Numerical Evaluation of MADE in Ultra High Dimensional Poisson Regression

UMBC REU Site: Interdisciplinary Program in High Performance Computing
Ely Biggs[1], Tessa Helble[2], George Jeffreys[3], Amit Nayak[4],
Graduate assistant: Elias Al-Najjar[2], Faculty mentor: Kofi Adragni[2], Client: Andrew M. Raim[6]
[1]Wentworth Institute of Technology, [2]Math & Stat, UMBC, [3]Rutgers U., [4]GWU, [5]U.S Census Bureau

Master Address File (MAF)

The United States Census Bureau maintains a massive list of information about all habitable addresses in the U.S. called the Master Address File (MAF). Verifying all of the information in the MAF via a process called Address Canvassing (AdCan) was the second most expensive part of the 2010 census. Address canvassing involves recording all addresses that must to be added to the MAF (new habitable addresses) and all addresses that must be deleted (no longer habitable).

MADE

MADE (Minimum Average Deviance Estimation) is a statistical methodology being developed by [1] for the purpose of analyzing and making predictions using Poisson counts of deletions and additions to the MAF with the goal of using these predictions to reduce the cost of future AdCans.
The difficulty in developing a methodology is the size of the data set, as it can have hundreds of millions of observations and thousands of variables for each observation, thus having incredibly high dimensions. MADE embeds a sufficient dimension reduction procedure, along with a local linear regression [2].
A sufficient reduction of X is $B^T X$, $B \in R^{p\times d}, d < p$, so that $B^T X$ replaces X in the regression of y on X.

Project Work

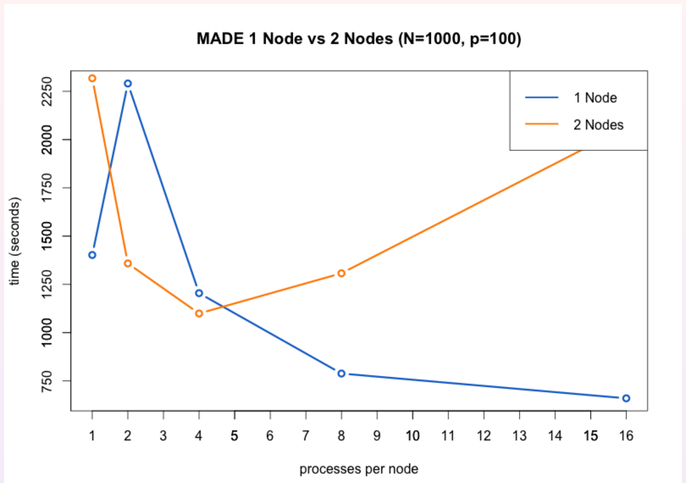
Our project had three main components:

▶ Analyze the existing code for MADE to determine what its limitations are in terms of speed and memory capacity.

▶ Parallelize the existing code for MADE so as to allow for faster computation on larger datasets.

▶ Provide recommendations for further improvements of the current code.

Results: Parallelization

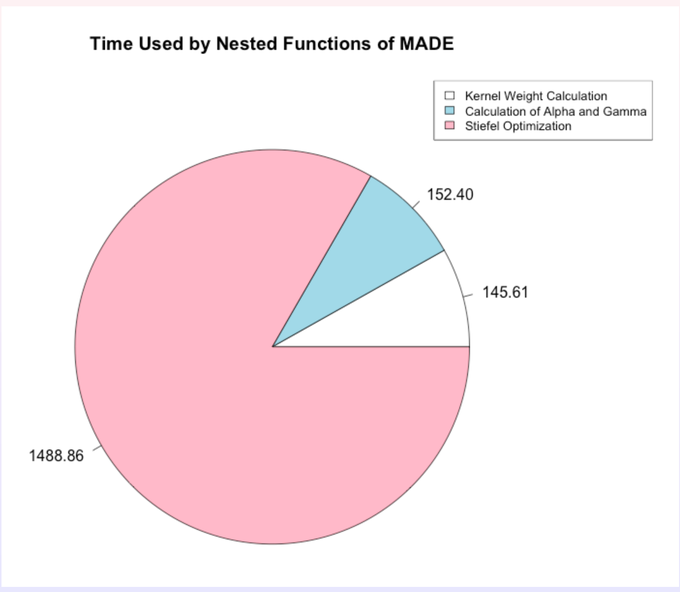


Optimization Function

$$Q(B) = \sum_{j=1}^{n}\sum_{i=1}^{n} w_i(B^T X_j)[y_i(\alpha_j+\gamma_j^T B^T(x_i-X_j))-\exp\{\alpha_j+\gamma_j^T B^T(X_i-X_j)\}-k_0 y_i]$$

The optimization of $Q(B)$ is carried out on a Stiefel manifold [3].

Function Timings



Conclusions

▶ Within MADE the optimization on the Stiefel manifold took by far the longest to do, followed by the kernel weights calculations.

▶ The optimal combination was one node and 16 ppn with a speedup of about twice the serial version.

▶ The upper boundary of the dimension of the dataset (in terms of memory capacity) was determined to be between 5000 by 100 and 10000 by 100.

▶ Future research might be best spent trying to optimize and further parallelize the Stiefel Optimization.

References and Acknowledgments

[1] K.P. Adragni, A.M. Raim, E. Al-Najjar, Unpublished Work (2015)
[2] R.D. Cook, Stat. Sc. (2007)
[3] A. Edelman, T. Arias, S. Smith, SIAM J. Matrix An. & App. (1998)
Full technical report: HPCF–2015–26
REU Site: `hpcreu.umbc.edu`
NSF, NSA, DOD, UMBC, HPCF, CIRC