Comparing Perturbation Models for Evaluating Stability of Post-Processing Pipelines in Neuroimaging

Gregory Kiar[1], Pablo de Oliveira Castro[2], Pierre Rioux[1], Eric Petit[3],
Shawn T. Brown[1], Alan C. Evans[1], Tristan Glatard[4]
[1]McGill University, Montreal, Canada; [2]University of Versailles, Versailles, France;
[3]Exascale Computing Lab, Intel, Paris, France; [4]Concordia University, Montreal, Canada.

I. Introduction

A lack of software reproducibility [1] has become increasingly apparent in the last several years, calling into question the validity of scientific findings affected by published tools. Reproducibility issues may have numerous sources of error, including undocumented system or parameterization differences and the underlying numerical stability of algorithms and implementations employed. As neuroimaging has evolved into a computational field, it has suffered from the same questions of numerical reproducibility as many other domains [2].

Various forms of instability have been observed in neuroimaging, including across operating system versions [3], minor noise injections [4], and implementation of theoretically equivalent algorithms [5]. In this paper we explore the effect of various perturbation methods on a typical neuroimaging pipeline through the use of i) near-epsilon noise injections, ii) Monte Carlo Arithmetic (MCA), and iii) varying operating systems to identify the quality and severity of their impact. All code for performing the experiments and creating associated figures are available on GitHub at https://github.com/gkiar/stability and https://github.com/gkiar/stability-mca, respectively.

II. Methods

A structural connectome estimation pipeline, providing a measure of anatomical connectivity between brain regions derived from multiple imaging modalities, was developed using Dipy [6], FSL [7] and an 83-region cortical and subcortical parcellation [8]. The pre-processing (alignment, denoising, and segmentation) and modeling (tensor estimation, tractography) components of this pipeline were performed separately, and perturbations were only introduced for the modeling components which were developed in Python using both Numpy and Cython. This pipeline was used to process 10 participants from the Nathan Kline Institute – Rockland Sample dataset (NKI-RS) [9].

Near-epsilon and Monte Carlo perturbation modes were tested 100x per image. Noise was represented by percent deviation of the Frobenius norm of a resulting connectome from the corresponding reference (no noise injection). A deviation of 50% indicates that the norm of the difference between the noisy and reference networks is 50% the size of the norm of the reference graph. This is formalized below in Eq. (1):

$$\%Dev(A,B) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}-b_{ij}|^2} \Big/ \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}, \quad (1)$$

where A is the reference graph, B is the perturbed graph, and $\square_{ij}$ is an element therein at row i and column j.

The type of near-epsilon noise used here will be referred to as 1-voxel noise and is similar to the one employed in Lewis et al. [4], where "independent" and "single" refer to injecting noise into a single voxel per 3D or 4D volume, respectively. Monte Carlo simulations were introduced within the processing tools using Verificarlo [10], an extension of the LLVM compiler which automatically instruments floating point operations at build-time for software written in C, C++, and Fortran. Noise through Verificarlo can be injected as Precision Bounded, simulating floating point cancellations, "Random Rounding" (RR), simulating rounding errors on computation, and "MCA", which includes both of these modes.

III. Results

Introduced perturbation showed highly-variable changes in resulting connectomes across both the perturbation model and subject, ranging from no change to deviations equivalent to typically observed differences across subjects, as shown in Fig. 1. For the 10 subjects tested, we see that instrumenting only Python-based libraries with MCA resulted in the largest deviation from the reference connectome. In these cases we also see that the results are modal, where each subject has discrete states that may be settled in, some of which result in deviations comparable to subject-level noise. This modality is likely due to minor differences introduced at crucial branch-points which then cascaded throughout the pipeline. This hypothesis is supported by observing that the fully instrumented pipeline with RR implementation shows a continuous distribution of differences. The 1-voxel independent mode

This work was funded by the Natural Sciences and Engineering Research Council of Canada (CGSD3 - 519497 - 2018).

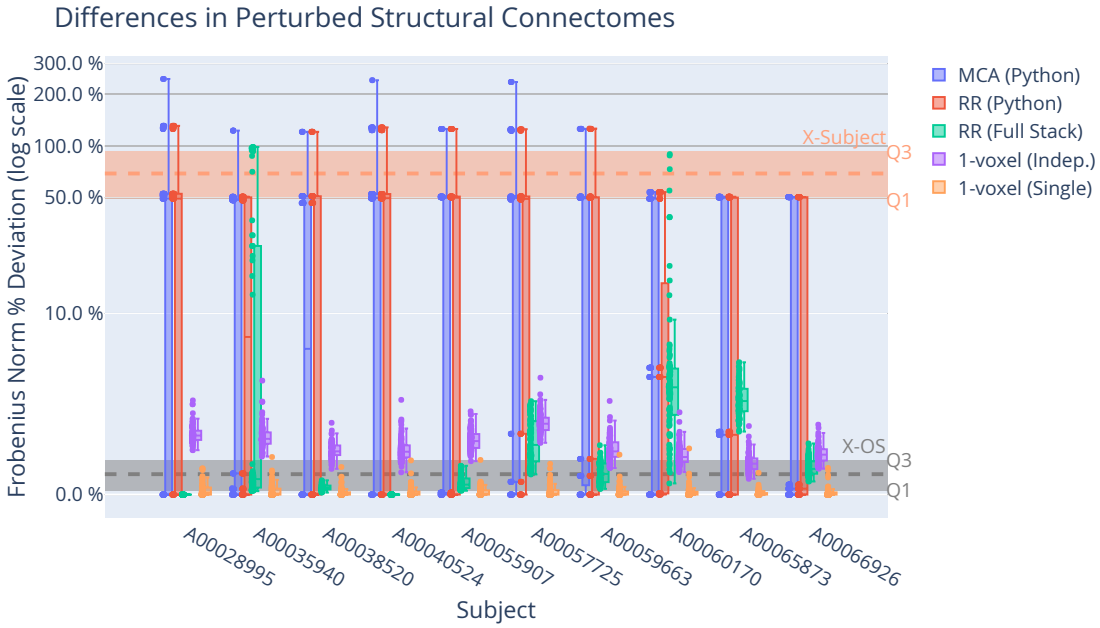

Fig. 1. Comparison of the relative impact of distinct perturbation modes for 10 subjects.

unsurprisingly produces larger changes than the 1-voxel single mode, but in both cases these differences are relatively minor in comparison to the extremes observed with Monte Carlo Arithmetic. Operating system deviations are very low or even zero in some cases.

IV. Discussion & Conclusion

We have demonstrated through the application of multiple perturbation methods how noise can be effectively injected into neuroimaging pipelines enabling the exploration and evaluation of the stability of resulting derivatives. These methods operate either by perturbing the datasets and tools used in processing, resulting in a range of structurally distinct noise profiles and distributions which each may provide value when exploring the stability of analyses. While 1-voxel noise is injected directly into the datasets prior to analysis, MCA and RR methods iteratively add significantly smaller amounts of noise to each operation performed.

The structure of the introduced deviations leads to several applications of perturbation models in neuroimaging. While the modes observed in several cases can serve as the basis for generating synthetic datasets, the high across-subject variability in perturbation-induced noise seen with the Full Stack RR setting suggests a possible application in quality assurance. Due to the highly controllable nature of 1-voxel perturbations, these could be used to test either the regional sensitivity of a particular method or the global stability to small quantifiable variations (i.e., conditioning).

The work presented here demonstrates that even low order computational models such as the connectome estimation pipeline that we used are susceptible to noise. This suggests that stability is a relevant axis upon which tools should be compared, developed, or improved, alongside more commonly considered axes such as accuracy/biological feasibility or performance. The heterogeneity observed across participants clearly illustrates that stability is a property of not just the data or tools independently, but their interaction. Characterization of stability should therefore be evaluated for specific analyses and performed on a representative set of subjects for consideration in subsequent statistical testing. Additionally, identifying how this relationship scales to higher-order models is an exciting next step which will be explored. Finally, the joint application of perturbation methods with post-processing approaches such as bagging or signal normalization may lead to the development of more numerically stable analyses while maintaining sensitivity to meaningful variation.

References

[1] R. D. Peng, "Reproducible research in computational science," *Science*, vol. 334, no. 6060, pp. 1226–1227, Dec. 2011.

[2] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, no. 7604, pp. 452–454, May 2016.

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

[4] 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.

[5] A. Bowring, C. Maumet, and T. E. Nichols, "Exploring the impact of analysis software on task fMRI results," *bioRXiv*, Mar. 2018.

[6] E. Garyfallidis *et al.*, "Dipy, a library for the analysis of diffusion MRI data," *Front. Neuroinform.*, vol. 8, p. 8, Feb. 2014.

[7] M. Jenkinson *et al.*, "FSL," *Neuroimage*, vol. 62, no. 2, pp. 782–790, Aug. 2012.

[8] L. Cammoun *et al.*, "Mapping the human connectome at multiple scales with diffusion spectrum MRI," *J. Neurosci. Methods*, vol. 203, no. 2, pp. 386–397, Jan. 2012.

[9] K. B. Nooner *et al.*, "The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry," *Front. Neurosci.*, vol. 6, p. 152, Oct. 2012.

[10] C. Denis, P. de Oliveira Castro, and E. Petit, "Verificarlo: Checking floating point accuracy through monte carlo arithmetic," 2016.