Covariance Scattering Transforms

Andrea Cavallo[1], **Ayushman Raghuvanshi**[2], **Sundeep Prabhakar Chepuri**[2], **Elvin Isufi**[1]

[1]Delft University of Technology, Delft, Netherlands
[2]Indian Institute of Science, Bangalore, India
{a.cavallo, e.isufi-1}@tudelft.nl, {ayushmanr, spchepuri}@iisc.ac.in

Abstract

Machine learning and data processing techniques relying on covariance information are widespread as they identify meaningful patterns in unsupervised and unlabeled settings. As a prominent example, Principal Component Analysis (PCA) projects data points onto the eigenvectors of their covariance matrix, capturing the directions of maximum variance. This mapping, however, falls short in two directions: it fails to capture information in low-variance directions, relevant when, e.g., the data contains high-variance noise; and it provides unstable results in low-sample regimes, especially when covariance eigenvalues are close. CoVariance Neural Networks (VNNs), i.e., graph neural networks using the covariance matrix as a graph, show improved stability to estimation errors and learn more expressive functions in the covariance spectrum than PCA, but require training and operate in a labeled setup. To get the benefits of both worlds, we propose Covariance Scattering Transforms (CSTs), deep untrained networks that sequentially apply filters localized in the covariance spectrum to the input data and produce expressive hierarchical representations via nonlinearities. We define the filters as covariance wavelets that capture specific and detailed covariance spectral patterns. We improve CSTs' computational and memory efficiency via a pruning mechanism, and we prove that their error due to finite-sample covariance estimations is less sensitive to close covariance eigenvalues compared to PCA, improving their stability. Our experiments on age prediction from cortical thickness measurements on 4 datasets collecting patients with neurodegenerative diseases show that CSTs produce stable representations in low-data settings, as VNNs but without any training, and lead to comparable or better predictions w.r.t. more complex learning models.

Code — https://github.com/andrea-cavallo-98/CST
Extended version — https://arxiv.org/abs/2511.08878

Introduction

Covariance information captures relevant data characteristics and is widely used to gain insights about data interdependencies, find latent relations and increase data processing performance in unsupervised settings. For example, correlations in brain activity recordings and cortical thickness measures are of high importance to identify interactions among brain regions and co-activation patterns, which lead to deeper understanding of neural dynamics and better neurodegenerative disease prediction (Yin et al. 2023; Bessadok, Mahjoub, and Rekik 2022; Sihag et al. 2024a; Bashyam et al. 2020). Often, the covariance information is accessed via the spectrum of the covariance matrix, which characterizes the concentration of variance along different directions. Principal Component Analysis (PCA), for example, projects the data on the covariance eigenvectors to maximize variance and potentially reduce dimensionality by selecting the directions corresponding to the largest eigenvalues (Jolliffe 2002) – this can be seen as an ideal high-pass filtering in the covariance spectrum, cf. Figure 1. However, data might exhibit complex patterns in the covariance spectrum that lead to relevant information localized in low-variance directions, which PCA filtration loses. Furthermore, PCA is heavily affected by estimation errors in low-sample regimes, causing its output to significantly diverge when the covariance estimation is not reliable (Jolliffe 2002; Jolliffe and Cadima 2016). To mitigate these problems, the work in (Sihag et al. 2022) considers each feature of a data sample as a node of a graph and the covariances among two features as weighted edges, and introduces coVariance Neural Networks (VNNs), graph neural networks operating on this graph. VNNs learn polynomial filtering functions in the covariance eigenvalues, which gives them the flexibility to identify complex variance patterns, and are stable to finite-sample estimation errors. Such properties made VNNs effective in a variety of settings (Sihag et al. 2024a,b; Cavallo, Sabbaqi, and Isufi 2024; Cavallo et al. 2025a). However, VNNs rely on labeled data for training, and their stability and expressivity depend on their learning dynamics, which are difficult to control.

To merge the benefits of both approaches – unsupervised and untrained nature of PCA as well as the stability and expressivity of VNNs – we propose Covariance Scattering Transforms (CSTs), deep architectures that process covariance information in a fully untrained manner with stability guarantees under finite-sample covariance estimation errors. The building block of CSTs are covariance wavelets, localized functions that capture specific patterns in the covariance eigenvalues, as shown in Figure 1, for which we propose three distinct implementations. CSTs sequentially apply banks of wavelet filters followed by nonlinearities and, optionally, low-pass aggregations to decrease representation

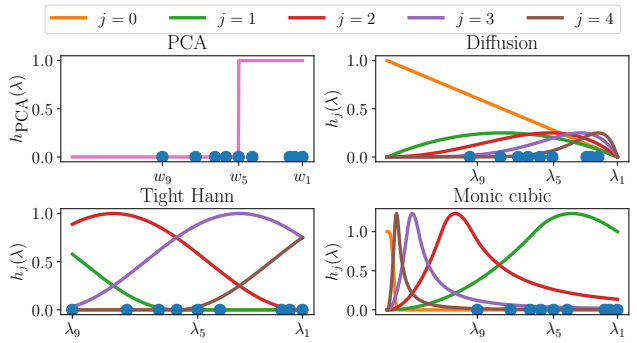

Figure 1: Filters on the covariance eigenvalues w and their scaled versions λ (see Method section for details). PCA acts as a high-pass filter selecting only the top k eigenvalues (here $k = 5$), whereas covariance wavelets provide more complex and localized filter shapes.

size. To reduce the computation and coefficients of CSTs and improve stability, we apply a pruning mechanism that removes wavelet coefficients carrying low energy, deemed irrelevant for efficient representations. Our contributions are summarized as follows.

- We coin the concept of CSTs, untrained deep networks for expressive covariance-based data representation.
- We study the stability of CSTs to finite-sample covariance estimates and input noise. Our results show that CSTs are less affected by close covariance eigenvalues than PCA and their error decreases as more samples T are observed with rate $\mathcal{O}(T^{-1/2})$, extending the advantages of VNNs to the untrained setting.
- We show that CSTs produce stable representations on 4 datasets of cortical thickness measurements which can be used by downstream linear regressors to predict patients' age with a performance that matches or beats more complex non-linear methods like VNNs with few labels.

Background

Consider a data matrix $\mathbf{X} \in \mathbb{R}^{N \times T}$ containing T observations of a random variable $\mathbf{x} \in \mathbb{R}^N$. The variable $\mathbf{x}$ has covariance $\mathbf{C} = \mathbb{E}[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^\mathsf{T}]$, which is estimated from the T samples as $\hat{\mathbf{C}} = \sum_{t=1}^{T}(\mathbf{x}_t - \hat{\boldsymbol{\mu}})(\mathbf{x}_t - \hat{\boldsymbol{\mu}})^\mathsf{T}/T$ where $\hat{\boldsymbol{\mu}} = \sum_{t=1}^{T} \mathbf{x}_t/T$ is the sample mean. The covariance matrix admits the eigendecomposition $\mathbf{C} = \mathbf{V}\mathbf{W}\mathbf{V}^\mathsf{T}$ where $\mathbf{V}$ contains the orthogonal eigenvectors in its columns and $\mathbf{W} = \text{diag}(w_1, \dots, w_N)$ contains the ordered eigenvalues $w_1 \geq w_2 \geq \cdots \geq w_N$. We denote the corresponding sample estimates as $\hat{\mathbf{C}} = \hat{\mathbf{V}}\hat{\mathbf{W}}\hat{\mathbf{V}}^\mathsf{T}$.

PCA transform. The PCA transform projects the data onto the eigenvectors of the covariance matrix, i.e., $\tilde{\mathbf{x}} = \hat{\mathbf{V}}^\mathsf{T}\mathbf{x}$, where each eigenvector captures a portion of variance of the data measured by its corresponding eigenvalue. Often, a subset with the largest k eigenvectors is used to represent the data, which leads to PCA for dimensionality reduction: $\tilde{\mathbf{x}}_{(k)} = [\hat{\mathbf{V}}]^\mathsf{T}_{1,\dots,k}\mathbf{x}$, where $[\cdot]_{1,\dots,k}$ selects the first k columns. This can be written as $\tilde{\mathbf{x}}_{(k)} = [\text{diag}(h_{\text{PCA}}(\hat{w}_1), \dots, h_{\text{PCA}}(\hat{w}_N))\hat{\mathbf{V}}^\mathsf{T}\mathbf{x}]_{1,\dots,k}$, where $h_{\text{PCA}}(w) = \mathbf{1}[w \geq \hat{w}_k]$ is a high-pass filter in the covariance eigenvalues (see Figure 1). While this allows for dimensionality reduction that retains high-variance information, it fails to represent information that is localized in low-variance directions, which might still be relevant for the task at hand. Moreover, the PCA transform and its successive filtrations are unstable to covariance estimation errors. Specifically, the projection error of a data sample $\mathbf{x}$ on the true and perturbed covariance eigenvectors $\mathbf{V}, \hat{\mathbf{V}}$ is bounded with high probability as (Cavallo, Sabbaqi, and Isufi 2024, Proposition 1):

$$\|[\hat{\mathbf{V}}]^\mathsf{T}_{1,\dots,k}\mathbf{x} - [\mathbf{V}]^\mathsf{T}_{1,\dots,k}\mathbf{x}\| \leq \mathcal{O}((\min_{\substack{i,j=1,\dots,k,\\ i\neq j}} |w_i - w_j|)^{-1}) \tag{1}$$

where $\|\cdot\|$ denotes the 2-norm for vectors and spectral norm for matrices throughout the paper. That is, if the covariance matrix has close distinct eigenvalues $w_i \approx w_j, i \neq j$, then the principal component estimation becomes difficult and requires large amounts of observations to be reliable.

Covariance filters. We provide a different interpretation for the PCA transform by building a weighted graph with N nodes where the weight of the edge (i, j) is the covariance value $[\hat{\mathbf{C}}]_{ij}$ and the i-th node has as signal the i-th entry of the vector $\mathbf{x}$ (cf. Figure 6 in the Appendix). The graph Fourier transform of $\mathbf{x}$ is defined as its projection on the graph eigenvectors, i.e., $\tilde{\mathbf{x}} = \hat{\mathbf{V}}^\mathsf{T}\mathbf{x}$, which coincides with the space of the PCA transform. To increase expressivity, covariance filters (Sihag et al. 2022) define a general function $h(w)$ computed on each distinct covariance eigenvalue to modulate the corresponding eigenvector. This function is instantiated via a polynomial $h(w) = \sum_{k=0}^{K} h_k w^k$, where the coefficients h_k are learned to optimize a task-specific loss. The processing of a signal $\mathbf{x}$ via the polynomial covariance filter can be performed directly in the covariance space as $\mathbf{H}(\hat{\mathbf{C}})\mathbf{x} = \hat{\mathbf{V}}\,\text{diag}(h(\hat{w}_1), \dots, h(\hat{w}_N))\hat{\mathbf{V}}^\mathsf{T} = \sum_{k=0}^{K} h_k \hat{\mathbf{C}}^k\mathbf{x}$. Moreover, covariance filters can be assembled into sequential filterbanks interleaved with nonlinearities to define coVariance Neural Networks (VNNs), which learn hierarchical representations $\mathbf{z}_\ell = \sigma(\mathbf{H}_\ell(\hat{\mathbf{C}})\mathbf{z}_{\ell-1})$ with $\mathbf{z}_0 = \mathbf{x}$, $\ell = 1, \dots, L$. The last layer output $\mathbf{z}_L$ represents the VNN output and is often fed to a task-specific readout function. VNNs can learn a large class of functions in the covariance eigenspace, extending the PCA transform, and achieve better stability in low-sample regimes (Sihag et al. 2022). However, they require training and labeled data to optimize the parameters h_k.

Problem statement. Aiming to retain the expressivity and stability of VNNs as well as the untrained/unsupervised nature of the PCA transform, we aim to develop a neural architecture with the following desiderata: D1) use the sample covariance matrix as inductive bias; D2) be expressive and flexible to handle information across all the covariance spectrum; D3) be stable to finite-sample estimation errors in the covariances and input signal; D4) be untrained.

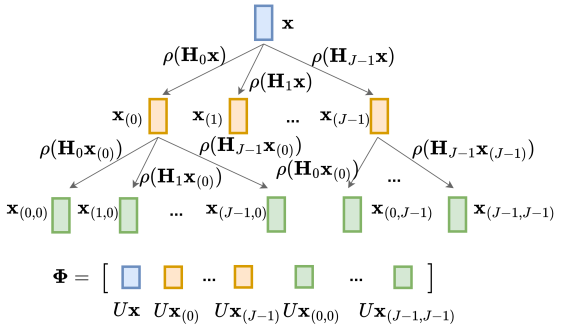

Figure 2: CSTs features are obtained via sequential application of wavelets at different scales, interleaved with nonlinear activations ρ and aggregation operators U to produce the final coefficients collected in $\mathbf{\Phi}$.

Method

We define the Covariance Scattering Transform (CST), a hierarchical untrained network that manipulates the covariance spectrum and meets our desiderata. The building block of CSTs are covariance wavelets, spectrally localized filters that are applied sequentially with nonlinearities in between.

Covariance Wavelets

Covariance wavelets are functions that span the spectrum of the covariance matrix of the data. Given a covariance matrix $\mathbf{C} = \mathbf{V}\mathbf{W}\mathbf{V}^\mathsf{T}$, covariance wavelets operate on a wavelet operator matrix $\mathbf{T} = \mathbf{V} f(\mathbf{W})\mathbf{V}^\mathsf{T} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}$ where $f(\cdot)$ preserves the positive semi-definiteness of $\mathbf{T}$. In particular, we propose two implementations for $\mathbf{T}$: the covariance matrix with normalized eigenspectrum, i.e., $\mathbf{T} = \mathbf{C}_N = \gamma\mathbf{C}/w_1$, where w_1 is the largest covariance eigenvalue and parameter $\gamma \in \mathbb{R}_+$ controls the domain of the spectrum of $\mathbf{T}$ (i.e., the interval $[0, \gamma]$); and $\mathbf{T} = \mathbf{C}_I = \gamma(\mathbf{I} - \mathbf{C}/w_1)$ (where $\mathbf{I}$ is the identity matrix), which reverses the order of the covariance eigenvalues in the interval $[0, \gamma]$, bringing benefits for wavelets that are more discriminative at lower frequencies. We denote with $\hat{\mathbf{T}} = \hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}^\mathsf{T}$ the sample estimate of $\mathbf{T}$ from $\hat{\mathbf{C}}$ and corresponding eigendecomposition.

Covariance wavelets are characterized by a function $h_j(\lambda)$ that acts as a band-pass filter (i.e., $h_j(0) = 0$ and $\lim_{\lambda\to\infty} h_j(\lambda) = 0$, cf. the wavelet admissibility criteria in (Mallat 1999)) and is instantiated at different *scales* $j \geq 1$, i.e., different localizations in the covariance eigenvalues. To also account for content at $\lambda = 0$, we define an additional function $h_0(\lambda)$ such that $h_0(0) > 0$. The collection of J wavelet functions $\{h_j(\lambda)\}_{j=0,\dots,J-1}$ is a multiscale filterbank that spans the covariance spectrum. For a covariance operator $\mathbf{T}$, the application of the wavelet function $h_j(\lambda)$ on all its eigenvalues $\lambda_1, \dots, \lambda_N$ produces the wavelet matrix $\mathbf{H}_j(\mathbf{T}) = \mathbf{V}\,\mathrm{diag}(h_j(\lambda_1), \dots, h_j(\lambda_N))\mathbf{V}^\mathsf{T}$. The *wavelet coefficients* (or *wavelet features*) of a signal $\mathbf{x}$ are its projection on the corresponding wavelet matrix, i.e., $\mathbf{x}_{(j)} = \mathbf{H}_j(\mathbf{T})\mathbf{x}$. We propose three implementations of covariance wavelets: diffusion, Hann and monic. From Figure 1, diffusion wavelets are more localized at high frequencies as the scales increase; Hann wavelets are localized on the specific eigenvalues; monic wavelets are sharper at low frequencies and capture higher frequencies as the scale increases. We define diffusion wavelets next, while we defer Hann and monic wavelets to the Appendix.

Covariance diffusion wavelets. In analogy with graph diffusion wavelets (Gama, Bruna, and Ribeiro 2019), we design the covariance diffusion wavelet function as:

$$h_j(\lambda) = \lambda^{2^{j-1}} - \lambda^{2^j}, \quad j \geq 1 \tag{2}$$

and $h_0(\lambda) = 1 - \lambda$. Given a covariance operator $\mathbf{T}$, the graph diffusion wavelet can be computed without taking the eigendecomposition as $\mathbf{H}_j(\mathbf{T}) = \mathbf{T}^{2^{j-1}} - \mathbf{T}^{2^j}$. This reduces the computational complexity, as eigendecomposition is of order $\mathcal{O}(N^3)$ while the recursive computation of $\mathbf{T}^k\mathbf{x} = \mathbf{T}(\mathbf{T}^{k-1}\mathbf{x})$ is of order $\mathcal{O}(kN^2)$. To maximize the expressiveness of diffusion wavelets, given a wavelet filterbank with J scales (i.e., $j = 0, \dots, J-1$), we rescale the covariance eigenvalues by setting $\gamma = (1/2)^{1/2^{J-2}}$, such that the J-th wavelet reaches its maximum on the largest covariance eigenvalue. Since this rescaling ensures that $\lambda_1 \leq 1$, $h_j(\lambda)$ behaves as a bandpass filter as $h_j(0) = h_j(1) = 0$.

Properties of covariance wavelets. The covariance wavelets enjoy the following properties.

Property 1 (Frame)**.** *The covariance wavelets conform a frame, i.e.,* $A^2\|\mathbf{x}\| \leq \sum_{j=1}^J \|\mathbf{H}_j\mathbf{x}\|^2 \leq B^2\|\mathbf{x}\|^2$ *with* $0 < A \leq B < \infty$.

Property 2 (Lipschitz)**.** *The covariance wavelets* $h_j(\lambda)$ *are Lipschitz, i.e.,* $|h_j(\lambda_k) - h_j(\lambda_l)|/|\lambda_k - \lambda_l| \leq P$ *for two covariance eigenvalues* λ_k, λ_l *and a constant* $P > 0$.

Property 3 (Localization)**.** *Covariance wavelets are localized both in frequency and covariance space.*

Property 1 characterizes the spread of energy of a wavelet filterbank through the constants A, B. Property 2 describes the variability of the wavelet h_j by limiting its derivative through the constant P. Property 3 extends the joint spectral and spatial localization of existing wavelets to the covariance case. Spectral localization corresponds to concentration of $h_j(\lambda)$ around a specific eigenvalue. Covariance space localization, instead, corresponds to the fact that wavelets centered on a feature assume smaller values on features that are distant from the center one, where the distance depends on the strength of the covariance among the features (see Appendix for a formal definition and analysis). For diffusion wavelets, the frame bounds are $B = 1$ and $A = 1 - \gamma$, i.e., using more scales J leads to larger γ and more spread-out eigenvalues, but may cause larger energy loss as A becomes smaller. The Lipschitz constant of a diffusion wavelet at scale j for $\lambda \in [0, 1]$ is $P = 2^{j-1}$, i.e., a larger scale leads to sharper variations in the wavelet functions and consequently a larger Lipschitz constant.

Covariance Scattering Transforms

Covariance Scattering Transforms (CSTs) are deep architectures defined by L layers and a bank of multiresolution covariance wavelets $\{\mathbf{H}_j\}_{j=0}^{J-1}$. Given an input $\mathbf{x}$, the CST produces a set of scattering features and concatenates them. At

layer $\ell = 0$, the scattering features are the input features $\mathbf{x}$. At layer $\ell = 1$, the input $\mathbf{x}$ is projected on all J wavelets and processed by nonlinearity ρ, i.e., $\mathbf{x}_{(j_1)} = \rho(\mathbf{H}_{j_1}\mathbf{x})$ for $j_1 = 0, \dots, J-1$. This process is repeated recursively, i.e., $\mathbf{x}_{(j_\ell, j_{\ell-1}\dots,j_1)} = \rho(\mathbf{H}_{j_\ell}\mathbf{x}_{(j_{\ell-1}\dots,j_1)})$ for $j_\ell = 0, \dots, J-1$ at every layer $\ell = 1, \dots, L-1$, creating a hierarchical structure of scattering representations (cf. Figure 2 and Algorithm 1 in the Appendix). By expanding the recursion, the scattering features at the ℓ-th layer are

$$\mathbf{x}_{(j_\ell, j_{\ell-1},\dots,j_1)} = \rho(\mathbf{H}_{j_\ell}\rho(\mathbf{H}_{j_{\ell-1}} \dots \rho(\mathbf{H}_{j_1}\mathbf{x}))). \quad (3)$$

The scattering features $\mathbf{x}_{(j_\ell, j_{\ell-1},\dots,j_1)}$ can be further processed by an operator U (e.g., mean for dimensionality reduction, identity for no processing), and the resulting $(J^L - 1)/(J-1)$ coefficients $\phi_{j_\ell, j_{\ell-1}\dots j_1}(\mathbf{x}) = U\mathbf{x}_{(j_\ell, j_{\ell-1},\dots,j_1)}$ computed at each layer $\ell = 0, \dots, L-1$ and for all scales $j_\ell = 0, \dots, J-1$ are concatenated to define the CST $\mathbf{\Phi}(\mathbf{T}, \mathbf{x})$.

Pruning

CSTs' number of coefficients grows exponentially with the increasing scales and number of layers, leading to large-dimensional representations to capture deep encodings. To counteract this, we consider a pruning strategy to sparsify the CST tree and only explore the branches that have larger potential to provide meaningful representations. Following (Ioannidis, Chen, and Giannakis 2020), given a the representation at the ℓ-th scattering transform layer $\mathbf{x}_{(j_\ell,\dots,j_1)}$, its projection on the i-th wavelet $\mathbf{x}_{(i,j_\ell,\dots,j_1)}$ (and the subsequent projections at deeper layers) is discarded if its normalized energy is lower than a predefined threshold τ, i.e., $\|\mathbf{x}_{(i,j_\ell,\dots,j_1)}\|/\|\mathbf{x}_{(j_\ell,\dots,j_1)}\| \leq \tau$. We denote with F_ℓ the number of scattering features selected at layer ℓ out of the J^ℓ available. This pruning strategy reduces the search space, leading to more parameter-efficient representations and making the CST a feasible technique for dimensionality reduction. Moreover, the amount of pruning affects the stability of CSTs as we shall elaborate in Theorem 3.

Theoretical Analysis

We characterize theoretically the CST by studying its permutation equivariance and its stability to perturbations in the covariance matrix. We discuss the stability to signal perturbations in the Appendix.

Permutation Equivariance

Since covariance information captures pairwise relations among data features that do not depend on the ordering in which features are observed, it is of interest that the CST is *permutation equivariant*, i.e., if the order of its input is permuted, its output is permuted likewise. Furthermore, if the CST's representations are used to produce a unique label or regression target that does not depend on feature ordering, it is desirable that the CST output is *permutation invariant*, i.e., its output does not change regardless of the input order. We first define the permutation operation for a CST.

Definition 1. *Consider a signal $\mathbf{x} \in \mathbb{R}^N$ and a CST $\mathbf{\Phi} = [U\mathbf{x}||U\mathbf{x}_{(1)}|| \dots ||U\mathbf{x}_{(J,\dots,J)}]$, where $||$ is the concatenation operation and U preserves the dimension N. Let $\mathbf{\Pi}$ be a permutation matrix. The CST permutation operator is*

$$\mathrm{Perm}(\mathbf{\Phi}, \mathbf{\Pi}) = [\mathbf{\Pi}U\mathbf{x}||\mathbf{\Pi}U\mathbf{x}_{(1)}|| \dots ||\mathbf{\Pi}U\mathbf{x}_{(J,\dots,J)}]. \quad (4)$$

The following theorem shows that the CST is permutation equivariant, and can be made permutation invariant by an appropriate choice of aggregation function U.

Theorem 1. *Consider a CST $\mathbf{\Phi}$ computed from a dataset $\mathbf{X} \in \mathbb{R}^{N\times T}$, and a CST $\hat{\mathbf{\Phi}}$ computed from a dataset $\hat{\mathbf{X}} = \mathbf{\Pi}\mathbf{X}$ given permutation matrix $\mathbf{\Pi} \in \mathbb{R}^{N\times N}$. If U is permutation equivariant (e.g., identity), then $\hat{\mathbf{\Phi}} = \mathrm{Perm}(\mathbf{\Phi}, \mathbf{\Pi})$. If U is permutation invariant (e.g., average), then $\mathbf{\Phi} = \hat{\mathbf{\Phi}}$.*

Theorem 1 establishes that CSTs provide permutation equivariant or invariant representations depending on the design of U, which extends analogous results in other scattering transforms (Gama, Ribeiro, and Bruna 2019; Bruna and Mallat 2013) and respects the domain requirements.

Stability to Covariance Perturbations

In practice, the CST is instantiated on a sample covariance $\hat{\mathbf{C}}$ that represents a perturbed version of the true one, i.e., $\hat{\mathbf{C}} = \mathbf{C} + \mathbf{E}_C$ where $\mathbf{E}_C$ collects the estimation error. We study here how the output of the CST changes in relation to this finite-sample error. We begin by bounding the output difference of covariance wavelets, i.e., given the wavelet $\mathbf{H}_j(\cdot)$ instantiated on true and sample operators $\mathbf{T}$ and $\hat{\mathbf{T}}$, we are interested in the quantity

$$\|\mathbf{H}(\mathbf{T}) - \mathbf{H}(\hat{\mathbf{T}})\| = \min\{c \geq 0 : \|\mathbf{H}(\mathbf{T})\mathbf{x} - \mathbf{H}(\hat{\mathbf{T}})\mathbf{x}\| \leq c\|\mathbf{x}\|\}. \quad (5)$$

Our analysis requires the following assumptions.

Assumption 1. *(Vershynin 2018, Theorem 5.6.1) Given a random variable $\mathbf{x}$ and constants $G \geq 1$, $\delta \approx 0$, it holds:*

$$\mathbb{P}\left(\|\mathbf{x}\| \leq G\sqrt{\mathbb{E}[\|\mathbf{x}\|^2]}\right) \geq 1 - \delta.$$

Assumption 2. *(Loukas 2017, Theorem 4.1) The eigenvalues $\{w_i\}_{i=1}^N$ and $\{\hat{w}_i\}_{i=1}^N$ of the true and sample covariance matrix, respectively, satisfy for each pair (w_i, w_j), $i \neq j$,*

$$\mathrm{sign}(w_i - w_j)2\hat{w}_i > \mathrm{sign}(w_i - w_j)(w_i + w_j).$$

Assumption 1 quantifies the variance of the data distribution via the constant G, which is higher for data with higher variance. Assumption 2 considers the estimation error of sample covariance eigenvalues compared to the true ones, and holds for each eigenvalue pair (w_i, w_j) with probability at least $1 - 2k_i^2/(N|w_i - w_j|)$, where $k_i = \left(\mathbb{E}[\|\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{v}_i\|^2] - w_i^2\right)^{1/2}$ is a term related to the kurtosis of the data distribution (Loukas 2017, Corollary 4.2).

We now provide the covariance wavelet stability result.

Theorem 2. *Consider a covariance wavelet $\mathbf{H}_j(\cdot)$ with Lipschitz constant P and let Assumptions 1,2 hold. The output difference of the wavelet computed on the true and perturbed covariance wavelet operators $\mathbf{T}$ and $\hat{\mathbf{T}}$, respectively, is bounded with probability at least $(1 - e^{-\epsilon})(1 - 2e^{-u})$ as*

$$\|\mathbf{H}_j(\hat{\mathbf{T}}) - \mathbf{H}_j(\mathbf{T})\| \leq$$

$$\frac{PN}{\sqrt{T}}(k_{\max}e^{\frac{\epsilon}{2}} + \frac{2QG\gamma\|\mathbf{C}\|}{w_1}\sqrt{\log N + u}) + \mathcal{O}\left(\frac{1}{T}\right) := \Delta$$

where Q is an absolute constant, $k_{\max} = \max_j k_j$ with $k_j = \left(\mathbb{E}[\|\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{v}_j\|^2] - w_j^2\right)^{1/2}$ related to the kurtosis of the data distribution, and $\epsilon, u > 0$ are arbitrarily large constants.

This result provides three main insights. First, the bound increases with the sample dimension N, as larger covariance matrices are more difficult to estimate, and decreases with the number of observed samples T with the rate $\mathcal{O}(T^{-1/2})$, as the sample covariance gets closer to the true one the more samples are available. This is in contrast with the stability bound of graph wavelets (Gama, Ribeiro, and Bruna 2019), where the graph perturbation does not reduce with the number of samples. Second, compared to the PCA bound in (1), the covariance wavelet stability does not depend on the covariance eigengap, which is absorbed by the Lipschitz constant P. This is crucial because the constant P bounds the variability of the wavelet function $h_j(\lambda)$ w.r.t. the covariance eigenvalues, such that a smaller P corresponds to more slowly-varying $h_j(\lambda)$ and, consequently, better stability at the cost of lower discriminability for close eigenvalues, as common in graph and covariance neural networks (Gama, Bruna, and Ribeiro 2020; Sihag et al. 2022). The constant P depends on the wavelet definition and its parameters. For diffusion wavelets, P increases as the scale j increases, since a larger scale introduces sharper transitions in $h_j(\lambda)$ (see the Appendix for details on P for Hann and monic wavelets). Third, the bound is modulated by the largest covariance eigenvalue w_1 and the parameter γ due to the covariance normalization. This is beneficial when $w_1 > 1$ and $\gamma < 1$, as the bound gets lower and errors are reduced.

After establishing the stability of covariance wavelets, we proceed to provide a condition under which the pruned branches of the CST are the same on the perturbed and true covariance in the following proposition.

Proposition 1. *Consider a CST $\mathbf{\Phi}$ instantiated on true and sample covariance wavelet operators $\mathbf{T}$ and $\hat{\mathbf{T}}$, respectively, with T observations. Let the covariance wavelet $\mathbf{H}_j$ be Lipschitz with constant P and form a frame with bound B. The pruned trees with threshold τ of $\mathbf{\Phi}(\mathbf{T}, \mathbf{x})$ and $\mathbf{\Phi}(\hat{\mathbf{T}}, \mathbf{x})$ for a generic signal $\mathbf{x}$ are identical if, for all representations $\mathbf{x}_{(j_\ell,\ldots,j_1)}$ at layer ℓ and $j = 0, \ldots, J-1$, it holds:*

$$\left|\|\mathbf{H}_j(\mathbf{T})\mathbf{x}_{(j_\ell,\ldots,j_1)}\|^2 - \tau\|\mathbf{x}_{(j_\ell,\ldots,j_1)}\|^2\right| > (\Delta B^{\ell-1}\|\mathbf{x}\|)^2((\ell+1)B + \ell\tau).$$

where Δ is defined in Theorem 2.

This condition depends on the design of the CST via τ, B, P and the wavelet filter $\mathbf{H}_j(\cdot)$, on data characteristics via $\mathbf{C}$ and N and on the number of observed samples T. In particular, the condition becomes more likely as T increases, since the covariance estimation improves and the perturbation affects the pruning less. Moreover, the condition becomes more likely for smaller τ as the left-hand term increases while the right-hand term decreases.

With this in place, we investigate the stability to covariance estimation errors of CST.

Theorem 3. *Consider a CST $\mathbf{\Phi}(\cdot)$ with L layers and J scales operating on a true covariance operator $\mathbf{T}$ and sample operator $\hat{\mathbf{T}}$ estimated from T samples. Let Δ and B be the largest stability and frame bounds, respectively, among the wavelets in the CST, and $\|U\| \leq B_U$, and let the condition in Proposition 1 hold. The distance between the CST representations operating on the true and estimated operators $\mathbf{T}$, $\hat{\mathbf{T}}$ can be upper-bounded as*

$$\|\mathbf{\Phi}(\mathbf{T}, \mathbf{x}) - \mathbf{\Phi}(\hat{\mathbf{T}}, \mathbf{x})\| \leq B_U\Delta\|\mathbf{x}\|\sqrt{\sum_{\ell=1}^{L-1} \ell^2 B^{2\ell-2} F_\ell}.$$

where $F_\ell \leq J^\ell$ is the number of selected scattering features at layer ℓ.

Theorem 3 proves that CSTs are stable to covariance perturbations proportionally to the stability of a single wavelet Δ. The stability bound increases with the total number of active scattering features (such that more pruning leads to better stability), number of layers L and frame bound B. F_ℓ increases with increasing J and decreasing τ (e.g., $F_\ell = J^\ell$ for $\tau = 0$), making the CST less stable while improving its expressivity via more coefficients at different scales. A proper choice of pruning threshold τ can help in this tradeoff by allowing for wavelets at larger scales j that carry relevant information, while pruning less informative wavelets at smaller scales. The number of scales J also increases the wavelet bound Δ for diffusion wavelets, as more scales lead to sharper wavelet transitions and larger P. The CST compensates the reduced expressivity of a single wavelet by the cascade of wavelet filterbanks interleaved with nonlinearities, which spread information across the frequency spectrum and increase the discriminability at deeper layers, reiterating the advantage of using deep architectures (Isufi et al. 2024). Compared to VNNs (Sihag et al. 2022, Theorems 1-2), this bound can be made smaller via a larger τ, whereas VNNs do not have any pruning mechanism. Furthermore, the Lipschitz constant of VNNs depends on the training dynamics, whereas CSTs are untrained and their Lipschitz constant depends on the choice of wavelet functions and scales, thus it can be defined a priori. VNNs also achieve a tradeoff between stability and expressivity thanks to their hierarchical deep architecture, but they require labeled data for parameter training while CSTs achieve the same result without any training.

Numerical Results

We evaluate CSTs with the following objectives: **(O1)** validate their stability to covariance estimation errors; **(O2)** show the effectiveness of their representations for downstream tasks; **(O3)** assess the impact of pruning on performance and time and parameter efficiency.

Setup

Datasets. We consider four datasets containing cortical thickness measurements extracted from MRI scans for patients with a specific disease and healthy control patients: **ADNI1** and **ADNI2** (Jack Jr et al. 2008), **PPMI** (Marek et al. 2011) and **Abide** (Craddock et al. 2013) (details in Table 1). Cortical thickness is the thickness of the cerebral cortex in various regions of brain and is correlated to a patient's age, as it tends to thin when patients get older. We

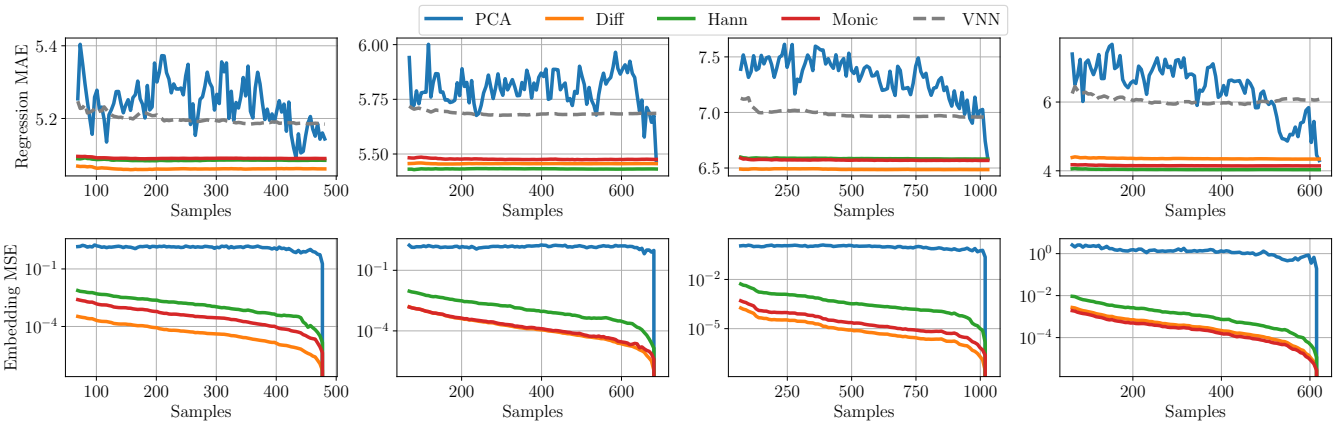

Figure 3: Age prediction Mean Average Error (MAE) and embedding Mean Squared Error (MSE) for increasing number of samples for CSTs, VNN and PCA on, from left to right, ADNI1, ADNI2, PPMI and Abide.

	ADNI1	ADNI2	PPMI	Abide
Patients (T)	801	1142	1704	1035
Brain areas (N)	68	68	68	62

Table 1: Dataset characteristics.

model the brain areas as N nodes and the thickness measures as node signals. Each of the T patients is an observation and the covariance among thickness measures constitutes our graph. Given this setup, our downstream regression task is to predict the patient's age, which is of high interest to identify neurodegenerative diseases or accelerated brain aging corresponding to a gap between predicted brain age and chronological age (Sihag et al. 2024a; Bashyam et al. 2020; Yin et al. 2023).

Models and baselines. We compare the 3 proposed implementations of CSTs (diffusion, Hann and monic) to (i) PCA, which processes covariance information in an unsupervised manner, but lacks stability and expressivity; (ii) VNN (Sihag et al. 2022), which achieves stability and expressive covariance manipulation via supervised training; (iii) a ridge regressor on the raw features, which does not consider covariance information. For the downstream task, we feed the representations of CSTs and PCA to a ridge regressor. We optimize all hyperparameters via grid search on a validation set. We report the grids and the final choices in Table 2 in the Appendix. We repeat every experiment over 10 different splits. We report additional experiments on controlled setups, more baselines and ablations on the real datasets in the Appendix. For all regression tasks, standard deviations are of the order 10^{-1}. We do not plot them for visual clarity, but we report them in Table 4 in the Appendix.

Stability and Regression Performance

Experimental setup. We keep 50% of the data as unlabeled (i.e., we do not use the age information of these patients for the downstream task), and we split the remaining as 10% for training, 20% for validation and 20% for testing. Let $\mathcal{U}$ be the union of the unlabeled and training sets. We compute the CST on $\mathcal{U}$ and we use it to produce representations for the test set, which we feed to a ridge regressor for the downstream task. To investigate the role of finite-sample covariance estimation errors, we recompute the CST using a covariance estimated from a subset of $\mathcal{U}$. With this perturbed CST, we produce new representations for the test samples, which we feed to the previously trained regressor for the downstream task. We do not perform thresholding.

Discussion. Figure 3 shows that CSTs maintain a consistent regression performance under covariance perturbations, demonstrating the empirical advantages of their stability **(O1)**. This significantly improves over PCA-based preprocessing, which suffers large instabilities in low-sample regimes, ultimately leading to significantly worse performance. The same observations stem from the embedding MSE (i.e., the quantity bounded in Theorem 3), which is contained for CSTs, while it grows larger for PCA. Moreover, CSTs' representations generally achieve better performance than PCA, VNN and raw features, demonstrating the informativeness of such embeddings and the usefulness of spectrally-localized information **(O2)**. VNN, while stable to covariance perturbations, achieves overall worse performance than CSTs, likely due to its higher training complexity which requires larger quantities of labeled data, whereas CSTs can exploit the unlabeled samples to produce meaningful representations.

Additional Results

We assess the impact of pruning, labeled data size and aggregation on ADNI1. The results for the other datasets are presented in the Appendix. We keep the same hyperparameter configuration as in Figure 3 except for the aggregation experiments where we re-optimize parameters on splits with 59.4% unlabeled - 0.6% train - 20% valid - 20% test sizes.

Impact of pruning. We evaluate the impact of pruning on model performance, time efficiency and number of selected features **(O3)**. Figure 4 shows that increasing τ does not lead to drastic changes in regression MAE, whereas it de-

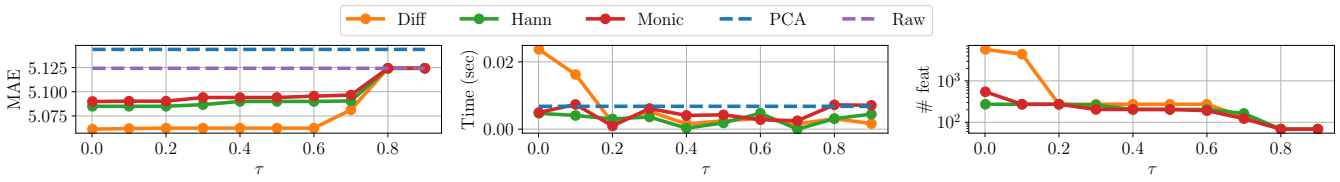

Figure 4: Impact of different thresholds τ for pruning on regression MAE, execution time and number of features on ADNI1.

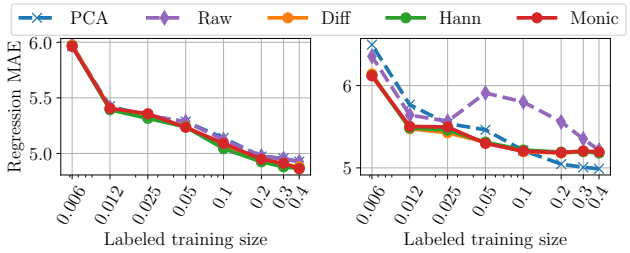

Figure 5: MAE for different labeled data sizes on ADNI1 for U as identity operator (left) and U as mean operator (right).

creases more significantly the number of features and execution time, leading to faster computations for CSTs compared to PCA. Therefore, the adopted pruning approach can significantly impact the time and memory requirements of the CST, while guaranteeing expressive data representations.

Impact of size of labeled dataset. We evaluate the CST's performance when the size of labeled data for training varies from 0.6% to 40% of the total dataset, while validation and test sets remain fixed to 20% each and the remaining portion of data is unlabeled. Figure 5 (left) shows that CSTs perform similarly or slightly better than PCA and raw features for all labeled data sizes. This corroborates the capability of CSTs to capture relevant information from unlabeled data, which is not exploited by pure learning approaches.

Dimensionality reduction. Finally, we evaluate the CST's performance when aggregating the scattering features via a mean operation for varying size of training data. Figure 5 (right) shows that reducing the data dimensions leads to advantages for low-training-data settings, where CSTs outperform raw features and PCA. For larger training data size, however, PCA leads to better results, as the regressor can effectively exploit its extensive information which is lost during the averaging aggregation of CST.

Related Works

Covariance-based learning. Covariance information is at the basis of several data processing techniques. In the unsupervised domain, PCA (Jolliffe 2002) and factor analysis (Child 2006), among others, are popular as they represent data in a low-rank space, where spurious correlations are removed and data dimension can be reduced. However, PCA is generally unstable to finite-sample covariance estimation errors, leading to unreliable component estimation in low-data regimes. This issue has been tackled by coVariance Neural Networks (VNNs) (Sihag et al. 2022), which rely on labeled data to estimate robust representations even in low-data regimes. This characteristic has made VNNs successful in a variety of settings, ranging from interpretable brain age estimation (Sihag et al. 2024a, 2022) to temporal data (Cavallo, Sabbaqi, and Isufi 2024), sparse covariances (Cavallo, Gao, and Isufi 2024; Cavallo et al. 2025b) and biased datasets (Cavallo et al. 2025a). Despite their success, VNNs and extensions need large quantities of labeled data for training, which may be unfeasible in practice. In this work, we provide a more flexible and robust framework to process data via their covariance information in an untrained manner.

Wavelets and scattering transforms. Wavelet transforms are popular tools in time, image and graph signal processing due to their space and frequency localization that allows for efficient signal representation and processing (Mallat 1999; Hammond, Vandergheynst, and Gribonval 2011; Shuman et al. 2015). Scattering transforms are cascades of wavelets interleaved with nonlinearities that achieve untrained deep hierarchical representations, and have been shown successful in a variety of domains ranging from images (Bruna and Mallat 2013) to graphs (Gama, Bruna, and Ribeiro 2019, 2020; Koke and Kutyniok 2022), audio (Andén and Mallat 2011) and simplicial complexes (Madhu, Gurugubelli, and Chepuri 2024). Their main advantages are their stability to domain and signal perturbations as well as their capability to extract expressive frequency patterns in an untrained way. In this work, we build on this literature to propose scattering transforms on covariance matrices, study their link with PCA and VNNs via covariance spectrum processing and their increased stability to finite-sample estimation errors.

Conclusions

We introduced Covariance Scattering Transforms (CSTs), untrained hierarchical deep networks that generate expressive representations for data by manipulating their covariance matrix. We proved that CSTs are permutation equivariant, achieve stability to covariance perturbations and produce rich data embeddings that lead to good performance on an age prediction task from cortical thickness measurements in four different datasets. CSTs present a tradeoff between increasing the sample dimension for improved expressivity or reducing it via low-pass aggregations that lose information. Future work will address this aspect and propose more effective aggregation functions. Moreover, the application of this framework to other settings where covariance information plays a crucial role, such as financial data and sensor measurements, represents a promising extension.

Acknowledgements

Part of this work was funded by the TU Delft AI Labs program, the NWO OTP GraSPA proposal #19497, the NWO VENI proposal 222.032.

References

Andén, J.; and Mallat, S. 2011. Multiscale Scattering for Audio Classification. In *ISMIR*, 657–662. Miami, Florida.

Bashyam, V. M.; Erus, G.; Doshi, J.; Habes, M.; Nasrallah, I. M.; Truelove-Hill, M.; Srinivasan, D.; Mamourian, L.; Pomponio, R.; Fan, Y.; et al. 2020. MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain*, 143(7): 2312–2324.

Bessadok, A.; Mahjoub, M. A.; and Rekik, I. 2022. Graph neural networks in network neuroscience. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5833–5848.

Bruna, J.; and Mallat, S. 2013. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1872–1886.

Cavallo, A.; Gao, Z.; and Isufi, E. 2024. Sparse Covariance Neural Networks. arXiv:2410.01669.

Cavallo, A.; Navarro, M.; Segarra, S.; and Isufi, E. 2025a. Fair covariance neural networks. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Cavallo, A.; Rey, S.; Marques, A. G.; and Isufi, E. 2025b. Precision neural networks: Joint graph and relational learning. *arXiv preprint arXiv:2509.14821*.

Cavallo, A.; Sabbaqi, M.; and Isufi, E. 2024. Spatiotemporal Covariance Neural Networks. In Bifet, A.; Davis, J.; Krilavičius, T.; Kull, M.; Ntoutsi, E.; and Žliobaitė, I., eds., *Machine Learning and Knowledge Discovery in Databases. Research Track*, 18–34. Cham: Springer Nature Switzerland. ISBN 978-3-031-70344-7.

Child, D. 2006. *The essentials of factor analysis*. A&C Black.

Craddock, C.; Benhajali, Y.; Chu, C.; Chouinard, F.; Evans, A.; Jakab, A.; Khundrakpam, B. S.; Lewis, J. D.; Li, Q.; Milham, M.; et al. 2013. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 7(27): 5.

Gama, F.; Bruna, J.; and Ribeiro, A. 2019. Diffusion scattering transforms on graphs. In *7th International Conference on Learning Representations, ICLR 2019*.

Gama, F.; Bruna, J.; and Ribeiro, A. 2020. Stability properties of graph neural networks. *IEEE Transactions on Signal Processing*, 68: 5680–5695.

Gama, F.; Ribeiro, A.; and Bruna, J. 2019. Stability of graph scattering transforms. *Advances in Neural Information Processing Systems*, 32.

Hammond, D. K.; Vandergheynst, P.; and Gribonval, R. 2011. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2): 129–150.

Ioannidis, V. N.; Chen, S.; and Giannakis, G. B. 2020. Pruned graph scattering transforms. In *International Conference on Learning Representations*.

Isufi, E.; Gama, F.; Shuman, D. I.; and Segarra, S. 2024. Graph Filters for Signal Processing and Machine Learning on Graphs. *IEEE Transactions on Signal Processing*, 1–32.

Jack Jr, C. R.; Bernstein, M. A.; Fox, N. C.; Thompson, P.; Alexander, G.; Harvey, D.; Borowski, B.; Britson, P. J.; L. Whitwell, J.; Ward, C.; et al. 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4): 685–691.

Jolliffe, I. 2002. *Principal component analysis*. New York: Springer Verlag.

Jolliffe, I. T.; and Cadima, J. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374.

Koke, C.; and Kutyniok, G. 2022. Graph scattering beyond wavelet shackles. *Advances in Neural Information Processing Systems*, 35: 30219–30232.

Loukas, A. 2017. How Close Are the Eigenvectors of the Sample and Actual Covariance Matrices? In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 2228–2237. PMLR.

Madhu, H.; Gurugubelli, S.; and Chepuri, S. P. 2024. Unsupervised Parameter-free Simplicial Representation Learning with Scattering Transforms. In *Forty-first International Conference on Machine Learning*.

Mallat, S. 1999. *A wavelet tour of signal processing*. Elsevier.

Marek, K.; Jennings, D.; Lasch, S.; Siderowf, A.; Tanner, C.; Simuni, T.; Coffey, C.; Kieburtz, K.; Flagg, E.; Chowdhury, S.; et al. 2011. The Parkinson progression marker initiative (PPMI). *Progress in neurobiology*, 95(4): 629–635.

Shuman, D. I.; Wiesmeyr, C.; Holighaus, N.; and Vandergheynst, P. 2015. Spectrum-adapted tight graph wavelet and vertex-frequency frames. *IEEE Transactions on Signal Processing*, 63(16): 4223–4235.

Sihag, S.; Mateos, G.; McMillan, C.; and Ribeiro, A. 2022. coVariance neural networks. *Advances in neural information processing systems*, 35: 17003–17016.

Sihag, S.; Mateos, G.; McMillan, C.; and Ribeiro, A. 2024a. Explainable brain age prediction using covariance neural networks. *Advances in Neural Information Processing Systems*, 36.

Sihag, S.; Mateos, G.; McMillan, C.; and Ribeiro, A. 2024b. Transferability of coVariance Neural Networks. *IEEE Journal of Selected Topics in Signal Processing*, 18(2): 199–215.

Vershynin, R. 2018. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. ISBN 9781108415194.

Yin, C.; Imms, P.; Cheng, M.; Amgalan, A.; Chowdhury, N. F.; Massett, R. J.; Chaudhari, N. N.; Chen, X.; Thompson, P. M.; Bogdan, P.; et al. 2023. Anatomically interpretable deep learning of brain age captures domain-specific cognitive impairment. *Proceedings of the National Academy of Sciences*, 120(2): e2214634120.