Ph.D. Thesis Pre-Submission Outline : Characterizing and Leveraging Numerical Instabilities in Neuroimaging Through Perturbation Analysis

Gregory Kiar
September 18th, 2020

Overview

Thesis Outline 1

Summary of Original Contributions 4

Summary of Work Completed to Date 5

Summary of Remaining Work 5

List of Publications 6

Schedule 7

References 8

The objective of this document (and associated meeting) is to summarize the outline of my thesis, highlight and outline both my contributions to date as well as those in progress, and layout a plan for the remainder of my degree. Please do not hesitate to reach out if for any questions or clarifications.

Thesis Outline

The proposed thesis will consist of work spread across four chapters, spanning the creation of infrastructures for perturbation analysis, a method for the identification of instabilities in neuroimaging pipelines, the evaluation of the analytical impact of numerical instabilities within neuroimaging, and the possible dual application of the stability evaluation experiments for data augmentation. The thesis will follow a manuscript-based model, as described on McGill's thesis guidelines, in which each chapter is composed of the contents of a single manuscript and are surrounded by a discussion and background section which situates the chapters with respect to one another. The following section outlines each of the core chapters with the titles and abstracts for the associated papers (for those completed) or outlines.

Chapter 1: A Serverless Tool for Platform Agnostic Computational Experiment Management (Frontiers in Neuroinformatics)

Neuroscience has been carried into the domain of big data and high performance computing (HPC) on the backs of initiatives in data collection and increasingly compute-intensive tools. While managing HPC experiments requires considerable technical acumen, platforms, and standards have been developed to ease this burden on scientists. While web-portals make resources widely accessible, data organizations such as the Brain Imaging Data Structure and tool description languages such as Boutiques provide researchers with a foothold to tackle these problems using their own datasets, pipelines, and environments. While these standards lower the barrier to adoption of HPC and cloud systems for neuroscience applications, they still require the consolidation of disparate domain-specific knowledge. We present Clowdr, a lightweight tool to launch experiments on HPC systems and clouds, record rich execution records, and enable the accessible sharing and re-launch of experimental summaries and results. Clowdr uniquely sits between web platforms and bare-metal applications for experiment management by preserving the flexibility of do-it-yourself solutions while providing a low barrier for developing, deploying and disseminating neuroscientific analysis.

Chapter 2: Comparing perturbation models for evaluating stability of neuroimaging pipelines (The International Journal of High Performance Computing Applications)

With an increase in awareness regarding a troubling lack of reproducibility in analytical software tools, the degree of validity in scientific derivatives and their downstream results has become unclear. The nature of reproducibility issues may vary across domains, tools, data sets, and computational infrastructures, but numerical instabilities are thought to be a core contributor. In neuroimaging, unexpected deviations have been observed when varying operating systems, software implementations, or adding negligible quantities of noise. In the field of numerical analysis, these issues have recently been explored through Monte Carlo Arithmetic, a method involving the instrumentation of floating-point operations with probabilistic noise injections at a target precision. Exploring multiple simulations in this context

allows the characterization of the result space for a given tool or operation. In this article, we compare various perturbation models to introduce instabilities within a typical neuroimaging pipeline, including (i) targeted noise, (ii) Monte Carlo Arithmetic, and (iii) operating system variation, to identify the significance and quality of their impact on the resulting derivatives. We demonstrate that even low-order models in neuroimaging such as the structural connectome estimation pipeline evaluated here are sensitive to numerical instabilities, suggesting that stability is a relevant axis upon which tools are compared, alongside more traditional criteria such as biological feasibility, computational efficiency, or, when possible, accuracy. Heterogeneity was observed across participants which clearly illustrates a strong interaction between the tool and data set being processed, requiring that the stability of a given tool be evaluated with respect to a given cohort. We identify use cases for each perturbation method tested, including quality assurance, pipeline error detection, and local sensitivity analysis, and make recommendations for the evaluation of stability in a practical and analytically focused setting. Identifying how these relationships and recommendations scale to higher order computational tools, distinct data sets, and their implication on biological feasibility remain exciting avenues for future work.

Chapter 3: Numerical Instabilities in Analytical Pipelines Compromise the Reliability of Network Neuroscience (preprint, to be submitted to Nature)

The analysis of brain-imaging data requires complex and often non-linear transformations to support findings on brain function or pathologies. And yet, recent work has shown that variability in the choices that one makes when analyzing data can lead to quantitatively and qualitatively different results, endangering the trust in conclusions [1]–[3]. Even within a given method or analytical technique, numerical instabilities could compromise findings [4]–[7]. We instrumented a structural-connectome estimation pipeline with Monte Carlo Arithmetic [8], [9], a technique to introduce random noise in floating-point computations, and evaluated the stability of the derived connectomes, their features [10], [11], and the impact on a downstream analysis [12], [13]. The stability of results was found to be highly dependent upon which features of the connectomes were evaluated, and ranged from perfectly stable (i.e. no observed variability across executions) to highly unstable (i.e. the results contained no trustworthy significant information). The extreme range and variability in results presented here could severely hamper our understanding of brain function in brain-imaging studies. However, it also highlights potential paths forward, such leveraging this variance to reduce bias in estimates of brain connectivity. This paper demonstrates that stability evaluations are necessary as a core component of typical analytical workflows.

Chapter 4: Augmenting Connectomics Datasets Through Aggregation of Numerically Unstable Derivatives (targeted for IEEE Transactions in Medical Imaging)

This manuscript will explore the aggregation of MCA-perturbed outputs towards improving the performance of classifiers seeking to link brain structure and phenotypic data. Using the dataset generated for the manuscript presented in Chapter 3, individuals will be classified into "high" or

"low" fitness categories based on thresholds of Body Mass Index, Cholesterol, and VO2max. Several classifiers (Logistic Regression, Support Vector Machine, and Random Forest) will be tasked with learning a relationship between connectivity and these classifications using a) jackknife resampling of the perturbed dataset, the edgewise b) mean and c) median within individuals, as well as d) a distance dependent consensus average. These approaches will also be compared to e) a meta-analytic approach in which the jackknife-trained classifiers are aggregated, and a f) mega-analytic approach in which all perturbed outputs are used to train the classifiers. The effect of these 6 forms of aggregation enabled through the MCA-perturbed execution of the pipelines will be compared to the benefits afforded through a dataset with repeated measurements. The consistency of these results will be studied across the definition of class membership (i.e. the measure of fitness). The core contribution of this work will be to leverage the induced variability to reduce the bias in the downstream construction of brain-behaviour relationships, ultimately leading to more generalizable machine learning models and scientific claims. Additionally, the aggregation of MCA-perturbed results may present an alternative to the expensive repeated cross-sectional measurements of individuals, having significant implications on the design of large data collection consortia. A bullet point form outline for this paper, including the specific experiment definitions and comparisons can be found here.

Summary of Original Contributions

Throughout my Ph.D. I have developed and contributed to several tools which increase the accessibility of deploying and evaluating neuroimaging pipelines at scale. I have demonstrated the utility of these tools to explore and characterize the stability of neuroimaging pipelines. Below, I summarize original contributions in each of these areas.

Software Contributions

Extending the Boutiques command-line descriptive framework [14], for which I am a co-maintainer, I developed Clowdr [15] to enable the rapid deployment of scientific workflows across cloud and cluster resources. This tool is publicly available and has effectively been used on Compute Canada, XSEDE, and Dell EMC resources, orchestrating decades of compute cycles over the resources in a matter of hours. As workflows in neuroimaging often rely on prebuilt and containerized dependencies through Docker or Singularity, I created Fuzzy (https://github.com/gkiar/fuzzy), a curated collection of scientific libraries which were recompiled and instrumented to allow the stability evaluation of the contained tools. These environments use Verificarlo [9] to instrument libraries with Monte Carlo Arithmetic. The precompiled libraries include: Python, Cython, BLAS, LAPACK, Libmath. The efficacy of the perturbations induced through these tools has been demonstrated for neuroimaging applications through several experiments mentioned in the following paragraph.

Scientific Contributions

The software contributions above were developed out of necessity for the exploration of the stability of neuroimaging analyses. First, I created and demonstrated the Fuzzy environments as an effective method for inducing instabilities in neuroimaging pipelines, and situated these instabilities with respect to other forms of perturbation [4]. In this paper, I demonstrated the considerable variability present in a structural connectome estimation pipeline solely due to numerical uncertainty. I applied this technique to study the stability of a set of typical network neuroscience experiments and characterized the effect of instabilities on each of a test-retest, network topology, and phenotypic classification setting (preprint attached). This work illustrates the significant impact that numerical instabilities play in all levels of downstream analysis, spanning the reliability of comparisons across subjects, to the lack of reliability in individual network features, and ultimately the modelling of relationships between connectivity and phenotypic information (in this case, Body Mass Index).The final chapter of my thesis will include the evaluation of various aggregation methods in a classification setting, and will compare the change in performance induced through MCA executions to that of repeated measurements. The findings of this paper will potentially have large implications on data collection and may provide an alternative to the collection of repeated measurements in neuroimaging experiments.

Summary of Work Completed to Date

Each of the first two chapters of my thesis, as outlined in the previous sections, have been completed and published. The titles and venues for these manuscripts are in a following section. The third chapter of my thesis has been completed and the manuscript will be submitted for publication in Nature shortly.

The fourth and final chapter has been outlined in depth, including implementation-level design for all experiments. Additionally, the datasets which will be analyzed in this chapter have been generated, and target physiological measurements have been selected. A working prototype of the classifiers employed in this analysis has already been developed and tested as a part of the manuscript created for the third chapter.

Summary of Remaining Work

The remaining work for the completion of my Ph.D. thesis, as outlined above, is the final implementation and analysis of the manuscript outlined for Chapter 4, along with an estimated duration for each task. Specifically, this includes the following:

- constructing desired set of classifiers in Python (~1 day),
- constructing routines for each aggregation method (~1 week),
- small-scale testing and validation of the experimental setup (~1 week),
- executing and evaluating each classification task and aggregation method (~1 week),
- summarizing and visualizing the results (~1 week), and
- writing the paper (~1 month).
- **Total estimated time:** 8 weeks

In total, I anticipate the remaining work taking approximately two months of my attention. I believe this timeline is generous given both the fact that the dataset is ready for analysis and that the prototypical version of the described experiments was constructed in approximately 2 days. Given this timeline, my thesis work should be completed by December 1st, 2020.

List of Publications

In the following lists of publications, **blue** refers to submitted or submission-ready work, and **orange** indicates that the project is in progress. **Black** indicates published work.

First Author

Ch. 1. A Serverless Tool for Platform Agnostic Computational Experiment Management. G. Kiar, S. T. Brown, T. Glatard, A. C. Evans. Frontiers in Neuroinformatics 13 (Mar. 2019).

Ch. 2. Comparing perturbation models for evaluating stability of neuroimaging pipelines. G. Kiar, P. de Oliveira Castro, P. Rioux, E. Petit, S. T. Brown, A. C. Evans, T. Glatard. The International Journal of High Performance Computing Applications (Mar. 2020).

Ch. 3. Numerical Instabilities in Analytical Pipelines Compromise the Reliability of Network Neuroscience. G. Kiar, Y. Chatelain, P. de Oliveira Castro, E. Petit, A. Rokem, G. Varoquaux, B. Misic, T. Glatard*, A. C. Evans*. Work in progress, for Nature (Sept. 2020).

Ch. 4. Augmenting Connectomics Datasets Through Aggregation of Numerically Unstable Outcomes. G. Kiar, A. C. Evans, T. Glatard. Work in progress, for IEEE Transactions in Medical Imaging (Nov. 2020).

Contributing Author

1. Membrane Protein Classification: a Reproducibility Study. H. Heidarzadeh, **G. Kiar**, T. Glatard. Work in progress, for PLoS CB. (2020).
2. An epidemic spreading model to simulate A spread in familial Alzheimer's disease. E. Levitis, ..., **G. Kiar**, ..., A. C. Evans. Work in progress, for JAMA Neurology. (2020).
3. Exploring the Relationship Between Early Psychosis Verbal Memory Deficits and White Matter Integrity. C. Henri-Bellemare, ..., **G. Kiar**, ..., M. Lepage. Submitted, Schiz. (2020).
4. File-based localization of numerical perturbations in data analysis pipelines. A. Salari, **G. Kiar**, ..., T. Glatard. GigaScience. (2020).
5. Neural correlates of polygenic risk score for autism spectrum disorders in the general population. B. Khundrakpam, ..., **G. Kiar**, ..., A. C. Evans. Brain Communications (2020).
6. qEEG toolbox for the MNI Neuroinformatics ecosystem: normative SPM of EEG source spectra. J. Bosch-Bayard, ..., **G. Kiar**, ..., P. Valdes-Sosa. Front. in Neuroinf. (2020).
7. Deploying Large Fixed File Datasets with SquashFS and Singularity. P. Rioux, **G. Kiar**, ..., S. T. Brown. PEARC '20, Association for Computing Machinery. (2020).
8. Brain status modeling with non-negative projective dictionary learning. M. Zhang, ..., **G. Kiar**, ..., A. C. Evans. Neuroimage. (Oct. 2019).
9. PyBIDS: Python tools for BIDS datasets. T. Yarkoni, ..., **G. Kiar**, ..., R. Blair. JOSS. (2019).
10. Boutiques: a flexible framework to integrate command-line applications in computing platforms. T. Glatard, **G. Kiar**, ..., A. C. Evans. GigaScience 7.5. (2018).

Schedule

Degree years	Milestone
0 – 0.5	Ph.D: Initial Meeting Ch. 1: Clowdr started
0.5 – 1	Ch. 1: Boutiques published Ch. 1: Clowdr submitted
1 – 1.5	Ph.D: Comprehensive Exam Ch. 1: Clowdr published Ch. 2: Perturbation Methods started
1.5 – 2	Ch. 2: Perturbation Methods awarded talk and submitted Ch. 3: Analytical Impact started Ch. 4: Aggregation started
2 – 2.5	Ch. 2: Perturbation Methods published Ch. 3: Analytical Impact started
2.5 – 3	Ch. 3: Analytical Impact written (soon to be submitted to Nature) Ch. 4: Aggregation started Ph.D: Thesis Pre-Submission
current date	all following milestones and dates are "planned"
3 – 3.5	Ch. 3: Analytical Impact published Ch. 4: Aggregation submitted Ph.D: Thesis written & submitted (December – February)
3.5 – 4	Ch. 4: Aggregation published Ph.D: Thesis Defended (February – April)

References

[1] R. Botvinik-Nezer *et al.*, “Variability in the analysis of a single neuroimaging dataset by many teams,” *Nature*, vol. 582, no. 7810, pp. 84–88, Jun. 2020.

[2] C. M. Bennett, M. B. Miller, and G. L. Wolford, “Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction,” *Neuroimage*, vol. 47, no. Suppl 1, p. S125, 2009.

[3] A. Eklund, T. E. Nichols, and H. Knutsson, “Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 28, pp. 7900–7905, Jul. 2016.

[4] G. Kiar *et al.*, “Comparing perturbation models for evaluating stability of neuroimaging pipelines,” *The International Journal of High Performance Computing Applications*. p. 109434202092623, 2020, doi: 10.1177/1094342020926237.

[5] A. Salari, G. Kiar, L. Lewis, A. C. Evans, and T. Glatard, “File-based localization of numerical perturbations in data analysis pipelines,” *arXiv [q-bio.QM]*, Jun. 03, 2020.

[6] L. B. Lewis *et al.*, “Robustness and reliability of cortical surface reconstruction in CIVET and FreeSurfer,” *Annual Meeting of the Organization for Human Brain Mapping*, 2017, [Online]. Available: http://www.bic.mni.mcgill.ca/users/llewis/CIVET_vs_FS_HBM2017_final.pdf.

[7] T. Glatard *et al.*, “Reproducibility of neuroimaging analyses across operating systems,” *Front. Neuroinform.*, vol. 9, p. 12, Apr. 2015.

[8] D. S. Parker, *Monte Carlo Arithmetic: exploiting randomness in floating-point arithmetic*. University of California (Los Angeles). Computer Science Department, 1997.

[9] C. Denis, P. de Oliveira Castro, and E. Petit, “Verificarlo: Checking Floating Point Accuracy through Monte Carlo Arithmetic,” *2016 IEEE 23nd Symposium on Computer Arithmetic (ARITH)*. 2016, doi: 10.1109/arith.2016.31.

[10] M. Rubinov and O. Sporns, “Complex network measures of brain connectivity: uses and interpretations,” *Neuroimage*, vol. 52, no. 3, pp. 1059–1069, Sep. 2010.

[11] R. F. Betzel, A. Griffa, P. Hagmann, and B. Misic, “Distance-dependent consistency thresholds for generating group-representative structural brain networks,” *bioRxiv*, 2018, [Online]. Available: https://www.biorxiv.org/content/10.1101/412346v1.abstract.

[12] B.-Y. Park, J. Seo, J. Yi, and H. Park, “Structural and Functional Brain Connectivity of People with Obesity and Prediction of Body Mass Index Using Connectivity,” *PLoS One*, vol. 10, no. 11, p. e0141376, Nov. 2015.

[13] A. Gupta *et al.*, “Patterns of brain structural connectivity differentiate normal weight from overweight subjects,” *Neuroimage Clin*, vol. 7, pp. 506–517, Jan. 2015.

[14] T. Glatard *et al.*, “Boutiques: a flexible framework to integrate command-line applications in computing platforms,” *Gigascience*, vol. 7, no. 5, May 2018, doi: 10.1093/gigascience/giy016.

[15] G. Kiar, S. T. Brown, T. Glatard, and A. C. Evans, “A Serverless Tool for Platform Agnostic Computational Experiment Management,” *Front. Neuroinform.*, vol. 13, p. 12, Mar. 2019.