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 xxxx

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/© 2018 Elsevier B.V. All rights reserved.

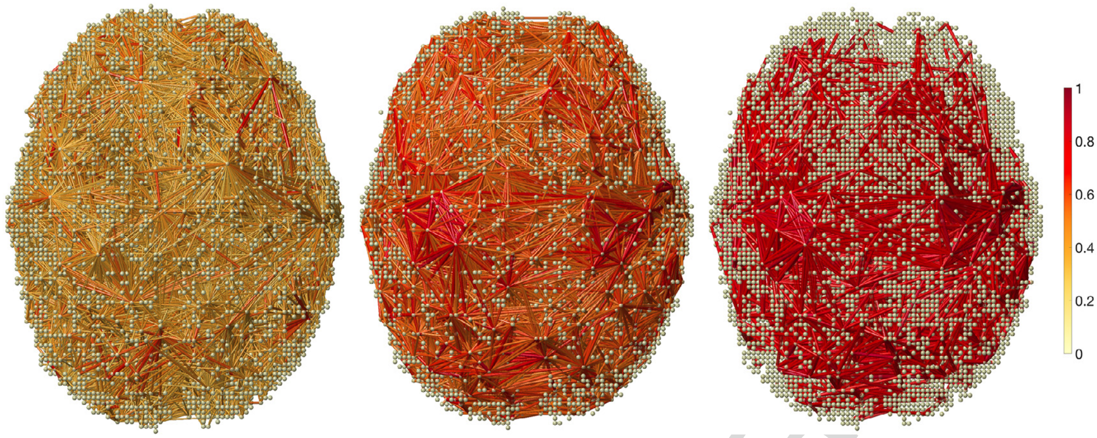

Fig. 1. Dense resting-state fMRI correlation network consisting of 25 000 nodes obtained from HCP. The network is so dense, simply displaying all the nodes and edges of the network is not very informative. It is necessary to represent such dense network more sparsely. The sparse correlation network model 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

Please cite this article in press as: Chung M.K., Statistical challenges of big brain network data. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.020.