Published in final edited form as:

Stat Biosci. 2017 December ; 9(2): 543–558. doi:10.1007/s12561-017-9195-y.

Big Data and Neuroimaging

Yenny Webb-Vargas, Shaojie Chen, Aaron Fisher, Amanda Mejia, Yuting Xu, Ciprian Crainiceanu, Brian Caffo, and Martin A. Lindquist

Abstract

Big Data are of increasing importance in a variety of areas, especially in the biosciences. There is an emerging critical need for Big Data tools and methods, because of the potential impact of advancements in these areas. Importantly, statisticians and statistical thinking have a major role to play in creating meaningful progress in this arena. We would like to emphasize this point in this special issue, as it highlights both the dramatic need for statistical input for Big Data analysis and for a greater number of statisticians working on Big Data problems. We use the field of statistical neuroimaging to demonstrate these points. As such, this paper covers several applications and novel methodological developments of Big Data tools applied to neuroimaging data.

Keywords

Big data; neuroimaging; high dimensional computation; high dimensional inference; high dimensional causal inference; data fusion; shrinkage; dynamic networks

1 Introduction

Quantitative neuroscience is a flourishing discipline. Understanding the brain is arguably among the most complex, important and challenging issues in science today. For these endeavors to be successful, a new generation of “neuro-quants” (biostatisticians working in neuroscience) is needed to make sense of the massive amounts of data being generated. This article serves as a brief discussion to some of the issues involved in the analysis of neuroimaging data. Neuroimaging is an umbrella term for an ever-increasing number of minimally invasive techniques designed to study the brain. These include a variety of rapidly evolving technologies for measuring brain properties, such as structure, function and disease pathophysiology. These technologies are applied in a vast collection of medical and scientific areas of inquiry. Neuroimaging is relevant in nearly every disease or disorder of the brain, such as Alzheimer’s disease, autism, lead exposure and multiple sclerosis, to name only a few.

Each imaging technique and application area results in enormous amounts of data. With new studies collecting repeated measurements on thousands of subjects over multiple years, the size of data sets are becoming unimaginably large and, more importantly, complex. This complexity arises because of the complicated correlation structure across space, time,

sessions and sites, coupled with a relatively weak signal that in many cases is an indirect measure of the process being studied. Complexity goes hand in hand with the modern Big Data problem. In the absence of complexity, Big Data applications are analytically routine. Given the complexity and size of neuroimaging data, simple reproducible data-analytic methods, causal thinking, data exploration, hypothesis confirmation and careful design of experiments will become increasingly important. In other words, it will require the expertise of statisticians!

Imaging techniques such as functional Magnetic Resonance Imaging (fMRI), structural Magnetic Resonance Imaging (MRI), Diffusion Tensor Imaging (DTI), Computed Tomography (CT), Dynamic Contrast Enhancement MRI (DCE-MRI), and Positron Emission Tomography (PET) have facilitated major advances in our understanding of human brain structure and function. The development of these techniques has led to an explosion in the number of neuroimaging studies performed annually. Moreover, the amount of available data produced by each modality in any given experiment is staggering. For example, a standard fMRI study is performed on a number of subjects (e.g., 10–50) brought in for multiple scanning sessions, each consisting of a number of runs (replications of a certain task), where each run consists of several hundred three-dimensional (3D) brain volumes, each made up of roughly 100, 000 voxels. Standard scanning protocols also include collection of multiple high-resolution structural and diffusion tensor images for presentation and research purposes. Though consisting of billions of measurements, we will refer to such a data set as “small-scale” imaging studies to differentiate them from “large-scale” imaging data sets that are currently being assembled and consist of thousands of subjects measured either cross-sectionally or longitudinally.

Recently there has been a great deal of discussion in the literature [1,2] regarding how studies with small sample sizes undermine the reliability of neuroscience research. The controversy centers around the focus on small-sampled studies (i.e., those consisting of 10–30 subjects), which generally have very low statistical power to detect true effects and reduced likelihood that statistically significant results represent a true effect. This has led to a situation where a large number of published results in the field are not reproducible and likely false. This problem is confounded by the fact that there exists almost as many unique data analysis pipelines as there are studies [3], and it is not well understood how the pipeline choices at various stages affect one another. Hence, there is critical need for empirical studies of the manner in which choices at each stage affect the results of all subsequent stages. This, in turn, would allow for a more evidence-based approach to pipeline creation.

Concurrent with these discussions, there have been several efforts to construct large-scale imaging databases by combining data from multiple studies. These endeavors have been performed on both a grass-roots (e.g., the 1000 Functional Connectomes Project [4]) and institutional level (e.g., the Human Connectome Project [5]), and databases consisting of more than 1, 000 subjects are becoming increasingly available, with more in the works.

In general, the largest imaging data sets tend to be collected for reasons other than targeted scientific hypothesis. That is, they tend to be either collected as part of a patient’s medical record in the course of routine medical care, or as part of large collective studies,

aggregating many smaller studies. In these cases, data quality, sampling bias, missing data and covariate or outcome availability tend to be problematic. This is in stark contrast with small data (those with a small number of subjects), where tight experimental control leads to direct measures for testing hypotheses of interest, but at the expense of lower power. Thus, speaking broadly, current large data sets used in isolation are useful for exploratory or predictive exercises, while small data sets are useful for knowledge creation and confirmatory analyses, but are critically hampered by low power.

As most researchers' ultimate goal is creating new scientific knowledge, it would be ideal to use information from large data sets to inform the analysis of small sample data. We refer to the process of merging big data sets and small data sets for the purpose of knowledge creation as "Fusion Science" or "Data Fusion". Note the distinction between knowledge creation as a goal versus translation as a goal. Often, machine learning approaches can produce useful translational tools out of big data sets, such as prediction models. However, these tools do not tend to produce the necessary interpretable parsimonious descriptions of phenomena under study, thus never creating the sort of parsimonious knowledge that could exist in a book or as scientific theory. To use an example, consider the fact that using Google searches to predict flu outbreaks [6], however accurate, yields little in the way of a mechanistic understanding of flu trends. Fusion Science has the potential to break this Exploratory/Predictive versus Confirmatory divide. Large data sets can be explored with the goal of establishing norms and priors for the use in small, more targeted data sets.

To achieve this goal we believe that work must be done on three related areas in neuroimaging. First, the analysis of small-sample and individual subject data must be improved and their limitations better understood. Second, methods must be developed for combining information from multiple imaging modalities and/or a large number of studies. Third, new methods need to be designed for harnessing this ensemble information and fusing information from large public databases with data obtained from single subject/small studies. In the absence of carefully designed Fusion Science implemented by technology savvy experts in translational science, the promise of Big Data is very likely to remain what it currently is, a *great promise* for the improvement of public health and knowledge discovery.

With this emergence in Big Data, statistics and statisticians promise to play an increasingly important role. Statistics provides rigorous methods for assessing sources of variability, reproducibility, and evaluating the performance of estimation methods. In this paper we highlight some of the areas in the cross-section of Big Data and statistics in which we work.

2 Statistics in Neuroimaging

It is becoming increasingly clear that the introduction of Big Data makes many standard statistical approaches infeasible and new paradigms must be explored. Despite the large amounts of data available, we have developed fast and (memory) scalable statistical procedures that can handle the enormous size of data sets to answer specific scientific problems. In our experience, while analytic and scientific problems are amplified by the size of the data, the major limiting factors are complexity, variability, and the heterogeneity of

technology-specific artifacts. This is particularly true in neuroimaging data, which exhibit significant correlations in space and time, low signal-to-noise ratios, and complex noise structures induced by technological, biological, temporal, and spatial variability [7]. This variability conspires to make the analysis of neuroimaging data challenging and thus requires deep statistical knowledge to avoid analytic traps.

In the following we illustrate five areas in the intersection of Big Data and neuroimaging where we are actively working. With these five examples, we seek to illustrate new approaches and tools that we utilize in order to make sense of massive amounts of neuroimaging data that we interact with on a daily basis.

2.1 High-dimensional computation

One of the most immediate problems for dealing with Big Data is the accompanying computational issues that are introduced. Recent advances in technology and software make it possible to both manage and process high-dimensional data efficiently. Hence, it is imperative that statisticians interested in Big Data expand their toolboxes to include modern data management and processing techniques. Here we describe some examples of these tools, and illustrate how they can be applied to neuroimaging data.

A number of NoSQL (Not only SQL) databases, such as MongoDB and Cassandra, have been developed specifically for managing Big Data [8]. These databases allow users to efficiently store and retrieve large amounts of data, and provide an invaluable resource for handling Big Data. In addition, as the size of data grow, parallel computing is also becoming increasingly attractive. This can be performed on multi-threads, multiple processors, network connected commodity computers, and general purpose computation with graphics processing units (GPGPU). For instance, Java provides the *Runnable* template and *Thread* class to support multi-threading. The Message Passing Interface (MPI) is a standardized and portable system for communication among multiple processes [9]. It is popular on both multi-core personal computers and large clusters. The Map-Reduce model could run tasks in parallel in a distributed system with high fault tolerance. It can be used on clusters, as well as grids. GPGPU use graphics processing units to perform computation traditionally handled by central processing units (CPU) [10]. It works best when we have numerous simple tasks, such as random number generation. Popular frameworks include the open-source OpenCL and proprietary CUDA.

Those technologies and software are becoming increasingly available in the statistics community. For instance, almost all the approaches mentioned above have been implemented in the R statistical software. The *sqldf* package implements the major features of relational database and fits naturally with the *data frame* in R [15]. Other RDMS interfaces include *RMySQL*, *RPostgreSQL*, among others [14, 16]. NoSQL interfaces in R are also evolving, for instance, the *rmongodb* package connects R with MongoDB and is a great tool for managing Big Data [17]. For parallel computing, the *snow* package supports both multithreading and multi-process programming [18]. It extends the *apply* function in R and do parallel computing in one line. The *RHadoop* package bridges R and Hadoop, making it possible to write Map/Reduce functions in R [19]. Many packages are also

developed for GPU computing, such as *gputools*, *rpud*, and *RCUDA* [20–22]. Some selected tools are shown in Figure 1.

Our group works extensively on problems that necessitate high-dimensional computation. For example, Chen and colleagues [23] have developed a parallel group independent component analysis (PGICA) algorithm that can be used to analyze fMRI images from thousands of subjects. The performance of the algorithm is greatly improved by using parallel computing on the most time consuming tasks. This extension is important, as it now allows researchers with the means to fit ICA using the complete data set from one of the large-scale imaging databases (e.g., the 1000 Functional Connectomes Project [4] or the Human Connectome Project [5]) that have recently become available. In addition, Chen and colleagues [24] have designed an efficient algorithm to solve high dimensional linear dynamical systems. Our algorithm scales to over 10, 000 dimensions without notable loss of efficiency, thereby allowing the method to be applied to whole-brain fMRI data. Shrinkage penalties on parameters and efficient storage of large sparse matrices are used to guarantee this scalability. Importantly, neither of these techniques would be feasible without using the tools outlined in this section.

2.2 High-dimensional bootstrap inference

We often analyze high-dimensional multi-subject imaging data with the ultimate goal of either making inferences about the population from which the subjects were drawn, or making out-of-sample predictions. However, for many algorithms commonly applied to imaging data the sampling distribution of the algorithm's output is either complex or unknown, which can make population inference difficult to perform using classical methods. One flexible solution that allows for the creation of confidence regions is to use the bootstrap procedure [25], but this carries the high computational cost of repeatedly fitting the algorithm many times. In addition, cross-validation procedures used to estimate prediction accuracy can similarly carry a high computational cost.

To address this computational challenge, we have developed a fast, exact method for bootstrapping and cross-validating high dimensional image data [26], where the number of voxels (or features) is far greater than the number of subjects (or observations). This fast method is directly applicable to any analysis that involves a singular value decomposition (SVD) of the sample data matrix, such as principal component analysis, independent component analysis, and penalized regression [27,28]. The method hinges on the fact that, by taking the SVD of the original sample, we can determine a lossless, low-dimensional representation of any bootstrap sample. This shortcut allows analysts to drastically simplify the calculation of any linear function of the bootstrap data, or of the bootstrap principal components (PCs). Such operations can be applied directly to low-dimensional representations of the bootstrap sample, without ever needing to store the high-dimensional bootstrap samples in memory. Computational improvements are also available for cross-validation, as this can be seen as a special case of bootstrap resampling.

As a specific illustration, we outline the steps for calculating bootstrap standard errors for principal components (PCs). Let $\mathbf{Y}$ denote a $p \times n$ data matrix of p -dimensional brain images from n subjects. Let $\mathbf{VDU}'$ denote the SVD of $\mathbf{Y}$, with the sample PCs being contained in

the columns of $\mathbf{V}$. Conditional on $\mathbf{V}$, the so-called matrix of scores ($\mathbf{DU}'$) can be used as a low-dimensional representation of the data. Let $\mathbf{V}^*\mathbf{D}^*\mathbf{U}^{*'}'$ denote the SVD of the bootstrap sample $\mathbf{Y}^*$. Finally, let the indexing notation $\mathbf{X}_{[i,j]}$ denote the element in the i^{th} row and j^{th} column of any matrix $\mathbf{X}$, let $\mathbf{X}_{[i,.]}$ denote the i^{th} row vector of $\mathbf{X}$, and let $\mathbf{X}_{[.,j]}$ denote the j^{th} column vector of $\mathbf{X}$.

It can be shown that $\mathbf{V}^*$, the matrix of bootstrap PCs, can also be calculated as $\mathbf{VA}^*$, where $\mathbf{A}^*$ is the $n \times n$ matrix of left singular vectors of the resampled scores [26]. Thus, conditional on $\mathbf{V}$, the $\mathbf{A}^*$ matrices contains all of the information regarding the bootstrap PCs. Bootstrap standard errors for the PCs can then be calculated from

$\text{Var}(\mathbf{V}_{[i,j]}^*) = \text{Var}(\mathbf{V}_{[i,.]}' \mathbf{A}_{[.,j]}^*) = \mathbf{V}_{[i,.]}' \text{Cov}(\mathbf{A}_{[.,j]}^*) \mathbf{V}_{[i,.]}$. Here, the covariance operator refers to the covariance over the bootstrap distribution.

We applied this method to quickly and efficiently calculate bootstrap standard errors and confidence regions for the first three PCs of a dataset of 352 brain MRIs. Each MRI consisted of roughly 3 million voxels. Standard errors were calculated based on 1000 bootstrap samples (see results Figure 2). Relative to standard methods, our method reduced the computation time for this procedure from approximately 4 days to only 47 minutes. Hence, the proposed method provides a scalable approach to perform inference on high-dimensional image data.

2.3 High-dimensional causal inference

The ability to make causal statements about brain function is becoming increasingly important to many neuroscientists conducting fMRI experiments. For example, researchers commonly pursue inferences about effective connectivity, i.e. the causal relationship between different brain regions [29]. In addition, they are often interested in determining how stimuli produce different brain responses, and how the brain mediates the relationship between a given stimuli and some response. However, fMRI experiments pose particular challenges to the application of conventional methods for causal inference, due to their size, indirect nature of the observations, low temporal resolution, and lack of readily available methods for disruption of neural activity. Furthermore, the blood oxygenation level dependent (BOLD) response (the signal of interest) is contaminated by measurement error, both systematic and random, that can further influence results. In order to circumvent these obstacles, we believe it is critical to have a principled framework for making causal inferences using fMRI data [30,31]. Our framework builds on the idea that causal relationships sustain counterfactual conditionals and is formalized using the potential outcomes notation commonly used in statistics [32,33].

To illustrate, let us consider a study of the effects of thermal pain on brain function. In a parallel randomized trial, we would randomly assign a portion of subjects to receive a non-painful thermal stimulation, whereas the rest of the subjects would receive a stimulus near their pain tolerance. Let $Y_i(0)$ denote subject i 's brain response if they receive the non-painful stimulus, and let $Y_i(1)$ denote their response if they receive the painful stimulus. In the language of causal inference, the pair $Y_i(1), Y_i(0)$ are called 'potential outcomes'. For each subject, the unit causal effect $Y_i(1) - Y_i(0)$ compares i 's potential outcomes under the

two conditions. This effect cannot be observed as we can only see at most one potential outcome per subject. However, under additional assumptions, the average causal effect across the population of subjects, $E[Y_X(1) - Y_X(0)]$, can be estimated. Causal assumptions help bridge the gap between potential outcomes and the outcomes that we actually observe. For example, one assumption needed to make causal conclusions is that of ‘no unmeasured confounding of the relation between the stimulus and the outcome’ [34,35]. This assumption allows us to take the average observed response of subjects under the non-painful stimuli as a proxy for the average of missing responses $Y_X(0)$ among subjects that received the painful stimuli. Randomizing the assignment of stimulus makes this assumption more plausible in practice.

In the neuroimaging context, we cannot readily make these assumptions because the standard design of an fMRI experiment is not a parallel randomized trial. Instead, the experimental design has a hierarchical structure in which different interventions are applied to the same subject over one or more sessions, the measured signal is a function of time, and the measurement error can depend on the stimuli received. In the standard statistical literature on causal inference, the assumption is that the potential outcomes are measured without systematic error. However, in practice the potential BOLD responses are measured with stimulus dependent systematic error. Thus we define unit and average causal effects that are free of systematic error [36]. Using this framework we have found that, in order to estimate unbiased causal effects, one needs to adjust for the systematic measurement error, which we do using measured covariates that model head motion, and physiological effects such as respiration and heart-rate. If instead one assumes that the units are the individual trials within the sequence for a single subject, the units later in time may depend on the treatment applied previously, and Luo and colleagues [37] developed a method applicable to this case of interference between units.

We are currently extending this approach to the case of brain regions consisting of many voxels. Although interest typically centers on the effects of stimuli on activity in brain regions (aggregates of voxels), we believe it is meaningful to consider the effect of a stimulus on the BOLD contrast in individual voxel elements, and necessary to develop the framework at this level, then build up from there. Thus we begin with a single voxel for one subject in one session, with the idea of extending the framework to multiple subjects, multiple sessions, and multiple brain regions. As each 3D brain volume consists of roughly 10^5 voxels, this entails modeling up to 10^{10} pairwise relationships. The ultimate goal is to assess and establish causality within the high-dimensional complex system that is the human brain.

We have also worked extensively on mediation analysis, which is typically used in the behavioral sciences to investigate the role of intermediate variables that lie on the causal path between a randomized treatment and an outcome variable. In our work we have extended mediation analysis to the functional data analysis (FDA) setting that allows the mediating variable to be a continuous function rather than a single scalar measure, thus providing the opportunity to study the functional effects of the mediator on the outcome [38]. This extension is particularly useful for analyzing mediators that consists of time series or high-dimensional images. We are currently working on extending these methods to

accommodate within-subject experimental design, systematic measurement error in the BOLD response, and the ability to estimate the effects of one brain region on another. For the latter, we propose experimental designs that use transcranial magnetic stimulation (TMS) in order to randomize the brain response of participants. TMS provides a means for disruption of neural activity within a specific brain region, and thus promises to allow researchers to randomize brain activation in different regions in a hypothesized brain network. This should allow for better understand causal links in the brain, by eliminating the possibility of confounding in the mediator-outcome response and providing a clearer picture of the effects of a particular brain region.

2.4 Data fusion/shrinkage

Data Fusion allows us to harness large data sets with the goal of establishing norms and priors for the use in small, more targeted data sets. In general, we have found shrinkage estimators to be a critical tool for achieving this goal. To illustrate consider the problem of using resting-state fMRI (rsfMRI) to understand brain connectivity and parcellate the brain into functionally distinct subregions. rsfMRI measures spontaneous (task-free) activation, and voxels that exhibit coordinated patterns of spontaneous activation are considered functionally connected. The motivation to create functional parcellations is multi-fold. First, such parcellations can be used to incorporate functional features into spatial normalization procedures [39]. Second, they can be used as a form of dimension reduction by grouping together functionally similar voxels and treating them as a single quantity. Such an approach can greatly reduce computational burdens and the need for overly stringent or complex multiple comparisons corrections [40]. Third, functional parcellations can be used to define functionally meaningful brain network nodes for graph analysis [41]. Finally, features of the functional parcellations themselves, such as the number or size of parcels in a pre-defined region of interest (ROI), can be studied to better understand differences in brain connectivity and organization across individuals and populations [42].

While parcellation has traditionally been performed at a group level, there is increasing interest in producing subject-specific parcellations (e.g. [43–45]). Such parcellations have the potential to increase understanding of group-level analyses by incorporating subject-level covariates. For instance, previous research has established differences in group-level parcellations of the motor cortex in children with autism spectrum disorder (ASD) and typically developing (TD) children [42]. Further analysis incorporating subject-level parcellations and covariates has the potential to shed light on the drivers of this effect, such as a developmental delay in children with ASD.

However, subject-level parcellations must be reliable to provide accurate and useful information. Unfortunately, this is often not the case due to several factors. Many methods for functional parcellation employ clustering (e.g. [46–49,42,45]), which requires a measure of similarity between voxels. The accuracy of the clustering results depend upon the accuracy of the estimate of this measure, which in turn depends upon both the noise levels in the original data and the quantity of data available. The low signal-to-noise ratio (SNR) of rsfMRI data means that long scans (30–60 minutes) have typically been required to produce accurate single-subject parcellations [43,45,50]. However, for many populations, such as

children, the elderly, and certain patient populations, long scans may not be feasible due to motion and other subject-related factors.

We recently proposed a method to improve the reliability of subject-level parcellations using rsfMRI scans of short or moderate length ([51]). Specifically, we proposed estimating the inter-voxel connectivity matrix for each subject using shrinkage estimators, which can then be used in clustering. Shrinkage estimators, such as empirical Bayes and James-Stein estimators, borrow strength from the larger population to produce more accurate subject-level estimates [52,53]. By shrinking towards the group mean, a small amount of bias is introduced in exchange for a greater reduction in variance, thereby reducing the overall error. The amount of shrinkage is determined by the ratio of noise variance to total (signal plus noise) variance and ranges from 0% to 100%. For example, 40% shrinkage produces an estimate that gives 40% weight to the group mean and 60% weight to the subject-level observation.

As a proof of concept, we performed empirical Bayes shrinkage on the inter-voxel correlation matrix within the primary motor cortex for twenty subjects. Each subject received two seven-minute rsfMRI scans, and the inter-voxel correlation matrix was estimated within each session for each subject. Spectral clustering ([54]) was performed to produce a parcellation for each session and subject. Shrinkage estimates of the subjects' correlation matrices were generated using the first scanning session for each subject. At each cell of the correlation matrix, the within-subject (noise) variance and across-subject (signal) variance were estimated and used to determine the degree of shrinkage at each cell. Using the second session as a test set, we found that shrinkage estimators were 25–30% more reliable, in terms of mean squared error. More importantly, we found that the parcellations produced using those estimates were also 25–30% more reliable, in terms of parcellation overlap as measured by the Dice coefficient. Parcellations from three example subjects are shown in Figure 3. For each subject, the parcellations produced using the first scanning session, with and without shrinkage, are shown, as well as the parcellation produced using the second scanning session without shrinkage.

However, if we are interested in features of individual subjects, an improvement in overall reliability does not necessarily indicate improved estimation of subject-level features. Therefore, we also conducted a simulation study, in which the true parcellation was known and allowed to vary across subjects within certain regions. Using simulated data from 20 'subjects', we computed both raw and shrinkage estimates of the correlation matrix, and performed spectral clustering using each to generate a subject-specific parcellation. Using the true subject-level parcellations as a reference, we computed the Dice overlap within two regions: the region that was shared across all subjects and the region where subject-level differences were allowed to occur. Somewhat surprisingly, we found that shrinkage estimates resulted in more accurate parcellations in both regions. While the improvement in accuracy was greater in the common region, with an improvement of approximately 30%, we also observed a 10–20% improvement in accuracy in the region containing subject-level differences.

While further validation is needed, this work suggests that shrinkage estimates of functional connectivity have the potential to produce reliable subject-level parcellations using short rsfMRI scans, thus achieving our goal of data fusion.

Fortunately, the increased interest in subject-level parcellation parallels a surge in the quantity of publically available rsfMRI data. The Human Connectome Project (HCP) is a large, publicly available dataset that includes 60 minutes of rsfMRI data for over 400 subjects [55]. As it will not always be feasible or desirable to collect large amounts of rsfMRI data for each subject in every study, such datasets can be used through shrinkage to improve the reliability of subject-level functional parcellations using shorter scan lengths. Datasets like the HCP can be leveraged in several ways. First, for a given ROI and connectivity measure, a “prior” population connectivity matrix can be estimated for use in shrinkage. Second, the matrix of between-subject (signal) variance can be estimated, and the within-subject (noise) variance can be estimated for different scan lengths, which can be used to determine the optimal degree of shrinkage towards the population mean in a given setting. Using a large dataset like the HCP, these quantities (the group mean and variance components) can be estimated. Finally, datasets like the HCP Lifespan Pilot Project, which includes several cohorts of differing age, and ABIDE (the Autism Brain Imaging Data Exchange) [56], can be used to produce priors for specific populations, which may be more representative of those populations than a prior generated from healthy adults. By borrowing strength from the growing amount of rsfMRI data available, we can produce more reliable functional parcellations, even when the amount of data for a given subject is limited.

2.5 Dynamic networks

Recently, there has been an increased interest in modeling brain connectivity using graphical models [57]. Here nodes can be defined according to anatomical region of interests, single voxels or sub-components from independent component analysis; while edges can be defined according to structural or functional connections. These graphical models provide a relatively simple and straightforward framework for understanding and visualizing brain activity across various regions. The rapid development of probabilistic graphical models and large-scale multivariate statistical methods provides the foundation of modeling the complex interactions within the human brain. With graph theory we can build the structural or functional network according to the following general procedure: (i) define the network nodes; (ii) estimate a continuous measure of connectivity between nodes; (iii) threshold the connectivity level to make a binary adjacency matrix for identifying the edges; (iv) calculate the network parameters of interest; and (v) compare them with the null distribution of corresponding parameters in similar random networks [58].

A complex brain network can be built on various experimental modalities, such as structural and functional MRI, DTI, MEG, or EEG. Our work focuses primarily on undirected graphs constructed from functional MRI time series, with edges computed according to the similarity of dynamic patterns between nodes over time. As each 3D brain volume contains roughly 10^5 voxels, repeatedly measured over several hundred time points, a typical fMRI experiment can result in up to 10^{10} pairwise relationships. To analyze a full voxel-wise correlation matrix thus poses significant statistical challenges. Calculation involving such

large matrices has been prohibitively slow in the past, so the ‘traditional’ way of investigating connectivity only considers a small subset of data. However, with high-performance computing technology and well-designed algorithms, the research into connectivity can extend to increasingly larger networks.

To date most rsfMRI studies have assumed a static network with constant functional connectivity between time series. However, evidence from various experimental studies suggest that there exist dynamic changes in functional connectivity, ranging from seconds to minutes [59], even as a subject ‘rests’. While exciting, this has led to further computational complications as separate connectivity matrices must now be estimated for every time point.

As researchers have developed an interest in time-varying networks, a simple and straight forward idea, the sliding-window analysis, is commonly used [60]. There are many ways to improve and utilize the sliding window idea and obtain smoothly varying networks. For example, Lindquist et al. [61] developed a hierarchical exponentially weighted moving average (EWMA) approach which produced reasonable and novel results on empirical data. Qiu and Han et al. [62] proposed a kernel based method for jointly estimating multiple graphical models with high dimensions. Moreover, recently Lindquist et al. [63] suggested to use model-based approach in estimating dynamic correlations, namely the dynamic conditional correlation (DCC) model [64], which is well-studied in finance literature. It achieves better balance between overall sensitivity and specificity compared to EWMA and other sliding window techniques, and scales well to high dimensional time series.

Yet another alternative approach is to detect the state-related changes in fMRI activity and estimate parameters for different states [65]. For example, dynamic connectivity regression (DCR), and its extensions, is an effective data-driven technique that partitions a time course into distinct segments based on differences in connectivity patterns [66–68]. The study of dynamic networks promises to be an active area in coming years.

3 Conclusions

These five examples are illustrations of problems facing researchers working in the intersection of Big Data and neuroimaging. The list is not meant to be exhaustive, but rather is meant to illustrate important problems that require new paradigms and new tools. There is an enormous need for new fast, scalable statistical procedures that can handle the enormous size of data sets, and be used to answer important scientific problems.

Working on the biostatistical aspects of neuroimaging is a tremendous opportunity for students who have rigorous quantitative training in statistics, mathematics, computer science or engineering. It is a wonderful time—in the midst of a data revolution—to be a statistician. Quantitative neuroscience provides the opportunity for statisticians to be at the forefront of one of the most exciting challenges in science today by providing the critical skills needed to address the questions posed by big data properly. We hope the discipline is up for the task and the coming generation of biostatisticians embraces the challenges of neuroimaging in the same manner previous generations embraced the genomics revolution.

Acknowledgments

The projects described were supported by the NIH grants R01 EB012547 and R01 EB016061 from the National Institute of Biomedical Imaging And Bioengineering, and R01 NS060910 from the National Institute of Neurological Disorders and Stroke.

References

1. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR. *Nature Reviews Neuroscience*. 2013; 14(5):365. [PubMed: 23571845]
2. Munafò M, Noble S, Browne WJ, Brunner D, Button K, Ferreira J, Holmans P, Langbehn D, Lewis G, Lindquist M, et al. *Nature biotechnology*. 2014; 32(9):871.
3. Carp J. *Neuroimage*. 2012; 63(1):289. [PubMed: 22796459]
4. Biswal BB, Mennes M, Zuo XN, Gohel S, Kelly C, Smith SM, Beckmann CF, Adelstein JS, Buckner RL, Colcombe S, et al. *Proceedings of the National Academy of Sciences*. 2010; 107(10):4734.
5. Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens T, Bucholz R, Chang A, Chen L, Corbetta M, Curtiss SW, et al. *Neuroimage*. 2012; 62(4):2222. [PubMed: 22366334]
6. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. *Nature*. 2008; 457(7232):1012.
7. Lindquist MA, et al. *Statistical Science*. 2008; 23(4):439.
8. Cattell R. *ACM SIGMOD Record*. 2011; 39(4):12.
9. Snir, M. *MPI—the Complete Reference: The MPI core*. Vol. 1. MIT press; 1998.
10. Luebke, D., Harris, M., Govindaraju, N., Lefohn, A., Houston, M., Owens, J., Segal, M., Papakipos, M., Buck, I. *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*. ACM; 2006. p. 208
11. Bock T. *R4couchdb: Collection of r functions for couchdb access*.
12. B. Lewis, M.B. Lewis, (2014)
13. James DA, Falcon S. *R package version 0.11*. 2011; 1
14. James DA, DebRoy S. *R package version 0.9-3*. 2012
15. Grothendieck G. *sqldf: Perform sql selects on r data frames*. 2012
16. Conway J, Eddelbuettel D, Nishiyama T, Prayaga S, Tiffin N. *Rpostgresql: R interface to the postgresql database system (2010)*. *r package version 0.1-7*. 2012
17. Lindsly G. *rmongodb: R-mongodb driver*. 2012
18. Tierney, L., Rossini, A., Li, N., Sevcikova, H. *R package version 0.3-3*. 2008. URL <http://CRAN.R-project.org/package=snow>
19. A.M.S. Pathak, T. Bannard,
20. Buckner J, Wilson J, Seligman M, Athey B, Watson S, Meng F. *Bioinformatics*. 2010; 26(1):134. [PubMed: 19850754]
21. D. Eddelbuettel, (2014)
22. Duato, J., Pena, AJ., Silla, F., Mayo, R., Quintana-Ortí, ES. *High Performance Computing and Simulation (HPCS), 2010 International Conference on*. IEEE; 2010. p. 224-231.
23. Chen S, Huang L, Qiu H, Nebel MB, Mostofsky S, Pekar J, Eloyan A, Caffo B. *A parallel group independent component analysis algorithm*. 2014
24. Chen S, Vogelstein J, Lee S, Lindquist M, Caffo B. *High dimensional state space model with l-1 and l-2 penalties*. 2014
25. Efron, B., Tibshirani, R. *An introduction to the bootstrap*. Vol. 57. CRC press; 1993.
26. Fisher A, Caffo B, Schwartz B, Zipunnikov V. *arXiv preprint arXiv:1405.0922*. 2014
27. Jolliffe, I. *Principal component analysis*. Wiley Online Library; 2005.
28. Koch, I. *Analysis of Multivariate and High-Dimensional Data*. Cambridge University Press; 2013. URL <http://dx.doi.org/10.1017/CBO9781139025805>. Cambridge Books Online
29. Stephan KE, Roebroeck A. *NeuroImage*. 2012; 62(2):856. [PubMed: 22248576]

30. Lindquist MA, Sobel ME. *NeuroImage*. 2011; 57(2):334. [PubMed: 20970507]
31. Lindquist MA, Sobel ME. *NeuroImage*. 2013; 76:446. [PubMed: 22119004]
32. Splawa-Neyman J, Dabrowska D, Speed T, et al. *Statistical Science*. 1990; 5(4):465.
33. Rubin DB. *Journal of educational Psychology*. 1974; 66(5):688.
34. Rubin DB. *Journal of Educational Psychology*. 1974; 66(5):688.
35. Greenland S, Robins JM, Pearl J. *Statistical Science*. 1999; 14(1):29.
36. Sobel ME, Lindquist MA. *Journal of the American Statistical Association*. 2014; 109(507):967. [PubMed: 25506108]
37. Luo X, Small DS, Li CSR, Rosenbaum PR. *Journal of the American Statistical Association*. 2012; 107(498):530. URL <http://www.tandfonline.com/doi/abs/10.1080/01621459.2012.655954>. doi: 10.1080/01621459.2012.655954 [PubMed: 26504255]
38. Lindquist MA. *Journal of the American Statistical Association*. 2012; 107(500):1297. [PubMed: 25076802]
39. Thirion B, Flandin G, Pinel P, Roche A, Ciuciu P, Poline JB. *Human brain mapping*. 2006; 27(8): 678. [PubMed: 16281292]
40. Lindquist M, Mejia A. *Psychosomatic Medicine*. 2014 In press.
41. Sporns O, Tononi G, Kötter R. *PLoS computational biology*. 2005; 1(4):e42. [PubMed: 16201007]
42. Nebel MB, Joel SE, Muschelli J, Barber AD, Caffo BS, Pekar JJ, Mostofsky SH. *Human brain mapping*. 2012
43. Cohen AL, Fair DA, Dosenbach NU, Miezin FM, Dierker D, Van Essen DC, Schlaggar BL, Petersen SE. *Neuroimage*. 2008; 41(1):45. [PubMed: 18367410]
44. Ryali S, Chen T, Supekar K, Menon V. *NeuroImage*. 2012
45. Blumensath T, Jbabdi S, Glasser MF, Van Essen DC, Ugurbil K, Behrens TE, Smith SM. *Neuroimage*. 2013; 76:313. [PubMed: 23523803]
46. Cordes D, Haughton V, Carew JD, Arfanakis K, Maravilla K. *Magnetic resonance imaging*. 2002; 20(4):305. [PubMed: 12165349]
47. Salvador S, Brovelli A, Longo R. *Magnetic resonance imaging*. 2002; 20(2):207. [PubMed: 12034342]
48. Kim JH, Lee JM, Jo HJ, Kim SH, Lee JH, Kim ST, Seo SW, Cox RW, Na DL, Kim SI, et al. *Neuroimage*. 2010; 49(3):2375. [PubMed: 19837176]
49. Craddock RC, James GA, Holtzheimer PE, Hu XP, Mayberg HS. *Human brain mapping*. 2011; 33(8):1914. [PubMed: 21769991]
50. Wig GS, Laumann TO, Cohen AL, Power JD, Nelson SM, Glasser MF, Miezin FM, Snyder AZ, Schlaggar BL, Petersen SE. *Cerebral Cortex*. 2013:bht056.
51. Mejia AF, Nebel MB, Shou H, Crainiceanu CM, Pekar JJ, Mostofsky S, Caffo B, Lindquist MA. *arXiv preprint arXiv:1409.5450*. 2014
52. James W, Stein C. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. 1961; 1:361–379.
53. Efron B, Morris C. *Journal of the American Statistical Association*. 1975; 70(350):311.
54. Ng, AY., Jordan, MI., Weiss, Y. *Proceedings of Advances in Neural Information Processing Systems*. Vol. 14. Cambridge, MA: MIT Press; 2001. p. 849
55. Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K. *Neuroimage*. 2013; 80:62. [PubMed: 23684880]
56. Di Martino A, Yan C, Li Q, Denio E, Castellanos F, Alaerts K, Anderson J, As-saf M, Bookheimer S, Dapretto M, et al. *Molecular psychiatry*. 2014; 19(6):659. [PubMed: 23774715]
57. Bullmore ET, Bassett DS. *Annual review of clinical psychology*. 7:113. (cloak).
58. Bullmore E, Sporns O. *Nature Reviews Neuroscience*. 2009; 10(3):186. [PubMed: 19190637]
59. Chang C, Glover GH. *Neuroimage*. 2010; 50(1):81. [PubMed: 20006716]
60. Nakai T, Bagarinao E, Matsuo K, Ohgami Y, Kato C. *Journal of neuroscience methods*. 2006; 157(1):158. [PubMed: 16765449]
61. Lindquist MA, Waugh C, Wager TD. *Neuroimage*. 2007; 35(3):1125. [PubMed: 17360198]
62. Qiu H, Han F, Liu H, Caffo B. *arXiv preprint arXiv:1311.0219*. 2013

63. Lindquist MA, Xu Y, Nebel MB, Caffo BS. NeuroImage. 2014
64. Engle R. Journal of Business & Economic Statistics. 2002; 20(3):339.
65. Robinson LF, Wager TD, Lindquist MA. Neuroimage. 2010; 49(2):1581. [PubMed: 19733671]
66. Cribben I, Haraldsdottir R, Atlas LY, Wager TD, Lindquist MA. Neuroimage. 2012; 61(4):907. [PubMed: 22484408]
67. Cribben I, Wager TD, Lindquist MA. Frontiers in computational neuroscience. 2013; 7
68. Xu Y, Lindquist MA. Frontiers in neuroscience. 2015; 9

Fig. 1.
R toolkit for analysing Big Data [11–14]

Fig. 2.

Bootstrap standard errors for MRI principal components (PCs) - The left column of plots shows mid-sagittal slices from the first three PCs of the dataset. The right column shows voxel-wise standard errors for these PCs, calculated via fast bootstrap.

Fig. 3.
Subject-level parcellations of the motor cortex from 3 example subjects.