H^1 \in \mathbb{R}^{N_{1,H}}$ is the harmonic embedding [7].

The projection of a specific component onto other Hodge subspaces is zero due to the orthogonality between different subspaces. This implicit and apparently simple property will play a major role in developing an MSD theory for topological signals. Two significant properties, which are common for real-world signals, stem from these three components:

- *Curl-free*: $\text{curl}(\mathbf{s}^1) = \mathbf{B}_2^\top \mathbf{s}^1 \in \mathbb{R}^{N_2}$ is the curl operator which measures the curl of a 1-signal (edge signal) $\mathbf{s}^1$. The i th element of this vector represents the sum of the total flow circulating along the i th triangle. An edge signal is curl-free if $\text{curl}(\mathbf{s}^1) = \mathbf{0}$. By definition, the gradient and harmonic components are curl-free. For instance, the currency exchange flow satisfying the arbitrage free condition is curl-free [20].
- *Divergence-free*: $\text{div}(\mathbf{s}^1) = \mathbf{B}_1 \mathbf{s}^1 \in \mathbb{R}^{N_0}$ is the divergence operator which measures the divergence of an edge signal $\mathbf{s}^1$. The i th element of this vector represents the difference between the inflow and outflow at the i th node. An

¹Note that, as indicated by the use of a different notation, the induced signals $\underline{\mathbf{s}}^0$ and $\bar{\mathbf{s}}^2$ in (4) and (6) are not the simplicial signals $\mathbf{s}^k$ that form the simplicial complex signal (2).

edge signal is divergence-free if $\text{div}(s^1) = \mathbf{0}$. By definition, the curl and harmonic components are divergence-free. For example, the Lastfm player transition flow is approximately divergence-free since the player is always switching between different artists [18].

D. Dirac Decomposition

The Hodge decomposition limits the spectral processing to individual level simplicial signals. That is, it focuses on processing the k -signal using the spectrum of Laplacian $\mathbf{L}_k$, without taking into account the interrelationship between signals of varying orders. For a comprehensive approach to processing signals across all simplicial levels and utilizing their inter-simplicial connections, we can turn to the Dirac operator [17], [26]. Specifically, given a simplicial complex $\mathcal{P}^K$ of order K , the Dirac operator $\mathbf{D} \in \mathbb{R}^{N \times N}$ is defined as

$$\mathbf{D} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{B}_1^\top & \mathbf{0} & \mathbf{B}_2 & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2^\top & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{B}_{K-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{B}_{K-1}^\top & \mathbf{0} & \mathbf{B}_K \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}_K^\top & \mathbf{0} \end{bmatrix}. \quad (7)$$

The square of Dirac operator is a block diagonal matrix of the form $\mathbf{D}^2 = \mathcal{L} = \text{blkdiag}(\{\mathbf{L}_k\}_{k=0}^K)$, where blkdiag represents the block diagonal matrix whose diagonal is formed by the matrices $\{\mathbf{L}_k\}_{k=0}^K$.

To facilitate explanation, we focus next on simplicial complexes with an order of $K = 2$. The Dirac operator $\mathbf{D}$ can be broken down into $\mathbf{D} = \mathbf{D}_l + \mathbf{D}_u$, where

$$\mathbf{D}_l = \begin{bmatrix} \mathbf{0} & \mathbf{B}_1 & \mathbf{0} \\ \mathbf{B}_1^\top & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{D}_u = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_2 \\ \mathbf{0} & \mathbf{B}_2^\top & \mathbf{0} \end{bmatrix}. \quad (8)$$

This implies that the space of the simplicial complex signals $s = [s^0 \| s^1 \| s^2] \in \mathbb{R}^N$ can be decomposed into three orthogonal subspaces, mirroring the scenario with k -signals and the Hodge Laplacian

$$\mathbb{R}^N \equiv \text{span}(\mathbf{D}_l) \oplus \text{span}(\mathbf{D}_u) \oplus \text{kernel}(\mathbf{D}) \quad (9)$$

where $\text{span}(\mathbf{D}_l)$ is the Dirac (or joint) gradient subspace considering the node potentials and the gradient flows jointly with dimension N_G ; $\text{span}(\mathbf{D}_u)$ is the Dirac (or joint) curl subspace considering the curl flows and the triangle potentials jointly with dimension N_C ; and $\text{kernel}(\mathbf{D})$ is the Dirac (or joint) harmonic subspace with dimension N_H . Thus, any simplicial complex signal s of order 2 can be expressed as a sum of three orthogonal signals

$$s = s_G + s_C + s_H \quad (10)$$

where $s_G \in \text{span}(\mathbf{D}_l)$, $s_C \in \text{span}(\mathbf{D}_u)$ and $s_H \in \text{kernel}(\mathbf{D})$. Therefore, the matrix of eigenvectors of $\mathbf{D}$ can be rearranged as

$$\mathbf{U}_\mathcal{P} = [\mathbf{U}_{\mathcal{P}_G}, \quad \mathbf{U}_{\mathcal{P}_C}, \quad \mathbf{U}_{\mathcal{P}_H}] \quad (11)$$

where $\mathbf{U}_{\mathcal{P}_G} \in \mathbb{R}^{N \times N_G}$ and $\mathbf{U}_{\mathcal{P}_C} \in \mathbb{R}^{N \times N_C}$ contain the non-zero eigenvectors of $\mathbf{D}_l$ and $\mathbf{D}_u$, respectively, and the columns of $\mathbf{U}_{\mathcal{P}_H} \in \mathbb{R}^{N \times N_H}$ span $\text{kernel}(\mathbf{D})$. These matrices of eigenvectors can be computed from the singular vectors of the incidence matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ [26].

III. PROBLEM FORMULATION

In practical scenarios, topological signals are often confined to specific topological subspaces, as indicated in equations (3) or (9). This is particularly true for signals that are curl-free or divergence-free. However, anomalies do not follow this pattern; their signals typically span multiple subspaces. Consequently, identifying the subspaces to which a signal belongs is crucial for detecting anomalies or patterns in simplicial complex signals. In this work, we assume that the underlying simplicial topology (i.e., the incidence matrices) is known and fixed. Our primary objective is to determine whether a simplicial complex signal s resides in certain topological subspaces, even when only noisy (and possibly incomplete) measurements are available.

More formally, let $x = \Theta(s + n) \in \mathbb{R}^{N_o}$ be the subset of noisy measurements, where $\Theta \in \{0, 1\}^{N_o \times N}$ is a sampling matrix with one 1 per row, selecting the N_o available entries. Here, s denotes the noise-free and complete simplicial complex signal, and n is a zero-mean Gaussian noise vector $n \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$ with variance σ^2 that we assume known (if unknown, σ^2 can be estimated using a maximum likelihood estimator (MLE), leading to a generalized likelihood ratio test with nuisance parameters, or via a plug-in estimator using signal-free training data). Our problem of interest can be formulated as the following hypothesis test:

$$\begin{aligned} \mathcal{H}_0: & \text{ } s \text{ resides within a specific topological subspace } \mathcal{S}_\mathcal{P} \\ \mathcal{H}_1: & \text{ } s \text{ does not belong to } \mathcal{S}_\mathcal{P}. \end{aligned} \quad (12)$$

We address the hypothesis testing problem (12) using noisy and potentially incomplete data $x \in \mathbb{R}^{N_o}$. When $\Theta = \mathbf{I}$, we deal with complete data, as discussed in Section IV. Otherwise, we handle missing data, addressed in Section V.

IV. DETECTION WITH COMPLETE SIGNAL

In this section, we consider the simplicial detection task with complete data, i.e., $\Theta = \mathbf{I}$. We begin by describing the Hodge subspace detection problem in Section IV-A. We then extend our approach to the Dirac subspace detector in Section IV-B and, in Section IV-C, explore the relationships between the two.

A. Hodge Subspace Detector

Consider that the k -simplicial signal s^k resides in a specific Hodge subspace, which can be written as a linear combination of the following eigenvectors:

$$\mathbf{U}_\Delta \in \{\mathbf{U}_G, \mathbf{U}_C, \mathbf{U}_H, [\mathbf{U}_G, \mathbf{U}_C], [\mathbf{U}_G, \mathbf{U}_H], [\mathbf{U}_C, \mathbf{U}_H]\}. \quad (13)$$

The columns of $\mathbf{U}_\Delta \in \mathbb{R}^{N_k \times N_\Delta}$ span the subspace of interest, which can be a combination of two of the Hodge subspaces

eigenvectors such as $[\mathbf{U}_G, \mathbf{U}_H]$. If $\mathbf{s}^k \in \text{span}(\mathbf{U}_\Delta)$, it is possible to write $\mathbf{s}^k = \mathbf{U}_\Delta \hat{\mathbf{s}}_\Delta^k$, with $\hat{\mathbf{s}}_\Delta^k \in \mathbb{R}^{N_\Delta}$ containing the coefficients associated with each of the N_Δ vectors in the columns of $\mathbf{U}_\Delta$.

Likewise, we consider the complement (orthogonal) eigenvectors to $\mathbf{U}_\Delta$ which are the corresponding element of

$$\mathbf{U}_{\bar{\Delta}} \in \{[\mathbf{U}_C, \mathbf{U}_H], [\mathbf{U}_G, \mathbf{U}_H], [\mathbf{U}_G, \mathbf{U}_C], \mathbf{U}_H, \mathbf{U}_C, \mathbf{U}_G\}. \quad (14)$$

whose $N_{\bar{\Delta}}$ columns span a complement Hodge subspace for $\mathbf{s}^k$. For instance, the foreign currency exchange rate flow tends to be curl-free and should align with the column space of $\mathbf{U}_\Delta = [\mathbf{U}_G, \mathbf{U}_H]$, while the complement subspace would be $\mathbf{U}_{\bar{\Delta}} = \mathbf{U}_C$. Under this setting, the hypothesis testing problem is

$$\begin{aligned} \mathcal{H}_0 : \mathbf{x}^k &= \mathbf{U}_\Delta \hat{\mathbf{s}}_\Delta^k + \mathbf{n}^k \\ \mathcal{H}_1 : \mathbf{x}^k &= \mathbf{U} \hat{\mathbf{s}} + \mathbf{n}^k, \end{aligned} \quad (15)$$

where $\mathbf{U} = [\mathbf{U}_G, \mathbf{U}_C, \mathbf{U}_H]$ is the full matrix of eigenvectors of the corresponding Hodge Laplacian –e.g. for the edge space, $\mathbf{U} = \mathbf{U}_1$ defined in (5)– and $\hat{\mathbf{s}}^k \in \mathbb{R}^{N_k}$ ($\hat{\mathbf{s}}_\Delta^k \in \mathbb{R}^{N_\Delta}$) contains the coefficients associated with each of the eigenvectors of $\mathbf{U}$ ($\mathbf{U}_\Delta$). In essence, we assess whether the simplicial signal of interest $\mathbf{x}^k$ can be expressed as a linear combination of the columns of $\mathbf{U}_\Delta$ or if it contains a component beyond the subspace defined by those columns.

Multiplying both sides of (15) by $\mathbf{U}_{\bar{\Delta}}^\top$ yields the projection of $\mathbf{x}^k$ onto the complement subspace $\hat{\mathbf{x}}_{\bar{\Delta}}^k = \mathbf{U}_{\bar{\Delta}}^\top \mathbf{s}^k + \mathbf{U}_{\bar{\Delta}}^\top \mathbf{n}^k = \hat{\mathbf{s}}_{\bar{\Delta}}^k + \hat{\mathbf{n}}_{\bar{\Delta}}^k$, with $\hat{\mathbf{s}}_{\bar{\Delta}}^k$ and $\hat{\mathbf{n}}_{\bar{\Delta}}^k$ representing the projections of the clean signal and noise onto the complement subspace, respectively. The projected noise satisfies $\hat{\mathbf{n}}_{\bar{\Delta}}^k \sim \mathcal{N}(\mathbf{0}_{N_{\bar{\Delta}}}, \sigma^2 \mathbf{I}_{N_{\bar{\Delta}}})$.

Under hypothesis $\mathcal{H}_0$, the signal $\mathbf{s}^k$ lives in the Hodge subspace spanned by the columns of $\mathbf{U}_\Delta$ and the projection $\mathbf{U}_{\bar{\Delta}}^\top \mathbf{s}^k$ is $\mathbf{0}$ due to the orthogonality between the eigenvectors $\mathbf{U}_{\bar{\Delta}}^\top \mathbf{U}_\Delta = \mathbf{0}$. Thus, the projection of $\mathbf{x}^k$ onto the complement subspace under $\mathcal{H}_0$ is only noise $\hat{\mathbf{n}}_{\bar{\Delta}}^k$. Differently, under hypothesis $\mathcal{H}_1$, the projection is not only noise. Therefore, the hypothesis test takes the form

$$\begin{aligned} l\mathcal{H}_0 : \hat{\mathbf{x}}_{\bar{\Delta}}^k &= \hat{\mathbf{n}}_{\bar{\Delta}}^k \\ \mathcal{H}_1 : \hat{\mathbf{x}}_{\bar{\Delta}}^k &= \mathbf{U}_{\bar{\Delta}}^\top \mathbf{s}^k + \hat{\mathbf{n}}_{\bar{\Delta}}^k. \end{aligned} \quad (16)$$

This is a classical matched subspace detection problem when a signal is corrupted by noise [21], in which we have to decide whether the projection onto the orthogonal subspace has a signal component or it is just noise. The problem of detecting deterministic signals with unknown parameters can be solved by the standard GLRT

$$T(\hat{\mathbf{x}}_{\bar{\Delta}}^k) = \frac{p(\hat{\mathbf{x}}_{\bar{\Delta}}^k; \hat{\mathbf{s}}_{\bar{\Delta}1}^{k*}, \mathcal{H}_1)}{p(\hat{\mathbf{x}}_{\bar{\Delta}}^k; \hat{\mathbf{s}}_{\bar{\Delta}0}^{k*}, \mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma \quad (17)$$

where $p(\hat{\mathbf{x}}_{\bar{\Delta}}^k; \hat{\mathbf{s}}_{\bar{\Delta}j}^{k*}, \mathcal{H}_j)$ is the probability density function (pdf) of $\hat{\mathbf{x}}_{\bar{\Delta}}^k$, $\hat{\mathbf{s}}_{\bar{\Delta}j}^{k*}$ is the MLE of $\hat{\mathbf{s}}_{\bar{\Delta}}^k$ under hypothesis $\mathcal{H}_j$, $j \in \{0, 1\}$ and γ is the decision threshold. When the test statistic $T(\hat{\mathbf{x}}_{\bar{\Delta}}^k)$ exceeds (is below) the threshold γ , the detector determines that

hypothesis $\mathcal{H}_1$ ($\mathcal{H}_0$) is true. Therefore, γ controls the false-alarm and detection probabilities (the lower this threshold, the fewer times we will decide $\mathcal{H}_0$ and viceversa).

Under a zero-mean Gaussian noise, the probability density function is

$$p(\hat{\mathbf{x}}_{\bar{\Delta}}^k; \hat{\mathbf{s}}_{\bar{\Delta}j}^{k*}, \mathcal{H}_j) = \mathcal{N}(\hat{\mathbf{x}}_{\bar{\Delta}}^k; \hat{\mathbf{s}}_{\bar{\Delta}j}^{k*}, \sigma^2 \mathbf{I}_{N_{\bar{\Delta}}}). \quad (18)$$

Clearly, the MLE $\hat{\mathbf{s}}_{\bar{\Delta}j}^{k*}$ is $\hat{\mathbf{s}}_{\bar{\Delta}}^{k*} = \hat{\mathbf{x}}_{\bar{\Delta}}^k$ under hypothesis $\mathcal{H}_1$ and is $\hat{\mathbf{s}}_{\bar{\Delta}}^{k*} = \mathbf{0}$ under hypothesis $\mathcal{H}_0$. Thus the Hodge subspace detector becomes

$$T(\hat{\mathbf{x}}_{\bar{\Delta}}^k) = \|\hat{\mathbf{x}}_{\bar{\Delta}}^k\|_2^2 / \sigma^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \quad (19)$$

which compares the SNR of the signal projected onto the column space of $\mathbf{U}_{\bar{\Delta}}$ with the threshold γ .

Given the Gaussian distribution of $\hat{\mathbf{x}}_{\bar{\Delta}}^k$, the test statistic $T(\hat{\mathbf{x}}_{\bar{\Delta}}^k)$ in (19) has a well-known a Chi-square distribution

$$T(\hat{\mathbf{x}}_{\bar{\Delta}}^k) \sim \begin{cases} \chi_{N_{\bar{\Delta}}}^2 & \text{under } \mathcal{H}_0 \\ \chi_{N_{\bar{\Delta}}}^2(\delta) & \text{under } \mathcal{H}_1 \end{cases} \quad (20)$$

where $N_{\bar{\Delta}}$ are the degrees of freedom and δ is a noncentrality parameter satisfying $\delta = \|\hat{\mathbf{s}}_{\bar{\Delta}}^k\|_2^2 / \sigma^2$. The higher the noncentrality parameter, the further apart from each other the distributions, and the easier the detection task. In the next section, we introduce the Dirac detector. We will show that when a system exhibits signals across multiple orders (e.g., nodes and triangles coupled with edges), the Dirac detector allows us to harvest the energy distributed across these levels. This increases the non-centrality parameter and enhances detection performance compared to analyzing the edge signal in isolation.

Before that, we are in a position to characterize the performance of the Hodge detector. Given the distribution of the test statistic, the probability of false alarm is

$$\text{P}_{\text{FA}} \triangleq \Pr \left\{ T(\hat{\mathbf{x}}_{\bar{\Delta}}^k) > \gamma; \mathcal{H}_0 \right\} = Q_{\chi_{N_{\bar{\Delta}}}^2}(\gamma), \quad (21)$$

and the probability of detection as

$$\text{P}_D \triangleq \Pr \left\{ T(\hat{\mathbf{x}}_{\bar{\Delta}}^k) > \gamma; \mathcal{H}_1 \right\} = Q_{\chi_{N_{\bar{\Delta}}}^2(\delta)}(\gamma), \quad (22)$$

where $Q_{\chi_{N_{\bar{\Delta}}}^2}(\cdot)$ is the right-tail probability function of the Chi-square distribution. We note that all calculations are performed under the assumption of Gaussian noise. If the noise is non-Gaussian, the ℓ_2 -norm detector in (19) is no longer the GLRT (optimal), and the test statistic does not follow a Chi-square distribution. Consequently, the theoretical thresholds for a desired P_{FA} will be inaccurate, and the detector may be sensitive to heavy-tailed noise (outliers).

B. Dirac Subspace Detector

Now, our objective is to determine whether the simplicial complex signal $\mathbf{s}$ lies in certain Dirac subspaces spanned by any of the following eigenvectors:

$$\begin{aligned} \mathbf{U}_{\mathcal{P}\Delta} \in \{ & \mathbf{U}_{\mathcal{P}G}, \mathbf{U}_{\mathcal{P}C}, \mathbf{U}_{\mathcal{P}H}, [\mathbf{U}_{\mathcal{P}G}, \mathbf{U}_{\mathcal{P}C}], \\ & [\mathbf{U}_{\mathcal{P}G}, \mathbf{U}_{\mathcal{P}H}], [\mathbf{U}_{\mathcal{P}C}, \mathbf{U}_{\mathcal{P}H}] \}. \end{aligned} \quad (23)$$

or any subcombination thereof². Analogous to the Hodge setting, let $\mathbf{U}_{\mathcal{P}\bar{\Delta}}$ be the complement eigenvectors.

As in (16), the hypothesis test for the Dirac setting can be restated as:

$$\begin{aligned}\mathcal{H}_0: \hat{\mathbf{x}}_{\bar{\Delta}} &= \hat{\mathbf{n}}_{\bar{\Delta}} \\ \mathcal{H}_1: \hat{\mathbf{x}}_{\bar{\Delta}} &= \mathbf{U}_{\mathcal{P}\bar{\Delta}}^\top \mathbf{s} + \hat{\mathbf{n}}_{\bar{\Delta}}.\end{aligned}\quad (24)$$

With similar derivations, the Dirac subspace detector becomes:

$$T(\hat{\mathbf{x}}_{\bar{\Delta}}) = \|\hat{\mathbf{x}}_{\bar{\Delta}}\|_2^2 / \sigma^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma. \quad (25)$$

Once again, we measure the SNR energy in the orthogonal subspace $\text{span}(\mathbf{U}_{\mathcal{P}\bar{\Delta}})$.

As in (19), the test statistic follows a Chi-square distribution, so the false alarm and detection probabilities match those in (21) and (22), respectively. In this case, under $\mathcal{H}_1$, the distribution's non-centrality parameter is $\delta = \|\hat{\mathbf{s}}_{\bar{\Delta}}\|_2^2 / \sigma^2$, where $\hat{\mathbf{s}}_{\bar{\Delta}} = \mathbf{U}_{\mathcal{P}\bar{\Delta}}^\top \mathbf{s} \in \mathbb{R}^{N_{\mathcal{P}\bar{\Delta}}}$, and $N_{\mathcal{P}\bar{\Delta}}$ is the dimension of the Dirac complement subspace. Because a simplicial complex signal has a higher dimensionality than a k -simplicial signal ($N > N_k$), in the Dirac setting the energy of $\hat{\mathbf{s}}_{\bar{\Delta}}$ is typically larger, leading to a higher non-centrality parameter and improved detection performance.

The detectors in (19) and (25) are energy detectors: they evaluate the signal energy in the orthogonal subspace, namely $\text{span}(\mathbf{U}_{\bar{\Delta}})$ versus $\text{span}(\mathbf{U}_{\mathcal{P}\bar{\Delta}})$. Given the probabilities of false alarm [cf. (21)] and detection [cf. (22)], we can characterize the asymptotic behavior of (25) following [27], as stated next.

Proposition 1 (Asymptotic performance): For a large dimension of complement subspace $N_{\mathcal{P}\bar{\Delta}}$, the detection probability of the energy detector in (25) is approximated by

$$P_D \approx Q\left(Q^{-1}(P_{FA}) - \sqrt{d^2}\right) \quad (26)$$

where $Q(\cdot)$ is the right-tail probability function of the standard normal distribution and d^2 is the deflection coefficient defined as $d^2 = (\|\hat{\mathbf{s}}_{\bar{\Delta}}\|_2^2 / \sigma^2)^2 / 2N_{\mathcal{P}\bar{\Delta}}$.

Proof: See Appendix A. $\square$

Equation (26) shows that the detection probability rises with the deflection coefficient d^2 , given that the Q function is monotonically decreasing. In turn, d^2 depends on: (i) the SNR of the projection of the signal onto the orthogonal subspace $\|\hat{\mathbf{s}}_{\bar{\Delta}}\|_2^2 / \sigma^2$; and (ii) the dimensionality of the orthogonal subspace $N_{\mathcal{P}\bar{\Delta}}$. Consequently, the higher the projection energy under hypothesis $\mathcal{H}_1$, the greater the detection probability. Finally, note that the asymptotic performance in the Hodge scenario [c.f. (19)] follows (26) by using $\hat{\mathbf{s}}_{\bar{\Delta}}^k$ instead of $\hat{\mathbf{s}}_{\bar{\Delta}}$ in the deflection coefficient.

C. Connections Between Hodge and Dirac Detectors

Understanding how these two detectors relate enables us to exploit their connections and enhance the detection task. Specifically, we compare them to quantify the performance

²For example, if the simplicial signal has a sparse representation in the joint gradient subspace and in the joint curl subspace, then $\mathbf{U}_{\mathcal{P}\Delta}$ could be built using only those eigenvectors.

gain achieved by jointly processing the full simplicial signal versus processing a single level (Hodge) and discarding the information from other orders. We summarize the connection in the following propositions.

Proposition 2: Let $\mathbf{s}^1$ be a 1-signal (edge signal). Also, let $\mathbf{D}$ denote the Dirac operator defined in (7), whose decomposition as in (8) is $\mathbf{D}_l$ and $\mathbf{D}_u$. Finally, let $\mathbf{B}_1$, $\mathbf{B}_2$ be the node-to-edge and edge-to-triangle incidence matrices, respectively, and let $\mathbf{L}_1$ be the Laplacian matrix. Then, let $\mathbf{s} = [\mathbf{0} \parallel \mathbf{s}^1 \parallel \mathbf{0}]$ be the corresponding simplicial complex signal. It holds that

$$\mathbf{s} \in \text{span}(\mathbf{D}_l) \Leftrightarrow \mathbf{s}^1 \in \text{span}(\mathbf{B}_1^\top) \quad (27a)$$

$$\mathbf{s} \in \text{span}(\mathbf{D}_u) \Leftrightarrow \mathbf{s}^1 \in \text{span}(\mathbf{B}_2) \quad (27b)$$

$$\mathbf{s} \in \text{kernel}(\mathbf{D}) \Leftrightarrow \mathbf{s}^1 \in \text{kernel}(\mathbf{L}_1) \quad (27c)$$

where $\Leftrightarrow$ denotes necessary and sufficient conditions.

Proof: See Appendix B. $\square$

While the proof for Proposition 2 follows directly from the definitions of the Dirac operator and the incidence and Hodge Laplacian matrices. This result indicates that detecting whether an edge signal $\mathbf{s}^1$ belongs to a particular Hodge subspace (or some subset thereof) is equivalent to detecting whether the simplicial complex signal whose node and triangle signals are zero-padded, $\mathbf{s} = [\mathbf{0} \parallel \mathbf{s}^1 \parallel \mathbf{0}]$, lies in the corresponding Dirac subspace.

Proposition 3: Let $\mathbf{s}^0$, $\mathbf{s}^1$, and $\mathbf{s}^2$ represent a 0-signal (node signal), 1-signal (edge signal), and 2-signal (triangle signal), respectively, and let $\mathbf{s} = [\mathbf{s}^0 \parallel \mathbf{s}^1 \parallel \mathbf{s}^2]$ be the corresponding simplicial complex signal. Also, let $\mathbf{D}$ denote the Dirac operator defined in (7), whose decomposition as in (8) is $\mathbf{D}_l$ and $\mathbf{D}_u$. Finally, let $\mathbf{B}_1$, $\mathbf{B}_2$ be the node-to-edge and edge-to-triangle incidence matrices, respectively, and let $\mathbf{L}_1$ be the Laplacian matrix. Then, it holds that

$$\mathbf{s} \in \text{span}(\mathbf{D}_l) \Rightarrow \mathbf{s}^1 \in \text{span}(\mathbf{B}_1^\top) \quad (28a)$$

$$\mathbf{s} \in \text{span}(\mathbf{D}_u) \Rightarrow \mathbf{s}^1 \in \text{span}(\mathbf{B}_2) \quad (28b)$$

$$\mathbf{s} \in \text{kernel}(\mathbf{D}) \Rightarrow \mathbf{s}^1 \in \text{kernel}(\mathbf{L}_1) \quad (28c)$$

where $\Rightarrow$ denotes sufficient conditions.

Proof: See Appendix C. $\square$

Proposition 3 indicates that simplicial signals of different orders can help detect the edge signal more effectively. To provide deeper insight, we consider the following simplified scenario. First, examine the Hodge setting under $\mathcal{H}_0$. The expected value of the test statistic in (19) (assuming σ^2 is absorbed into γ) is

$$\mathbb{E}[\|\hat{\mathbf{x}}^k\|_2^2] = \mathbb{E}[\|\hat{\mathbf{n}}_{\bar{\Delta}}\|_2^2] = N_{\bar{\Delta}}\sigma^2. \quad (29)$$

In the Dirac setting, the expected value of the test statistic in (25) is $N_{\mathcal{P}\bar{\Delta}}\sigma^2$. Conversely, under $\mathcal{H}_1$, we have

$$\mathbb{E}[\|\hat{\mathbf{x}}^k\|_2^2] = \mathbb{E}[\|\mathbf{U}_{\bar{\Delta}}^\top \mathbf{s}^k + \hat{\mathbf{n}}_{\bar{\Delta}}\|_2^2] = \mathbb{E}[\|\mathbf{U}_{\bar{\Delta}}^\top \mathbf{s}^k\|_2^2] + N_{\bar{\Delta}}\sigma^2, \quad (30)$$

If we assume the signal energy is proportional to its dimensionality, i.e., $\mathbb{E}[\|\mathbf{U}_{\bar{\Delta}}^\top \mathbf{s}^k\|_2^2] = N_{\bar{\Delta}}\eta$, where η is a constant, then

$$\mathbb{E}[\|\hat{\mathbf{x}}^k\|_2^2] = N_{\bar{\Delta}}(\eta + \sigma^2). \quad (31)$$

Assuming that η remains the same in the Dirac setting, it follows that $\mathbb{E}[\|\hat{\mathbf{x}}\|_2^2] = N_{\mathcal{P}\bar{\Delta}}(\eta + \sigma^2)$.

When comparing the test statistic to a threshold, a useful measure of performance is the expected difference between the test statistic under $\mathcal{H}_1$ and $\mathcal{H}_0$. A larger difference implies an easier detection. In the Hodge case, this difference is

$$\underbrace{N_{\overline{\Delta}}(\eta + \sigma^2)}_{\mathcal{H}_1} - \underbrace{N_{\overline{\Delta}}\sigma^2}_{\mathcal{H}_0} = N_{\overline{\Delta}}\eta, \quad (32)$$

whereas in the Dirac case, the difference is $N_{\mathcal{P}\overline{\Delta}}\eta$. Since $N_{\mathcal{P}\overline{\Delta}} \geq N_{\overline{\Delta}}$, assuming identical noise power and signal energy under $\mathcal{H}_1$, the expected difference in the Dirac setting is larger, thereby facilitating detection.

D. Computational Complexity

The computational complexity of the proposed framework is dominated by the eigenvalue decomposition (EVD) of the Hodge Laplacian or Dirac operator, which scales as $\mathcal{O}(N^3)$, where N is the number of simplices in the simplicial complex. For large-scale problems, this can become computationally prohibitive, particularly when extracting eigenvectors associated with gradient, curl, or harmonic subspaces. To mitigate this, sparse matrix techniques can be leveraged for simplicial complexes with sparse connectivity, reducing the cost of matrix-vector multiplications during iterative decomposition. Furthermore, randomized EVD approaches provide an efficient alternative by approximating the dominant eigenvectors with a complexity of approximately $\mathcal{O}(Nk^2 + k^3)$, where k is the number of dominant eigenvectors required. These approaches significantly decrease computational and memory demands, facilitating the use of our approach in real-world datasets with large simplicial complexes.

V. DETECTION WITH MISSING DATA

In the presence of missing values, the previous detectors do not hold because it is unclear whether the observed signal resides in the subspace of interest. In this section, we discuss the topological matched subspace detector for incomplete signals. For simplicity, we will focus on the Dirac subspace detection problem, as extending it to the Hodge setting is straightforward.

More formally, we have access only to a subset of entries selected by the sampling matrix $\Theta \neq \mathbf{I}$. The observed signal is defined as $\mathbf{x} = \Theta(\mathbf{s} + \mathbf{n}) \in \mathbb{R}^{N_o}$. The hypothesis testing problem can be reformulated as

$$\begin{aligned} \mathcal{H}_0 : \mathbf{x} &= \mathbf{U}_{\mathcal{P}\Delta\Theta}\hat{\mathbf{s}}_0 + \mathbf{n}_\Theta \\ \mathcal{H}_1 : \mathbf{x} &= \mathbf{U}_{\mathcal{P}\Theta}\hat{\mathbf{s}}_1 + \mathbf{n}_\Theta \end{aligned} \quad (33)$$

where $\mathbf{U}_{\mathcal{P}\Delta\Theta} = \Theta\mathbf{U}_{\mathcal{P}\Delta} \in \mathbb{R}^{N_o \times N_{\mathcal{P}\Delta}}$, $\mathbf{U}_{\mathcal{P}\Theta} = \Theta\mathbf{U}_{\mathcal{P}} \in \mathbb{R}^{N_o \times N}$ (i.e., the rows of the eigenvector matrices corresponding to the elements chosen by Θ), $\hat{\mathbf{s}}_0$ and $\hat{\mathbf{s}}_1$ are the coefficients corresponding to the eigenvectors in $\mathbf{U}_{\mathcal{P}\Delta}$ and $\mathbf{U}_{\mathcal{P}}$, respectively, that construct the signal of interest. As before, we consider Gaussian noise $\mathbf{n}_\Theta = \Theta\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{N_o})$.

Note that projecting onto the orthogonal subspace $\text{span}(\mathbf{U}_{\mathcal{P}\overline{\Delta}})$ is not feasible, as $\mathbf{U}_{\mathcal{P}\overline{\Delta}\Theta}^\top \mathbf{U}_{\mathcal{P}\Delta\Theta} = \mathbf{U}_{\mathcal{P}\overline{\Delta}}^\top \Theta^\top \Theta \mathbf{U}_{\mathcal{P}\Delta} \neq \mathbf{0}_N$, where $\mathbf{0}_N$ is the $N \times N$ all-zero matrix. We therefore formulate the GLRT by considering the

distribution of $\mathbf{x}$ under each hypothesis, and by using the MLE of $\hat{\mathbf{s}}_j$, $j \in \{0, 1\}$. The distribution of $\mathbf{x}$ under $\mathcal{H}_j$ is $\mathbf{x} \sim \mathcal{N}(\mathbf{U}_{\mathcal{P}\Theta, j}\hat{\mathbf{s}}_j^*, \sigma^2\mathbf{I}_{N_o})$, where $\mathbf{U}_{\mathcal{P}\Theta, 0} = \mathbf{U}_{\mathcal{P}\Delta\Theta}$ under $\mathcal{H}_0$ and $\mathbf{U}_{\mathcal{P}\Theta, 1} = \mathbf{U}_{\mathcal{P}\Theta}$ under $\mathcal{H}_1$. Hence, the detector becomes

$$T(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{U}_{\mathcal{P}\Delta\Theta}\hat{\mathbf{s}}_0^*\|_2^2 - \|\mathbf{x} - \mathbf{U}_{\mathcal{P}\Theta}\hat{\mathbf{s}}_1^*\|_2^2}{\sigma^2} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma. \quad (34)$$

This detector measures the difference between the residual energies of the signals in the respective subspaces for each hypothesis. Specifically, the numerator in (34) is the difference between two distinct terms, each capturing the energy of the discrepancy between the observed signal and its reconstruction via the eigenvectors of each hypothesis. Consequently, the difference between the missing-data detector (34) and the Dirac subspace detector (25) is that (34) does not explicitly include the complement subspace but instead considers the subspace of interest for each hypothesis.

Finally, the observed signal also affects the MLE of $\hat{\mathbf{s}}_j$ for $j \in \{0, 1\}$. This MLE depends on the relationship between the number of observed samples N_o and the dimension of the subspace $N_{\mathcal{P}\Delta}$. If $N_o > N_{\mathcal{P}\Delta}$, we are in the overdetermined case; if $N_o \leq N_{\mathcal{P}\Delta}$, we are in the underdetermined case. These two scenarios are detailed in the following sections.

A. Overdetermined Case

For the overdetermined case, we have $N_o > N_{\mathcal{P}\Delta}$, and thus we find the MLE of $\hat{\mathbf{s}}_j$ by solving

$$\hat{\mathbf{s}}_j^* = \underset{\hat{\mathbf{s}}_j}{\text{argmin}} \|\mathbf{x} - \mathbf{U}_{\mathcal{P}\Theta, j}\hat{\mathbf{s}}_j\|_2^2. \quad (35)$$

Next, we substitute this value into the test statistic in (34), but before doing so, it is necessary to analyze the solution of this problem under both $\mathcal{H}_0$ and $\mathcal{H}_1$.

For the null hypothesis $\mathcal{H}_0$, the number of observations satisfies $N_o > N_{\mathcal{P}\Delta}$, so $\mathbf{U}_{\mathcal{P}\Delta\Theta} = \Theta\mathbf{U}_{\mathcal{P}\Delta} \in \mathbb{R}^{N_o \times N_{\mathcal{P}\Delta}}$ is a tall matrix. The MLE of $\hat{\mathbf{s}}_0$ is thus given by the left pseudoinverse: $\hat{\mathbf{s}}_0^* = (\mathbf{U}_{\mathcal{P}\Delta\Theta})^\dagger \mathbf{x}_\Theta$. Since $\mathbf{U}_{\mathcal{P}\Delta\Theta}(\mathbf{U}_{\mathcal{P}\Delta\Theta})^\dagger \neq \mathbf{I}_{N_o}$, we have $\|\mathbf{x} - \mathbf{U}_{\mathcal{P}\Delta\Theta}\hat{\mathbf{s}}_0^*\|_2^2 \neq 0$.

Under the alternative hypothesis $\mathcal{H}_1$, $\mathbf{U}_{\mathcal{P}\Theta} = \Theta\mathbf{U}_{\mathcal{P}} \in \mathbb{R}^{N_o \times N}$ is a full row-rank, fat matrix, as it is formed by choosing $N_o \leq N$ rows from the full-rank $N \times N$ matrix $\mathbf{U}_{\mathcal{P}}$. One of the infinitely many solutions is obtained via the right pseudoinverse of $\mathbf{U}_{\mathcal{P}\Theta}$, yielding $\hat{\mathbf{s}}_1^* = (\mathbf{U}_{\mathcal{P}\Theta})^\dagger \mathbf{x}$. Note that, because in this case $\mathbf{U}_{\mathcal{P}\Theta}(\mathbf{U}_{\mathcal{P}\Theta})^\dagger = \mathbf{I}$, it follows that $\|\mathbf{x} - \mathbf{U}_{\mathcal{P}\Theta}(\mathbf{U}_{\mathcal{P}\Theta})^\dagger \mathbf{x}\|_2^2 = 0$.

Substituting the estimates $\hat{\mathbf{s}}_0^*$ and $\hat{\mathbf{s}}_1^*$ back into (34) results in the simplified detector

$$T(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{U}_{\mathcal{P}\Delta\Theta}(\mathbf{U}_{\mathcal{P}\Delta\Theta})^\dagger \mathbf{x}\|_2^2}{\sigma^2} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma. \quad (36)$$

Here, the matrix $\mathbf{P}_{\Delta\Theta} = \mathbf{U}_{\mathcal{P}\Delta\Theta}(\mathbf{U}_{\mathcal{P}\Delta\Theta})^\dagger$ is a projection operator onto the column space $\text{span}(\mathbf{U}_{\mathcal{P}\Delta\Theta})$. Consequently, the proposed test statistic $T(\mathbf{x})$ measures the difference between $\mathbf{x}$ and its projection onto $\text{span}(\mathbf{U}_{\mathcal{P}\Delta\Theta})$.

The transformed variable $\mathbf{x} - \mathbf{P}_{\Delta\Theta}\mathbf{x} = (\mathbf{I} - \mathbf{P}_{\Delta\Theta})\mathbf{x} = \mathbf{P}_{\overline{\Delta}\Theta}\mathbf{x}$ is the projection of $\mathbf{x}$ onto the orthogonal subspace

of $\text{span}(\mathbf{U}_{\mathcal{P}\Delta\Theta})$ ³. This variable still follows a Gaussian distribution with covariance matrix $\sigma^2 \mathbf{P}_{\Delta\Theta}^\top \mathbf{P}_{\Delta\Theta} = \sigma^2 \mathbf{P}_{\Delta\Theta}$ (since the projection matrix is symmetric and idempotent). Its norm follows a Chi-square distribution with $\text{tr}(\mathbf{P}_{\Delta\Theta}) = N - \text{rank}(\mathbf{U}_{\mathcal{P}\Delta\Theta})$ degrees of freedom [28], where tr is the trace operator and rank returns the rank of the matrix. Using the fact that a projection matrix has exactly one eigenvalue per dimension of the subspace it projects onto (and zeros for the rest), if $\mathbf{U}_{\mathcal{P}\Delta\Theta}$ is full column rank (i.e., $\text{rank}(\mathbf{U}_{\mathcal{P}\Delta\Theta}) = N_{\mathcal{P}\Delta}$), the degrees of freedom of the Chi-square distribution become $N - N_{\mathcal{P}\Delta} = N_{\mathcal{P}\Delta}$. The associated false alarm and detection probabilities are given by (21) and (22), respectively, with the non-centrality parameter under $\mathcal{H}_1$ being $\delta = \|\mathbf{P}_{\Delta\Theta} \mathbf{U}_{\mathcal{P}\Theta} \hat{\mathbf{s}}_1\|_2^2 / \sigma^2$.

Remark 1: When no data is missing, the sampling matrix Θ is the identity. In this case, detector (36) simplifies to $T(\mathbf{x}) = (\|\mathbf{x}\|_2^2 - \|\mathbf{U}_{\mathcal{P}\Delta}^\top \mathbf{x}\|_2^2) / \sigma^2$, which is equivalent to detector (25). Indeed, $\|\mathbf{x}\|_2^2 - \|\mathbf{U}_{\mathcal{P}\Delta}^\top \mathbf{x}\|_2^2$ is the energy of the projection of the signal onto the complement subspace, $\|\mathbf{U}_{\mathcal{P}\Delta}^\top \mathbf{x}\|_2^2$, as dictated by Parseval's theorem.

Connections to projection detectors. The GLRT topological detector in (36) is equivalent to the projection detector proposed in [29, Section 5] when $\mathbf{U}_{\mathcal{P}\Delta\Theta}$ is a full column rank matrix, as stated in Proposition 4. Under these conditions, we can adapt the results of [29] to probabilistically characterize the performance of (36) in comparison to the scenario with no missing values. Let $\text{span}(\mathbf{U}_{\mathcal{P}\Delta})$ denote the subspace spanned by the columns of $\mathbf{U}_{\mathcal{P}\Delta}$ (of dimension $N_{\mathcal{P}\Delta}$), and let $\text{span}(\mathbf{U}_{\mathcal{P}\Delta})^\perp$ denote the orthogonal subspace spanned by the columns of $\mathbf{U}_{\mathcal{P}\Delta}$ (of dimension $N_{\mathcal{P}\Delta}$).

Proposition 4: Let $\mathbf{U}_{\mathcal{P}\Delta\Theta} = \Theta \mathbf{U}_{\mathcal{P}\Delta}$ be the rows of the eigenvectors matrix $\mathbf{U}_{\mathcal{P}\Delta}$ selected by the sampling matrix Θ . Also, let $\mathbf{x}$ be the elements of the sampled simplicial complex signal. Finally, let $\mathbf{P}_{\Delta\Theta} = \mathbf{U}_{\mathcal{P}\Delta\Theta} (\mathbf{U}_{\mathcal{P}\Delta\Theta}^\top \mathbf{U}_{\mathcal{P}\Delta\Theta})^{-1} \mathbf{U}_{\mathcal{P}\Delta\Theta}^\top$ be the projection operator onto the subspace $\text{span}(\mathbf{U}_{\mathcal{P}\Delta\Theta})$. Assuming that $\mathbf{U}_{\mathcal{P}\Delta\Theta}$ is full column rank, the detector given in (36) is equivalent to the detector

$$T(\mathbf{x}) = \|\mathbf{x} - \mathbf{P}_{\Delta\Theta} \mathbf{x}\|_2^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \quad (37)$$

proposed in [29, Section 5].

Proof: By using the definition of the left pseudoinverse for a full column rank matrix $\mathbf{A}^\dagger = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top$, we have that

$$\begin{aligned} \mathbf{U}_{\mathcal{P}\Delta\Theta} (\mathbf{U}_{\mathcal{P}\Delta\Theta})^\dagger &= \mathbf{U}_{\mathcal{P}\Delta\Theta} (\mathbf{U}_{\mathcal{P}\Delta\Theta}^\top \mathbf{U}_{\mathcal{P}\Delta\Theta})^{-1} \mathbf{U}_{\mathcal{P}\Delta\Theta}^\top \\ &= \mathbf{U}_{\mathcal{P}\Delta\Theta} (\mathbf{U}_{\mathcal{P}\Delta\Theta}^\top \mathbf{U}_{\mathcal{P}\Delta\Theta})^\dagger \mathbf{U}_{\mathcal{P}\Delta\Theta}^\top = \mathbf{P}_{\Delta\Theta}, \end{aligned}$$

where in the second equality we used the fact that, as $\mathbf{U}_{\mathcal{P}\Delta\Theta}^\top \mathbf{U}_{\mathcal{P}\Delta\Theta}$ is a full rank square matrix (owing to $\mathbf{U}_{\mathcal{P}\Delta\Theta}$ being full column rank), its pseudo-inverse and inverse coincide. Moreover, the noise power σ^2 in the denominator of (36) is absorbed into the threshold γ on the right-hand side, making the two detectors equivalent. $\square$

³In a slight abuse of notation, we let $\mathbf{P}_{\Delta\Theta} = \mathbf{I} - \mathbf{P}_{\Delta\Theta}$ denote the projection operator onto the orthogonal subspace of $\text{span}(\mathbf{U}_{\mathcal{P}\Delta\Theta})$, even though it is not strictly a projection onto $\text{span}(\mathbf{U}_{\mathcal{P}\Delta\Theta})^\perp$. Its role depends on the entries chosen by Θ . The two subspaces (orthogonal to $\text{span}(\mathbf{U}_{\mathcal{P}\Delta\Theta})$ and $\text{span}(\mathbf{U}_{\mathcal{P}\Delta\Theta})$) coincide only when no data is missing, i.e., $\Theta = \mathbf{I}$.

We are ready to probabilistically characterize the performance of the detector, but first we define the coherence of a subspace:

Definition 1 (Subspace coherence [30]): The coherence of an R -dimensional subspace $\mathcal{S}$ is defined as

$$\mu(\mathcal{S}) := \frac{N}{R} \max_j \|\mathbf{P}_{\mathcal{S}} \mathbf{e}_j\|_2^2, \quad (38)$$

where $\mathbf{P}_{\mathcal{S}}$ is the projection operator onto $\mathcal{S}$, $\mathbf{e}_j$ is the j th standard basis element, and N is the signal dimension.

For a vector $\mathbf{v}$, $\mu(\mathbf{v})$ denotes the coherence of the subspace spanned by $\mathbf{v}$. We now claim the following about detector (36).

Corollary 1: Define the decomposition of the signal $\mathbf{x}$ as $\mathbf{x} = \mathbf{x}_\Delta + \mathbf{x}_{\Delta^\perp} \in \mathbb{R}^N$, where $\mathbf{x}_\Delta \in \text{span}(\mathbf{U}_{\mathcal{P}\Delta})$ and $\mathbf{x}_{\Delta^\perp} \in \text{span}(\mathbf{U}_{\mathcal{P}\Delta})^\perp$. Let $\epsilon > 0$ be a constant and assume $N_o \geq \frac{8}{3} N_{\mathcal{P}\Delta} \mu(\text{span}(\mathbf{U}_{\mathcal{P}\Delta})) \log(\frac{2N_{\mathcal{P}\Delta}}{\epsilon})$. Then, with probability at least $1 - 4\epsilon$,

$$\alpha \|\mathbf{x} - \mathbf{P}_{\Delta} \mathbf{x}\|_2^2 \leq \|\mathbf{x} - \mathbf{P}_{\Delta\Theta} \mathbf{x}\|_2^2 \leq (1 + \beta) \frac{N_o}{N} \|\mathbf{x} - \mathbf{P}_{\Delta} \mathbf{x}\|_2^2 \quad (39)$$

where $\delta = \sqrt{\frac{8N_{\mathcal{P}\Delta} \mu(\text{span}(\mathbf{U}_{\mathcal{P}\Delta}))}{3N_o} \log(\frac{2N_{\mathcal{P}\Delta}}{\epsilon})}$, $\gamma = \sqrt{2\mu(\mathbf{x}_{\Delta^\perp}) \log(\frac{1}{\epsilon})}$, $\alpha = \frac{N_o(1-\beta) - N_{\mathcal{P}\Delta} \mu(\text{span}(\mathbf{U}_{\mathcal{P}\Delta})) \frac{(1+\gamma)^2}{(1-\delta)}}{N}$, and $\beta = \sqrt{\frac{2\mu(\mathbf{x}_{\Delta^\perp})^2}{N_o} \log(\frac{1}{\epsilon})}$.

Proof: The probabilistic bounds in (39) follow by applying the proof of [29, Theorem 1]. $\square$

The result provided in Proposition 4 indicates that, under the assumption that $\mathbf{U}_{\mathcal{P}\Delta\Theta}$ is full column rank, the GLRT-based detector for simplicial complex signals is equivalent to the detector proposed for missing data in [29]. This requirement is the same as the one in [31, Th. 1] for perfect recovery under sampling, essentially stipulating that the N_o observed rows of $\mathbf{U}_{\mathcal{P}\Delta}$ span $\mathbb{R}^{N_\Delta}$. Otherwise, the problem falls into the underdetermined setting, discussed in the next section.

When β , γ , and δ are close to zero, the lower bound in $\|\mathbf{x} - \mathbf{P}_{\Delta\Theta} \mathbf{x}\|_2^2$ is approximately

$$\frac{N_o - N_{\mathcal{P}\Delta} \mu(\text{span}(\mathbf{U}_{\mathcal{P}\Delta}))}{N} \|\mathbf{x} - \mathbf{P}_{\Delta} \mathbf{x}\|_2^2. \quad (40)$$

This arises when, for instance, N_o is large or the subspace dimension $N_{\mathcal{P}\Delta}$ is small. Because the coherence of $\text{span}(\mathbf{U}_{\mathcal{P}\Delta})$ is bounded by $1 \leq \mu(\text{span}(\mathbf{U}_{\mathcal{P}\Delta})) \leq \frac{N}{N_{\mathcal{P}\Delta}}$, if $N_o \leq N_{\mathcal{P}\Delta}$, the lower bound might always be zero or negative even if $\|\mathbf{x} - \mathbf{P}_{\Delta} \mathbf{x}\|_2^2 \geq 0$. Hence, the performance of the detector in (36) will be poor with high probability. This underscores that, for the proposed detector to function effectively, we need at least $N_{\mathcal{P}\Delta}$ observations—that is, the dimension of the subspace we aim to detect.

B. Underdetermined Case

Now we deal with the case of having fewer observations than the subspace dimension, i.e., $N_o \leq N_{\mathcal{P}\Delta}$. The fact that, for $\mathcal{H}_1$, $\mathbf{U}_{\mathcal{P}\Theta} (\mathbf{U}_{\mathcal{P}\Theta})^\dagger = \mathbf{I}_{N_o}$ and thus $\|\mathbf{x} - \mathbf{U}_{\mathcal{P}\Theta} (\mathbf{U}_{\mathcal{P}\Theta})^\dagger \mathbf{x}\|_2^2 = 0$ still holds if we estimate the MLE $\hat{\mathbf{s}}_1^*$ by solving (35). However, under $\mathcal{H}_0$, $\mathbf{U}_{\mathcal{P}\Theta,0} = \mathbf{U}_{\mathcal{P}\Delta\Theta}$ is a fat matrix. If this matrix is full

TABLE I
PROPERTIES OF THE DATASETS

Datasets	Nodes	Edges	Triangles	Edge Signal Property	N_{Δ}	$N_{\overline{\Delta}}$	$N_{\mathcal{P}\Delta}$	$N_{\mathcal{P}\overline{\Delta}}$
Forex [20]	25	300	2300	Curl-free	24	276	48	2577
Lastfm [20]	657	1997	1276	Divergence-free	1341	656	2618	1312
Cherry hills [11]	36	40	2	Curl-free	38	2	74	4
Football	11	55	165	Divergence-free	45	10	211	20

row rank and we obtain the MLE of $\hat{s}_0$ via (35), the solution is non-unique, and setting $\hat{s}_0^* = (\mathbf{U}_{\mathcal{P}\Delta\Theta})^\dagger \mathbf{x}$ makes detection impossible. This arises because, if $\mathbf{U}_{\mathcal{P}\Delta\Theta}$ is full row rank, its columns span $\mathbb{R}^{N_o}$, implying that $\mathbf{x} \in \text{span}(\mathbf{U}_{\mathcal{P}\Delta\Theta}) \equiv \mathbb{R}^{N_o}$ in every scenario. Consequently, for our proposed detector, $\mathbf{U}_{\mathcal{P}\Delta\Theta} (\mathbf{U}_{\mathcal{P}\Delta\Theta})^\dagger = \mathbf{I}_{N_o}$, which always yields $T(\mathbf{x}) = \|\mathbf{x} - \mathbf{U}_{\mathcal{P}\Delta\Theta} (\mathbf{U}_{\mathcal{P}\Delta\Theta})^\dagger \mathbf{x}\|_2^2 = 0$ in (36).

For this more challenging case, we can employ a regularized version of the detector and obtain the MLEs of both $\hat{s}_0$ and $\hat{s}_1$ by solving

$$\hat{s}_j^* = \underset{\hat{s}_j}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{U}_{\mathcal{P}\Theta,j} \hat{s}_j\|_2^2 + \lambda_j \Omega(\hat{s}_j) \quad (41)$$

where $\Omega(\hat{s}_j)$ leverages prior information about $\hat{s}_j$, such as it being low-pass or sparse. For instance, if the simplicial embedding $\hat{s}_j$ is low-pass, then we can set the regularizer $\lambda_j \Omega(\hat{s}_j)$ in (41) to $\lambda_j \|\mathbf{R}_j \hat{s}_j\|_2^2$, where $\mathbf{R}_j$ is a diagonal matrix with decreasing diagonal entries. A closed-form solution may exist depending on the regularization Ω ; otherwise, a numerical estimate can be obtained.

VI. NUMERICAL RESULTS

We corroborate the proposed detectors with numerical experiments on four different real-world datasets. In Sec. VI-A, we introduce the datasets, whereas in Sec. VI-B, we evaluate the performance of the Hodge subspace detectors (HSD). In Sec. VI-C, we evaluate the Dirac subspace detectors (DSD). Finally, in Sec. VI-D, we assess the impact of having incomplete data.

A. Datasets

We use four datasets, summarized in Table I:

1) *Forex* [20]: This dataset represents foreign currency exchanges, where each currency is a node, pairwise exchanges between two currencies are treated as edges, and any three currencies form a triangle. The edge signal is the logarithm of the exchange rate. To ensure the exchange rates are arbitrage-free—meaning no profit can be obtained by trading currencies in a loop—the rates must balance in any cyclical exchange. For example, starting with currency A, converting to B, then to C, and finally back to A, should yield no net gain. Denoting the exchange rate between A and B as $r^{A/B}$, the arbitrage-free condition can be written as $r^{A/B} r^{B/C} = r^{A/C}$. Taking the logarithm of the exchange rate, defined as $\hat{r}^{A/B} = \log(r^{A/B})$, we obtain $\hat{r}^{A/B} + \hat{r}^{B/C} = \hat{r}^{A/C}$, indicating that the edge signal is curl-free.

2) *Lastfm* [20]: This dataset records instances when a user switches from one artist to another on a music player. Each artist

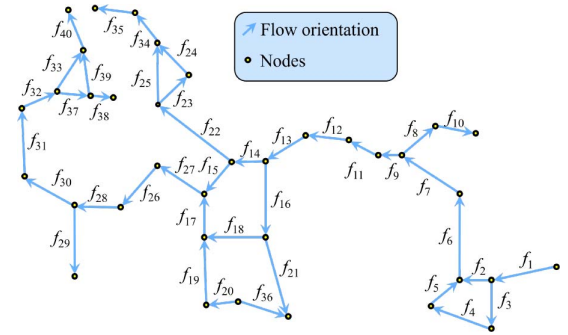


Fig. 2. Cherry hills water network which has 36 nodes represent tanks, 40 edges represent the pipes and 2 triangles represents the areas enclosed by three pipes. Different water demands generates different water flow rate over the edges and water pressure over the nodes.

is represented as a node, and an edge is created between two artists whenever a user switches from one artist to the other. Any triangle formed by three edges is treated as filled. The edge signal is built as follows: each time a user switches from artist A to B, a unit is added to the edge signal from A to B. Since users consistently switch to another artist after listening to one, except for the initial and terminal nodes, the divergence at other nodes is zero. Consequently, the edge signal is approximately divergence-free.

3) *Cherry hills* [11]: This dataset represents a water distribution network, where each node corresponds to a tank, each edge to a pipe transporting water, and each triangle to an area enclosed by three pipes (see Fig. 2). The node signal is the water pressure at each tank (in pounds per square inch, scaled by 10^{-4}), the edge signal captures water flow rate through each pipe (in cubic feet per second), and the triangle signal is the sum of the water demand across the three nodes forming the triangle (in cubic feet per second). Although water demand is physically a nodal quantity, summing the demands of the constituent nodes provides a proxy for the “regional consumption intensity” within the triangular area. This definition captures the interplay between localized demand clusters and the circulating flows (curl) surrounding them, thereby exploiting the higher-order connections in the signals. The edge flow signals are generated with the EPANET software under a demand-driven model [11], where varying demands lead to different flow rates. The dataset comprises 55 hours of recorded edge signal and node pressure data, sampled hourly, with hourly averages used as experimental data.

4) *Football*: This dataset considers the passing data from the German national team, collected during their match against England in the 2020 European Championship. Each player is

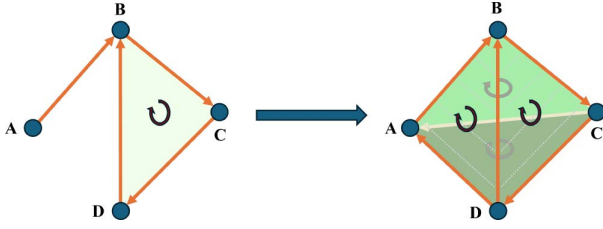


Fig. 3. Illustration of the football dataset with four players A, B, C, and D. Suppose the ball is passed along the path $A \rightarrow B \rightarrow C \rightarrow D \rightarrow B$. There is a single passing loop with no passing error. This scenario can be modeled by a simplicial complex with four nodes, six edges, and four triangles. The edge signal is 1 on the edges $\{A, B\}$, $\{B, C\}$, $\{C, D\}$, and $\{D, B\}$, and 0 on the remaining edges. Notice that, except for the start node A and the end node B, all other nodes have zero divergence. Since no passing error occurs, all node signals are zero. Consequently, there is one passing loop $B \rightarrow C \rightarrow D \rightarrow B$, and only the triangle $\{B, C, D\}$ carries a value of 1, while the other triangles have zero signal.

TABLE II
EXPERIMENTAL SETUP FOR THE HODGE SUBSPACE DETECTION CASE

Dataset	$\mathcal{H}_0$	$\mathcal{H}_1$	SNR
Forex [20]	Curl-free flow	$\mathbf{s}^1 = \mathbf{B}_2 \bar{\mathbf{s}}^2$, $\bar{\mathbf{s}}^2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$	-10 dB
Lastfm [20]	Div-free flow	$\mathbf{s}^1 = \mathbf{B}_1^\top \bar{\mathbf{s}}^0$, $\bar{\mathbf{s}}^0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$	-10 dB
Cherry [11]	Curl-free flow	Flow with curl component	20 dB
Football	Div-free flow	Non-div-free flow	0 dB

represented as a node. An edge exists between any two players, and a triangle represents the passing loop among three players. The method for acquiring simplicial complex signals is as follows: the node signal corresponds to the total number of passing errors each player made throughout the game; the edge signal reflects the number of passes between two players; and the triangle signal indicates the number of passing loops among three players. When players A, B, and C form a passing loop, we add a unit to the triangle signal formed by those three players. This construction ensures that the edge signals are approximately divergence-free if no passing error occurs. Since a player receiving the ball passes it on without holding it, only the first and last players have non-zero divergence, as illustrated in Fig. 3.

B. Hodge Subspace Detector

Experimental setup. The experimental setup for the HSD is summarized in Table II, while the signal energy projection onto the Hodge subspaces is shown in Fig. 4. Here, we focus on detecting edge signals without considering node and triangle signals.

Forex: under hypothesis $\mathcal{H}_0$, the edge signal represents a foreign exchange rate flow that is curl-free. Under hypothesis $\mathcal{H}_1$, we generate the flow as $\mathbf{s}^1 = \mathbf{B}_2 \bar{\mathbf{s}}^2$, where $\bar{\mathbf{s}}^2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_2})$, placing it in the curl subspace. This setup reflects a scenario in which the arbitrage-free condition is violated. We then add zero-mean Gaussian noise at an SNR of -10 dB. The goal is to detect whether the foreign exchange rate satisfies the arbitrage-free condition so that the edge signal is curl-free.

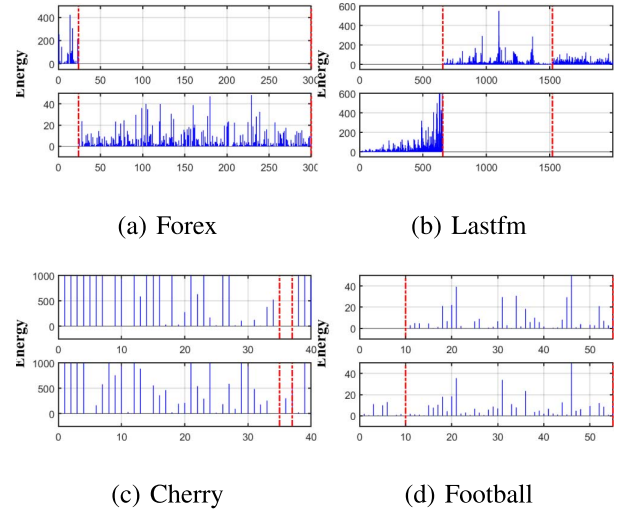


Fig. 4. Energy of projection onto the Hodge subspace for different datasets. In (a), (b), (c), and (d), the upper and lower subgraphs are the energy of the edge signals projection under hypothesis $\mathcal{H}_0$ and $\mathcal{H}_1$ in the Hodge subspaces, respectively. The regions divided by the red lines represent, from left to right, the Hodge gradient, curl and harmonic subspaces.

Lastfm: under hypothesis $\mathcal{H}_0$, the edge signal captures a user transition flow, which is divergence-free. Under $\mathcal{H}_1$, we synthetically generate flows as $\mathbf{s}^1 = \mathbf{B}_1^\top \bar{\mathbf{s}}^0$, where $\bar{\mathbf{s}}^0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_0})$, placing it in the gradient subspace. We then add zero-mean Gaussian noise at an SNR of -10 dB. The objective is to determine whether the edge signal is divergence-free.

Cherry hills: edge signals under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ correspond to two distinct demand conditions, referred to as demand-0 and demand-1. By controlling demand-0, we ensure that the edge signals under $\mathcal{H}_0$ are curl-free. We set the water demands to remain constant. Under hypothesis $\mathcal{H}_0$, the edge signal denotes the flow rate in the pipes under a specified water demand-0. Under hypothesis $\mathcal{H}_1$, the edge signal corresponds to the flow rate under a different specified water demand-1. As shown in Fig. 4(c), the flow rate generated by demand-0 resides in the gradient and harmonic subspaces, exhibiting nearly zero projection energy onto the curl subspace. In contrast, the flow rate produced by demand-1 has a non-zero curl component. We then add zero-mean Gaussian noise at an SNR of 20 dB. The task is to identify the demand pattern of the water flow rate by detecting whether the flow is curl-free.

Football: under hypothesis $\mathcal{H}_0$, we construct the edge signal without incorporating passing errors. For example, if a pass from player A is intercepted and ultimately returned to player B, we treat it as a direct pass from A to B, making the passing flow approximately divergence-free and thus modeling a scenario where passing is uninterrupted. Under hypothesis $\mathcal{H}_1$, we consider the true passing process, where the existence of passing errors results in an edge signal that is not divergence-free. We then add zero-mean Gaussian noise at an SNR of 0 dB to simulate real-world uncertainties and imperfections in the observation or measurement of the passing process. The objective is to determine whether passing interruptions occur by checking whether the edge signal is divergence-free.

TABLE III

AREA UNDER THE CURVES (AUC) FOR THE COMPLETE DATA. -TH. AND -EXP. REPRESENT THE THEORETICAL AND EMPIRICAL RESULTS

Method	Forex	Lastfm	Cherry	Football
HSD-Th.	0.80	1.00	0.82	0.73
HSD-Exp.	0.80 ± 0.01	1.00 ± 0.00	0.82 ± 0.01	0.73 ± 0.01
DSD-Th.	0.99	1.00	1.00	0.95
DSD-Exp.	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.95 ± 0.01
B-SMSD [32]	0.57 ± 0.02	0.67 ± 0.02	0.76 ± 0.18	0.53 ± 0.01

For a fair comparison, we set the same energy level for the edge signals under both hypotheses. Our results are averaged over 1×10^3 independent noisy realizations of a single sample. We select the area under the curves (AUCs) of the receiver operating characteristics (P_D vs P_{FA}) as our evaluation metric. For each dataset, we adjust the SNR to emphasize performance differences.

We compare our method with the blind simple matched subspace detector (B-SMSD) [32], originally designed for graph signals. To adapt it for edge signals, we first map edges to nodes by constructing the line-graph [19]. This detector assumes the observed signal is bandlimited with respect to the graph Fourier transform of the line-graph, and then compares the out-of-band SNR with a threshold γ . For the (bandwidth of the) B-SMSD, we select 95% of the line-graph eigenvectors corresponding to the smallest eigenvalues.

Results. The outcomes of the HSD are presented in Table III, where theoretical and experimental results align closely, corroborating the proposed theory. Although the Cherry dataset has the highest SNR (see Table II), its detection performance for the HSD is not optimal due to the small dimension N_{Δ} of the complement subspace, making the energy detector in (19) more noise-sensitive. In contrast, despite having the same SNR as the Forex dataset, Lastfm achieves better detection performance because its larger complement subspace dimension provides greater noise robustness. The baseline B-SMSD method is not effective at distinguishing between hypotheses because its design principle does not match the higher-order detection task. Despite relatively acceptable performance on the Cherry dataset—likely because its underlying topology closely resembles a path graph (see Fig. 2), which approximates its line-graph—the high standard deviation across multiple runs reveals instability. Moreover, on other datasets, the B-SMSD clearly falls short, indicating that a line-graph approach is unsuitable for this task.

C. Dirac Subspace Detector

Experimental setup. The experimental setup is summarized in Table V, and the signal projection energy onto the Dirac subspaces is shown in Fig. 5. We focus on the detection task that involves node and triangle signals. The edge signals s^1 to be detected are the same as those used in the Hodge-based experiments, with the key distinction being the inclusion of node and triangle signals to evaluate their impact on detection performance. Specifically, for the Forex and Lastfm datasets, we generate synthetic node and triangle signals to investigate

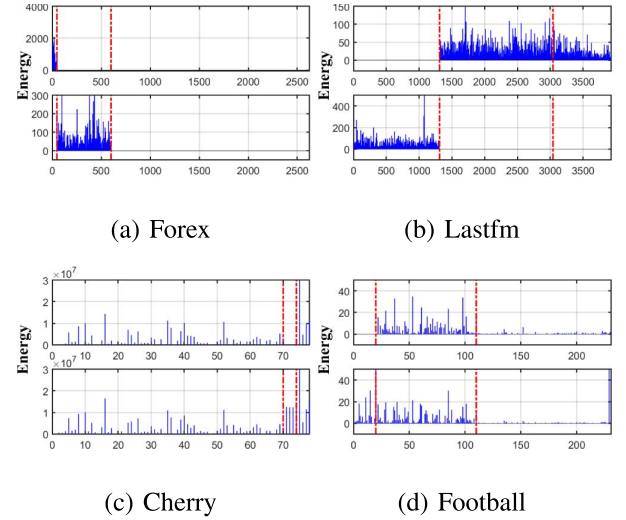


Fig. 5. Energy of the projection onto the Dirac subspace for different datasets. In (a), (b), (c) and (d), the upper and lower subgraphs are the energy of the edge signals projection under hypothesis $\mathcal{H}_0$ and $\mathcal{H}_1$ in the Dirac subspaces, respectively. The regions divided by the red lines represent, in order: the Dirac gradient, curl and harmonic subspace.

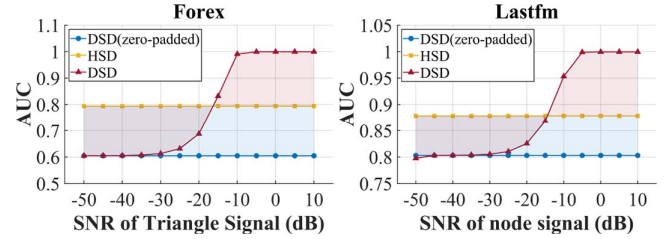


Fig. 6. AUC for different detectors. The yellow line is the HSD. The blue line is the DSD without considering the information in the node and triangle signals (zero-padded). the red line is the DSD considering the node and triangle signals. The SNRs of the edge signal are -10 and -15 dBs, for the Forex and Lastfm datasets, respectively.

their influence on the detector's performance; for the Cherry and Football datasets, we employ real signals. The DSD (zero-padded) in Fig. 6 represents a special case in which the node and triangle signals are padded with zeros.

Forex: under hypothesis $\mathcal{H}_0$, the node signal is $s^0 = \mathbf{B}_1 \bar{s}^1$, where $\bar{s}^1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_1})$, and the triangle signal is $s^2 = \mathbf{0}$. Accordingly, the constructed signal s lies in the Dirac gradient subspace, as illustrated in Fig. 5(a). Under hypothesis $\mathcal{H}_1$, the node signal is $s^0 = \mathbf{0}$, and the triangle signal is $s^2 = \mathbf{B}_2^\top \bar{s}^1$, where $\bar{s}^1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_1})$. The resulting signal s lies in the Dirac curl subspace.

Lastfm: under hypothesis $\mathcal{H}_0$, the node signal is $s^0 = \mathbf{0}$, and the triangle signal is $s^2 = \mathbf{B}_2^\top \bar{s}^1$, where $\bar{s}^1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_1})$. Hence, the constructed signal s occupies the Dirac curl and harmonic subspace without the Dirac gradient component, as shown in Fig. 5(b). Under hypothesis $\mathcal{H}_1$, the node signal is $s^0 = \mathbf{B}_1 \bar{s}^1$, where $\bar{s}^1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_1})$, and the triangle signal is $s^2 = \mathbf{0}$. As a result, the constructed signal s resides in the Dirac gradient subspace.

Cherry: under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, the node signal s^0 represents the water pressure at each node, while the triangle signal s^2 is the sum of the water demands in each triangular area (i.e., over the three nodes forming a triangle). Consequently, under $\mathcal{H}_0$, the constructed signal s remains in the Dirac gradient and harmonic subspaces, containing no Dirac curl component. Under $\mathcal{H}_1$, however, there is a curl component, as illustrated in Fig. 5(c). Here, the priors for $\mathcal{H}_0$ and $\mathcal{H}_1$ stem from observations in specific experimental results.

Football: the node and triangle signals capture the passing errors of each player and the passing loops among three players, respectively, as shown in Fig. 3. The resulting signal s spans the Dirac curl and harmonic subspaces. Under $\mathcal{H}_1$, however, s adds a non-zero Dirac gradient component, as depicted in Fig. 5(d).

For a fair comparison, we set the energy of the signal s to be equal under both hypotheses. We then add zero-mean Gaussian noise that preserves the same edge-signal SNR used in the Hodge-based experiments.

Results. The results for the controlled setting on the Forex and Lastfm datasets are shown in Fig. 6. They indicate that, as the SNR of the node or the triangle signal increases, the detection performance improves gradually. When the energy of the node or triangle signal becomes sufficiently large, the AUC approaches one. The yellow line representing the HSD lies above the blue line for the DSD, which implies that if the node or triangle signal is unknown and therefore zero-padded, the HSD performs better than the DSD. This occurs because the dimension $N_{\mathcal{P}\Delta}$ in the Dirac-based experiments is larger than the Hodge-based N_{Δ} , while the node or triangle signal does not contribute to the detection when it is zero-padded. Consequently, the deflection coefficient d^2 is smaller for the DSD when only the edge signal is considered, resulting in poorer detection performance (c.f. (26)). However, as the energy of the node or triangle signals increases, the DSD's performance rises and eventually surpasses that of the HSD. The performance of the DSD is heavily influenced by the properties of the node and triangle signals. Specifically, when these signals reside in certain Dirac subspaces, the Dirac detector can more effectively capture the structural characteristics and outperform the HSD. In practical scenarios, however, node and triangle signals may deviate from these assumptions; under such circumstances, the DSD might no longer surpass the HSD's performance.

From Table III, we see that the DSD outperforms every other alternative, including real data node and triangle signals. As noted, adding node and triangle signals enlarges the energy gap in the complement Dirac subspace between hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, whereas the HSD considers only the edge signal information.

As for the computational complexity, the runtime of the EVD and detection of HSD and DSD is shown in Table IV. As the data dimensionality increases, computational complexity and runtime rise significantly, which is consistent with expectations. Notably, the DSD requires the full EVD of the Dirac operator, leading to higher runtimes than the HSD, yet the detection step remains efficient once the subspace basis is available.

TABLE IV
RUNTIME COMPARISON OF HSD AND DSD FOR EVD AND DETECTION

Category	Method	Forex	Lastfm	Cherry	Football
EVD	HSD	0.066s	5.481s	0.002s	0.002s
	DSD	8.787s	51.542s	0.006s	0.016s
Detection	HSD	0.005s	0.028s	0.001s	0.001s
	DSD	0.034s	0.056s	0.002s	0.004s

D. Incomplete Data

In this subsection, we address missing data by varying the sampling rate from 0.1 to 1. We compare the performance with that of two baselines: subspace-based and smooth Laplacian interpolation (inpainting). In both approaches, we first reconstruct the incomplete data based on prior information and then perform detection on the recovered signal.

Subspace-based Interpolation. For this baseline, we adopt the interpolation method from [19] using $\mathcal{H}_0$ as the underlying assumption. The challenge is that the signals under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ have different priors. Consequently, we leverage the subspace prior of the signal under hypothesis $\mathcal{H}_0$ for the interpolation task, and also under hypothesis $\mathcal{H}_1$ because the exact origin of the noisy signal is unknown. Concretely, we solve problem

$$\underset{\hat{x}}{\operatorname{argmin}} \|\mathbf{Q}\hat{x}\|_2^2 \quad \text{subject to} \quad \Theta\hat{x} = x, \quad (42)$$

where the matrix $\mathbf{Q}$ is $\mathbf{U}_{\Delta}^{\top}$ or $\mathbf{U}_{\mathcal{P}\Delta}^{\top}$ for the Hodge- and Dirac-based experiments, respectively. The matrix $\Theta \in \{0, 1\}^{N_o \times N}$ is the sampling matrix, and x is the observation. The objective is to minimize the energy of the interpolated signals in the complement subspaces, given that this energy should be zero under hypothesis $\mathcal{H}_0$. We solve this problem via ADMM.

Laplacian smoothing. In this method, which is an agnostic baseline not tied to $\mathcal{H}_0$, we reconstruct the incomplete data by solving

$$\underset{\hat{x}}{\operatorname{argmin}} \|\Theta\hat{x} - x\|_2^2 + \mu\hat{x}^{\top}\mathbf{L}\hat{x}, \quad (43)$$

where $\mu > 0$ and the Laplacian is defined as $\mathbf{L} = \mathbf{L}_1$ under the Hodge setting and $\mathbf{L} = \operatorname{blkdiag}(\mathbf{L}_0, \mathbf{L}_1, \mathbf{L}_2)$ under the Dirac setting. The closed form solution is

$$\hat{x} = (\Theta^{\top}\Theta + \mu\mathbf{L})^{\dagger}\Theta^{\top}x. \quad (44)$$

Overdetermined case. Fig. 7 presents the results, where the proposed GLRT-based detector proves effective for both the HSD and DSD. The DSD consistently performs better because the information contributed by the node or triangle signals bolsters the detection task, even when some of the data is missing. Across different datasets, the performance of the HSD and DSD varies, largely due to the node or triangle signals' differing power levels. When these signals hold greater energy (e.g., in the Football dataset), the DSD's improvement over the HSD becomes more pronounced. It is noted that the error bars for the Cherry Hills and Football datasets are wider than those for the Forex and Lastfm. This is due to their complement subspace

TABLE V
EXPERIMENTAL SETUP FOR THE DIRAC SUBSPACE DETECTOR

Dataset	$\mathcal{H}_0 - \text{node}$	$\mathcal{H}_0 - \text{edge}$	$\mathcal{H}_0 - \text{triangle}$	$\mathcal{H}_1 - \text{node}$	$\mathcal{H}_1 - \text{edge}$	$\mathcal{H}_1 - \text{triangle}$
Forex	$\mathbf{B}_1 \bar{\mathbf{s}}^1, \bar{\mathbf{s}}^1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$	Curl-free flow	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{B}_2 \bar{\mathbf{s}}^2, \bar{\mathbf{s}}^2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$	$\mathbf{B}_2^\top \bar{\mathbf{s}}^1, \bar{\mathbf{s}}^1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Lastfm	$\mathbf{0}$	Div-free flow	$\mathbf{B}_2^\top \bar{\mathbf{s}}^1, \bar{\mathbf{s}}^1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$	$\mathbf{B}_1 \bar{\mathbf{s}}^1, \bar{\mathbf{s}}^1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$	$\mathbf{B}_1^\top \bar{\mathbf{s}}^0, \bar{\mathbf{s}}^0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$	$\mathbf{0}$
Cherry	Pressure-0	Non-curl flow	Area demand-0	Pressure-1	Flow with curl component	Area demand-1
Football	Passing errors-0	Div-free flow	Passing loops-0	Passing errors-1	Non-div-free flow	Passing loops-1

TABLE VI
EXPERIMENTAL SETUP FOR THE EDGE SIGNALS IN THE UNDERDETERMINED EXPERIMENT

Dataset	$\mathcal{H}_0$	$\mathcal{H}_1$	$[\lambda_0 \mathbf{R}_0]_{ii}$	$[\lambda_1 \mathbf{R}_1]_{ii}$	SNR
Forex [20]	$[\hat{\mathbf{s}}_0]_i \sim \mathcal{N}(\exp(-i/20)^\top, 10^{-3})$	$[\hat{\mathbf{s}}_1]_i \sim \mathcal{N}(\exp(-i/1000)^\top, 10^{-3})$	$0.01 * \exp(i/50)$	$\exp(i/2000)$	-10 dB
Lastfm [20]	$[\hat{\mathbf{s}}_0]_i \sim \mathcal{N}(\exp(-i/1000)^\top, 10^{-3})$	$[\hat{\mathbf{s}}_1]_i \sim \mathcal{N}(\exp(-i/3000)^\top, 10^{-3})$	$0.01 * \exp(i/50)$	$\exp(i/2000)$	-10 dB
Cherry [11]	$[\hat{\mathbf{s}}_0]_i \sim \mathcal{N}(\exp(-i/20)^\top, 10^{-3})$	$[\hat{\mathbf{s}}_1]_i \sim \mathcal{N}(\exp(-i/30)^\top, 10^{-3})$	$0.01 * \exp(i/2)$	$\exp(i/100)$	20 dB
Football	$[\hat{\mathbf{s}}_0]_i \sim \mathcal{N}(\exp(-i/100)^\top, 10^{-3})$	$[\hat{\mathbf{s}}_1]_i \sim \mathcal{N}(\exp(-i/200)^\top, 10^{-3})$	$0.01 * \exp(i/5)$	$\exp(i/50)$	0 dB

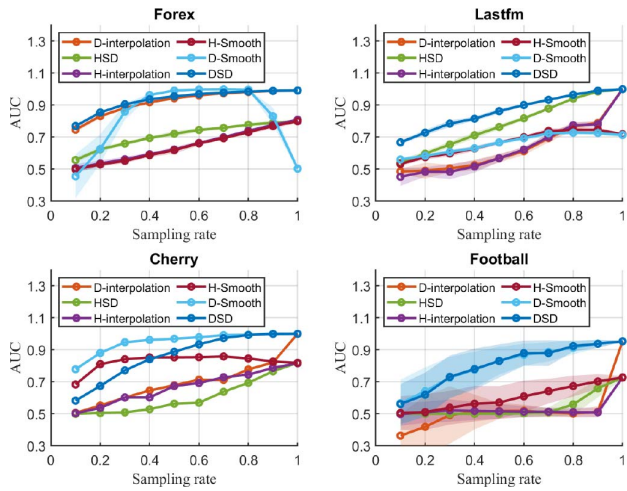


Fig. 7. AUC for the incomplete data experiment. The percentage of the missing data is ranging from 0% to 90%. D-interpolation and H-interpolation refer to the subspace-based interpolation baselines in (42) under the Dirac and Hodge frameworks, respectively. D-smooth and H-smooth denote the Laplacian smoothing baselines in (43). Error bars for the Cherry Hills dataset are omitted to enhance visibility, as the low dimensionality of its complement subspace induces high variance and visual clutter.

$\mathbf{U}_{\Delta}$ having smaller dimensions, which makes the detection more sensitive to noise.

The performance of the interpolation detector differs significantly among datasets because it is not an optimal solution, and its effectiveness depends heavily on the specific properties of the underlying signals. The interpolation baseline does not perform well here because the regularizer in (42) imposes an inaccurate prior for the signal under $\mathcal{H}_1$, reducing its efficacy. Moreover, because the matrix $\mathbf{Q}$ is fat, the solution to the interpolation problem is non-unique, and ADMM-based outcomes lack stability. This instability is one of the primary causes of the interpolation detector's poor performance.

As for the Laplacian smoothing baseline, its performance varies across datasets because it acts as a suboptimal detector. The regularizer $\mu \hat{\mathbf{x}}^\top \mathbf{L} \hat{\mathbf{x}}$ imposes a generic smoothness prior on

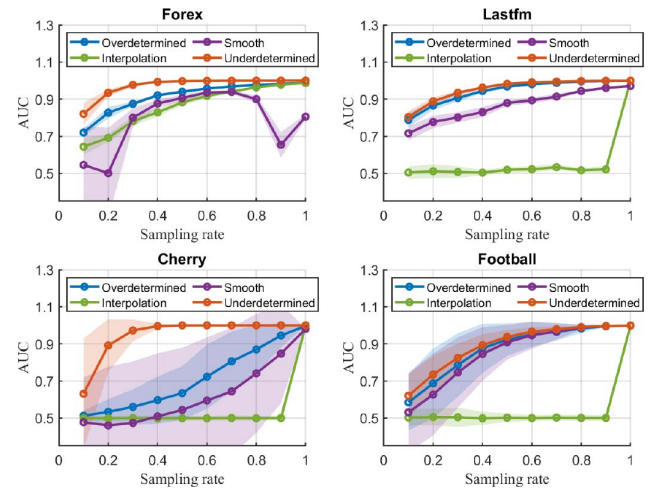


Fig. 8. AUC for overdetermined and underdetermined cases. We assume that $\hat{\mathbf{s}}_0$ and $\hat{\mathbf{s}}_1$ are both low-pass. Smooth and Interpolation denote the Laplacian smoothing and subspace-based interpolation baselines, respectively.

the signal. For example, in the Hodge setting, the regularizer can be written as

$$\mu \hat{\mathbf{x}}^\top \mathbf{L}_1 \hat{\mathbf{x}} = \mu (\|\mathbf{B}_1 \hat{\mathbf{x}}\|_2^2 + \|\mathbf{B}_2^\top \hat{\mathbf{x}}\|_2^2), \quad (45)$$

which encourages edge flows with low divergence and curl [8], [18]. While this property holds for many physical signals, it is not tied to the specific subspace detection task. Consequently, this bias does not consistently aid in distinguishing between $\mathcal{H}_0$ and $\mathcal{H}_1$, leading to performance that is generally inferior to the proposed matched subspace detectors.

Underdetermined case. Finally, we assess the underdetermined scenario in (34). We set the regularizer in (41) based on the prior knowledge of $\hat{\mathbf{s}}_j$. Employing synthetic data on the topologies of these four datasets allows us to verify that incorporating the prior information on the signal can enhance detection performance. For simplicity, we only examine the underdetermined cases in the Dirac setting, as described in Table VI, where i indexes the vector. We have prior knowledge that both simplicial embeddings $\hat{\mathbf{s}}_0$ and $\hat{\mathbf{s}}_1$ are low-pass. Thus,

in (41), we set $\lambda_j \Omega(\hat{s}_j)$ as $\lambda_j \|\mathbf{R}_j \hat{s}_j\|_2^2$, where $\mathbf{R}_j$ is diagonal with decreasing diagonal entries.

As shown in Fig. 8, the underdetermined detector that incorporates prior knowledge of the signal outperforms approaches lacking this information. By contrast, the subspace-based interpolation and Laplacian smoothing detectors yield suboptimal performance because they fail to effectively utilize accurate prior knowledge.

VII. CONCLUSION

This paper proposed an MSD framework to determine whether a simplicial complex signal resides in a specific subspace of interest via hypothesis testing. We first applied the methodology to k -signals (node, edge, or triangle signals) to detect membership in gradient, curl, harmonic, or combined subspaces of the Hodge Laplacian. We then extended our approach to simplicial complex signals using the Dirac operator and its associated subspaces, thereby establishing a theoretical link between the Hodge and Dirac frameworks. The resulting detector, which is optimal under a GLRT perspective, leverages the signal's energy in the orthogonal complement of the target subspace. Recognizing the prevalence of missing data in real-world signals, we also developed an optimal detector for incomplete observations.

We evaluated our proposed MSD on four real-world simplicial complexes, two of which include real simplicial signals residing in Dirac subspaces. The results demonstrated superior performance by (i) considering the entire simplicial signal and (ii) employing the GLRT-optimal detector.

Future work will focus on extending this framework to other topological spaces—such as cell complexes or hypergraphs—and on exploring the task of jointly detecting and localizing anomalies in the simplicial subspaces. Another promising topic is developing a variant of topological MSD under uncertain topology. This is particularly relevant because the detection performance relies on the accurate identification of the subspaces spanned by the eigenvectors of the topological operators. As shown in the graph domain [33], [34], topological perturbations can lead to significant deviations in the spectral basis. In simplicial complexes, even minor structural noise (e.g., a missing triangle) can alter the dimensions and orientations of the gradient or curl subspaces, necessitating the development of robust detectors that can tolerate such spectral mismatch.

APPENDIX

A. Proof of Proposition 1

Adapting the derivations provided in [27], we have the following proof. To determine the asymptotic performance of an energy detector, we need to solve for its first second-order moments. The Chi-square distribution with the non-centrality parameter $\|\hat{s}_\Delta\|_2^2/\sigma^2$ and N_Δ degrees of freedom:

$$\begin{cases} E(T(\hat{x}_\Delta); \mathcal{H}_0) = N_\Delta \\ E(T(\hat{x}_\Delta); \mathcal{H}_1) = \|\hat{s}_\Delta\|_2^2/\sigma^2 + N_\Delta \\ \text{var}(T(\hat{x}_\Delta); \mathcal{H}_0) = 2N_\Delta \\ \text{var}(T(\hat{x}_\Delta); \mathcal{H}_1) = 4\|\hat{s}_\Delta\|_2^2/\sigma^2 + 2N_\Delta \end{cases} \quad (46)$$

Thus, the false alarm and the detection probability of the energy detector can be expressed respectively as

$$P_{\text{FA}} = Q\left(\frac{\gamma - N_\Delta}{\sqrt{2N_\Delta}}\right) \quad (47)$$

$$P_{\text{D}} = Q\left(\frac{\gamma - \|\hat{s}_\Delta\|_2^2/\sigma^2 - N_\Delta}{\sqrt{4\|\hat{s}_\Delta\|_2^2/\sigma^2 + 2N_\Delta}}\right) \quad (48)$$

Following a standard routine, the detection probability can be written into a function of the false alarm probability as

$$P_{\text{D}} = Q\left(\frac{Q^{-1}(P_{\text{FA}}) - \sqrt{\frac{N_\Delta}{2} \frac{\|\hat{s}_\Delta\|_2^2/\sigma^2}{N_\Delta}}}{\sqrt{1 + 2 \frac{\|\hat{s}_\Delta\|_2^2/\sigma^2}{N_\Delta}}}\right) \quad (49)$$

When the number of degrees of freedom N_Δ is large, the term $\|\hat{s}_\Delta\|_2^2/(\sigma^2 N_\Delta) \approx 0$. Hence, by expanding the argument of the Q function using a first-order Taylor expansion, the detection probability is approximated as

$$P_{\text{D}} \approx Q\left(Q^{-1}(P_{\text{FA}}) - \sqrt{\frac{(\|\hat{s}_\Delta\|_2^2/\sigma^2)^2}{2N_\Delta}}\right), \quad (50)$$

which concludes the proof.

B. Proof of Proposition 2

We start by proving (27a). The condition $[0\|\mathbf{s}^1\|0] \in \text{span}(\mathbf{D}_l)$ indicates that there exists a nonzero simplicial signal $[\tilde{s}^0\|\tilde{s}^1\|\tilde{s}^2]$ that satisfies

$$\begin{bmatrix} 0 \\ \mathbf{s}^1 \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_1 & \mathbf{0} \\ \mathbf{B}_1^\top & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{s}^0 \\ \tilde{s}^1 \\ \tilde{s}^2 \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 \tilde{s}^1 \\ \mathbf{B}_1^\top \tilde{s}^0 \\ \mathbf{0} \end{bmatrix}. \quad (51)$$

This means that $\mathbf{B}_1 \tilde{s}^1 = \mathbf{0}$ and $\mathbf{s}^1 = \mathbf{B}_1^\top \tilde{s}^0 \in \text{span}(\mathbf{B}_1^\top)$, which proves that $[0\|\mathbf{s}^1\|0] \in \text{span}(\mathbf{D}_l) \Rightarrow \mathbf{s}^1 \in \text{span}(\mathbf{B}_1^\top)$.

Next, we show that the converse, i.e., $\mathbf{s}^1 \in \text{span}(\mathbf{B}_1^\top) \Rightarrow [0\|\mathbf{s}^1\|0] \in \text{span}(\mathbf{D}_l)$ is also true. The condition $\mathbf{s}^1 \in \text{span}(\mathbf{B}_1^\top)$ implies that there exists a 0-signal $\tilde{s}^0$ satisfying that $\mathbf{s}^1 = \mathbf{B}_1^\top \tilde{s}^0$. We can write this as a function of $\mathbf{D}_l$ as

$$\begin{bmatrix} \mathbf{0} & \mathbf{B}_1 & \mathbf{0} \\ \mathbf{B}_1^\top & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{s}^0 \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{B}_1^\top \tilde{s}^0 \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{s}^1 \\ \mathbf{0} \end{bmatrix} \quad (52)$$

which indicates that $[0\|\mathbf{s}^1\|0] \in \text{span}(\mathbf{D}_l)$. The proof of $[0\|\mathbf{s}^1\|0] \in \text{span}(\mathbf{D}_l) \Rightarrow \mathbf{s}^1 \in \text{span}(\mathbf{B}_1^\top)$ completes, therefore completing the proof of (27a). The proof of (27b) is equivalent to that of (27a).

The proof of (27c) is provided next: the condition $[0\|\mathbf{s}^1\|0] \in \text{kernel}(\mathbf{D})$ indicates that $\mathbf{D} [0\|\mathbf{s}^1\|0] = \mathbf{0}$. We can multiply both sides of the equation by $\mathbf{D}$ to write $\mathbf{D}^2 [0\|\mathbf{s}^1\|0] = \mathbf{D}\mathbf{0} = \mathbf{0}$ so that we are able to use the definition of $\mathbf{D}^2$ as the block diagonal matrix with the Hodge Laplacians and write

$$\begin{bmatrix} \mathbf{L}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{L}_2 \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{s}^1 \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{L}_1 \mathbf{s}^1 \\ \mathbf{0} \end{bmatrix} = \mathbf{0} \quad (53)$$

which indicates $\mathbf{L}_1 \mathbf{s}^1 = \mathbf{0} \Rightarrow \mathbf{s}^1 \in \text{kernel}(\mathbf{L}_1)$ and proves that $[\mathbf{0} \parallel \mathbf{s}^1 \parallel \mathbf{0}] \in \text{kernel}(\mathbf{D}) \Rightarrow \mathbf{s}^1 \in \text{kernel}(\mathbf{L}_1)$. Regarding the converse, the condition $\mathbf{s}^1 \in \text{kernel}(\mathbf{L}_1)$ indicates that $\mathbf{s}^1$ is a solution to $\mathbf{L}_1 \mathbf{s}^1 = \mathbf{0}$. Considering that $\mathbf{D}$ and $\mathbf{D}^2 = \mathcal{L}$ share the same kernel space (as they share eigenvectors and squaring the eigenvalues do not change which of them are 0), $[\mathbf{0} \parallel \mathbf{s}^1 \parallel \mathbf{0}]$ lives in the kernel space of $\mathbf{D}$ [26], therefore proving that $\mathbf{s}^1 \in \text{kernel}(\mathbf{L}_1) \Rightarrow [\mathbf{0} \parallel \mathbf{s}^1 \parallel \mathbf{0}] \in \text{kernel}(\mathbf{D})$. This completes the proof of (27c) and therefore of Proposition 2.

C. Proof of Proposition 3

We start with (28a). The condition $[\mathbf{s}^0 \parallel \mathbf{s}^1 \parallel \mathbf{s}^2] \in \text{span}(\mathbf{D}_l)$ indicates that there exists a nonzero simplicial signal $[\tilde{\mathbf{s}}^0 \parallel \tilde{\mathbf{s}}^1 \parallel \tilde{\mathbf{s}}^2]$ that satisfies

$$\begin{bmatrix} \mathbf{s}^0 \\ \mathbf{s}^1 \\ \mathbf{s}^2 \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_1 & \mathbf{0} \\ \mathbf{B}_1^\top & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{s}}^0 \\ \tilde{\mathbf{s}}^1 \\ \tilde{\mathbf{s}}^2 \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 \tilde{\mathbf{s}}^1 \\ \mathbf{B}_1^\top \tilde{\mathbf{s}}^0 \\ \mathbf{0} \end{bmatrix}. \quad (54)$$

This means that $\mathbf{s}^0 = \mathbf{B}_1 \tilde{\mathbf{s}}^1$, $\mathbf{s}^1 = \mathbf{B}_1^\top \tilde{\mathbf{s}}^0 \in \text{span}(\mathbf{B}_1^\top)$ and $\mathbf{s}^2 = \mathbf{0}$, which proves (28a) completes.

The proof of (28b) is analogous to that of (28a).

To prove (28c), we note that the condition $[\mathbf{s}^0 \parallel \mathbf{s}^1 \parallel \mathbf{s}^2] \in \text{kernel}(\mathbf{D})$ implies that $\mathbf{D} [\mathbf{s}^0 \parallel \mathbf{s}^1 \parallel \mathbf{s}^2] = \mathbf{0}$. Multiplying both sides of this equality by $\mathbf{D}$ yields $\mathbf{D}^2 [\mathbf{s}^0 \parallel \mathbf{s}^1 \parallel \mathbf{s}^2] = \mathbf{D}\mathbf{0} = \mathbf{0}$. Next, use the definition of $\mathbf{D}^2$ and write

$$\begin{bmatrix} \mathbf{L}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{L}_2 \end{bmatrix} \begin{bmatrix} \mathbf{s}^0 \\ \mathbf{s}^1 \\ \mathbf{s}^2 \end{bmatrix} = \begin{bmatrix} \mathbf{L}_0 \mathbf{s}^0 \\ \mathbf{L}_1 \mathbf{s}^1 \\ \mathbf{L}_2 \mathbf{s}^2 \end{bmatrix} = \mathbf{0}. \quad (55)$$

This implies that $\mathbf{L}_1 \mathbf{s}^1 = \mathbf{0} \Rightarrow \mathbf{s}^1 \in \text{kernel}(\mathbf{L}_1)$, completing the proof of (28c) and, as a result, the proof of Proposition 3.

REFERENCES

- [1] C. Liu and E. Isufi, "Hodge-aware matched subspace detectors," in *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*, Piscataway, NJ, USA: IEEE Press, 2024, pp. 817–821.
- [2] C. Bick, E. Gross, H. A. Harrington, and M. T. Schaub, "What are higher-order networks?" *SIAM Rev.*, vol. 65, no. 3, pp. 686–731, 2023.
- [3] V. Salnikov, D. Cassese, and R. Lambiotte, "Simplicial complexes and complex systems," *Eur. J. Phys.*, vol. 40, no. 1, 2018, Art. no. 014001.
- [4] E. Isufi, G. Leus, B. Beferull-Lozano, S. Barbarossa, and P. Di Lorenzo, "Topological signal processing and learning: Recent advances and future challenges," *Signal Process.*, vol. 233, 2025, Art. no. 109930.
- [5] S. Barbarossa and S. Sardellitti, "Topological signal processing over simplicial complexes," *IEEE Trans. Signal Process.*, vol. 68, pp. 2992–3007, 2020.
- [6] M. T. Schaub, Y. Zhu, J.-B. Seby, T. M. Roddenberry, and S. Segarra, "Signal processing on higher-order networks: Livin' on the edge... and beyond," *Signal Process.*, vol. 187, 2021, Art. no. 108149.
- [7] M. Yang, E. Isufi, M. T. Schaub, and G. Leus, "Simplicial convolutional filters," *IEEE Trans. Signal Process.*, vol. 70, pp. 4633–4648, 2022.
- [8] M. Yang and E. Isufi, "Simplicial trend filtering," in *Proc. 56th Asilomar Conf. Signals, Syst., Comput.*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 930–934.
- [9] M. Yang, E. Isufi, and G. Leus, "Simplicial convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 8847–8851.
- [10] C. Battiloro, L. Testa, L. Giusti, S. Sardellitti, P. Di Lorenzo, and S. Barbarossa, "Generalized simplicial attention neural networks," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 10, pp. 833–850, 2024.
- [11] J. Krishnan, R. Money, B. Beferull-Lozano, and E. Isufi, "Simplicial vector autoregressive model for streaming edge flows," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Piscataway, NJ, USA: IEEE Press, 2023, pp. 1–5.
- [12] S. Reddy and S. P. Chepuri, "Recovery of signals on a simplicial complex from subsampled neighbourhood aggregation," *IEEE Signal Process. Lett.*, vol. 32, pp. 536–540, 2024.
- [13] M. T. Schaub, A. R. Benson, P. Horn, G. Lippner, and A. Jadbabaie, "Random walks on simplicial complexes and the normalized Hodge 1-Laplacian," *SIAM Rev.*, vol. 62, no. 2, pp. 353–391, 2020.
- [14] L.-H. Lim, "Hodge Laplacians on graphs," *SIAM Rev.*, vol. 62, no. 3, pp. 685–715, 2020.
- [15] E. Isufi and M. Yang, "Convolutional filtering in simplicial complexes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 5578–5582.
- [16] G. Bianconi, "The topological Dirac equation of networks and simplicial complexes," *J. Phys.: Complexity*, vol. 2, no. 3, 2021, Art. no. 035022.
- [17] F. Baccini, F. Geraci, and G. Bianconi, "Weighted simplicial complexes and their representation power of higher-order network data and topology," *Phys. Rev. E*, vol. 106, no. 3, 2022, Art. no. 034319.
- [18] C. Liu, G. Leus, and E. Isufi, "Unrolling of simplicial elasticnet for edge flow signal reconstruction," *IEEE Open J. Signal Process.*, vol. 5, pp. 186–194, 2023.
- [19] M. T. Schaub and S. Segarra, "Flow smoothing and denoising: Graph signal processing in the edge-space," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 735–739.
- [20] J. Jia, M. T. Schaub, S. Segarra, and A. R. Benson, "Graph-based semi-supervised & active learning for edge flows," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2019, pp. 761–771.
- [21] L. L. Scharf and B. Friedlander, "Matched subspace detectors," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 2146–2157, Aug. 1994.
- [22] M. L. McCloud and L. L. Scharf, "Interference estimation with applications to blind multiple-access communication over fading channels," *IEEE Trans. Inf. Theory*, vol. 46, no. 3, pp. 947–961, May 2000.
- [23] M. Rangaswamy, F. C. Lin, and K. R. Gerlach, "Robust adaptive signal processing methods for heterogeneous radar clutter scenarios," *Signal Process.*, vol. 84, no. 9, pp. 1653–1665, 2004.
- [24] D. W. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 58–69, Jan. 2002.
- [25] S. P. Chepuri and G. Leus, "Subgraph detection using graph signals," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Piscataway, NJ, USA: IEEE Press, 2016, pp. 532–534.
- [26] L. Calmon, M. T. Schaub, and G. Bianconi, "Higher-order signal processing with the Dirac operator," in *Proc. 56th Asilomar Conf. Signals, Syst., Comput.*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 925–929.
- [27] S. Kay, *Fundamentals of Statistical Signal Processing: Detection theory* (Fundamentals of Statistical Si), Prentice-Hall PTR, Englewood Cliffs, NJ, vol. 10, no. 151045, p. 148, 1993.
- [28] I. Olkin, A. M. Mathai, and S. B. Provost, "Quadratic forms in random variables: Theory and applications," *J. Am. Stat. Assoc.*, vol. 87, p. 1244, 1992.
- [29] L. Balzano, B. Recht, and R. Nowak, "High-dimensional matched subspace detection when data are missing," in *Proc. IEEE Int. Symp. Inf. Theory*, Piscataway, NJ, USA: IEEE Press, 2010, pp. 1638–1642.
- [30] E. Candes and B. Recht, "Exact matrix completion via convex optimization," *Commun. ACM*, vol. 55, no. 6, pp. 111–119, 2012.
- [31] S. Chen, R. Varma, A. Sandryhaila, and J. Kovačević, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6510–6523, Dec. 2015.
- [32] E. Isufi, A. S. Mahabir, and G. Leus, "Blind graph topology change detection," *IEEE Signal Process. Lett.*, vol. 25, no. 5, pp. 655–659, May 2018.
- [33] E. Ceci and S. Barbarossa, "Graph signal processing in the presence of topology uncertainties," *IEEE Trans. Signal Process.*, vol. 68, pp. 1558–1573, 2020.
- [34] S. Rey, V. M. Tenorio, and A. G. Marques, "Robust graph filter identification and graph denoising from signal observations," *IEEE Trans. Signal Process.*, vol. 71, pp. 3651–3666, 2023.