Latent Variable Graphical Model Selection using Harmonic Analysis: Applications to the Human Connectome Project (HCP)

Won Hwa Kim[†1]   Hyunwoo J. Kim[†1]   Nagesh Adluru[‡]   Vikas Singh[§†]

[†]Dept. of Computer Sciences, University of Wisconsin, Madison, WI, U.S.A.
[§]Dept. of Biostatistics & Med. Informatics, University of Wisconsin, Madison, WI, U.S.A.
[‡]Waisman Center, Madison, WI, U.S.A.

Abstract

A major goal of imaging studies such as the (ongoing) Human Connectome Project (HCP) is to characterize the structural network map of the human brain and identify its associations with covariates such as genotype, risk factors, and so on that correspond to an individual. But the set of image derived measures and the set of covariates are both large, so we must first estimate a 'parsimonious' set of relations between the measurements. For instance, a Gaussian graphical model will show conditional independences between the random variables, which can then be used to setup specific downstream analyses. But most such data involve a large list of 'latent' variables that remain unobserved, yet affect the 'observed' variables sustantially. Accounting for such latent variables is not directly addressed by standard precision matrix estimation, and is tackled via highly specialized optimization methods. This paper offers a unique harmonic analysis view of this problem. By casting the estimation of the precision matrix in terms of a composition of low-frequency latent variables and high-frequency sparse terms, we show how the problem can be formulated using a new wavelet-type expansion in non-Euclidean spaces. Our formulation poses the estimation problem in the frequency space and shows how it can be solved by a simple sub-gradient scheme. We provide a set of scientific results on ∼500 scans from the recently released HCP data where our algorithm recovers highly interpretable and sparse conditional dependencies between brain connectivity pathways and well-known covariates.

1. Introduction

Consider a large scale neuroimaging study, e.g., the ongoing Human Connectome Project (HCP), where diffusion weighted magnetic resonance images (diffusion MRI) are acquired for a cohort of participants. Each subject provides a variety of clinical and cognitive measures in addition to the images, as well as demographic information such as age, gender, education status and so on. Such a rich data resource offers an unprecedented opportunity to answer many scientific questions. For instance, how do brain networks differ across gender, and does education or genotype have an association with structural brain connectivity *beyond* the expected effects of age? Until recently, the scientific community had limited means to answer such questions because public datasets were either small, not well curated or the imaging protocols used for acquisition were too heterogeneous. The recent public release of images (and covariates) from the HCP study makes such an analysis possible *if we can address the associated modeling issues that arise in performing inference on such a high dimensional dataset.*

A fundamental scientific goal in statistical analysis of HCP (and similar datasets) is to identify associations between the full set of variables and the entire spectrum of image-derived measurements [1, 30, 26, 24]. For example, are a subset of the clinical covariates highly predictive of the inter-regional connectivity derived from the images? The traditional approach here may proceed by estimating a graphical model that best explains the data: where the nodes correspond to the full set of covariates (image-derived measures *and* clinical/cognitive scores) as jointly Gaussian random variables. By estimating the inverse of the covariance matrix between the variables, we precisely recover the graphical model structure. This may then be used to setup hypothesis driven structural equation models (SEM) or simple regression model based experiments. The difficulty is that in many modern image analysis problems, the total number of such covariates, say p, is far larger than the number of samples (subjects) n in the study. Classical model selection is highly problematic in this high dimensional setting since the empirical statistics are often poorly behaved. The popular solution here is to impose a *sparsity* regularization on the inverse of the covariance matrix Σ^{-1}. Using a ℓ_1 penalty on the entries of this matrix, under mild conditions, one can guarantee that the maximum likelihood solution will recover the true model [45, 13]. In the last five years,

---

[1]Won Hwa Kim and Hyunwoo J. Kim are joint first authors.

this idea has been extensively used in a broad spectrum of applications in computer vision [15, 31], machine learning [2, 35, 29, 44] and medical imaging [38, 43, 19, 11].

The formulation above, given its broad applicability, has been heavily studied and we now have a comprehensive treatment of efficient optimization routines [3, 7, 36, 33] and regularization properties [35, 28]. These developments notwithstanding, there are various situations in medical image analysis, computational biology and other applications, which are not an ideal fit for the standard sparse inverse covariance matrix estimation model. For instance, in many real-world studies, there are a non-trivial number of *latent* variables that either *cannot* be directly observed or can only be measured at a high monetary cost or discomfort to the subject. The incorporation of such latent variables in the estimated structural relationship, generally called "latent variable graphical models", is not as extensively studied.

Related Work. There is some degree of consensus that a straightforward incorporation of such 'latent' variables in the default construction described above is problematic. Therefore, existing approaches [10] must pre-specify the number of such variables in an ad-hoc manner and proceed with a bi-level non-convex scheme to estimate the parameters. There are other combinatorial heuristics [12] which cluster the observed variables and assign them incrementally to a latent variable. The practical effectiveness of such algorithms varies and they offer few theoretical guarantees. An interesting recent paper [5] resolves many of these problems and presents an algorithm where all variables (observed and latent) are jointly Gaussian. The main idea is to approximate the sample covariance matrix Σ in a way where the corresponding Σ^{-1} is expressed as a sum of a sparse matrix and a low-rank matrix. This recovers the influence of the unobserved latent variables as well as the conditional graphical model, as desired. This strategy works well as long as the low rank requirement remains valid; however, as the number of latent variables grow, the data will deviate farther from the low rank assumption. Consequently, the sparse term must explain a larger 'mass' of the data and the estimated matrix becomes denser.

Motivating this paper. The above discussion suggests that the means of regularizing the degrees of freedom (i.e., the low rank term) for the *latent components* may be less than ideal from a numerical perspective as the number of latent variables grow. This issue is, to our knowledge, does not have a simple 'fix'. To put this in perspective, notice that the literature suggests that *high rank matrix completion* (columns of the matrix belong to a union of multiple low-rank subspaces) uses a set of concepts that are quite different from those used for completing low-rank matrices (e.g., nuclear norm). So, a potential solution in our graphical model setting must also look for alternatives to the algebraic characterization (used in [5]). Certain classical tools at

the high level, express a closely related intuition. Consider the following simple idea. If we think of the precision matrix as the composition of *low* and *high* frequency terms, the lower order terms may easily serve as a proxy for the latent components. Then, by asking that the remaining contribution should be sparse yields a similar overall effect as [5] but involves *no* spectral relaxations of the rank constraint. Harmonic analysis offers a natural tool for such needs via wavelets, which come with the full set of benefits of Fourier analysis but without the *global* bases — undesirable in this discrete setting. While wavelet expansions have traditionally been studied only for Euclidean spaces, recent results from [16, 6] provide mechanisms for wavelet (and Fourier) transform of discrete (non-Euclidean) spaces such as graphs as well. This discussion suggests that the potential ingredients for formulating our graphical model estimation problem in a dual space *are* available and provide at least a good starting point for the current problem.

The main **contribution** of this work is to demonstrate how this latent graphical model estimation problem can be viewed via the lens of harmonic analysis. We show that by operating on the inverse covariance matrix via its associated graph (actually a wavelet transform of this graph), it becomes an inference problem casted *in the frequency space*. The actual optimization requires no sophisticated solvers, we only need to perform a simple gradient descent on one variable that controls the band-pass filtering property of wavelets. Our motivating application is the analysis of the Human Connectome Project (HCP) dataset which includes more than ~ 350 covariates (and therefore, many latent variables) together with a rich set of imaging data. Here, we obtain neuroscientifically meaningful sparse models relating image-derived brain connectivity to covariates where alternative approaches yield uninterpretable results.

2. Multi-resolution Analysis of Euclidean/Non-Euclidean Spaces via Wavelets

Our proposed framework relies heavily on a novel multi-resolution perspective of the given data. Specifically, our model will reformulate the statistical model estimation problem utilizing the theory of wavelets to understand the underlying structural associations in the given data. We therefore briefly provide an overview of the wavelet transform in both Euclidean and non-Euclidean spaces.

2.1. Continuous Wavelet Transform

The wavelet transform is similar to the Fourier transform in that a function is decomposed as a linear combination of coefficients and certain basis functions. While the Fourier expansion uses $\sin()$ bases with infinite support, the wavelet expansion instead uses a mother wavelet basis ψ which is nicely localized in *both time and frequency*.

A description of the classical wavelet transform must start by defining a mother wavelet $\psi_{s,a}$ with a scale parameter t and a translation parameter a as

$$\psi_{s,a}(x) = \frac{1}{s}\psi(\frac{x-a}{s}) \qquad (1)$$

where s (and a) control the dilation (and localization) of $\psi_{s,a}$ respectively. Using $\psi_{s,a}$ as the bases, the wavelet transform of a function $f(x)$ is defined as the inner product of the ψ and f, represented as

$$W_f(s,a) = \langle f, \psi \rangle = \frac{1}{s}\int f(x)\psi^*(\frac{x-a}{s})\mathrm{d}x \qquad (2)$$

where $W_f(s,a)$ is the wavelet coefficient at scale s and at location a, and ψ^* is the complex conjugate of ψ. Together with the wavelet coefficients and the wavelet bases, we obtain a wavelet expansion. Observe the similarity of (2) to the Fourier expansion given as

$$\hat{f}(\omega) = \underbrace{\int f(x)e^{-j\omega x}\mathrm{d}x}_{\langle f, \mathrm{basis}\rangle} \qquad (3)$$

Interestingly, in the frequency domain, the mother wavelets ψ_s at multiple scales behave as band-pass filters corresponding to different bandwidths. Since these band-pass filters do *not* cover the low-frequency components, an additional low-pass filter is typically introduced: a scaling function ϕ. We will shortly discuss the significant benefits of this low-pass property to express the latent components in our formulation. A wavelet transform with the scaling function ϕ returns a smooth representation of the original function f. Due to this selective filtering property, wavelets offer a nice multi-resolution view of the given signal.

2.2. Wavelet Transforms in Non-Euclidean Spaces

The implementation of a mother wavelet and wavelet transform in the Euclidean space (i.e., represented as a regular lattice) is convenient since one can easily define the 'shape' of a mother wavelet. This concept has been extensively used within computer vision and image processing for nearly three decades. However, in a non-Euclidean setting where the domain is irregular (e.g., a graph), the notions of scale and translation are not as easy to conceptualize. For instance, in a graph, the distance between each vertex and the number of connected edges is not uniform, it is difficult to define a localized mother wavelet function at a specific scale. Due to this difficulty, the wavelet transform has not been traditionally suitable for analysis when the domain has an arbitrary structure until very recently when [6, 16] presented a result dealing with wavelet and Fourier transform of graphs (and other non-Euclidean spaces).

High level summary. The key idea behind the non-Euclidean wavelet transform in constructing a mother wavelet ψ on the nodes of a graph G is to utilize a kernel function and a set of orthonormal bases. Recall that the wavelet behaves as a band-pass filter in the frequency domain. So, if we construct a *band-pass* filter (a kernel function) in the frequency domain and then localize it in the original (i.e., native) domain using graph Fourier transform in [16], it will exactly implement a mother wavelet ψ on the original graph. The preferred representation of the graph is its Laplacian whose eigenvectors provide the 'bases' for transforming the graph to the frequency space.

Formally, a graph $G = \{V, E\}$ is defined by a vertex set V (where the number of vertices is N) and a edge set E. Such a graph G is generally represented as an adjacency matrix A of size $N \times N$ where the elements a_{ij} denote the edge weight between ith and jth vertices. Then a degree matrix D is defined as a diagonal matrix where the ith diagonal is the sum of edge weights connected to the ith vertex. From these two graph matrices, a graph Laplacian is defined as $L = D - A$. Here, L is self-adjoint and positive semi-definite, therefore the spectrum of L yields eigenvalues $\lambda_l \geq 0$ and corresponding eigenvectors χ_l where $l = 0, 1, \cdots N - 1$. The orthonormal bases χ allow one to setup the graph Fourier transform as

$$\hat{f}(l) = \sum_{n=1}^{N}\chi_l^*(n)f(n) \text{ and } f(n) = \sum_{l=0}^{N-1}\hat{f}(l)\chi_l(n) \qquad (4)$$

where $\hat{f}(l)$ is the graph Fourier coefficient. Interestingly, this transform offers a convenient means for transforming a signal/measurement on graph nodes/vertices to the frequency domain. Utilizing the graph Fourier transform, the mother wavelet ψ can be constructed by first defining a kernel function $g(\cdot)$ in the frequency domain and then localizing it by a delta function δ in the original graph via the inverse graph Fourier transform. Since $\langle \delta_n, \chi_l \rangle = \chi_l^*(n)$, the mother wavelet $\psi_{s,n}$ at vertex n at scale s is defined as

$$\psi_{s,n}(m) = \sum_{l=0}^{N-1}g(s\lambda_l)\chi^*(n)\chi_l(m). \qquad (5)$$

Notice that the scale s is defined *inside* g by the scaling property of Fourier transform [37] and the eigenvalues serve as the analogs of frequency. Using ψ, the wavelet transform of a function f at scale s can be expressed simply as

$$W_f(s,n) = \langle f, \psi_{s,n} \rangle = \sum_{l=0}^{N-1}g(s\lambda_l)\hat{f}(l)\chi_l(n) \qquad (6)$$

resulting in wavelet coefficients $W_f(s,n)$. Such a transform offers a multi-resolution view of signals on graphs [25]. *The most important fact relevant to our formulation is that the multi-resolution property can be easily captured by a single parameter s in the kernel function g which controls the low-pass properties of the transform entirely.*

3. A Harmonic Analysis of Latent Variable Graphical Models

With the aforementioned wavelet concepts in hand, we can now describe our formulation for estimating a precision matrix while concurrently taking into account the effect of an unknown but large number of latent components. Our procedure below will parameterize the to-be-inferred graphical model *not* in terms of its precision matrix directly, rather via its low and high frequency components. Operating on these latent (low-frequency) and sparse (high-frequency) pieces will model the structural associations within the graph. Recall that recent developments in wavelet analysis on discrete spaces such as graphs have overwhelmingly been used to analyze signals defined on the nodes where the graph has a "fixed" (known) structure. In order to apply wavelet analysis to our problem, we will need to introduce a few key technical results that are summarized below, and described in detail in this section. **(a)** First, we will introduce multi-resolution analysis for modeling the graph *structure* and *not* just the measurement at individual graph nodes. We will define a new set of basis functions for estimating the graph structures and provide theoretical conditions which guarantee its validity. **(b)** Second, we will introduce an information theoretic "closeness" measure for graph structure (i.e., precision matrices). Here, we will identify an additional condition which will yield a valid symmetric positive definite matrix at each scale s. **(c)** Finally, we will discuss our optimization scheme in the dual space (i.e., frequency domain) with a simple gradient descent method.

3.1. Multi-scale Analysis of a Precision Matrix

Let us assume we are given a positive definite covariance matrix Σ of size $n \times n$. Now, Σ can be easily decomposed in terms of its eigenvector and eigenvalues as,

$$\Sigma = V \Lambda V^T = \sum_{\ell=1}^{n} \lambda_\ell V_\ell V_\ell^T \qquad (7)$$

where the ℓth column vector of V is the ℓth eigenvector and the ℓth diagonal of Λ is the corresponding ℓth eigenvalue of Σ which are all positive. Then, the precision matrix Θ is given as the inverse of the covariance matrix as

$$\Theta = \sum_{\ell=1}^{n} \frac{1}{\lambda_\ell} V_\ell V_\ell^T = \sum_{\ell=1}^{n} \sigma_\ell V_\ell V_\ell^T \qquad (8)$$

where $\sigma = \frac{1}{\lambda}$ and σ are positive since λ are positive. Notice that both Σ and Θ are positive definite and self-adjoint, so their eigenvectors can be used for defining a Fourier type of transform which is analogous to the graph Fourier transform as in (4). For multi-resolution analysis of the precision matrix Θ, we first define our basis functions as

$$\psi_{\ell,s}(i,j) = g(s\sigma_\ell) V_\ell^*(i) V_\ell(j), \forall \ell \in \{1,\ldots,n\} \qquad (9)$$

at scale s and along the ℓth basis. Since we deal only with real valued functions, to avoid notational clutter, we will omit the conjugate operation for the eigenfunctions, i.e., $V^*(i) = V(i)$. These basis functions are analogous to mother wavelets and yield a nice result which we will present shortly. Now, we can easily setup a transform of the precision matrix using our basis above. This yields wavelet-like coefficients as

$$\begin{aligned} W_{\Theta,s}(\ell) &= \langle \Theta, \psi_{\ell,s} \rangle \qquad (10) \\ &= \sum_{i,j}^{n,n} \sum_{\ell'=1}^{n} \sigma_{\ell'} V_{\ell'}(i) V_{\ell'}(j) g(s\sigma_\ell) V_\ell(i) V_\ell(j) \\ &= \sigma_\ell g(s\sigma_\ell). \end{aligned}$$

Using $W_{\Theta,s}(\ell)$, the multi-resolution reconstruction with a non-constant weight ds/s is obtained by

$$\tilde{\Theta}(i,j) = \frac{1}{C_g} \int_0^\infty \frac{1}{s} \sum_{\ell=1}^{n} W_{\Theta,s}(\ell) \psi_{\ell,s}(i,j) ds. \qquad (11)$$

Roughly speaking, this can be viewed as the weighted average of multi-resolution reconstruction over scale s.

A natural question here is whether we can guarantee if the reconstruction in (11) is identical to the original precision matrix Θ. To address this issue, we define the *admissibility condition* for the function defined on the structure (or *edges*) of the graph. A kernel $g(x)$ is said to satisfy the *admissibility condition* if the following condition holds

$$\int_0^\infty \frac{g^2(x)}{x} dx =: C_g < \infty \qquad (12)$$

when the reconstruction is defined with a non-constant weight dx/x as (11). Lemma 1 below shows that using the bases we constructed in (9), if $g(x)$ satisfies *admissibility condition*, the matrix reconstruction in (11) is *identical*, namely, $\tilde{\Theta}(i,j) = \Theta(i,j)$.

Lemma 1 *If $\Theta \succ 0, \Theta = \Theta^T$ and kernel g satisfies the admissibility condition*

$$\int_0^\infty \frac{g^2(s\sigma)}{s} ds =: C_g < \infty \qquad (13)$$

then,

$$\frac{1}{C_g} \int_0^\infty \frac{1}{s} \sum_{\ell=1}^{n} W_{\Theta,s}(\ell) \psi_{\ell,s}(i,j) ds = \Theta(i,j) \qquad (14)$$

The full proof of this lemma is available in the supplement.

We can derive a stronger result showing that using the bases in (9), the admissibility condition holds for two parameter kernels as well, i.e., $g(s,\sigma)$. This allows defining a kernel, if desired, that separately handles the influence of the eigen value σ and a scale parameter s.

Lemma 2 *If kernel g satisfies the admissibility condition*

$$\int_0^\infty \frac{g^2(s,\sigma)}{s}ds =: C_g < \infty \qquad (15)$$

then,

$$\frac{1}{C_g}\int_0^\infty \frac{1}{s}\sum_{\ell=1}^n W_{\Theta,s}(\ell)\psi_{\ell,s}(i,j)ds = \Theta(i,j) \qquad (16)$$

This two parameter kernel result can be used for functions defined on either nodes (commonly used in non-Euclidean Wavelets) or edges (graph structure). The proof is available in the supplement. The admissibility condition for the classical SGWT (one parameter kernel for the functions defined on nodes) is studied in [16] and consistent with our result. Based on this harmonic analysis of graphical models, we next describe our main estimation algorithm to recover the sparse precision matrix by explicitly taking into account the contribution of the latent components.

4. Estimating the Optimal Scale for $\tilde\Theta$

In this subsection, we describe the optimization scheme to estimate $\tilde\Theta$ which satisfies two properties: **i)** it is consistent with the empirical Θ and **ii)** satisfies sparsity properties (in the sense of the multi-resolution characterization)

The reconstruction of Θ at level s is given by

$$\tilde\Theta = \sum_{\ell=1}^n \sigma_\ell g^2(s\sigma_\ell)V_\ell V_\ell^T = \sum_{\ell=1}^n K(s,\sigma_\ell)V_\ell V_\ell^T \qquad (17)$$

where $K(s,\sigma_\ell) := \sigma_\ell g^2(s\sigma_\ell)$. To keep notations concise, we will often use K as shorthand in this subsection. To perform the reconstruction at every level s, the kernel function g should satisfy the condition, $g^2(x) > 0, \forall x \geq 0$. Then, one can easily check that $\tilde\Theta$ is symmetric positive definite, i.e., $\tilde\Theta \in$ SPD, exactly as desired.

At a high level, we seek for a $\tilde\Theta$ which is similar to the empirical (potentially non-sparse) estimate, Θ. To do so, we need to define "closeness" between our estimate $\tilde\Theta$ and Θ. Instead of using a metric directly for SPD matrices, i.e., Riemannian metric used in [22, 23], in this paper, we regard them as two corresponding Gaussian distributions with zero mean, but with covariance matrices $\tilde\Sigma$ and Σ.

Using KL-divergence $\text{KL}(\cdot\|\cdot)$ between the two Gaussian densities, we can measure "closeness" by

$$\text{KL}(p(x;\tilde\Sigma)\|p(x;\Sigma)) = \frac{1}{2}\text{D}_{\text{ld}}(\Sigma,\tilde\Sigma) = \frac{1}{2}D_{\text{ld}}(\tilde\Theta,\Theta) \qquad (18)$$

The last two identities express closeness by Bregman divergence using the log determinant, $\text{D}_{ld}(\cdot\|\cdot)$ as in [8].

With this fidelity measure, our objective is to find the optimal scale s which minimizes the Bregman divergence using $\text{logdet}(\cdot)$ between the empirical precision matrix Θ and

the sparse reconstruction $\tilde\Theta$. We impose a sparsity penalty in the usual way using the ℓ_1-norm of the matrix. Then, our optimization problem is given as,

$$\max_{s\geq 0}\ \ \text{tr}(\tilde\Theta\Theta^{-1}) - \text{logdet}(\tilde\Theta\Theta^{-1}) - n + \gamma|\tilde\Theta|_1 \qquad (19)$$

$$\text{subject to }\tilde\Theta = \sum_{\ell=1}^n \sigma_\ell g^2(s\sigma_\ell)V_\ell V_\ell^T. \qquad (20)$$

Substituting in the identity from (7) for Θ, we obtain an almost unconstrained optimization model (which only involves one non-negativity constraint),

$$\max_{s\geq 0}\sum_{\ell=1}^n \lambda_\ell K(s,\sigma_\ell) - \sum_{\ell=1}^n \log(\lambda_\ell K(s,\sigma_\ell)) - n$$
$$+ \gamma\sum_{i=1}^n\sum_{j=1}^n\left|\sum_{\ell=1}^n K(s,\sigma_\ell)X_\ell(i,j)\right| \qquad (21)$$

where $X_\ell = V_\ell V_\ell^T$ and $X_\ell(i,j)$ is i,jth element in X. The optimal (sparse) precision matrix will then correspond to some s which minimizes (20) or (21).

Deriving the first derivative to optimize (21). To optimize (20), we compute the first derivative of D with respect to s, which can be written as

$$\frac{d}{ds}\text{tr}(\tilde\Theta\Theta^{-1}) - \frac{d}{ds}\text{logdet}(\tilde\Theta\Theta^{-1}) + \frac{d}{ds}\gamma|\tilde\Theta|_1 \qquad (22)$$

Here, we calculate $\frac{d}{ds}\text{tr}(\sum_{\ell=1}^n \lambda_\ell K(s,\sigma_\ell)V_\ell V_\ell^T)$ taking a derivative of each element and then taking the sum of the diagonal elements of $\tilde\Theta\Theta^{-1}$ and we obtain,

$$\frac{d}{ds}\text{tr}(\sum_{i=1}^n \lambda_\ell K(s,\sigma_\ell)V_\ell V_\ell^T) = \sum_{\ell=1}^n \lambda_\ell K'(s,\sigma_\ell) \qquad (23)$$

where $K'(s,\sigma_\ell) := \partial K/\partial s$. The derivative of the second term takes the form,

$$\frac{d}{ds}\text{logdet}(\tilde\Theta\Sigma) = \sum_{i=1}^n \frac{K'(s,\sigma_i)}{K(s,\sigma_i)} \qquad (24)$$

Notice that the third term involves the ℓ_1 norm which is not differentiable, so we approximate its search direction instead as

$$\sum_{i=1}^n\sum_{j=1}^n \text{sign}(\tilde\Theta(i,j))\sum_{\ell=1}^n K'(s,\sigma_\ell)X_\ell(i,j). \qquad (25)$$

Combining all three terms together yields a direction to optimize (20). The actual optimization then only involves a simple gradient descent-like method.

Remarks. Observe that a precision matrix always has non-zero diagonal elements. So, the sparsity regularization may not be meaningful for diagonal elements. One can impose sparsity for only the off-diagonal elements with minor changes in the third term (21) and its search direction (25), namely, $\sum_{i\neq j}|\tilde\Theta_{(i,j)}|$ and its search direction is

$$\sum_{i\neq j}^{n,n} \text{sign}(\tilde\Theta(i,j))\sum_{\ell=1}^n K'(s,\sigma_\ell)X_\ell(i,j). \qquad (26)$$
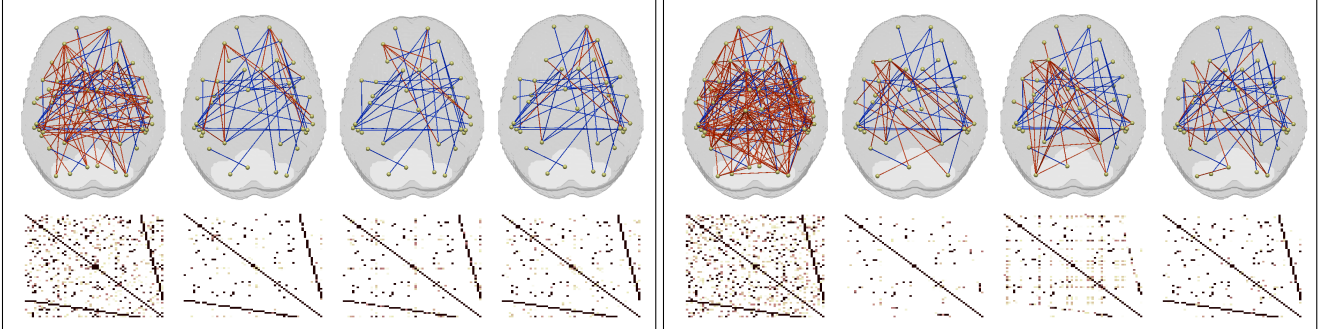
Figure 1. Comparison of results from estimation of statistical dependencies between observed variables (when there are at least a few latent components) using synthetic brain network data. Left/right blocks show results for 5 and 10 latent variables respectively, and the top/bottom rows show estimated dependencies in the data (correct estimation in blue and false positive in red) and corresponding precision matrices. First column: sample precision matrix, Second column: result using GLasso, Third column: result using [5], Fourth column: our result. We can observe that while the sample precision matrix is dense, the results in the second, third and fourth column show sparse and more accurate results.

5. Experimental Results

We demonstrate two sets of experiments, one on synthetic brain network data to validate our framework where the ground truth is available, and the other on the Human Connectome Project (HCP) data. The first experiment evaluates precision matrix estimation results using our framework by comparing it to the estimations from other methods and the ground truth. In the second experiment, we analyze an exquisite recently released imaging dataset of ~ 500 individuals from the Human Connectome Project. We obtain brain connectivity pathways by processing Diffusion Tensor Images (DTI) and analyze this connectivity data jointly with a rich set of covariates. Among the many inter-regional fiber bundles, we focus our analysis on 17 major connections and identify which of the covariates are statistically associated with these major pathways. In both experiments, the objective is to estimate true dependencies between the observed variables when the latent variables are unobserved.

5.1. Statistical Dependency Estimation on Synthetic Brain Connectivity Data

In this section, we demonstrate results of precision matrix estimation using synthetic brain connectivity data. Consider a case where we observe a set of $n_p + n_c$ random variables, i.e., a set of n_p structural brain connections (i.e., pathways) and n_c covariates. We add additional n_h number of latent variables that are assumed to be *unobserved* but statistically influence the full set of observed variables. Then, the statistical model estimation task is to find the true conditional dependencies between the observed variables alone, i.e., properly taking into account the effect of latent factors. In other words, we want to identify which brain connections are statistically associated with the covariates as well as how these pathways are related to one another.

Design. We set $n_p = 50$ and $n_c = 10$, so the total number of observed variables $n_o = 60$. We run multiple replicates, each for a different setting for the number of latent variables n_h. The dependencies between the brain connections are arbitrarily chosen such that 5% of the elements in the true precision matrix (i.e., ground truth) are non-zeros. We set each covariate to be dependent on the brain connections in a pattern (i.e., first five connections depend on the first covariate, next five connections depend on the second covariate, etc.). These dependencies between observed variables are the ground truth and can be visual checked to see if we estimate the same pattern. The latent variables are then connected to all observed variables with random weights; this ensures that our measurements of the observed variables include an effect from all latent variables. This yields a precision matrix Θ of size $n \times n$ and its corresponding covariance matrix Σ. Synthetic data are sampled from a multi-variate normal distribution using Σ. We draw samples only from the observed variables to construct a sample covariance matrix Σ_o, which serves as the input for estimating $\tilde{\Theta}_o$.

Experimental results with $n_h = 5$ and $n_h = 10$ latent variables are shown in the left/right blocks of Fig. 1 respectively. Fig. 1 (top row) shows the estimated dependencies between the full set of connections (covariates are not shown in the top row). The small spheres represent the physical centers of each brain connection and the edges in blue/red denote correct/incorrect estimation of the conditional dependency. The bottom row shows the estimated $\tilde{\Theta}_o$ including the covariates. In each block, the first column shows Σ_o^{-1} as the estimated precision matrix $\tilde{\Theta}_o$ (i.e., sample precision matrix). Here, both sample precision matrices are *dense* due to the effect from the latent variables, leading to a solution which is far from Θ_o. In the second and third column in both blocks, we include results from graphical Lasso [13] and the method from [5]. When $n_h = 5$, the sparsity pattern in the estimated precision matrices for both baselines and our algorithm are quite similar to the ground truth (few red edges). In the second row, we see that the oblique patterns expressing the relationship between the connectivity and the covariates is also recovered. When the number of latent variables grows, the low rank assumption
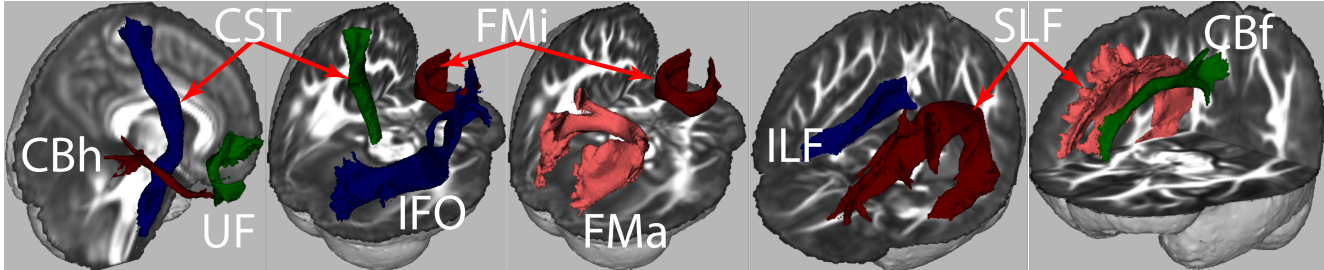
Figure 2. The subset of connections (from the seventeen presented in Table 1) that are statistically associated to non-imaging covariates.

in [5] becomes weaker and the data deviates from the assumptions of graphical lasso (which assumes all variables are observed). For $n_h = 10$, neither of the baselines are able to recover the conditional dependencies between the connections and covariates (oblique patterns in the precision matrix). On the other hand, the fourth columns (second row) shows that our algorithm recovers Θ_o with a sparsity pattern that is highly consistent with the ground truth Θ_o.

Limitations. These results suggest that our algorithm is effective in identifying the true precision matrix even when there are diffusive effects of unobserved variables. However, in situations where we have a large number of latent variables and each affects only a small number of observed variables (i.e., high-frequency effect), our algorithm may not be able to identify the correct associations. This situation is very challenging and we are not aware of any other methods that can model such data.

5.2. Experiments on Human Connectome Data

Dataset. The HCP[2] project recently made available high-quality imaging and clinical data ([40, 14]) for over 500 healthy adults [18]. We analyzed the high angular resolution diffusion MR imaging (dMRI) dataset, consisting of 489 images [39, 41]. We obtained DTI from the dMRI data via standard fitting procedures which were then spatially normalized [47]. Seventeen major white matter connectivity measures were obtained by registering (using ANTS) the publicly available IIT atlas [42] to the HCP template. The average fractional anisotropy (FA) in each pathway was a proxy for the connection strength. Table 1 lists the seventeen connectivity pathways.

Non-imaging covariates. Besides the imaging data, HCP provides several categories of non-imaging covariates for the subjects [17] covering factors such as cognitive function, demographic variables, education and so on. In our experiments, we chose 22 variables related to demographics, physical health, sleep, memory, cognitive flexibility and

Connection name	Description (count)
Forceps major (FMa)	inter-hemispheric (1)
Forceps minor (FMi)	inter-heispheric (1)
Fornix	inter-hemispheric (1)
Cingulum bundle frontal (CBf)	bi-lateral (2)
Cingulum bundle hippocampal (CBh)	bi-lateral (2)
Cortico-spinal tracts (CST)	bi-lateral (2)
Inferior fronto-occipital (IFO)	bi-lateral (2)
Inferior longitudinal fasciculus (ILF)	bi-lateral (2)
Superior longitudinal fasciculus (SLF)	bi-lateral (2)
Uncinate fasciculus (UF)	bi-lateral (2)

Table 1. These pathways span connections between all major lobes of the brain (frontal, parietal, occipital and temporal) with several important regions such as amygdala, hippocampus, pre-frontal cortex.

etc. These covariates span a wide range high-level human behavior and highly relevant physiological measurements and the full list is included in the supplement.

Experiment 1) Figures 2–4 summarize the results of our experiments on the HCP data. The matrix shown in Fig. 3 lists the full set of connections and covariates used in our analysis, along the axes. Our goal was to recover a sparse (and interpretable) precision matrix explaining the conditional dependencies among these variables. It is clear from the figure that our algorithm indeed finds a parsimonious set of statistical relations, among the non-imaging covariates, among the brain pathways as well as across these two groups of variables. As we can expect, several connectivity pathways seem to be involved in several different categories of behavioral measures. These many-to-many dependencies are shown in Fig. 4 and the major pathways that show strong associations with non-imaging measures are displayed visually (overlaid on a standard brain template) in Fig. 2. Note that similar to the simulation setup, in this case, results from the baseline algorithms were non-sparse and hence harder to interpret. Part of the reason is that none of the measurements were controlled for various (observed or unobserved) nuisance variables. One advantage of our algorithm is to precisely take into account the effect of such latent nuisance variables automatically.

Finally, since there is no 'ground truth' available for these results, we checked if our findings are corroborated by independent results in the literature. We found that many of the associations in Figs. 3–4 appear as standalone find-
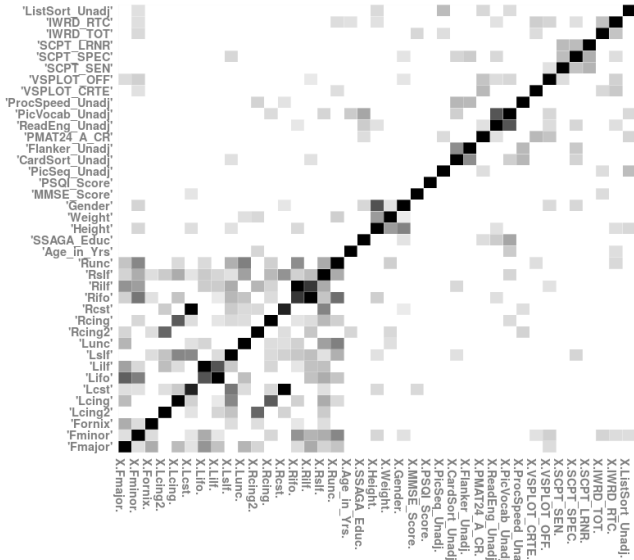
Figure 3. Estimated precision matrix on HCP dataset. Notice that the matrix shows sparse connections between the pathways and covariates, and those identified dependencies are demonstrated in Fig. 4.
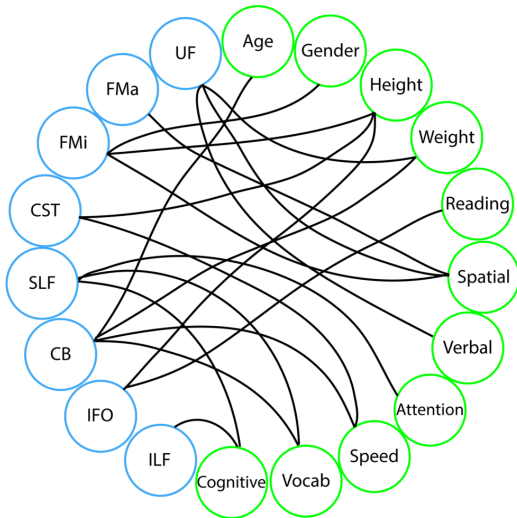

Figure 4. Discovered conditional dependencies between brain connectivity pathways and covariates. Blue circles are connectivity pathways and the green circles are non-imaging covariates.

ings in multiple papers [34, 4]. For example, the association between the cingulum bundle and processing speed was the focus of [32], whereas [21, 27] identified a relation between longitudinal fasciculus and cognitive/verbal ability and [46] demonstrated that forceps major and gender were related. Significant associations have also been found between integrity of the uncinate fasciculus and spatial working memory [9]. This is not definitive evidence that we identify the real underlying precision matrix, but promising that most of the identified associations have precedence in the literature.

Experiment 2) A recent result last year [20] demonstrated structural connectome differences across gender suggesting that men and women are 'wired' differently.

Since the data used in our experiments provides connectivity information jointly with a full set of covariates, it offers an opportunity to ask a similar question as [20] but instead analyze second order effects — that is, are associations of structural connectivity with cognitive/behavioral covariates different across genders on a large well characterized cohort? We used the aforementioned seventeen brain connection pathways, divided our dataset into male/female groups and estimated precision matrices for each group separately. Ideally, each precision matrix will automatically control for the latent factors that contribute to the connectivity measurement independent of gender, and comparisons between the two matrices will suggest how statistical associations between structural connections and cognitive scores vary between men and women. We found at least a few conditional associations that had large differences across the two groups. For example, the associations between the Left IFO/Right ILF, Left ILF/Right IFO and FMa/Left UF were stronger in men whereas the Right IFO/Right ILF, Left IFO/Left ILF and Left ILF/Right ILF were stronger in women. (A full table is included in the supplement).

6. Conclusion

Undirected graphical models are used to address a variety of needs in in computer vision and machine learning. While existing methods for estimating statistical conditional independence between a set of random variables are quite effective, this analysis becomes problematic when there are multiple latent (unobserved) variables that nontrivially affect our measurements of the observed variables. This situation is becoming more frequent in many modern medical image analysis and computer vision datasets, where the latent variables cannot be measured due to cost or privacy reasons. We propose a novel perspective on this sparse inverse covariance matrix estimation problem (which involves latent variables) using non-Euclidean wavelet analysis. The experimental results using synthetic brain network data demonstrates that our algorithm provides substantial improvement over other graphical model selection methods. We present an extensive set of results on the recently released HCP imaging data set showing statistical dependencies between brain connectivity pathways and cognitive/behavioral covariates. Our result are consistent with independent findings in the neuroscience literature.

7. Acknowledgement

References

[1] H. Akil, M. E. Martone, and D. C. Van Essen. Challenges and opportunities in mining neuroscience data. *Science (New York, NY)*, 331(6018):708, 2011. 1

[2] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *JMLR*, 9:485–516, 2008. 2

[3] O. Banerjee, L. E. Ghaoui, A. d'Aspremont, and G. Natsoulis. Convex optimization techniques for fitting sparse Gaussian graphical models. In *ICML*, pages 89–96. ACM, 2006. 2

[4] T. Booth, M. E. Bastin, L. Penke, et al. Brain white matter tract integrity and cognitive abilities in community-dwelling older people: the Lothian birth cohort, 1936. *Neuropsych.*, 27(5):595, 2013. 8

[5] V. Chandrasekaran, P. A. Parrilo, A. S. Willsky, et al. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012. 2, 6, 7

[6] R. Coifman and M. Maggioni. Diffusion wavelets. *Applied and Computational Harmonic Analysis*, 21(1):53 – 94, 2006. 2, 3

[7] A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008. 2

[8] J. V. Davis, B. Kulis, P. Jain, et al. Information-theoretic metric learning. In *ICML*, pages 209–216. ACM, 2007. 5

[9] S. W. Davis, N. A. Dennis, N. G. Buchler, et al. Assessing the effects of age on long white matter tracts using diffusion tensor tractography. *Neuroimage*, 46(2):530–541, 2009. 8

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *JRSS-B*, pages 1–38, 1977. 2

[11] L. Dodero, A. Gozzi, A. Liska, et al. Group-wise functional community detection through joint Laplacian diagonalization. In *MICCAI*, pages 708–715. Springer, 2014. 2

[12] G. Elidan, I. Nachman, and N. Friedman. Ideal parent structure learning for continuous variable Bayesian networks. *JMLR*, 2007. 2

[13] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical LASSO. *Biostatistics*, 9(3):432–441, 2008. 1, 6

[14] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, et al. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*, 80:105–124, 2013. 7

[15] L. Gu, E. P. Xing, and T. Kanade. Learning GMRF structures for spatial priors. In *CVPR*, pages 1–6, 2007. 2

[16] D. Hammond, P. Vandergheynst, and R. Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129 – 150, 2011. 2, 3, 5

[17] R. Herrick, M. McKay, T. Olsen, et al. Data dictionary services in XNAT and the Human Connectome Project. *Frontiers in neuroinformatics*, 8, 2014. 7

[18] M. R. Hodge, W. Horton, T. Brown, et al. ConnectomeDB–sharing human brain connectivity data. *NeuroImage*, 2015. 7

[19] S. Huang, J. Li, L. Sun, et al. Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3):935–949, 2010. 2

[20] M. Ingalhalikar, A. Smith, D. Parker, et al. Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences*, 111(2):823–828, 2014. 8

[21] K. H. Karlsgodt, T. G. van Erp, R. A. Poldrack, et al. Diffusion tensor imaging of the superior longitudinal fasciculus and working memory in recent-onset schizophrenia. *Bio. psych.*, 63(5):512–518, 2008. 8

[22] H. J. Kim, N. Adluru, M. D. Collins, et al. Multivariate general linear models on Riemannian manifolds with applications to statistical analysis of diffusion weighted images. In *CVPR*, 2014. 5

[23] H. J. Kim, J. Xu, B. C. Vemuri, et al. Manifold-valued Dirichlet processes. In *ICML*, 2015. 5

[24] W. H. Kim, N. Adluru, M. K. Chung, et al. Multi-resolution statistical analysis of brain connectivity graphs in preclinical Alzheimer's disease. *NeuroImage*, 118:103–117, 2015. 1

[25] W. H. Kim, D. Pachauri, C. Hatt, et al. Wavelet based multi-scale shape features on arbitrary surfaces for cortical thickness discrimination. In *NIPS*, pages 1250–1258, 2012. 3

[26] W. H. Kim, V. Singh, M. K. Chung, et al. Multi-resolutional shape features via non-Euclidean wavelets: Applications to statistical analysis of cortical thickness. *NeuroImage*, 93:107–123, 2014. 1

[27] M. Kubicki, C.-F. Westin, R. W. McCarley, et al. The application of DTI to investigate white matter abnormalities in schizophrenia. *Ann. of the NY Academy of Sci.*, 1064(1):134–148, 2005. 8

[28] C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. of stat.*, 37(6B):4254, 2009. 2

[29] H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. In *NIPS*, pages 1432–1440, 2010. 2

[30] D. S. Marcus, J. Harwell, T. Olsen, et al. Informatics and data mining tools and strategies for the Human Connectome Project. *Frontiers in neuroinformatics*, 5, 2011. 1

[31] B. M. Marlin and K. P. Murphy. Sparse Gaussian graphical models with unknown block structure. In *ICML*, pages 705–712, 2009. 2

[32] P. G. Nestor, M. Kubicki, K. M. Spencer, et al. Attentional networks and cingulum bundle in chronic schizophrenia. *Schizophrenia research*, 90(1):308–315, 2007. 8

[33] F. Oztoprak, J. Nocedal, S. Rennie, et al. Newton-like methods for sparse inverse covariance estimation. In *NIPS*, 2012. 2

[34] L. Penke, S. M. Maniega, M. Bastin, et al. Brain white matter tract integrity as a neural foundation for general intelligence. *Molecular psychiatry*, 17(10):1026–1030, 2012. 8

[35] G. Raskutti, B. Yu, M. J. Wainwright, et al. Model selection in Gaussian graphical models: High-dimensional consistency of ℓ_1-regularized mle. In *NIPS*, pages 1329–1336, 2008. 2

[36] K. Scheinberg, S. Ma, and D. Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In *NIPS*, pages 2101–2109, 2010. 2

[37] S.Haykin and B. V. Veen. *Signals and Systems*. Wiley, 2005. 3

[38] S. M. Smith, K. L. Miller, Salimi-Khorshidi, et al. Network modelling methods for fMRI. *Neuroimage*, 54(2):875–891, 2011. 2

[39] S. N. Sotiropoulos, S. Jbabdi, J. Xu, et al. Advances in diffusion MRI acquisition and processing in the Human Connectome Project. *Neuroimage*, 80:125–143, 2013. 7

[40] K. Uğurbil, J. Xu, E. J. Auerbach, et al. Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *Neuroimage*, 80:80–104, 2013. 7

[41] D. C. Van Essen, S. M. Smith, D. M. Barch, et al. The WU-Minn Human Connectome Project: an overview. *Neuroimage*, 80:62–79, 2013. 7

[42] A. Varentsova, S. Zhang, and K. Arfanakis. Development of a high angular resolution diffusion imaging human brain template. *NeuroImage*, 91:177–186, 2014. 7

[43] G. Varoquaux, A. Gramfort, J.-B. Poline, et al. Brain covariance selection: better individual functional connectivity models using population prior. In *NIPS*, pages 2334–2342, 2010. 2

[44] M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *JMLR*, 11:2261–2286, 2010. 2

[45] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007. 1

[46] M. Zarei, D. Mataix-Cols, I. Heyman, et al. Changes in gray matter volume and white matter microstructure in adolescents with obsessive-compulsive disorder. *Bio. psych.*, 70(11):1083–1090, 2011. 8

[47] H. Zhang, P. A. Yushkevich, D. C. Alexander, et al. Deformable registration of diffusion tensor MR images with explicit orientation optimization. *MIA*, 10(5):764–785, 2006. 7

Supplementary material: Latent Variable Graphical Model Selection using Harmonic Analysis: Applications to the Human Connectome Project (HCP)

1. Summary

We provide proofs of lemmas and technical details for implementation. Also, additional experimental results are provided which were not included in the main paper due to limited space.

2. Proofs of Lemmas

Lemma 1 *If $\Theta \succ 0, \Theta = \Theta^T$ and kernel g satisfies the admissibility condition*

$$\int_0^\infty \frac{g^2(s\sigma)}{s} ds =: C_g < \infty \tag{1}$$

then,

$$\frac{1}{C_g} \int_0^\infty \frac{1}{s} \sum_{\ell=1}^n W_{\Theta,s}(\ell) \psi_{\ell,s}(i,j) ds = \Theta(i,j) \tag{2}$$

Proof: Note that since $\int_0^\infty \frac{g^2(s\sigma)}{s} ds = \int_0^\infty \frac{g^2(x)}{x} dx =: Cg$, the admissibility condition can be written in both ways. Using the definition of $\psi_{\ell,s}$ and $W_{\Theta,s}(\ell)$ for the reconstruction of Θ,

$$\frac{1}{C_g} \int_0^\infty \frac{1}{s} \sum_{\ell=1}^n W_{\Theta,s}(\ell) \psi_{\ell,s}(i,j) ds$$

$$= \frac{1}{C_g} \int_0^\infty \frac{1}{s} \left(\sum_{\ell=1}^n \sigma_\ell g(s\sigma_\ell) g(s\sigma_l) V_\ell(i) V_\ell(j) \right) ds$$

$$= \frac{1}{C_g} \int_0^\infty \frac{1}{s} \left(\sum_{\ell=1}^n \sigma_\ell g^2(s\sigma_\ell) V_\ell(i) V_\ell(j) \right) ds$$

$$= \sum_{\ell=1}^n \left(\frac{1}{C_g} \int_0^\infty \frac{g^2(s\sigma_\ell)}{s} ds \right) \sigma_\ell V_\ell(i) V_\ell(j)$$

$$= \sum_{\ell=1}^n \sigma_\ell V_\ell(i) V_\ell(j) = \Theta(i,j) \text{ by the admissibility condition in (1)}$$

$\square$

Now, we introduce a result for our inverse covariance estimation with a two parameter kernel function $g(s,\sigma)$. Analogously, we can define a basis function $\psi_{\ell,s}$ and wavelet-like coefficients $W_{\Theta,s}$ as (9) and (10) in the main paper, namely,

$$\psi_{\ell,s} = g(s,\sigma_\ell) V_\ell^*(i) V_\ell(j) \text{ and } W_{\Theta,s}(\ell) = \sigma_\ell g(s,\sigma_\ell), \forall \ell \in \{1,\ldots,n\}. \tag{3}$$

Then, we can show the admissibility condition for two parameter kernels for the reconstructing the graph structure (a function on edges) represented by the precision matrix as in Lemma 2

1

Lemma 2 *If $\Theta \succ 0, \Theta = \Theta^T$ and kernel g satisfies the admissibility condition*

$$\int_0^\infty \frac{g^2(s,\sigma)}{s} ds =: C_g < \infty \tag{4}$$

then,

$$\frac{1}{C_g} \int_0^\infty \frac{1}{s} \sum_{\ell=1}^n W_{\Theta,s}(\ell) \psi_{\ell,s}(i,j) ds = \Theta(i,j) \tag{5}$$

This corollary can be proved in the same way as Lemma 1. This result shows that two parameter kernels can be used for the reconstruction of both the graph *structure* (functions on edges) as well as in the classical setting (functions on the nodes) with a well-defined admissibility condition. This gives the ability to define more flexible kernels if needed.

Similar to Lemma 2, we provide the admissibility condition for two parameter kernels for for the classical SGWT. We follow the notations of [1].

Lemma 3 *If the kernel g satisfies the admissibility condition*

$$\int_0^\infty \frac{g^2(s,\sigma)}{s} ds =: C_g < \infty \tag{6}$$

then,

$$\frac{1}{C_g} \int_0^\infty \frac{1}{s} \sum_{n=1}^N W_f(s,n) \psi_{s,n}(m) ds = f^\#(m) \tag{7}$$

where $f^\# = f - \langle \chi_0, f \rangle \chi_0$ if there exists a zero eigenvalue. When all eigenvalues are strictly positive, $f^\# = f$.

Proof: By the definition of $\psi_{s,n}$ and $W_f(s,n)$ with graph Fourier bases, the left hand side of (7) is given as

$$\frac{1}{C_g} \int_0^\infty \frac{1}{s} \sum_{n=1}^N \left(\sum_{l=1}^N g(s,\sigma_l) \chi_l(n) \hat{f}(l) \sum_{l'=1}^N g(s,\sigma_{l'}) \chi_{l'}^*(n) \chi_{l'}(m) \right) ds$$

$$= \frac{1}{C_g} \int_0^\infty \frac{1}{s} \left(\sum_{l,l'} g^2(s,\sigma_l) \hat{f}(l) \chi_{l'}(m) \sum_{n=1}^N \chi_{l'}^*(n) \chi_l(n) \right) ds$$

$$\overset{(a)}{=} \frac{1}{C_g} \int_0^\infty \frac{1}{s} \left(\sum_{l=1}^N g^2(s,\sigma_l) \hat{f}(l) \chi_l(m) \right) ds$$

$$= \frac{1}{C_g} \sum_{l=1}^N \left(\int_0^\infty \frac{g^2(s,\sigma_l)}{s} ds \right) \hat{f}(l) \chi_l(m)$$

$$= \sum_{l=1}^N \hat{f}(l) \chi_l(m) \text{ by (6)}$$

$$= f^\#,$$

Equality (a) holds since $\sum_n \chi_{l'}^*(n) \chi_l(n) = \delta_{l,l'}$ (Kronecker's delta) with orthonormal basis χ_l. $\square$

3. Choice of the Kernel Function

For the choice of kernel function $g()$, we used the popular Gaussian function $\exp(-\frac{1}{2}sx)$. This kernel function models diffusion or a random walk process, and is used to define diffusion type of wavelets. The kernel function itself may not satisfy the admissibility condition, which is not an issue because we work with a single scale. However, to work with all scales concurrently, we will be limited to only that class of kernels which directly satisfies the admissibility condition.

$$K(s, \sigma) = \sigma g^2(s\sigma) = \sigma e^{-s\sigma} \tag{8}$$

is able to perfectly reconstruct the original sample precision matrix Θ. Intuitively, in our optimization problem, we should be able to reconstruct the exact Θ in the non-regularized setting. This is because of σ in front of $g()$ in (8). That is, our estimation $\tilde{\Theta}$

$$\tilde{\Theta} = \sum_l \sigma_l g^2(s\sigma_l) V_l V_l' \tag{9}$$

becomes P when $s = 0$ as

$$\tilde{\Theta} = \sum_l \sigma_l e^{-s\sigma} V_l V_l' \tag{10}$$

$$= \sum_l \sigma_l V_l V_l' = \Theta. \tag{11}$$

4. Details of Our Experiments

4.1. Experimental Setup and Additional Analysis of Synthetic Pathways

To provide more details for the experiments with synthetic brain pathways, we first demonstrate how the data were generated as in Fig. 1. First, arbitrary relations between the 50 connectivity pathways were randomly generated assuming and their relation with the 10 covariates were defined as shown in Fig. 1 a). Then, 10 latent variables were added with random associations with the observed variables (i.e., pathways and covariates) as in Fig. 1 b), which is a precision matrix with all variables and demonstrates the effects from the latent variables to our observation. Finally, a covariance matrix was derived from the precision matrix as in Fig. 1 c). The covariance matrix was used to generate 2000 multi-variate Gaussian samples. For our algorithm, we chose the kernel function $g(\cdot)$ as in section 3, and $\gamma = 2$ for the sparsity parameter which yields a good

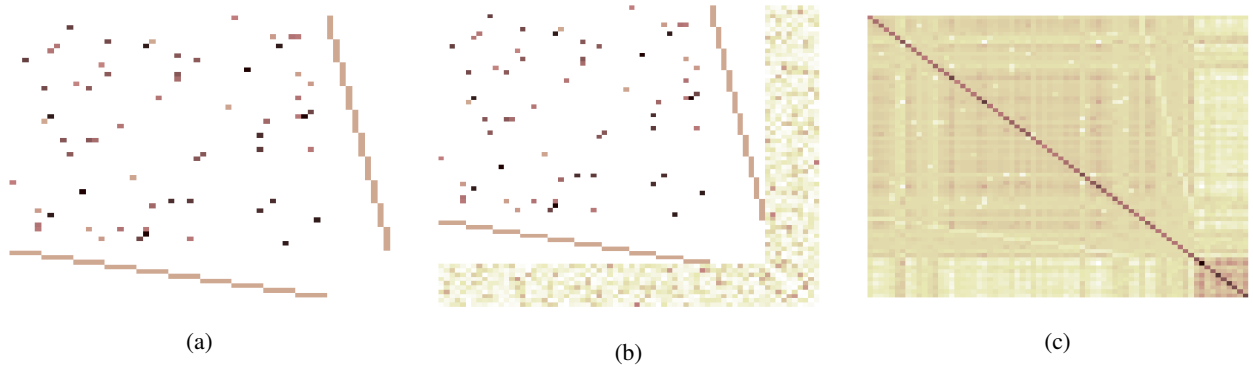|		
(a)	(b)	(c)

Figure 1: a) Precision with 50 observed variables only (Ground truth), b) Precision matrix with both 50 observed and 10 latent variables, c) Covariance matrix derived from b).

result as shown in the main paper.

In the following, we demonstate additional analysis of precision matrix estimation. In Fig. 2, results from experiments with different number of latent variables (i.e., 0, 5, 10, 20) are displayed. The top row in Fig. 2 shows the inverse covariance matrix and the bottom row shows our result. At a glance, we can easily see that the increase in the number of latent variables makes the inverse covariance matrix denser, on the other hand, our estimation results yield correct sparse graphical models despite the increase. In Fig. 3, we show the precision matrix estimation in different scales with 5 latent variables. Fig. 3 a), b) and c) are the estimation results with $s = 0, 0.5, 0.7$, and d) shows a result with the optimal scale $s = 0.2089$ obtained from our optimization scheme. The estimation result significantly varies depending on the scale parameter with dense non-zero elements, but our optimization is able to find the optimal scale that provides an estimation that is sparse and similar to the inverse covariance matrix.
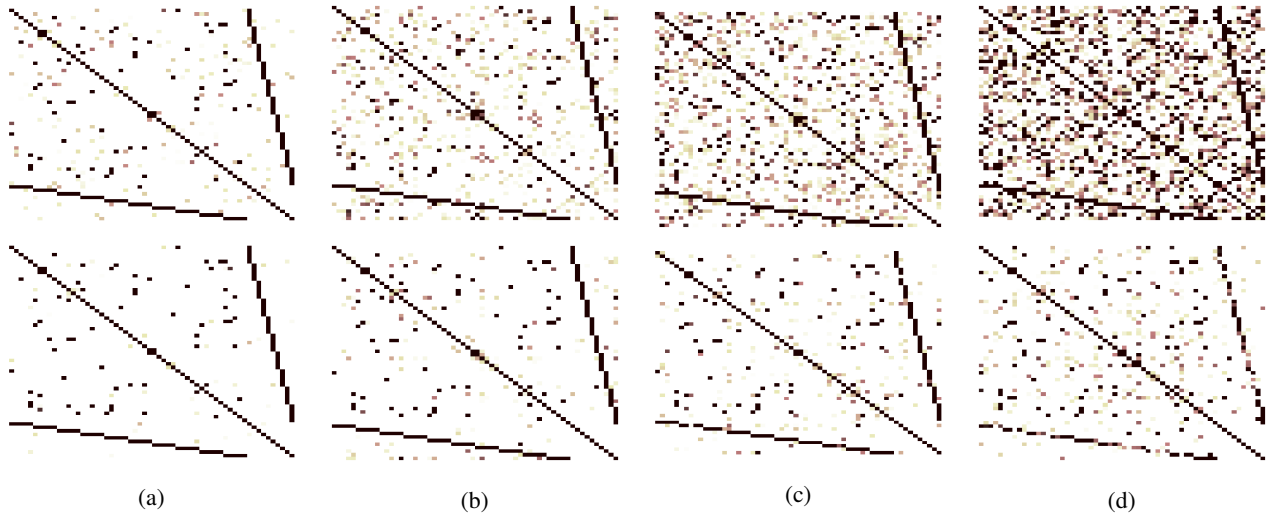
Figure 2: Precision matrix estimation with different numbers of latent variables. Top row shows the inverse covariance matrix, and the bottom row shows our estimation result. a) no latent variables, b) 5 latent variables, c) 10 latent variables, d) 20 latent variables. We can easily see that the inverse covariance matrix becomes denser as the number of latent variables increases, while our method yields good estimation of the sparse graphical model.
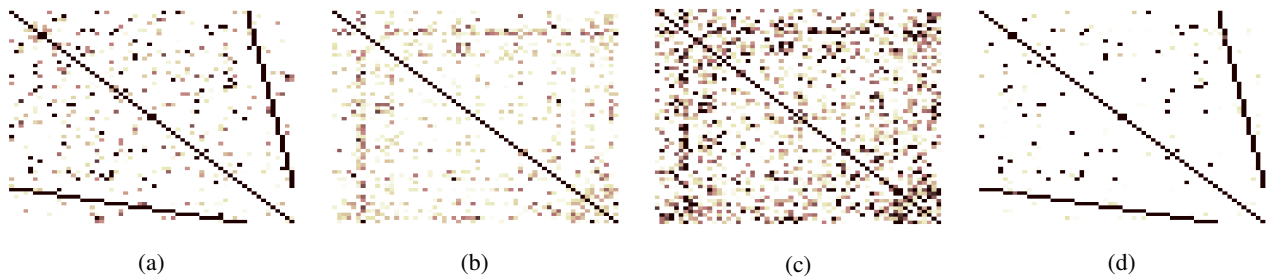

Figure 3: Precision matrix with 5 latent variables in different scales. a) estimation with $s = 0$, b) estimation with $s = 0.5$, c) estimation with $s = 0.7$, d) estimation with optimal scale $s = 0.2089$. Our optimization scheme find the optimal scale that gives a sparse graphical model.

4.2. Details of non-Imaging Covariates

In this section, we provide a full list of non-imaging covariates that we used in our HCP brain connectivity pathways analysis. The list consists of various measurements from physical, mental, and cognitive status of participants, and are commonly used for analysis of structural and functional behavior of brains.

References

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

Category	Covariate name
Demographics	Age, gender, years of education completed (Edu)
Physical health	Height, weight
Alertness	Mini mental status exam (MMSE)
Sleep	Pittsburgh sleep questionnaire (PSQI)
Episodic memory	Picture sequence recall (PicSeq)
Cognitive flexibility	Picture matching accuracy and reaction time (CardSort)
Inhibition	Flanking accuracy and reaction time (Flanker)
Fluid intelligence	Correct responses in Penn progressive matrices (PMA_CR)
Reading	NIH toolbox reading recognition test (ReadEng)
Vocabulary	NIH toolbox picture vocabulary (PicVocab)
Processing speed	NIH toolbox pattern comparison speed (ProcSpeed)
Spatial orientation	Expected number of correct clicks (VSPLOT_CRTE), total off positions (VSPLOT_OFF)
Sustained attention	Short Penn continuous performance test: sensitivity (SCPT_SEN), specificity (SCPT_SPEC), longest run of non-responses (SCPT_LRNR)
Episodic memory	Penn word memory test: total correct responses (IWRD_TOT), reaction time (IWRD_RTC)
Working memory	NIH toolbox sorting working memory (ListSort)

Table 1: Full list of non-imaging covariates used in our analysis spanning a wide range high-level human behavior and highly relevant physiological measurements.