NeuroData SIMPLEX Report: February 2017

The following report documents the progress made by the labs of PI Joshua T. Vogelstein and Co-PIs Randal Burns and Carey Priebe at Johns Hopkins University towards goals set by the DARPA SIMPLEX grant.

Contents

1 Data: What's in the Cloud

2 Statistical Theory and Methods

2.1 LOL @jovo

LOL took a backseat this month while jovo had a baby :)

2.2 meda @JesseLP

Updates in meda include the addition of plots that explore the clusters generated by hierarchical clustering. We are using mclust [1] in our hierarchical clustering function. At each level we use the Bayseian Information Criterion (BIC) to determine if the data should be split into two clusters or kept as one. The dendrogram (figure 1 left) shows the binary tree clustering structure with branch size denoting the size of the cluster. The stacked level means plot (figure 1 right) shows the means of features in each node of the tree.

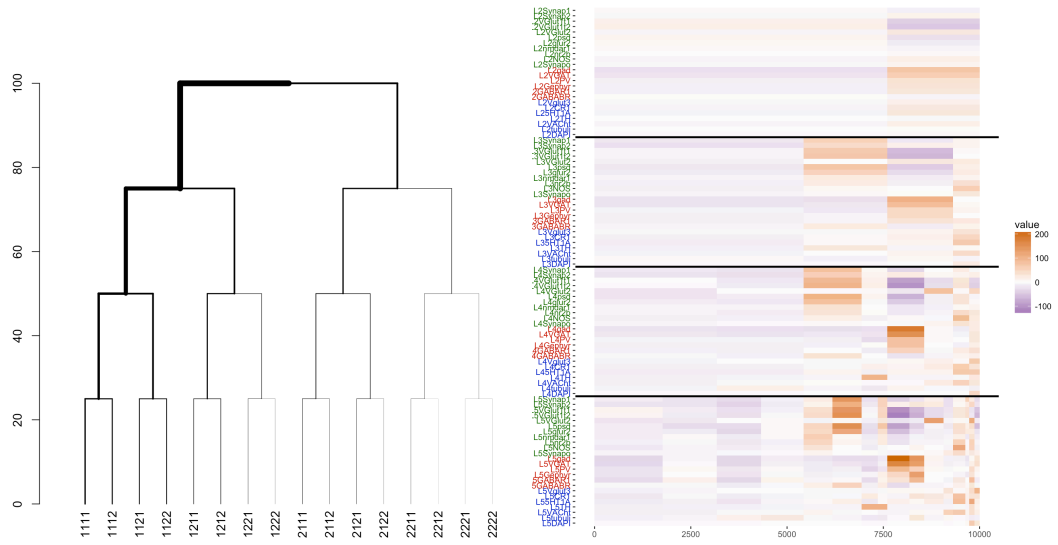

Figure 1: Left: A dendrogram showing the results of hierarchical mclust. The splits are constrained to be binary and branch sizes show relative cluster sizes. Right: A stacked level means plot showing for each node in the dendrogram the feature means.

2.3 Multiscale Generalized Correlation (MGC)

Your content here. Please make sure that it fits on one page.

Figure 2: Please provide a detailed caption for your figure.

2.4 RerF

Your content here. Please make sure that it fits on one page.
RerF is now working in pure R code. The running time of this implementation is roughly the same as the Matlab implementation for various sizes of input. We are re-implementing portions of the code in both C and FlashR in an attempt to reduce the running time of the algorithm.

Figure 3: Please provide a detailed caption for your figure.

2.5 Discriminability

Your content here. Please make sure that it fits on one page.

Figure 4: Please provide a detailed caption for your figure.

2.6 Law of Large Graphs

We showed that our estimates are better in terms of the MSE for the real data. To further illustrate the differences between the two estimators, now we examine the actual estimates to see how they look like. In order to emphasize the connectome smoothing effect, here we consider the same random sample of size $M = 1$ based on the Desikan atlas in Fig. 5 and plot the estimate $\hat{P}$ in the right panel. Note that compared to the sample mean $\bar{A}$, $\hat{P}$ has a finer gradient of values which in this case leads to a more accurate estimate of the true probability matrix P, especially for edges between the two hemispheres, in the upper right and corresponding lower left block.

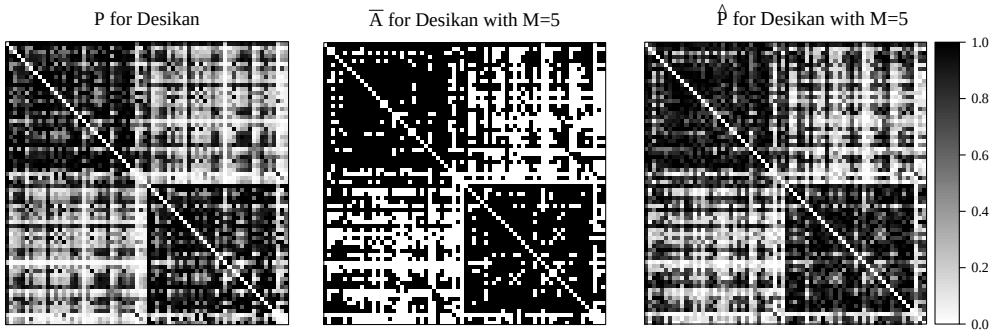

Figure 5: **Heat maps of the population mean, the sample mean, and the estimator $\hat{P}$.** These heat maps indicate the population mean for the 454 graphs (left), sample mean for the 1 sampled graph (center), and $\hat{P}$ for the 1 sampled graph with dimension $d = 12$ selected using the Zhu and Ghodsi method (right). Darker pixels indicate a higher probability of an edge between the given vertices. Note that $\hat{P}$ appears to better estimate the true probability matrix P, especially for edges between the two hemispheres, in the upper right and corresponding lower left block.

2.7 Robust Law of Large Graphs

We had theorems for the MLqE under the exponential distribution. Actually the results can be generalized to a broader class of distribution families, and even a different entry-wise robust estimator other than MLqE with the following conditions:

1. Let $A_{ij} \overset{ind}{\sim} (1-\epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}}$, then $E[(A_{ij} - E[\hat{P}_{ij}^{(1)}])^k] \leq \text{const} \cdot k!$, where $\hat{P}^{(1)}$ is the entry-wise MLE as defined before; This is to ensure the that observations will not deviate from the expectation too far away, such that the concentration inequality can apply.
2. There exists $C_0(P_{ij}, \epsilon) > 0$ such that under the contaminated model with $C > C_0(P_{ij}, \epsilon)$,

$$\lim_{m\to\infty} \left| E[\hat{P}_{ij}] - P_{ij} \right| < \lim_{m\to\infty} \left| E[\hat{P}_{ij}^{(1)}] - P_{ij} \right|;$$

   It requires the contamination of the model to be large enough (a restriction on the distribution) and $\hat{P}$ to be robust enough with respect to the contamination (a condition on the estimator).
3. $\hat{P}_{ij} \leq \text{const} \cdot \hat{P}_{ij}^{(1)}$; (This might be generalized to with high probability later)
   Since we use the results of $\hat{P}^{(1)}$ to bound $\hat{P}^{(q)}$, the proof can apply directly with this condition for an arbitrary $\hat{P}$.
4. $\text{Var}(\hat{P}_{ij}) = O(m^{-1})$, where m is the number of observations.
   We will get exactly the same results under this condition. However, even if the variance of the new estimator is not of order $O(m^{-1})$, we will get similar results with a different term related to m.

2.8 Nonparametric Network Dependence Test

Your content here. Please make sure that it fits on one page.

Figure 6: Please provide a detailed caption for your figure.

2.9 Batch effect removal in dimension reduction of multiway array data

Your content here. Please make sure that it fits on one page.

Figure 7: Please provide a detailed caption for your figure.

2.10 Reduced Dimension Clustering

Your content here. Please make sure that it fits on one page.

Figure 8: Please provide a detailed caption for your figure.

Table 1: The runtime and memory consumption of FlashR on the billion-scale datasets on the 48 CPU core machine. The runtime of iterative algorithms is measured when the algorithms converge. We run PageRank on the PageGraph dataset, run k-means on PageGraph-32ev and the remaining algorithms on Criteo.

	Runtime (s)	Memory (GB)
Correlation	91.23	1.5
PCA	136.71	1.5
NaiveBayes	76.55	3
LDA	2280	8
Logistic regression	4154.40	26
k-means	1110.82	28
PageRank	3900	135

3 Scalable Algorithm Implementations

3.1 FlashX

We use FlashR to process the billion-scale datasets to demonstrate its scalability (Table 1). We use three datasets here: (i) the Criteo dataset has over four billion data points with binary labels (click vs. no-click), used for advertisement click prediction; (ii) PageGraph is the adjacency matrix of a graph, which has 3.5 billion vertices and 128 billion edges; (iii) PageGraph-32ev are 32 singular vectors that we computed on the largest connected component of Pagegraph with the tools we built previously. In these experiments, we run the iterative algorithms (Logistic regression, k-means and PageRank) on the datasets until they converge.

Even though we process the billion-scale datasets in a single machine (with 48 CPU cores), none of the algorithms are prohibitively expensive. Simple algorithms, such as Naive Bayes and PCA, require one or two passes over the datasets and take only one or two minutes to complete. Logistic regression and k-means take about $10-20$ iterations to converge. Because the PageRank implementation uses the power method, it takes 100 iterations to converge. Nevertheless, all of the iterative algorithms take about one hour or less.

FlashR scales to datasets with billions of data points easily when running out of core. Most of the algorithms have negligible memory consumption. PageRank consumes more memory because the sparse matrix multiplication in PageRank keeps vectors in memory for semi-external memory computation. The scalability of FlashR is mainly bound by the capacity of SSDs. The functional programming interface generates a new matrix in each matrix operation, which potentially leads to high memory consumption. Thanks to lazy evaluation and virtual matrices, FlashR only needs to materialize the small matrices to effectively reduce memory consumption.

3.2 ndstore

Your content here. Please make sure that it fits on one page.

Figure 9: Please provide a detailed caption for your figure.

3.3 ndviz

Your content here. Please make sure that it fits on one page.

Figure 10: Please provide a detailed caption for your figure.

3.4 knor

Your content here. Please make sure that it fits on one page.

Figure 11: Please provide a detailed caption for your figure.

4 Scientific Pipelines: Infrastructure & Dataset Specific Progress

4.1 SIC

4.2 ndstore

4.3 ndmg

4.4 ndviz

4.5 MRI

4.6 CLARITY

4.7 Ophys

5 Bibliography

Manuscripts

[1] N. Peeps, "Paper with a cool title," 2017.

Invited Talks

[1] N. Peeps, "Talk with a cool title," SIAM, JSM, NIPS, ???, Jan 2017.

Conferences

[1] N. Peeps, "Poster with a cool title," SIAM, JSM, NIPS, ???, Jan 2017.

References

[1] C. Fraley and A. E. Raftery, “Model-based clustering, discriminant analysis, and density estimation,” Journal of the American statistical Association, vol. 97, no. 458, pp. 611–631, 2002.