Optimal Design for Discovery Science: Applications in Neuroimaging

Shangsi Wang[1], Zhi Yang[2], Xi-Nian Zuo[2], Michael Milham[3], Cameron Craddock[3], Carey E. Priebe[1], Joshua T. Vogelstein[3,4]
[1]Department of Applied Mathematics and Statistics, Johns Hopkins Univesity, Baltimore, Maryland [2]Insititute of Psychology, Chinese Academy of Sciences, Beijing, China
[3]Child Mind Institute, New York, New York [4]Department of Biomedical Engineering, Johns Hopkins Univesity, Baltimore, Maryland
Contact: Shangsi Wang: swang127@jhu.edu, Joshua T. Vogelstein: jovo@jhu.edu

Opportunity

- When collecting data for multiple potential uses, it is unclear both how to optimally (i) collect and (ii) process the data.
- Lots of entities are collecting and making data publicly available[1-4] for processing for multiple inference tasks.

Challenge

- Experimental design and generalizability theory make strong distributional assumptions.
- Processed data need to achieve optimal performance for diversified subsequent inference tasks.

Action

- We proposed a definition of reliability and designed an estimator to compute reliability.
- We proved that reliability bounds prediction error for all subsequent inference tasks.
- We demonstrate through simulation and real data experiments that maximizing reliability leads to optimal processing.

Conclusion

- We demonstrate utility of maximizing reliability in data processing and analysis for subsequent inference..
- We prove reliability bounds predictive accuracy.
- Open source implementation is available.

Opportunity

In many problems arising in the data science, data collection and processing is the first step toward statistical inference. However, these crucial beginning steps are sometimes done in an arbitrary or subjective fashion which lacks rigorous guidance and theoretical justification. This poses an important question to researchers:

- How to collect and process data for subsequent inference

Optimally addressing this question may dramatically improve inference performance and also reduce financial cost.

Challenge

Modern data comes in large volume and complex form which makes processing data computationally and financially expensive. Therefore, there is an urgent need for a unified framework which enable data to be processes optimally for a variety of inference task.

Traditional generalizability theory[5] tries to conceptualize, investigate and design reliable observations. Attempts to quantify reliability include intraclass correlation[6] and image intraclass correlation[7]. However, they makes strong distributional assumptions which are not suitable for high dimensional data.Motivated by this, we propose a new definition of reliability. By processing to maximize reliability of data, subsequent inference performance should beoptimal.

Action

Theory

- Reliability is defined to be the probability that the distance between two measurements of a fixed subject is smaller than distance between measurement from an another subject and the measurement from the fixed subject. Let O_{ij} denote the jth measure from subject i, then we define reliability as

$$R := P(\|O_{ij} - O_{pq}\| < \|O_{ij} - O_{ik}\|)$$

- Although true population reliability is unknown, we may compute sample reliability Rhat through formulas below:

$$\hat{R}_{ijk} := \frac{\sum_{p=1,p\neq i}^{n} \sum_{q=1}^{s} I\{\|O_{ij}-O_{pq}\|<\|O_{ij}-O_{ik}\|\}}{(n-1)s}$$

$$\hat{R} := \frac{\sum_{i=1}^{n} \sum_{j=1}^{s} \sum_{k=1,k\neq j}^{s} \hat{R}_{ijk}}{ns(s-1)}$$

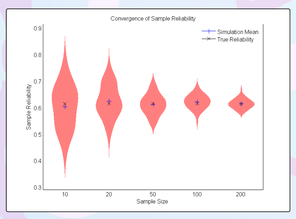
- The following two lemmas show that sample reliability provides unbiased and consistent estimate population reliability.

Lemma 1:

$$E(\hat{R}) = R$$

Lemma 2:

$$\hat{R} \to_p R$$



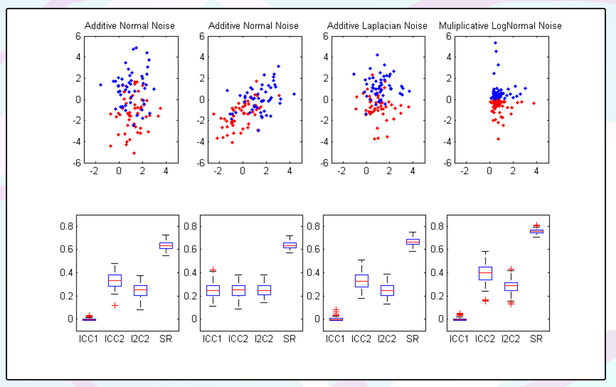
- If we assume each observation can be represented by the sum of true measure and an independent noise and we also observe a class label Y_i associated with each subject, under some regularity conditions we have the following theorem.

Theorem 1: There is a decreasing function g() upper bound L,

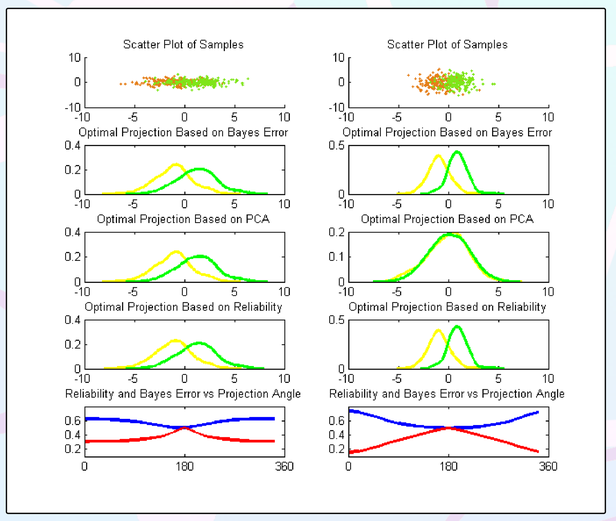
$$L \leq g(E(\hat{R}))$$

The theorem implies that optimizing reliability also maximizes a bound on performance of any inference task.

Simulation Experiment

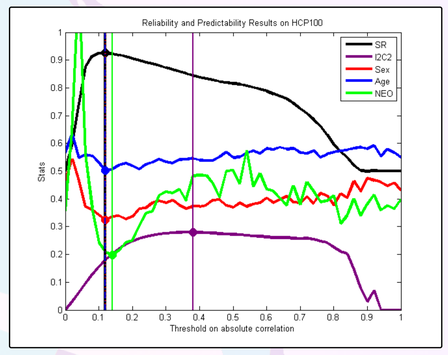


Experiment 1: Reliability of four different generated data set is computed. Bayes prediction errors of four data set are 0.2395, 0.2395, 0.1839 and 0 respectively. Intraclass correlation and Image intraclass correlation are also listed for reference. Reliability respects ordering of Bayes error, whereas the other statistics do not.

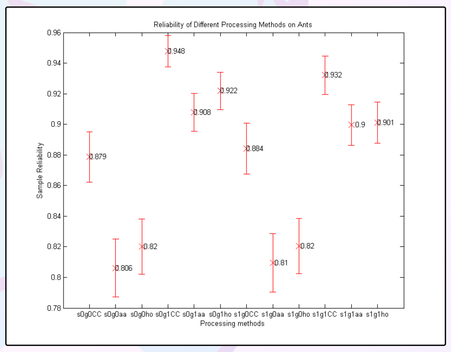


Experiment 2: 2D data is projected to 1D linearly. Reliability and PCA are used to select optimal projections. Reliability selects the linear projection which is Bayes optimal. The bottom plot shows the reliability (blue) and Bayes error (red) of different projections.

Real Data Experiments



Experiment 3: Reliability of fMRI processing is investigated. During processing fMRI data, correlations between different regions of brain can be thresholded to estimate human brain graphs. We investigate optimal value to threshold graphs based on reliability, gender prediction and age prediction on HCP100 data set. The result shows that when reliability is maximized, prediction errors are also close to its minimum.



Experiment 4: Reliability of several processing methods are compared. Scrubbing0_global1_CC200 yields most reliable data.

Experiment 5: Reliability of 7 different data sets are computed and compared. Among 7 data sets, NKI131 is the most reliable data set.

Data set	Reliability	Standard Error
NKI131	0.9849	0.0034
NKI165	0.9785	0.0044
HCP25	0.8607	0.0030
HCP50	0.8923	0.0029
HCP100	0.9247	0.0029
HCP200	0.9157	0.0026
HCP300	0.9056	0.0027

Conclusion

We demonstrate reliability can guide processing to yield data with optimal inference performance through simulation and real data experiments. Currently, we investigate only a few steps of processing fMRI image. In the future, we may apply our methodology to:

- The whole fMRI automatic pipeline
- Other methods to process fMRI data yielding different data structure
- Other neuroimaging data, for example, dMRI
- Non-neuroimaging data with multiple measurements per subject

In terms of theory, we show sample reliability provides consistent estimate of reliability which bounds the Bayes error of prediction. Our current results utilize Euclidean distance. In the future, we may try to prove results with other general distance under less restrictive setting. Nevertheless, our primary purpose is to demonstrate the utility of reliability in processing and the relationship between reliability and prediction.

References

[1]: Zuo X.N., et al., In Press
[2]: Milham M., Nathan Cline Institute-RS, CMI, 2012.
[3]: Cameron C., et al., Frontiers in Neuroinformatics, 2015.
[4]: Van Essen D.C. et al., NeuroImage, 2012.
[5]: Brennan, R. L., Springer, 2012.
[6]: Coch, Gary G., John Wiley & Sons, 1982.
[7]: Shou, H., et al, Cogn Affect Behav Neurosci, 2014.

Acknowledgements

This work was partially supported by the following awards: DARPA N66001-14-1-4028 (GRAPHS) and the Child Mind Institute's Endeavor Scholars program.