Adaptive Signal Recovery on Graphs via Harmonic Analysis for Experimental Design in Neuroimaging

Won Hwa Kim[1(✉)], Seong Jae Hwang[1], Nagesh Adluru[4], Sterling C. Johnson[3], and Vikas Singh[1,2]

[1] Department of Computer Sciences, University of Wisconsin, Madison, WI, USA
wonhwa@cs.wisc.edu
[2] Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA
[3] GRECC, William S. Middleton VA Hospital, Madison, WI, USA
[4] Waisman Center, Madison, WI, USA
http://pages.cs.wisc.edu/~wonhwa

Abstract. Consider an experimental design of a neuroimaging study, where we need to obtain p measurements for each participant in a setting where $p'(< p)$ are cheaper and easier to acquire while the remaining $(p - p')$ are expensive. For example, the p' measurements may include demographics, cognitive scores or routinely offered imaging scans while the $(p - p')$ measurements may correspond to more expensive types of brain image scans with a higher participant burden. In this scenario, it seems reasonable to seek an "adaptive" design for data acquisition so as to minimize the cost of the study without compromising statistical power. We show how this problem can be solved via harmonic analysis of a band-limited graph whose vertices correspond to participants and our goal is to fully recover a multi-variate signal on the nodes, given the full set of cheaper features and a partial set of more expensive measurements. This is accomplished using an adaptive query strategy derived from probing the properties of the graph in the frequency space. To demonstrate the benefits that this framework can provide, we present experimental evaluations on two independent neuroimaging studies and show that our proposed method can reliably recover the true signal with only partial observations directly yielding substantial financial savings.

1 Introduction

Consider an experimental design setting which involves a cohort $\mathcal{S}$ comprised of N individuals (or examples) in total. We are allowed to obtain a maximum of p

This research was supported by NIH grants AG040396, and NSF CAREER award 1252725, UW ADRC AG033514, UW ICTR 1UL1RR025011, UW CPCP AI117924, UW CIBM 5T15LM007359-14 and Waisman Core Grant P30 HD003352-45.

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46466-4_12) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part VI, LNCS 9910, pp. 188–205, 2016.
DOI: 10.1007/978-3-319-46466-4_12

measurements (or features) for each participant (or example) in $\mathcal{S}$. Depending on the application, these p measurements may be variously interpreted — for example, in a machine learning experiment, we may have p distinct numerical preferences a user assigns to each item whereas in computer vision, the measurements may reflect p specific requests for supervision or indication on each image in $\mathcal{S}$ [1–4]. In a neuroscience experiment, the cohort corresponds to individual subjects — the p measurements will denote various types of imaging and clinical measures we can acquire. Of course, independent of the application, the "cost" of measurements is quite variable: while features such as gender and age of a participant have negligible cost, requesting a user to rate an image in abstract terms, "How natural is this image on a scale of 1 to 5?", may be more expensive. In neuroimaging, acquiring some clinical and cognitive measures is cheap, whereas certain image scans can cost several thousands of dollars [5,6].

In the past, when datasets were smaller, these issues were understandably not very important. But as we move towards acquiring and annotating large scale datasets in machine learning and vision [7–9], the cost implications can be substantial. For instance, if the budget for a multi-modal brain imaging study involving several different types of image scans for ∼200 subjects is \$3M+ and we know *a priori* which type of inference models will finally be estimated using this data, it seems reasonable to ask if "adaptive" data acquisition can bring down costs by 25 % with negligible deterioration in statistical power. While experiment design concepts in classical statistics provide an excellent starting point, they provide little guidance in terms of practical technical issues one faces in addressing the question above. Outside of a few recent works [10–12], this topic is still not extensively studied within mainstream machine learning and vision.

In this paper, we study a natural form of the experimental design problem in the context of an important brain imaging application. Assume that we have access to a cohort $\mathcal{S}$ of n subjects. In principle, we can acquire p measurements for each participant. But all p measures are not easily available — say, we start only with a *default* set of p' measures for each subject which may be considered as "inexpensive". This yields a matrix of size $N \times p'$. We are also provided the remaining set of $(p - p')$ measurements but only for a small subset $\mathcal{S}'$ of n' subjects — possibly due to the associated expense of the measurement. We can, if desired, acquire these additional $(p - p')$ measures for each individual participant in $\mathcal{S} \backslash \mathcal{S}'$, but at a high per-individual cost. Our goal is to eventually estimate a statistical model that has high fidelity to the "true" model estimated using the full set of p measures/features for the full cohort $\mathcal{S}$. The key question is whether we can design an adaptive query strategy that minimizes the overall cost we incur and yet provides high confidence in the parameter estimates we obtain. The problem statement is quite general and models experimental design considerations in numerous scientific disciplines including systems biology and statistical genomics where an effective solution can drive improvements in efficiency.

1.1 Related Work

There are three distinct areas of the literature that are loosely related to the development described in this paper. At the high level, perhaps the most closely related to our work is *active learning* which is motivated by similar cost-benefit considerations, but in terms of minimizing the number of queries (seeking the label of an example) [13]. Here, one starts with a pool of unlabeled data and picks a few examples at random to obtain their labels. Then, we repeatedly fit a classifier to the labeled examples seen so far and query the unlabeled example that is most uncertain or likely to decrease overall uncertainty. This strategy is generally successful though may asymptotically converge to a sub-optimal classifier [14]. Adaptive query strategies have been presented to guarantee that the hypothesis space is fully explored and to obtain theoretically consistent results [15,16]. Much of active learning focuses on learning discriminative classifiers; while the Bayesian versions of active learning can, in principle, be applied to far more general settings, it is not clear whether such formulations can be adapted for the stratified cost structure we encounter in the motivating example above and for general parameter estimation problems where the likelihood expressions are not computationally 'nice'.

Within the statistics literature, the problem of experiment design has a rich history going back at least four decades [17–19], and seeks to formalize how one deals with the non-deterministic nature of physical experiments. In contrast to the basic setting here and even data-driven measures of merit such as D-optimality [20,21], experiment design concepts such as the Latin hypercube design [22] intentionally assume very little about the relationship between input features and the output labels. Instead, with d features, such procedures will generate a space-filling design so that each of the dimensions is divided into equal levels — the calculated configuration merely provides a selection of inputs at which to compute the output of an experiment to achieve specific goals. Despite a similar name, the goals of these ideas are quite different from ours.

Within machine learning and vision, papers related to collaborative filtering (and matrix completion) [23–26] share a number of technical similarities to the development in our work. For instance, one may assume that in a matrix of size $N \times p$ (subjects $\times$ measurements), the first p' columns are fully observed whereas multiple rows in the remaining $(p - p')$ columns are missing. This clearly yields a matrix completion problem; unfortunately, the setup lies far from incoherent sampling and the matrix versions of restricted isometry property (RIP) that make the low-rank completion argument work in practice [27,28]. This observation has been made in recent works where collaborative filtering was generalized to the graph domain [29] and where random sampling was introduced for graphs in [30]. However, these approaches, which will serve as excellent baselines, do not exploit the band-limited nature of measurements in frequency space. Separately, matrix completion within an adaptive query setting [31,32] yields important theoretical benefits but so far, no analogs for the graph setting exist.

The contribution of this paper is to provide a harmonic analysis inspired algorithm to estimate band-limited signals that are defined on graphs. It turns

out that such solutions directly yield an efficient procedure to conduct adaptive queries for designing experiments involving stratified costs of measurements, i.e., where the first subset of measures is free whereas the second set of $(p - p')$ measures is expensive and must be requested for a small fraction of participants. Our framework relies on the design of an efficient decoder to recover the band-limited original signal involving multiple channels which was only partially observed. In order to accomplish these goals, the paper makes the following contributions.

(i) We propose a novel sampling and signal recovery strategy on a graph that is derived via harmonic analysis of the graph.
(ii) We show how a band-limited multi-variate signal on a graph can be reconstructed with only a few observations via a simple optimization scheme.
(iii) We provide an extensive set of experiments on *two independent datasets* which demonstrate that our framework works well in estimating expensive image-derived measurements based on (a) a partial set of observations (involving less expensive image-scan data) and (b) a full set of measurements on only a small fraction of the cohort.

2 Preliminaries: Linear Transforms in Euclidean and Non-euclidean Spaces

Well known signal transforms in the forward/inverse directions such as the wavelet and Fourier transforms (in non-Euclidean space) are fundamental to our proposed framework. These transforms are well understood in the Euclidean setting, however, their analogues in non-Euclidean spaces have not been studied until recently [33]. We provide a brief overview of these transforms in both Euclidean and non-Euclidean spaces.

2.1 Continuous (Forward) Wavelet Transform

The Fourier transform is a fundamental tool for frequency analyses of a signal by transforming the signal $f(x)$ into the frequency domain as

$$\hat{f}(\omega) = \langle f, e^{j\omega x} \rangle = \int f(x) \mathrm{e}^{-j\omega x} \mathrm{d}x \tag{1}$$

where $\hat{f}(\omega)$ is the resultant Fourier coefficient. Wavelet transform is similar to the Fourier transform, but it uses a different type of oscillating basis function (i.e., mother wavelet). Unlike Fourier basis (i.e., sin()) with infinite support, a wavelet ψ is a localized function with finite support. One can define a mother wavelet $\psi_{s,a}(x) = \frac{1}{s}\psi(\frac{x-a}{s})$ with *scale* and *translation* properties, controlled by s and a respectively. Here, changing s controls the dilation and varying a controls the location of ψ. Using $\psi_{s,a}$ as bases, a wavelet transform of a function $f(x)$ results in wavelet coefficients $\mathcal{W}_f(s, a)$ at scale s and at location a as

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

where ψ^* is the complex conjugate of ψ [34].

Interestingly, ψ_s is localized not only in the original domain but also in the frequency domain. It behaves as a band-pass filter covering different bandwidths corresponding to scales s. These band-pass filters *do not* cover the low-frequency components, therefore an additional low-pass filter ϕ, a scaling function, is typically introduced. A transform with the scaling function ϕ results in a low-pass filtered representation of the original function f. In the end, filtering at multiple scales s of the wavelet offers a multi-resolution view of the given signal.

2.2 Wavelet Transform in Non-euclidean Spaces

Defining a wavelet transform in the Euclidean space is convenient because of the regularity of the domain (i.e., a regular lattice). In this case, one can easily define the shape of a mother wavelet in the context of an application. However, in non-Euclidean spaces (e.g., graphs that consists of a set of vertices and edges with arbitrary connections), an implementation of a mother wavelet becomes difficult due to the ambiguity of dilation and translation. Due to these issues, the classical definition of the wavelet transform has not been suitable for analyses of data in non-Euclidean spaces until recently when [33,35] proposed wavelet and Fourier transforms in non-Euclidean spaces.

The key idea in [33] for constructing a mother wavelet ψ on the nodes of a graph is simple. Instead of defining it in the original domain where the properties of ψ are ambiguous, we define a mother wavelet in a dual domain where its representation is clear and then transform it back to the original domain. The core ingredients for such a construction are (1) a set of "orthonormal" bases that provide the means to transform a signal between a graph and its dual domain (i.e., an analogue of the frequency domain) and (2) a kernel function $h()$ that behaves as a band-pass filter determining the shape of ψ. Utilizing these ingredients, a mother wavelet is first constructed as a kernel function in the frequency domain and then localized in the original domain using a δ function and the orthonormal bases. Such an operation will implement a mother wavelet ψ on the original graph. Defining a kernel function in the 1-D frequency domain is simple, and one can rely on spectral graph theory to obtain the orthonormal bases of a graph [33] which can be used for graph Fourier transform.

A graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ is formally defined by a vertex set $\mathcal{V}$ with N number of vertices and a edge set $\mathcal{E}$ with edges that connect the vertices. Such a graph is generally represented by an adjacency matrix $\mathcal{A}_{N\times N}$ where each element a_{ij} denotes the connection between ith and jth vertices by a corresponding edge weight. Another matrix that summarizes the graph, a degree matrix $\mathcal{D}_{N\times N}$, is a diagonal matrix where the ith diagonal is the sum of edge weights connected to the ith vertex. A graph Laplacian is then defined from these two matrices as $\mathcal{L} = \mathcal{D} - \mathcal{A}$, which is a self-adjoint and positive semi-definite operator. The matrix $\mathcal{L}$ can be decomposed into pairs of eigenvalues $\lambda_l \geq 0$ and corresponding eigenvectors χ_l where $l = 0, 1, \cdots, N-1$. The orthonormal bases χ can be used as analogues of Fourier bases in the Euclidean space to define the graph Fourier transform of a function $f(n)$ defined on the vertices n as

$$\hat{f}(l) = \sum_{n=1}^{N} \chi_l^*(n) f(n) \quad \text{and} \quad f(n) = \sum_{l=0}^{N-1} \hat{f}(l) \chi_l(n) \tag{3}$$

where the forward transform yields the graph Fourier coefficient $\hat{f}(l)$ and the inverse transform reconstructs the original function $f(n)$. If the signal $f(n)$ lies in the spectrum of the first k number of χ_l in the dual space, we say that $f(n)$ is k band-limited. Just like in the conventional Fourier transform, this graph Fourier transform offers a mechanism to transform a signal on graph vertices back and forth between the original and the frequency domain.

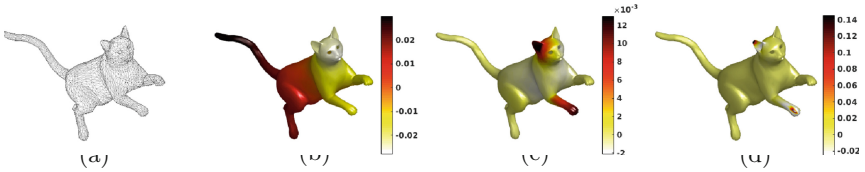


Fig. 1. Examples of bases functions on a graph. (a) Cat shaped graph, (b) A graph Fourier basis χ_2, (c) Graph wavelet bases ψ_1 at two different locations (ear and paw), (d) Graph wavelet basis ψ_4 as in (c). Notice that wavelet bases in (c) and (d) are localized while χ_2 is spread all over the mesh.

Using the graph Fourier transform, a mother wavelet ψ is implemented by first defining a kernel function $h()$ and then localizing it by a Dirac delta function δ_n in the original graph through 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

$$\psi_{s,n}(m) = \sum_{l=0}^{N-1} h(s\lambda_l) \chi_l^*(n) \chi_l(m). \tag{4}$$

Here, using the scaling property of Fourier transform [36], the scale s can be defined as a parameter in the kernel function $h()$ independent from the bases χ. Representative examples of a graph Fourier basis and graph wavelet bases are shown in Fig. 1. A cat shaped graph is given in Fig. 1(a), and one of its graph Fourier basis χ_2 is shown in (b). Also, graph wavelets at two different scales (i.e., dilation) at two different locations (ear and paw) are shown in Fig. 1(c) and (d). Notice that χ in Fig. 1(b) is diffused all over the graph, while the wavelet bases in (c) and (d) are localized with finite support.

Once the bases ψ are defined, the wavelet transform of a function f on graph vertices at scale s follows the classical definition of the wavelet transform:

$$\mathcal{W}_f(s,n) = \langle f, \psi_{s,n} \rangle = \sum_{l=0}^{N-1} h(s\lambda_l) \hat{f}(l) \chi_l(n) \tag{5}$$

resulting in wavelet coefficients $W_f(s,n)$ at scale s and location n. This transform offers a multi-resolution view of signals defined on graph vertices by

multi-resolution filtering. Our framework, to be described shortly, will utilize the definition of the mother wavelet in (4) for data sampling strategy on graphs as well as the graph Fourier transform for signal recovery.

3 Adaptive Sampling and Signal Recovery on Graphs

Suppose there exists a band-limited signal (of p channels/features) defined on graph vertices, and we have limited access to the observation on only a few of the vertices in the graph. Our goal is to estimate the entire signal using only the partial observations. Since the signal is band-limited, we do not need to sample every location in the native domain (i.e., Nyquist rate). Unfortunately, we do not have powerful sampling theorems for graphs. In this regime, in order to recover the original signal, we need an efficient sampling strategy for the data. In the following, we describe how the vertices should be selected for accurate recovery of the band-limited signal and propose a novel decoder working in a dual space that is more efficient than alternative techniques.

3.1 Graph Adaptive Sampling Strategy

In order to derive a random sampling of the data measurement on a graph (i.e., signal measurement on vertices), we first need to assign a probability distribution p on the graph nodes. This probability tells us which vertices are more likely to be sampled for measurements, and needs to satisfy the definition of a probability distribution as $\sum_{n=1}^{N} \mathsf{p}(n) = 1$ where $\mathsf{p} > 0$. The construction of p is based on how the energy spreads over the graph vertices, given the graph structure. It means that it is easier to reconstruct a given signal with limited number of bases at some vertices than other vertices, and prioritizing those vertices for sampling will yield better estimation of the original signal.

In order to define the probability distribution p over the vertices, we make use of the eigenvalues and eigenvectors from spectral graph theory to describe the energy propagation on the graph. In [30], the authors show how well a δ_n can be reconstructed at a vertex n with k number of eigenvectors and normalize them to construct a probability distribution as

$$\mathsf{p}(n) = \frac{1}{k} ||V_k^T \delta_n||_2^2 = \frac{1}{k} \sum_{l=0}^{k-1} \chi_l(n)^2 \tag{6}$$

where V_k is a matrix with column vectors as $V_k = [\chi_0 \ \cdots \chi_{k-1}]$. Their solution puts the same weight on each eigenvector to compute the distribution, assuming that the signal is uniformly distributed in the k-band (i.e., the spectrum of the first k eigenvectors). Such a strategy uses the graph Fourier bases to reconstruct a delta function, which typically is not desirable in many applications since Fourier bases suffers from ringing artifacts. Moreover, in many cases, the signal may be localized even within the k-band, and it necessitates a scaling (i.e., filtering) of the signal at multiple scales in the frequency domain.

Interestingly, it turns out that the definition of p above can be viewed entirely via a non-Euclidean wavelet expansion described in Sect. 2. Recall that a mother wavelet $\psi_{s,n}$ is implemented by localizing a wavelet operation at scale s as in (4). It constructs a mother wavelet at scale s localized at n as a unit energy propagating from n to neighboring vertices as a diminishing wave function. When we look at $\psi_{s,n}(n)$, the self-effect of a mother wavelet at vertex n is written as

$$\psi_n(s,n) = \sum_{l=0}^{N-1} h(s\lambda_l)\chi_l(n)^2. \tag{7}$$

At the high level, (7) tells us how much of the unit energy is maintained at n itself at scale s. Notice that (7) is a kernelized version of (6) using a kernel function $h()$. Depending on the design of the kernel function $h()$, we may interpret it as robust graph-based signatures such as heat-kernel signature (HKS) [37], wave kernel signature (WKS) [38], global point signature (GPS) [39] and wavelet kernel descriptor (WKD) [40], which were introduced in computer vision literature for detecting interest points on graphs and mesh segmentation.

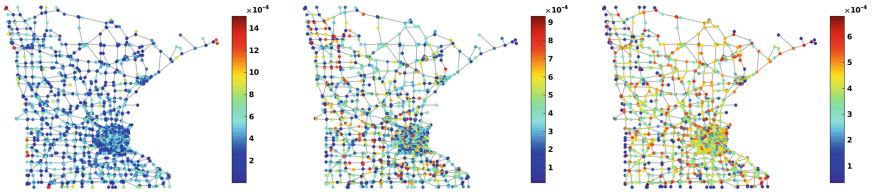


Fig. 2. Sampling probability distribution p_s in different scales derived from "Meyer" wavelet on Minnesota graph. Left: at scale $s = 1$, Middle: at scale $s = 2$, Right: at scale $s = 3$.

Our idea is to make use of the wavelet expansion to define a probability distribution at scale s as

$$\mathsf{p}_s(n) = \frac{1}{Z_s}\psi_n(s,n) = \frac{1}{Z_s}\sum_{l=0}^{N-1} h(s\lambda_l)\chi_l(n)^2 \tag{8}$$

where $Z_s = \sum_{n=1}^{N} \psi_n(s,n)$. Then p_s is used as a sampling probability distribution which drives how we adaptively query the measurements at the unobserved vertices. Depending on application purposes, $h()$ can be designed as any known filters for wavelets such as Morlet, Meyer, difference of Gaussians (DOG) and so on. Examples of p_s using Meyer wavelet are shown in Fig. 2.

Our formulation in (8) is especially useful when we know the distribution of λ prior to the analysis by imposing higher weights on the band where signal is concentrated. We also work with only k eigenvectors when a full diagonalization of $\mathcal{L}$ is expensive. We will see that this observation is important in the next Section, where we utilize a low dimensional space spanned by the k eigenvectors for an efficient solver, while other methods require the full eigenspectrum.

3.2 Recovery of a Band-Limited Signal in a Dual Space

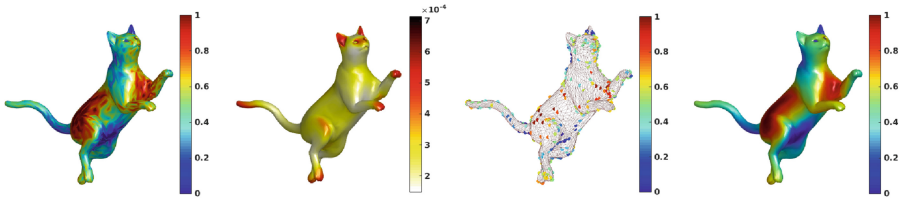
Consider a setting where we observe only a partial signal $y \in \mathbb{R}^{m \times p}$ of a full signal $f \in \mathbb{R}^{N \times p}$ where $m \ll N$, and our goal is to recover the original signal f given y. Suppose that our budget allows querying m vertices (to acquire measurements) in the setting phase. Let the locations where we observe the signal be denoted as $\Omega = \{\omega_1, \cdots, \omega_m\}$ yielding $y(i) = f(\omega_i), \quad \forall i \in \{1, 2, \cdots, m\}$. Now the question is how Ω should be selected for optimal (or high fidelity) recovery of f. Our framework uses the strategy described in Sect. 3.1 to sample data according to a sampling probability. Based on the m samples (observations), we can build a projection operator $M_{m \times N}$ (i.e., a sampling matrix) yielding $Mf = y$ as

$$M_{i,j} = \begin{cases} 1 & \text{if} \quad j = \omega_i \\ 0 & \text{o.w.} \end{cases} \tag{9}$$

Using the ideas described above, a typical decoder would solve for an estimation g of the original signal f using a convex problem as

$$g^* = \arg\min_{g \in \mathbb{R}^n} ||\mathcal{P}_\Omega^{-\frac{1}{2}}(Mg - y)||_2^2 + \gamma g^T h(\mathcal{L}) g \tag{10}$$

where $\mathcal{P}_\Omega = \text{diag}(\mathsf{p}(\Omega))$ and $h(\mathcal{L}) = \sum_{l=0}^{N-1} h(\lambda_l)\chi_l \chi_l^T$. Taking a close look at the formulation above, it prioritizes minimizing the error between an estimation at the sampled locations (with weights of $\frac{1}{\sqrt{\mathsf{p}_\Omega}}$), and the remaining missing elements are filled in by the regularizer representing graph smoothness. Such a recovery explained in [30] has three weaknesses. (1) It does not take into account whether the recovered signal is band-limited. (2) The main objective function (i.e., the first term) in (10) suggests that it does not matter whether the estimated elements in the unsampled locations are correct. (3) Finally, the analytic solution to the above problem is not easily obtainable without the regularizer or when the regularizer is not full rank. This becomes computationally problematic in real cases when the given graph is large, since the filtering operation in (10) requires a full eigendecomposition of the graph Laplacian $\mathcal{L}$.



(a) (b) (c) (d)

Fig. 3. A toy example of our framework on a cat mesh ($N = 3400$). (a) Band-limited random signal in $[0, 1]$ with noise, (b) Sampling probability p_1 derived from (8), (c) Sampled signal at $m = 340$ locations out of 3400, (d) Recovered signal using our method with only $k = 50$.

To deal with the problems above, we propose to encode the band-limited nature of the recovered signal as a constraint. Our framework solves for a solution to (10) entirely in a dual space by projecting the problem to a low dimensional space where we search for a solution of size $k \ll N$.

Let $\hat{g}(l) = \sum_{n=1}^{N} g(n)\chi_l(n)$ be the graph Fourier transform of a function g and $\hat{g}_k$ be the first k coefficients, then reformulating the model in (10) using $g = V_k\hat{g}_k$ (assuming that g is k-band limited) yields

$$\hat{g}_k^* = \arg\min_{\hat{g}_k \in \mathbb{R}^k} ||\mathcal{P}_\Omega^{-\frac{1}{2}}(MV_k\hat{g}_k - y)||_2^2 + \gamma(V_k\hat{g}_k)^T h(\mathcal{L})V_k\hat{g}_k. \tag{11}$$

An analytic solution to this problem can be achieved by taking the derivative of (11) and setting it to 0. The optimal solution $\hat{g}_k^*$ must satisfy the condition

$$(V_k^T M^T \mathcal{P}_\Omega^{-1} MV_k + \gamma V_k^T h(\mathcal{L})V_k)\hat{g}_k^* = V_k^T M^T \mathcal{P}_\Omega^{-1} y \tag{12}$$

which reduces to

$$(V_k^T M^T \mathcal{P}_\Omega^{-1} MV_k + \gamma h(\Lambda_k))\hat{g}_k^* = V_k^T M^T \mathcal{P}_\Omega^{-1} y \tag{13}$$

where Λ_k is a $k \times k$ diagonal matrix where the diagonals are the first k eigenvalues of $\mathcal{L}$. Using the optimal $\hat{g}_k^*$, we can easily recover a low-rank estimation $g^* = V_k\hat{g}_k^*$ that reconstructs f. Notice that we only need to find a solution of a much smaller dimension which is significantly more efficient. Moreover, the filtering operation $h()$ in the regularizer in (12) becomes much simpler, and concurrently the solution natively maintains the k-band limited property of the original signal.

A toy example demonstrating this idea is shown in Fig. 3. Given a cat mesh with $N = 3400$ vertices, we first define a random signal $f \in [0, 1]$ that is band-limited in the spectrum of $\mathcal{L}$ with Gaussian noise of $N(0, 0.1)$. We take p_1 for the sampling distribution and sample $m = 340$ (10 % of the total) vertices without replacement. Our estimation g using only $k = 50$ bases is shown in Fig. 3(d), where the error between the true f and g is extremely small despite using such little data to begin with. We also can see that our method is robust to noise.

4 Experiment Design in Neuroimaging

In this section, we present proof of principle experimental results on two different neuroimaging studies: **(1)** the Human Connectome Project (HCP) dataset and **(2)** Wisconsin Registry for Alzheimer's Prevention (WRAP) dataset. In both studies, we demonstrate the performance of our method in estimating expensive neuroimage-derived measurements at regions of interests (ROI) in the brain using **(1)** a set of p' less expensive measures of all p measures available to the full cohort $\mathcal{S}$ of N subjects and **(2)** a set of $(p - p')$ expensive measures available to a small cohort subset $\mathcal{S}'$ which includes m subjects. Given these datasets, the goal of these experiments is to see if we can get accurate estimates of the $(p - p')$ expensive measures of the *full cohort* $\mathcal{S}$ of N subjects in a way that statistical power for the follow-up analysis is not greatly compromised.

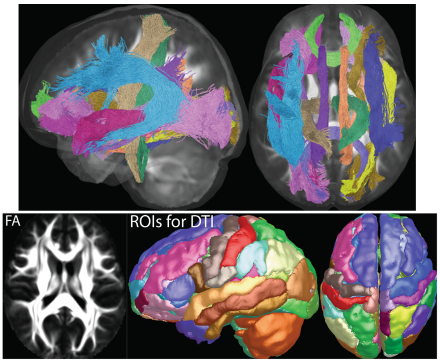
4.1 Experimental Setup

We compare the performance of our method with two other state-of-the-art methods, **(1)** Collaborative filtering by Rao et al. [29] and **(2)** Random sampling of band-limited signals by Puy et al. [30]. For all three methods: **(a)** We derived adjacency matrices $\mathcal{A}$ using data from the full set $\mathcal{S}$ of N samples and p' economical measures (i.e., more widely available and/or less expensive modalities) and the radial basis function $\exp(-||x-y||^2/\sigma^2)$. We then constructed normalized graph Laplacians $\mathcal{L} = \mathcal{D}^{-1/2}(\mathcal{D}-\mathcal{A})\mathcal{D}^{-1/2}$ used in our framework. **(b)** We set $h(\lambda_l) = \lambda_l^4$ for $h(\mathcal{L})$ for the filtering operation in the regularizer and set $\gamma = 0.01$ in (11). **(c)** We show estimation results of the $(p-p')$ expensive measures using $R \in \{20, 40, 60\}\%$ of total N samples for both studies and assess the ℓ_2-norm error of the difference between the estimated and observed measures. Because of the stochastic nature of the sampling step, we ran the estimation 100 times and use the average of the corresponding errors for comparisons. In addition, we also compare the predicted values of the $(p-p')$ neuroimaging measures at each ROI (averaged across subjects) against true values and the estimates of the other two baseline methods. For example, given a cohort of $N = 100$ subjects, suppose we have full data for $p' = 10$ low-cost measurements. Then, the goal is to acquire the $p - p' = 5$ measurements on only $m = 20$ subjects (i.e., 20 % of the cohort) and estimate the $(p - p')$ measurements on the remaining $N - m$ subjects.

4.2 Prediction on the Human Connectome Project

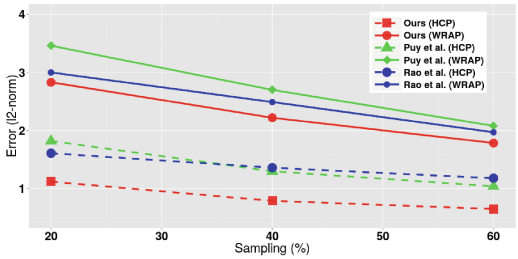
Dataset. The diffusion weighted MR images (DW-MRI) from HCP ([42]) were acquired on custom built hardware using advanced pulse sequences [43] and for a *lengthy* scan time (∼1 h). It allows estimating microstructural properties of the brain, accurate reconstruction of the white matter pathways ([44]) (e.g., see Fig. 4) which form a crucial component in mapping the structural connectome of the human brain [45–48]. Typically, such an acquisition of DW-MRI is not feasible in many research sites due to limitations of hardware and software. On the other hand, the set of non-imaging measurements are cheaper and easier to acquire. Hence the ability to predict such high quality diffusion metrics (e.g. fractional anisotropy (FA)) from only a small sample of the DW-MRI scans and the non-imaging measurements has value. HCP provides several categories of non-imaging covariates for the subjects [49] covering factors spanning several different categories. (The full list of covariates is given in the appendix.) We demonstrate the performance of our model on the task of FA prediction in 17 widely studied fiber bundles (shown in Fig. 4) [41,50] using 27 variables related to cognition, demographics, education and so on.

Results. Given the full cohort $\mathcal{S}$ of $N = 487$ subjects from the HCP dataset with the selected $p' = 27$ low-cost covariates, we recovered high-cost FA measures in $p - p' = 17$ ROIs (i.e. pathways) using p' covariates and the FA values from $m \ll N$ participants. The p' measures were used to construct $\mathcal{L}$ with $\sigma = 5$ and $k = 100$ for generating the sampling distribution p for our framework.



Connection name	Description(count)
Forceps Major (FMajor)	inter-hemispheric (1)
Forceps Minor (FMinor)	inter-hemispheric (1)
Fornix	inter-hemispheric (1)
Cingulum Bundle Frontal (CingAnt)	bi-lateral (2)
Cingulum Bundle Hippocampal (CingHipp)	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)

Fig. 4. Top: The 17 major white matter pathways analyzed in the HCP study [41], Bottom: ROIs and measures analyzed in the WRAP study (Left: A sample FA map and the 162 gray matter ROIs for DTI, Right: Sample ^{11}C PiB DVR map and the 16 gray matter ROIs).



Dataset	Method	20%	40%	60%
HCP	Ours	**2.83**	**2.22**	**1.79**
HCP	Puy et al.	3.46	2.70	2.08
HCP	Rao et al.	3.00	2.4	1.97
WRAP	Ours	**1.12**	**0.79**	**0.65**
WRAP	Puy et al.	1.82	1.30	1.06
WRAP	Rao et al.	1.61	1.36	1.18

Fig. 5. Sampling ratio versus error plot (left) on the HCP dataset (dashed lines) and the WRAP dataset (straight lines). The corresponding values are in the table on the right. (Color figure online)

We analyzed three cases by sampling 20 %, 40 %, 60 % of the total population according to p for m observations to predict FA on N subjects.

Figure 5 (dashed lines) summarizes the overall estimation errors using $R = \{20, 40, 60\}\%$ samples of the total population. For all three methods, the errors decreased with an increase in sample size, and our method (red) consistently outperformed the other two methods (blue and green). When we look at the distribution of errors, shown in the top row of Fig. 6, the center of the error distribution using our framework (red) is far lower than the other methods (blue and green). Anatomical specificity of the estimation measures (using 40 % samples) is illustrated on the top panel of Fig. 7 where the location of spheres represents the position of the ROIs and their sizes and colors correspond to the mean errors. As seen in Fig. 7, our method (top-left) clearly has smaller and blue spheres compared to the other methods (middle and right). The quantitative error for individual ROIs used for the spheres are provided in left table of Table 1, and

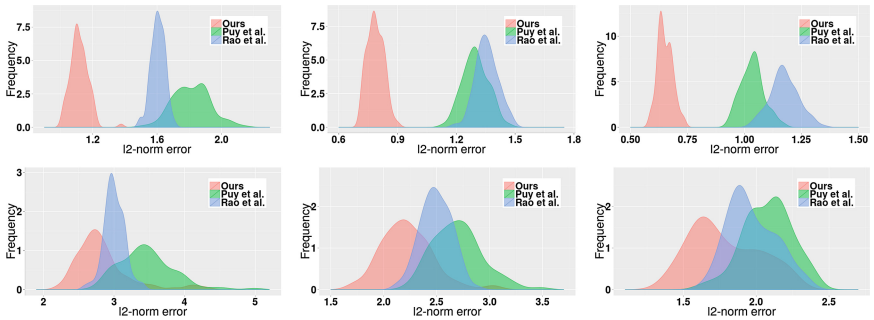


Fig. 6. Distribution of mean errors over the ROIs from 100 runs using 20 % (left column), 40 % (middle column) and 60 % (right column) samples on the HCP (top row) and the WRAP dataset (bottom row). Ours (red) show the lower errors than Puy et al. (green) and Rao et al. (blue). (Color figure online)

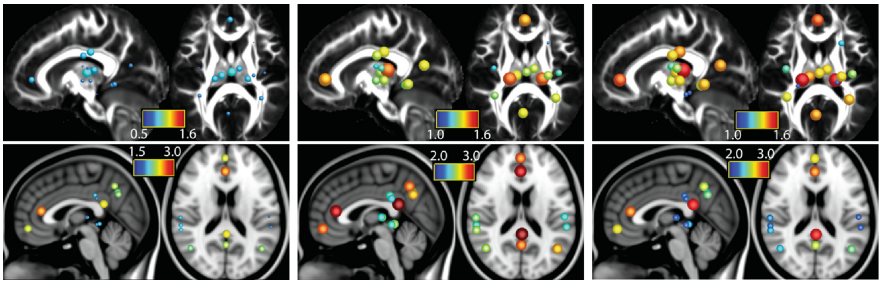


Fig. 7. Spherical representations of the prediction errors (ℓ_2-norm) in the HCP study (top) and in the WRAP study (bottom). Left: errors using Ours, Middle: errors using Puy et al., Right: errors using Rao et al. The spheres are centered at the center-of-mass of the specific bundle/regional volumes, and the radius of the spheres are proportional to the prediction error. (Color figure online)

the predicted FA for all ROIs (averaged across subjects) are presented in Fig. 8. For all 17 FA measures, with 40 % sampling, we see that our results (blue) are closest to the ground truth (red) while other methods under/over estimate. (Additional results shown in supplement.) When the ℓ_2-norm error is small, we expect results from downstream statistical analysis (e.g., p-values) will be accurate since the distributions of measurements are closer to the true sample distribution.

4.3 Prediction on a Preclinical Alzheimer's Disease Project

Dataset. Alzheimer's disease (AD) is known as a disconnection syndrome [51,52] because connectivity disruption can impede functional communication between brain regions, resulting in reduced cognitive performance [53,54]. Currently, positron emission tomography (PET) using radio-chemicals such as ^{11}C

Pittsburgh compound B (PiB) is important in mapping functional AD pathology. Distribution volume ratios (DVR) of PiB in the brain offer a good measure of the plaque pathology which is considered specific to AD. Unfortunately these PET scans are costly and involve lengthy procedures. WRAP dataset consists of participants in preclinical stages of AD [54,55] and contains 140 samples with both *low-cost* FA measures and *high-cost* PiB DVR (examples shown in Fig. 4). Utilizing the FA values over the entire set of subjects and a partial observation of the PiB measures from a fraction of the population, we investigate the performance of our model for the recovery of PiB measures.

Remark. From a neuroimaging perspective, predicting PiB measures accurately enough for actual scientific analysis is problematic. Utilizing a modality (e.g., cerebrospinal fluid) will be more appropriate for predicting PiB measures, and such results are available on the project homepage. The results below demonstrate that such a prediction task yields results numerically feasible compared to baseline strategies although not directly deployable for neuroscientific studies.

Results. For this set of experiments, we selected $p' = 17$ pathways with most reliable FA measures to construct a graph with $N = 140$ vertices (i.e., subjects). Utilizing the graph and a partial set of PiB DVR measurements from $m \ll N$ participants (20 %, 40 % and 60 % of the total population), we predicted the expensive PiB DVR values on 16 ROIs over the whole subjects. To define $\mathcal{L}$ and p, we used $\sigma = 3$ and $k = 50$. As shown in Fig. 5 in straight lines, our estimation (red) yields the smallest error compared to [30] (green) and [29] (blue) for all three sampling cases. The bottom row in Fig. 6 shows that the centers of error distribution using our algorithm (red) have lower errors than those of other methods (green and blue). As seen in the bottom panel of Fig. 7, similar to the HCP results in Sect. 4.2, we observe smaller errors in every ROI, where the actual region-wise errors are given in the right table of Table 1. Figure 8 presents the predicted regional PiB DVR values against the ground truth where our prediction in blue are consistently closer to the ground truth in red. Additional results using 20 % and 60 % of the subjects are presented in the appendix.

Table 1. Region-wise mean ℓ_2-norm of 100 runs of HCP-FA (left) and PiB DVR (right) with 40% samples. Errors from our method are the lowest shown in bold.

HCP ROIs	Ours	Puy et al.	Rao et al.
FMajor	**1.15**	1.93	1.70
FMinor	**1.21**	1.99	1.75
Fornix	**1.20**	1.84	1.65
LCingAnt	**0.95**	1.49	1.33
LcingHipp	**1.17**	1.87	1.65
LCST	**1.25**	2.06	1.82
LIFO	**1.14**	1.87	1.65
LILF	**1.16**	1.90	1.68
LSLF	**1.08**	1.77	1.56
LUnc	**0.99**	1.61	1.42
RCingAnt	**0.93**	1.48	1.32
RcingHipp	**1.20**	1.92	1.71
RCST	**1.25**	2.07	1.83
RIFO	**1.11**	1.82	1.61
RIFL	**1.12**	1.83	1.62
RSLF	**1.04**	1.69	1.50
RUnc	**1.05**	1.70	1.50

PiB ROIs	Ours	Puy et al.	Rao et al.
Angular_L	**2.89**	3.42	2.98
Angular_R	**2.73**	3.20	2.82
Cingulum_Ant_L	**3.19**	3.73	3.30
Cingulum_Ant_R	**3.18**	3.78	3.32
Cingulum_Post_L	**3.29**	4.10	3.49
Cingulum_Post_R	**3.20**	4.03	3.43
Frontal_Med_Orb_L	**2.90**	3.44	3.05
Frontal_Med_Orb_R	**3.08**	3.66	3.24
Precuneus_L	**2.88**	3.45	3.03
Precuneus_R	**3.03**	3.61	3.15
SupraMarginal_L	**2.43**	3.09	2.67
SupraMarginal_R	**2.51**	3.13	2.70
Temporal_Mid_L	**2.47**	3.13	2.68
Temporal_Mid_R	**2.59**	3.22	2.78
Temporal_Sup_L	**2.42**	3.14	2.68
Temporal_Sup_R	**2.52**	3.21	2.75

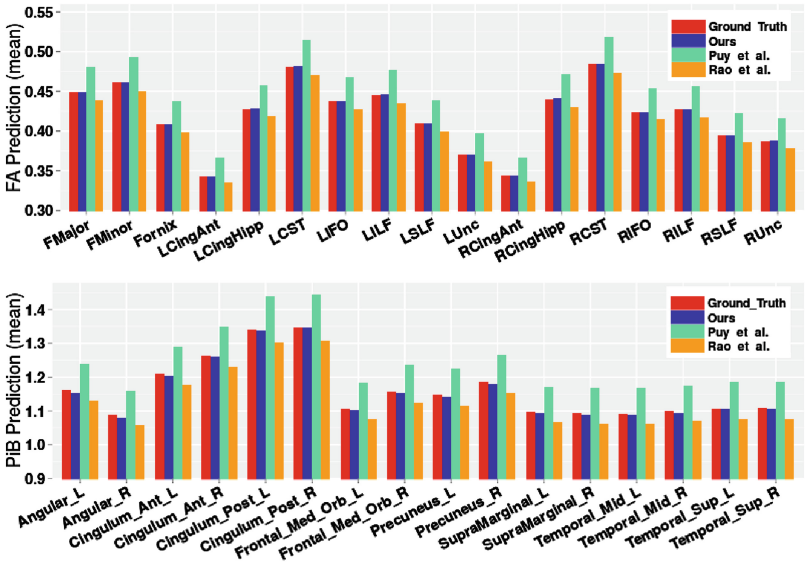


Fig. 8. Average estimations of the HCP-FA in the fiber bundle (top) and the WRAP PiB DVRs (bottom) using 40 % samples. For each measurement, the bars from the left to right are the measurements of the ground truth, ours, Puy et al. and Rao et al. Ours most closely estimate the actual ground truth values of all the measurements. (Color figure online)

5 Conclusion

In this paper, we presented an adaptive sampling scheme for signals defined on a graph. Using a dual space of these measurements obtained via a non-Euclidean wavelet transform, we show how signals can be recovered with high fidelity based on a stratified set of partial observations on the nodes of a graph. We demonstrated the application of this core technical development on accurately estimating diffusion imaging and PET imaging measures from two independent neuroimaging studies, so that one can perform standard analysis just as if the measurements were acquired directly. We presented experimental results demonstrating that our framework can provide accurate recovery using observations from only a small fraction of the full samples. We believe that this ability to estimate unobserved data based on a partial set of measurements can have impact in numerous computer vision and machine learning applications where acquisitions of large datasets often involve varying degrees of stratified human interaction. Many real experiments involve entities that have intrinsic relationships best captured as a graph. Mechanisms to exploit the properties of these graphs using similar formulations as those presented in this work may have important practical and immediate ramifications for many experimental design considerations in numerous scientific domains.

References

1. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. Artif. Intell. **97**(1), 245–271 (1997)
2. Biswas, A., Parikh, D.: Simultaneous active learning of classifiers & attributes via relative feedback. In: CVPR, pp. 644–651 (2013)
3. Jayaraman, D., Grauman, K.: Zero-shot recognition with unreliable attributes. In: NIPS, pp. 3464–3472 (2014)
4. Lughofer, E.: Hybrid active learning for reducing the annotation effort of operators in classification systems. Pattern Recognit. **45**(2), 884–896 (2012)
5. Hancock, C., Bernal, B., Medina, C., et al.: Cost analysis of diffusion tensor imaging and MR tractography of the brain. Open J. Radiol. 2014 (2014)
6. Saif, M.W., Tzannou, I., Makrilia, N., et al.: Role and cost effectiveness of PET/CT in management of patients with cancer. Yale J. Biol. Med. **83**(2), 53–65 (2010)
7. Prasad, A., Jegelka, S., Batra, D.: Submodular meets structured: finding diverse subsets in exponentially-large structured item sets. In: NIPS, pp. 2645–2653 (2014)
8. Deng, J., Dong, W., Socher, R., et al.: Imagenet: a large-scale hierarchical image database. In: CVPR, pp. 248–255 (2009)
9. Vijayanarasimhan, S., Grauman, K.: Large-scale live active learning: training object detectors with crawled data and crowds. IJCV **108**(1–2), 97–114 (2014)
10. Deng, J., Russakovsky, O., Krause, J., et al.: Scalable multi-label annotation. In: SIGCHI, pp. 3099–3102. ACM (2014)
11. Bragg, J., Weld, D.S., et al.: Crowdsourcing multi-label classification for taxonomy creation. In: AAAI (2013)
12. Read, J., Bifet, A., Holmes, G., et al.: Scalable and efficient multi-label classification for evolving data streams. Mach. Learn. **88**(1–2), 243–272 (2012)
13. Settles, B.: Active learning literature survey. University of Wisconsin, Madison vol. 52(55–66), p. 11 (2010)
14. Dasgupta, S.: Analysis of a greedy active learning strategy. In: NIPS, pp. 337–344 (2004)
15. Beygelzimer, A., Dasgupta, S., Langford, J.: Importance weighted active learning. In: ICML, pp. 49–56. ACM (2009)
16. Dasgupta, S., Hsu, D.: Hierarchical sampling for active learning. In: ICML, pp. 208–215. ACM (2008)
17. Winer, B.J., Brown, D.R., Michels, K.M.: Statistical Principles in Experimental Design. McGraw-Hill, New York (1971)
18. Lentner, M.: Generalized least-squares estimation of a subvector of parameters in randomized fractional factorial experiments. Ann. Math. Stat. **40**, 1344–1352 (1969)
19. Myers, J.L.: Fundamentals of Experimental Design. Allyn & Bacon, Boston (1972)
20. Mitchell, T.J.: An algorithm for the construction of D-optimal experimental designs. Technometrics **16**(2), 203–210 (1974)
21. De Aguiar, P.F., Bourguignon, B., Khots, M., et al.: D-optimal designs. Chemometr. Intell. Lab. Syst. **30**(2), 199–210 (1995)
22. Park, J.S.: Optimal Latin-hypercube designs for computer experiments. J. Stat. Plann. Infer. **39**(1), 95–111 (1994)
23. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artif. Intell. **2009**, 4: 2 (2009)
24. Dabov, K., Foi, A., Katkovnik, V., et al.: Image denoising by sparse 3-D transform-domain collaborative filtering. Image Process. **16**(8), 2080–2095 (2007)

25. Yu, K., Zhu, S., Lafferty, J., et al.: Fast nonparametric matrix factorization for large-scale collaborative filtering. In: SIGIR, pp. 211–218. ACM (2009)
26. Srebro, N., Salakhutdinov, R.R.: Collaborative filtering in a non-uniform world: learning with the weighted trace norm. In: NIPS, pp. 2056–2064 (2010)
27. Juditsky, A., Nemirovski, A.: On verifiable sufficient conditions for sparse signal recovery via $\ell 1$ minimization. Math. Program. **127**(1), 57–88 (2011)
28. Krahmer, F., Ward, R.: Stable and robust sampling strategies for compressive imaging. Image Process. **23**(2), 612–622 (2014)
29. Rao, N., Yu, H.F., Ravikumar, P.K., et al.: Collaborative filtering with graph information: consistency and scalable methods. In: NIPS (2015)
30. Puy, G., Tremblay, N., Gribonval, R., et al.: Random sampling of bandlimited signals on graphs. Appl. Comput. Harmonic Anal. (2016)
31. Kumar, S., Mohri, M., Talwalkar, A.: Sampling methods for the Nyström method. JMLR **13**(1), 981–1006 (2012)
32. Krishnamurthy, A., Singh, A.: Low-rank matrix and tensor completion via adaptive sampling. In: NIPS, pp. 836–844 (2013)
33. Hammond, D., Vandergheynst, P., Gribonval, R.: Wavelets on graphs via spectral graph theory. Appl. Comput. Harmonic Anal. **30**(2), 129–150 (2011)
34. Mallat, S.: A Wavelet Tour of Signal Processing. Academic press, San Diego (1999)
35. Coifman, R., Maggioni, M.: Diffusion wavelets. Appl. Comput. Harmonic Anal. **21**(1), 53–94 (2006)
36. Haykin, S., Veen, B.V.: Signals and Systems. Wiley, New York (2005)
37. Bronstein, M.M., Kokkinos, I.: Scale-invariant heat kernel signatures for non-rigid shape recognition. In: CVPR, pp. 1704–1711. IEEE (2010)
38. Aubry, M., Schlickewei, U., Cremers, D.: The wave kernel signature: a quantum mechanical approach to shape analysis. In: ICCV Workshops, pp. 1626–1633. IEEE (2011)
39. Rustamov, R.M.: Laplace-Beltrami eigen functions for deformation invariant shape representation. In: Eurographics Symposium on Geometry Processing, Eurographics Association, pp. 225–233 (2007)
40. Kim, W.H., Chung, M.K., Singh, V.: Multi-resolution shape analysis via non-euclidean wavelets: applications to mesh segmentation and surface alignment problems. In: CVPR, pp. 2139–2146. IEEE (2013)
41. Varentsova, A., Zhang, S., Arfanakis, K.: Development of a high angular resolution diffusion imaging human brain template. Neuroimage **91**, 177–186 (2014)
42. Van Essen, D.C., Smith, S.M., Barch, D.M., et al.: The WU-Minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)
43. Setsompop, K., Cohen-Adad, J., Gagoski, B., et al.: Improving diffusion MRI using simultaneous multi-slice echo planar imaging. Neuroimage **63**(1), 569–580 (2012)
44. Jbabdi, S., Sotiropoulos, S.N., Haber, S.N., et al.: Measuring macroscopic brain connections in vivo. Nature Neurosci. **18**(11), 1546–1555 (2015)
45. Sporns, O., Tononi, G., Kötter, R.: The human connectome: a structural description of the human brain. PLoS Comput. Biol. **1**(4), e42 (2005)
46. Van Essen, D.C., Ugurbil, K.: The future of the human connectome. Neuroimage **62**(2), 1299–1310 (2012)
47. Toga, A.W., Clark, K.A., Thompson, P.M., et al.: Mapping the human connectome. Neurosurgery **71**(1), 1 (2012)
48. Sporns, O.: The human connectome: origins and challenges. Neuroimage **80**, 53–61 (2013)
49. Herrick, R., McKay, M., Olsen, T., et al.: Data dictionary services in XNAT and the human connectome project. Front. Neuroinform. **8**, 65 (2014)

50. Kim, W.H., Kim, H.J., Adluru, N., et al.: Latent variable graphical model selection using harmonic analysis: applications to the human connectome project (HCP). In: CVPR. IEEE (2016)
51. Brier, M.R., Thomas, J.B., Ances, B.M.: Network dysfunction in Alzheimer's disease: refining the disconnection hypothesis. Brain connectivity **4**(5), 299–311 (2014)
52. Delbeuck, X., Van der Linden, M., Collette, F.: Alzheimer's disease as a disconnection syndrome? Neuropsychol. Rev. **13**(2), 79–92 (2003)
53. Geschwind, N.: Disconnexion syndromes in animals and man. In: Geschwind, N. (ed.) Selected Papers on Language and the Brain. Boston Studies in the Philosophy of Science, vol. 16, pp. 105–236. Springer, Amsterdam (1974)
54. Kim, W.H., Adluru, N., Chung, M.K., et al.: Multi-resolution statistical analysis of brain connectivity graphs in preclinical Alzheimer's disease. Neuroimage **118**, 103–117 (2015)
55. Kim, W.H., Singh, V., Chung, M.K., et al.: Multi-resolutional shape features via non-Euclidean wavelets: applications to statistical analysis of cortical thickness. Neuroimage **93**, 107–123 (2014)