Statistics and Probability Letters 136 (2018) 78–82

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Statistical challenges of big brain network data

Moo K. Chung

University of Wisconsin, Madison, WI, USA

ARTICLE INFO

Article history:
Available online 9 April 2018

Keywords:
Big brain network data
Sparsity
Hierarchy
Multiscale
Graph filtration

ABSTRACT

We explore the main characteristics of big brain network data that offer unique statistical challenges. The brain networks are biologically expected to be both sparse and hierarchical. Such unique characterizations put specific topological constraints onto statistical approaches and models we can use effectively. We explore the limitations of the current models used in the field and offer alternative approaches and explain new challenges.

1. Introduction

Wikipedia defines *big data* as datasets that are so large or complex that traditional data processing application software is inadequate to deal with them (en.wikipedia.org/wiki/Big_data). Big data is not just about the size of the data although that is the main obstacle of using traditional statistical approaches. Big data usually include datasets with sizes beyond the ability of standard software tools to process and analyze within a reasonable time limit. Even 100 MB of data can be big if existing computing resources can only handle 1 MB of data at a time. Thus, the size of the data is a *relative* quantity respect to the available computing resources.

If we pick any article in big data literature these days, chances are that we often encounter hardware solutions to solving big data problems. They often suggest increasing more central processing units (CPU) or graphical processing units (GPU) and emphasize the need for cluster or parallel computing. For instance, Boubela et al. (2016) suggests to use parallel computing as a way to compute large-scale Pearson correlation coefficients for 390 GB of data in the Human Connectome Project (HCP) but did not suggest any other simpler algorithmic approaches that can be implemented in a limited computing resource environment. Simply adding more hardware is not necessarily an effective but costly strategy for big data. Such hardware approaches often do not provide a venue for more interesting statistical problems. Further, the access to fast computational resources is not necessarily given to everyone. Many biological laboratories still do not have technical expertise of using cluster or parallel computing. Therefore, it is often necessary to develop more algorithmic and statistical approaches in addressing big data at least for biological sciences.

In this paper, we focus on the statistical challenges of big data in brain imaging and networks that are somewhat different from more traditional big data problems.

2. Large-scale brain imaging data

Many big datasets introduce unique computational and statistical challenges that include scalability, storage bottleneck, data representation visualization, and computation mostly related to sample sizes (Fan et al., 2014). However, the challenges in big brain imaging datasets such as HCP and Alzheimer's Disease Neuroimaging Initiative (ADNI; adni.loni.usc.edu) are

E-mail address: mkchung@wisc.edu.
URL: http://www.stat.wisc.edu/∼mchung.

https://doi.org/10.1016/j.spl.2018.02.020
0167-7152/

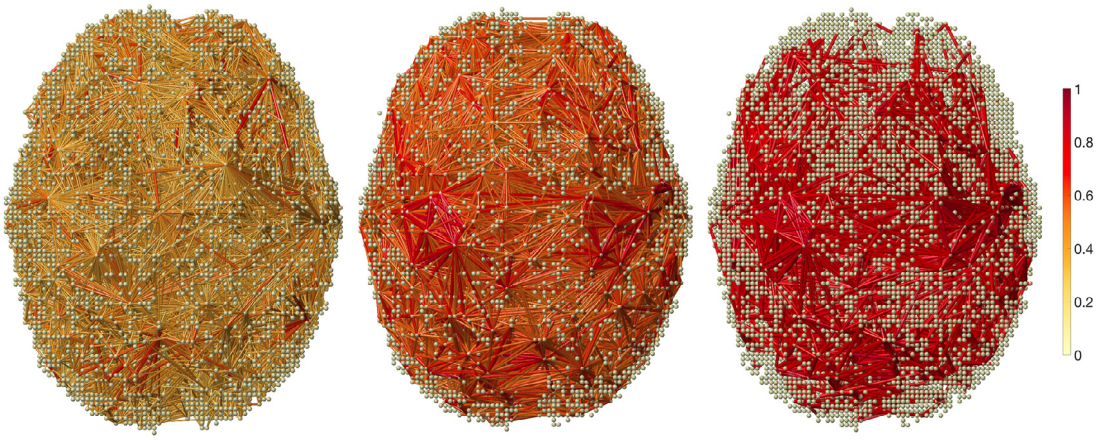

Fig. 1. Dense resting-state fMRI correlation network consisting of 25 000 nodes obtained from HCP. The network is so dense, simply displaying all the nodes and edges of the network is not very informative. It is necessary to represent such dense network more sparsely. The sparse correlation network model with sparse parameters $\lambda = 0, 3, 0.5, 0.7$ (Chung et al., 2017c). It can be shown that they form a nested hierarchy called the graph filtration.

slightly different. There are substantially more number of voxels (p) per image than the number of images (n) in the datasets. Even at 3 mm low resolution, functional magnetic resonance images (fMRI) has more than 25 000 voxels (Chung et al., 2017c). Unless the dataset consists of more than 25 000 images, brain imaging is often the problem of *small-n large-p*, which is different from the usual big data setting where n is often big. HCP and ADNI have n in the range of a thousands, far smaller than the number of voxels.

Traditionally, numerical accuracy has been less of concerns in brain imaging particularly due to spatial and temporal smoothing often done in images to smooth out various image processing artifacts and physiological noises. Due to the increased sample size and the central limit theorem, which is further reinforced by smoothing, the statistical distribution of the data might become less of a concern in big imaging data (Salmond et al., 2002).

In the traditional mass univariate approaches (Chung et al., 2015; Worsley et al., 1992), where statistical inference is done at each voxel, the problem of *small-n large- p* is not critical. Further, spatial smoothing has the effect of reducing the number of *resolution element* (RESEL), so we have far less number of effective p (Worsley et al., 1992). Smoothing also reduces the effect of image registration errors and high frequency noise. Gaussian kernel smoothing introduces continuous hierarchical structure through scale space (Worsley et al., 1996). However, small-n large-p problems become critical in brain network modeling, where we need to correlate different voxels. In the small-n large-p setting, the sample covariance and correlation matrices are no longer positive definite. Subsequently, up to $p - n$ nodes are statistically dependent although there might be *no* true dependency at all. Thus, there is need to constrain the covariance or correlation matrices by regularization methods such as sparse network models. Unfortunately, for large p, many sparse models have severe computational bottlenecks (Chung et al., 2015).

There begin to emerge large-scale brain networks with more than 25 000 nodes, where each voxel is taken as a network node (Fig. 1) (Chung et al., 2017c; Eguíluz et al., 2005; Hagmann et al., 2007; Taylor et al., 2017). The size of such large-scale brain networks can easily match publicly available network data such as Stanford Large Network Dataset (snap.stanford.edu/data). In such large-scale networks, the small-n large-p problem will be more severe.

3. Large-scale brain networks

Purely data-driven approaches for large-scale brain networks are not going to be computationally efficient or effective. It is often necessary to incorporate the first-order principles of brain networks into models to possibly reduce computational bottlenecks.

3.1. Sparsity

At the microscopic level, the activation of cortical neurons in the brain show *sparse* and widely distributed patterns (Histed et al., 2009). At the macroscopic level, diffusion tensor imaging (DTI) can produce up to a half million white matter fiber tracts per brain. Even then not every part of the brain is anatomically connected to other parts of the brain but sparsely connected (Chung et al., 2017b). This can be seen from Fig. 2, where the brain is parcellated into 116 disjoint regions and the number of white matter fiber tracts passing between the regions is used in constructing the structural connectivity matrix (Chung et al., 2017b). Even though the white matter fibers are very dense, the resulting connectivity matrix is sparse. For 116×116 connectivity matrix, 60% of entries are zeros. As we increases the number of parcellations, the sparsity increases

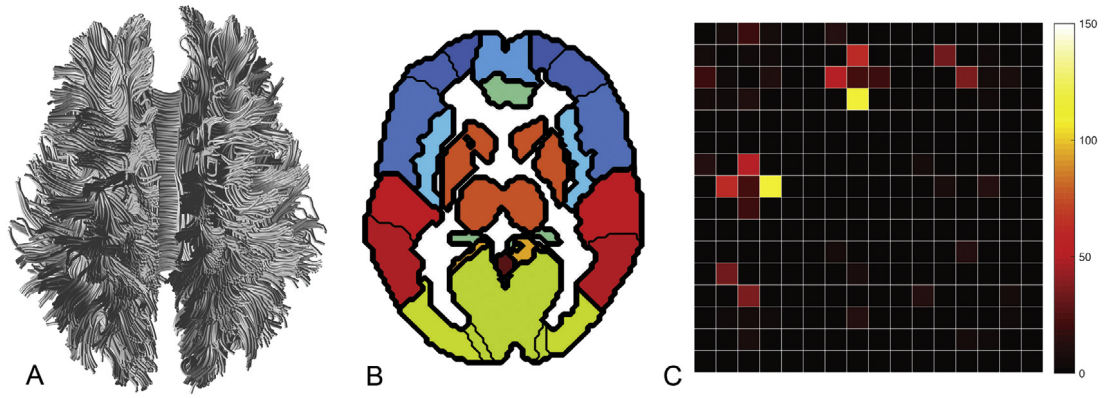

Fig. 2. A. White matter fiber tracts obtained from a tractography algorithm. B. The brain is parcellated into 116 disjoint regions. C. Connectivity matrix showing how each region is connected to other regions. Even though fiber tracts are very dense, the resulting connective matrix is always sparse since not every part of brain is connected to each other (Chung et al., 2017b).

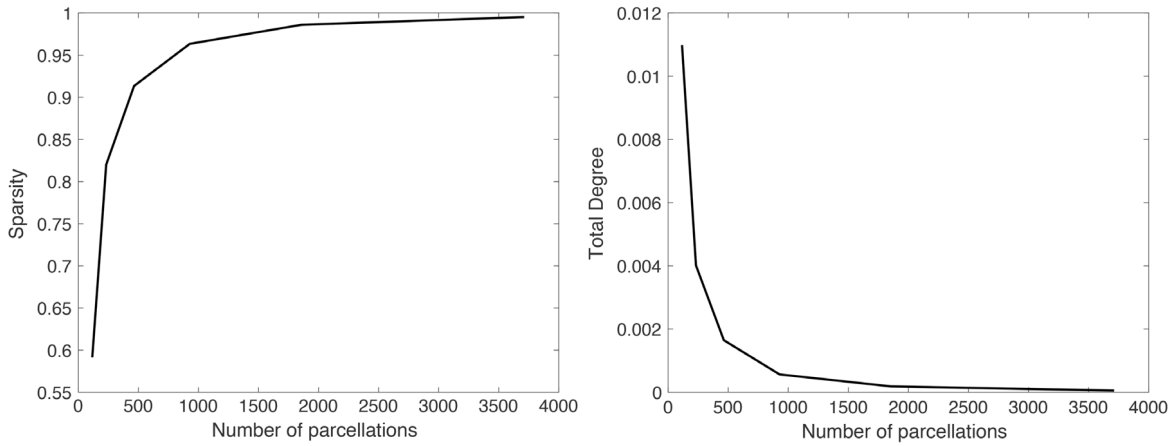

Fig. 3. Left: plot of sparsity over the number of parcellations. The sparsity is measured as the ratio of zero entries over all entries in the connectivity matrix. Right: plot of total degree of nodes over the number of parcellations. The vertical axis measures the ratio of the total number of connections over every possible connection. The plots all show the sparse nature of brain networks at any spatial scale.

while the total degree of all nodes decreases (Fig. 3). Note the degree of nodes counts the number of connections at a node. Thus, it also measures the sparsity of the network.

In fMRI studies, functional connectivity, which measures the dependency of brain activity in one region to another region, is often measured by correlation, covariance or spectral coherence of fMRI time series. Since the brain does not activate everywhere simultaneously (Chung et al., 2017c), functional connectivity is also expected to be not dense but sparsely clustered. It is reasonable to assume both functional and anatomical brain networks are sparsely connected at the both microscopic and macroscopic levels. Thus, there is strong biological justifications for modeling brain networks sparsely.

The small-n large-p problem in brain imaging often produces under-determined models with infinitely many possible solutions. Such problems are usually remedied by regularizing the systems with additional sparse penalties. Sparse models used in brain imaging include compressed sensing (CS) (Lee et al., 2011), sparse correlations (Chung et al., 2017c), least absolute shrinkage and selection operator (LASSO) (Huang et al., 2009; Lee et al., 2011), sparse canonical correlations (Avants et al., 2010) and graphical-LASSO (Chung et al., 2015; Huang et al., 2009). Most of these sparse models require optimizing $L1$-norm penalties, which has been the major computational bottleneck for solving large-scale problems in brain imaging. Thus, almost all sparse brain network models have been restricted to a few hundreds nodes or less. 2527 MRI features used in a LASSO model for Alzheimer's disease (Xin et al., 2015) is probably the largest number of features used in any sparse model in the brain imaging literature. Recently, a more scalable large-scale sparse brain network models, where each voxel is a network node, are begin to emerge (Chung et al., 2017c). For such large-scale network construction, faster scalable algorithms are needed. In Chung et al. (2017c), the computational bottleneck of $L1$-optimization is overcame by simplifying the sparse network problem into an orthogonal design. Other promising methods include a constrained L_1-minimization estimator (CLIME) (Wang et al., 2016) and faster computations for graphical-LASSO (Witten et al., 2011) although they were never applied to large-scale brain networks yet.

3.2. Hierarchy

Brain networks are fundamentally *multiscale*. Intuitive and palatable biological hypothesis is that brain networks are organized into *hierarchies* (Betzel and Bassett, 2017). A brain network at any particular sale might be subdivided into subnetworks, which can be further subdivided into smaller subnetworks in an iterative fashion. There have been various attempts at modeling brain networks at multiple scales (Betzel and Bassett, 2017; Chung et al., 2015, 2017c; Lee et al., 2012). Unfortunately, many multiscale models give raise to conflicting topological structures of the networks from one scale to the next. For instance, the estimated modular structure in the multiscale community detection problem usually do not have continuity over different resolution parameters (Betzel and Bassett, 2017).

Any sparse brain network model is usually parameterized by a tuning parameter that controls the sparsity of the solution. Increasing the sparse parameter makes the solution more sparse. Thus, sparse models are inherently multiscale, where the scale of the model is determined by the sparsity. Many existing sparse network models use a fixed parameter λ that may not be optimal in other datasets or studies. Depending on the choice of the sparse parameter, the final network structure will be different (Chung et al., 2015; Lee et al., 2012). There is a need to develop a multiscale sparse network model that provide a consistent analysis results and interpretation regardless of the choice of parameter (Chung et al., 2015, 2017c).

Persistent homology may offer an effective framework in addressing the topological inconsistency in multiscale models. Instead of studying images and networks at a fixed scale, as usually done in traditional approaches, persistent homology summarizes the changes of topological features over different scales and identifies the most persistent topological features that are robust under different scales. This robust performance under different scales is needed for network models that are parameter and scale dependent. Instead of building networks at one fixed parameter that may not be optimal, persistent homological approaches exploit the topological structure of the data and models. In doing so, topologically consistent nested hierarchical networks called the *graph filtration* is obtained (Lee et al., 2012; Chung et al., 2015). Such a nested hierarchical structure can further speed up various computations for even for large-scale networks with a billions of connections (Chung et al., 2017c).

4. Discussion

We have presented two main characterizations (sparsity and hierarchy) of brain networks that should be utilized even in big data environments. We have further explored various statistical challenges related to such characterizations.

The issue of sparsity and hierarchy is highly relevant to other types of big network data such as social networks (Christakis and Fowler, 2007), World Wide Web (WWW) (Adamic, 1999) and genomic regulatory networks (Luscombe et al., 2004). Given any type of real world network, it is unlikely that all the nodes are densely connected to each other. It is expected that the network to have sufficient sparsity. Many large scale networks such as social networks and WWW show scale-free characteristic, which is the main characteristic of hierarchical networks. Although we do not expect all networks to be hierarchical or sparse, these aspect of brain network should be applicable to other big network data.

In terms of computation, many existing brain image analysis software such as SPM (www.fil.ion.ucl.ac.uk/spm) and AFNI (afni.nimh.nih.gov) are not effective for big data. The general statistical premise of such mainstream tools is that all the image measurements are available in the computer memory and statistics are computed using all the data. However, in the big data setting, it may not be possible to fit all of the imaging data in a computer's memory, making it necessary to perform the analysis by adding one image at a time in a sequential manner. We need a way to incrementally update the statistical analysis results without repeatedly running the entire analysis whenever new images or parts of images are added.

An *online algorithm* is one that processes its inputted data in a sequential manner (Chung et al., 2017a). Instead of processing the entire set of imaging data from the start, an online algorithm processes one image at a time. That way, we can bypass the memory requirement, reduce numerical instability and increase computational efficiency. With the ever-increasing amount of large-scale brain imaging datasets such as ADNI and HCP, the development of various online statistical method is warranted (Chung et al., 2017a). Thus, here is an immediate need to develop the online version of sparse or hierarchical network models although there are no such available methods yet. Even large-scale Pearson correlation coefficients can be computed using an online algorithm.

Existing statistical analysis packages such as MATLAB and R also assume all measurements to be available in computer memory. Unless substantial modification to existing codes is made, we cannot even compute t-statistics for extremely large data that will not fit into the computer memory using the built-in functions. Thus, there is a strong need to develop online algorithms for big data beyond brain imaging.

Acknowledgments

This research was supported by NIH Brain Initiative Grant R01 EB022856. We would like to thank anonymous reviewer for constructive critiques that improved the paper. We would like to thank Gregory Kirk and Andrey Grisenko of University of Wisconsin-Madison for providing logistical support for generating Fig. 1. We would also like to thank John Aston and Eardi Lila of Cambridge University for explaining us the final structures of HCP.

References

Adamic, L., 1999. The small world web. In: ECDL, Vol. 99. pp. 443–452.

Avants, B., Cook, P., Ungar, L., Gee, J., Grossman, M., 2010. Dementia induces correlated reductions in white matter integrity and cortical thickness: a multivariate neuroimaging study with sparse canonical correlation analysis. NeuroImage 50, 1004–1016.

Betzel, R., Bassett, D., 2017. Multi-scale brain networks. Neuroimage 160, 73–83.

Boubela, R., Kalcher, K., Huf, W., Našel, C., Moser, E., 2016. Big data approaches for the analysis of large-scale fMRI data using apache spark and GPU processing: a demonstration on resting-state fMRI data from the human connectome project. Front. Neurosci. 9, 492.

Christakis, N., Fowler, J., 2007. The spread of obesity in a large social network over 32 years. N. Engl. J. Med. 2007, 370–379.

Chung, M., Chuang, Y., Vorperian, H., 2017a. Online statistical inference for large-scale binary images. In: MICCAI. In: Lecture Notes in Computer Science (LNCS), vol. 10434, pp. 729–736.

Chung, M., Hanson, J., Adluru, L., Alexander, A., Davidson, R., Pollak, S., 2017b. Integrative structural brain network analysis in diffusion tensor imaging. Brain Connect. 7, 331–346.

Chung, M., Hanson, J., Ye, J., Davidson, R., Pollak, S., 2015. Persistent homology in sparse regression and its application to brain morphometry. IEEE Trans. Med. Imaging 34, 1928–1939.

Chung, M., Vilalta-Gil, V., Lee, H., Rathouz, P., Lahey, B., Zald, D., 2017c. Exact topological inference for paired brain networks via persistent homology. In: Information Processing in Medical Imaging. (IPMI), In: Lecture Notes in Computer Science, vol. 10265, pp. 299–310.

Eguíluz, V., Chialvo, D., Cecchi, G., Baliki, M., Apkarian, A., 2005. Scale-free brain functional networks. Phys. Rev. Lett. 94 (1).

Fan, J., Han, F., Liu, H., 2014. Challenges of big data analysis. Natl. Sci. Rev. 1, 293–314.

Hagmann, P., Kurant, M., Gigandet, X., Thiran, P., Wedeen, V., Meuli, R., Thiran, J., 2007. Mapping human whole-brain structural networks with diffusion MRI. PLoS One 2 (7), e597.

Histed, M.H., Bonin, V., Reid, R., 2009. Direct activation of sparse, distributed populations of cortical neurons by electrical microstimulation. Neuron 63, 508–522.

Huang, S., Li, J., Sun, L., Liu, J., Wu, T., Chen, K., Fleisher, A., Reiman, E., Ye, J., 2009. Learning brain connectivity of Alzheimer's disease from neuroimaging data. In: Advances in Neural Information Processing Systems. pp. 808–816.

Lee, H., Kang, H., Chung, M., Kim, B.-N., Lee, D., 2012. Persistent brain network homology from the perspective of dendrogram. IEEE Trans. Med. Imaging 31, 2267–2277.

Lee, H., Lee, D., Kang, H., Kim, B.-N., Chung, M., 2011. Sparse brain network recovery under compressed sensing. IEEE Trans. Med. Imaging 30, 1154–1165.

Luscombe, N., Babu, M., Yu, H., Snyder, M., Teichmann, S., Gerstein, M., 2004. Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431, 308–312.

Salmond, C., Ashburner, J., Vargha-Khadem, F., Connelly, A., Gadian, D., Friston, K., 2002. Distributional assumptions in voxel-based morphometry. NeuroImage 17, 1027–1030.

Taylor, P., Wang, Y., Kaiser, M., 2017. Within brain area tractography suggests local modularity using high resolution connectomics. Sci. Rep. 7, 39859.

Wang, Y., Kang, J., Kemmer, P., Guo, Y., 2016. An efficient and reliable statistical method for estimating functional connectivity in large scale brain networks using partial correlation. Front. Neurosci. 10, 123.

Witten, D.M., Friedman, J.H., Simon, N., 2011. New insights and faster computations for the graphical LASSO. J. Comput. Graph. Statist. 20, 892–900.

Worsley, K., Evans, A., Marrett, S., Neelin, P., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. 12, 900–918.

Worsley, K., Marrett, S., Neelin, P., Evans, A., 1996. Searching scale space for activation in pet images. Hum. Brain Mapp. 4, 74–90.

Xin, B., Hu, L., Wang, Y., Gao, W., 2015. Stable feature selection from brain sMRI. In: Proceedings of the Twenty-Nineth AAAI Conference on Artificial Intelligence.