NeuroData SIMPLEX Report: Q1 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 Statistical Theory and Methods

1.1 LOL @jovo

This month we finalized the real data analysis using LOL. In particular, we considered four datasets, the Prostate and Colon datasets have extensively been studied in the sparse literature. LOL yields better perforance, and for Colon, with lower dimensionality. MNIST is an even more prominent dataset, LOL achieves the best performance for all dimensions. MRN is a new dataset that we generated; it has over 500 million features, and 112 samples. We subsampled to 100 samples for cross-validation purposes. To our knowledge, no other machine learning tool is capable of even operating on 500 million features. Moreover, we demonstrate that our implementation outperforms first doing PCA on the data, for any number of dimensions that we embed into. We then also investigate the amount of time it takes to run LOL on very wide datasets. For a 128 million dimensional dataset, with 2000 samples, requiring nearly half a terabyte of space just to store, we have an approximate implementation that runs on a single machine and only takes about 3 minutes.

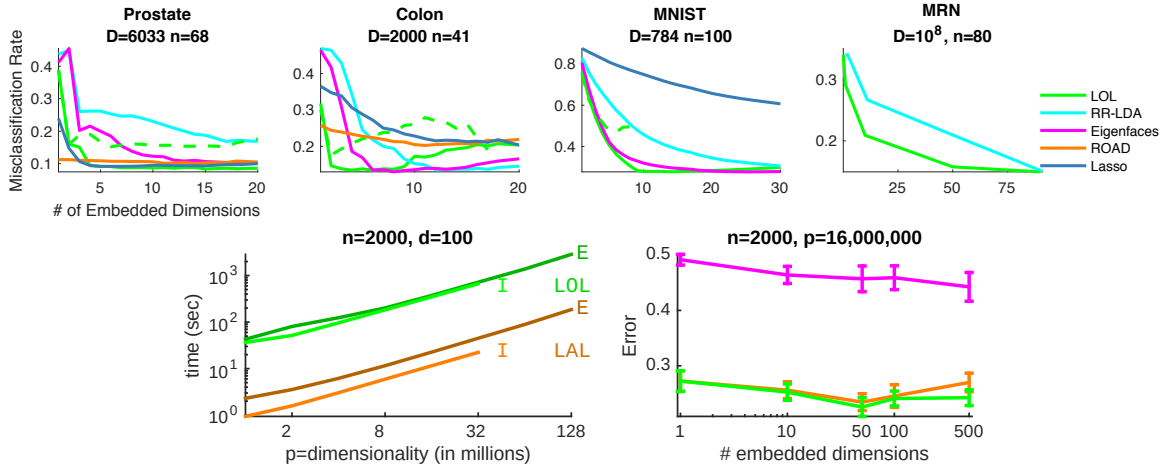

Figure 1: Top: LOL outperforms essentially all other methods for essentially all problems and number of dimensions to embed into. Bottom: LOL (green) and LAL (brown, using random projections to approximate) for both fully in memory (IE) and semi-external memory (E) implementations. Magenta compares performance to our scalable implementation of eigenfaces on the same problem.

1.2 meda @JesseLP

1.3 Multiscale Generalized Correlation (MGC)

We developed the Multiscale Generalized Correlation method to better detect associations between two datasets X and Y. We demonstrate that Oracle MGC is a consistent test statistic (power converge to 1 as sample size increases) under standard regularity conditions, is equivalently to the global correlation under linear dependency (i.e., each observation X_i is a linear transformation of Y_i), and can be strictly better than the global correlation under common nonlinear dependencies. Thus Oracle MGC dominates the global correlation, and the sample MGC (i.e., choose the optimal scale by p-value map approximation, as the testing power are not available in the absence of the true model and training data) also empirically dominates the global correlation. A flowchart to illustrate the advantage of MGC is shown in Figure 2.
Upon final draft polishing and addressing feedback from statisticians and biologists, the newest draft is updated to arXivand submitted for publication.

1.4 Network Dependence Test via Diffusion Maps and MGC

Deciphering the association between network structures and corresponding nodal attributes of interest is a core problem in network science. We propose a new nonparametric procedure for testing dependence between network topology and nodal attributes, via diffusion maps and `MGC`. Specifically, under an exchangeable graph, we verify that the diffusion maps provide a set of conditionally independent multivariate coordinates for the nodes, which can be combined with `MGC` (or in general, any distance-based correlation measures) to yield consistent statistic for network dependence testing. Moreover, our method is computationally inexpensive and robust against parameter mis-specifications, very efficient in capturing a wide variety of nonlinear and high-dimensional relationships, and readily extend-able to testing independence between two graphs.
Figure 3 illustrates the advantage of the proposed method on testing dependency between two graphs. The graphs are simulated by the random dot product graph, with the underlying latent variables being related by a quadratic function. By repeatedly generating dependent sample graphs, the testing power equals the percentage of rejection of the independence hypothesis. Although all methods are consistent (having power 1 as number of vertices increases), the proposed approach using `MGC` is able to achieve perfect testing power at a very small size, which is significantly better than other benchmarks.
An early draft is recently awarded the Best Student Paper Awards by the American Statistical Association Nonparametric Statistics Section, which will be presented in a special section in the Joint Statistical Meeting this year. We collected and addressed feedback from experts in graph inference, and submitted the complete manuscript this month.

1.5 Randomer Forest (RerF)

1.6 Non-Parametric Shape Clustering

1.7 Joint Embedding

1.8 Law of Large Graphs

1.9 Robust Law of Large Graphs

Although we only present the results under exponential distributions, 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 $E[(A_{ij} - E[\hat{P}_{ij}^{(1)}])^k] \leq \text{const}^k \cdot k!$, where $\hat{P}^{(1)}$ is the entry-wise MLE as defined before;
   This is to ensure that observations will not deviate from the expectation too far away, so that the concentration inequalities hold.
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. $\hat{P}_{ij} \leq \text{const} \cdot \hat{P}_{ij}^{(1)}$;
   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 based on this order. 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.

Previously we consider the model to be based on exponential distribution, which is continuous and monotone. Now we consider Poisson distribution instead. Poisson distribution is a commonly used distribution for nonnegative graphs with integer values. And we will prove that it satisfies the conditions for generalization and as a result all theories apply directly.
Let $A_{ij} \overset{ind}{\sim} (1-\epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}}$ with f to be Poisson, then we proved that $E[(A_{ij} - E[\hat{P}_{ij}^{(1)}])^k] \leq \text{const}^k \cdot k!$, where $\hat{P}^{(1)}$ is the entry-wise MLE as defined before. So Condition 1 is satisfied. Intuitively, since exponential distribution has a fatter tail compare to Poisson, we should have the bound for central moment of Poisson directly from the results for exponential distribution. Condition 2 is satisfied as long as the contamination is large enough while keep using the robust MLqE. For Condition 3, the extreme case happens when there are m data $x_1, \cdots, x_m$ with $0 \leq x_1 = \cdots = x_k \leq \bar{x} \leq x_{k+1} = \cdots = x_m \leq m\bar{x}/(m-k)$. In order to have ML$q$E larger than MLE $\bar{x}$, we need the weights of the first m data to be smaller than the weights of the rest $m-k$ data. So $e^{-\bar{x}} < \bar{x}^{x_m} e^{-\bar{x}}/x_m!$. Then $x_m! < \bar{x}^{x_m}$. By the lower bound in Stirling's formula, we have $x_m < e\bar{x}$ when $x_m > 0$. Note that if $x_m = 0$ then MLE equals MLqE since all data equals zero. Thus MLqE is bounded by $e\bar{x}$. As a result, $\hat{P}_{ij} \leq e\hat{P}_{ij}^{(1)}$ and Condition 3 is satisfied. At last, Condition 4 follows directly from theory of minimum contrast estimators.
So for all the theorems proved before, we can replace the exponential distribution by Poisson distribution and all the results still hold.

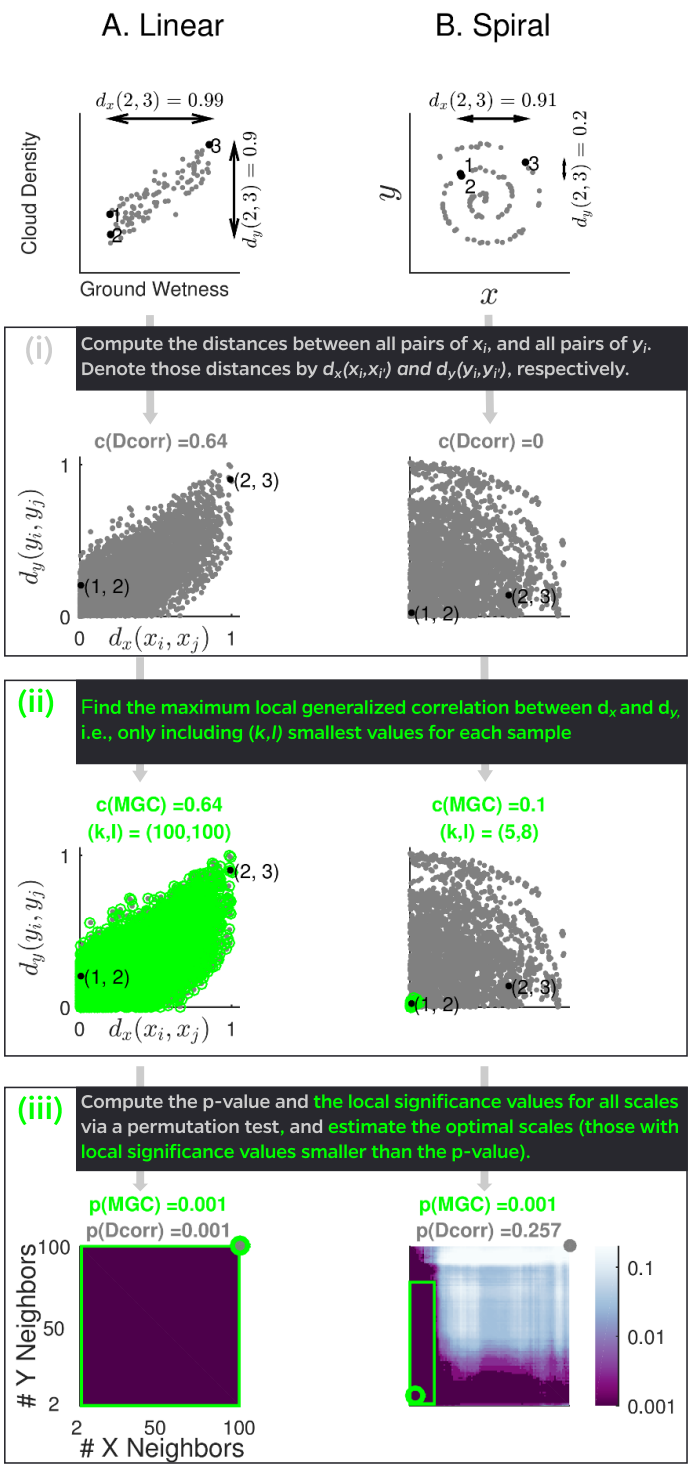

Figure 2: A flowchart to illustrate the advantages of MGC.

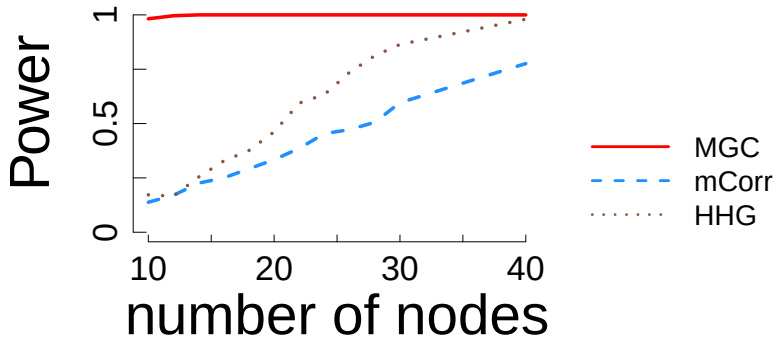

Figure 3: The power curve with respect to increasing number of vertices for the two-graph dependency testing simulation. The proposed approach quickly attains perfect power at a very small vertex size, while other benchmarks often require a much larger graph for perfect testing.