# Fine Tuning the Grouping Approach to Parallelization of Statistics/Machine Learning Methods

Norm Matloff
University of California at Davis

2022 SIAM Conference on Parallel Processing for Scientific Computing February 24, 2022

Norm Matloff University of California at Davis

# Grouping

# Grouping

- Parallel comp. for stat methods often uses grouping.
- Break data into chunks of rows.
- Apply stat method to each chunk.
- Average the results.
- Provably efficient for asympt. normal estimators.
- In some cases, superlinear speed up (Matloff, JSS, 2016).

Norm Matloff University of California at Davis

# "Old" Parallel Schemes Work on "Modern" Methods?

# "Old" Parallel Schemes Work on "Modern" Methods?

- SVM, NNs etc.: Can be parallelized well using grouping (Yancey and Matloff, 2018).
- Collaborative filtering: Most popular approach is SVD etc., well-known parallel tools.
- What about FOCI?

Norm Matloff University of California at Davis

# What Is Foci?

# What Is Foci?

- Azadkia and Chatterjee, 2018.
- Method for predictor/feature selection.
- Nonparametric, no tuning parameters.
- Motivation: Predicting Y from X. Should we add predictor Z? Measure have much less Var(Y | X,Z) is than Var(Y | X).
- Highly computationally intensive.

Norm Matloff University of California at Davis

# CRAN Package

# CRAN Package

- Azadkia, Chatterjee, Matloff
- They asked me to get involved because I complained FOCI was too slow :-) .
- One part of FOCI is "embarrassingly parallel." Most is NOT.

Norm Matloff University of California at Davis

# A Problem in Grouping "Modern" Methods

Norm Matloff University of California at Davis

# A Problem in Grouping "Modern" Methods

- Modern methods tend to have lots of tuning parameters.
- With $r$ groups, we are finding the best tuning par. set for data of size $n/r$, not $n$.
- FOCI has no tuning pars., but same issue. It finds a good set of predictor variables for data of size $n/r$.
- So: How should FOCI be parallelized?

Norm Matloff University of California at Davis

# Parallel FOCI

# Parallel FOCI

Case study of a "modern" stat method.

Issues:

- Say we take a grouping approach.
- As noted, the predictors chosen by a smaller dataset (size n/r) will generally be different from (and fewer in number than) those chosen on a larger set (n).

Norm Matloff University of California at Davis

# Possible Ways to Combine Group Outputs

Norm Matloff University of California at Davis

# Possible Ways to Combine Group Outputs

- Take the union of the $r$ predictor sets.
- Take the intersection of the $r$ predictor sets.
- Under assumption that the union set is "too much," prune by running FOCI on this set.

The intersection approach wasn't too promising–it often would be empty, especially for large $r$—and won't be pursued here.

Norm Matloff University of California at Davis

# Some (Small-Scale) Examples

# Some (Small-Scale) Examples

- African soil data: 1157 x 3579 (p >> n). Numeric X and Y. Predict pH.
- Million Song data: 515345 x 91. Numeric X and Y. Predict year of release. (50K subset used here.)
- Other datasets not shown.
- Simple quad core.
- Criterion: How well can the selected variables predict Y?
- Prediction models: Linear, polynomial, gradient boosting etc., from **qeML** package.

Norm Matloff University of California at Davis

# Results

# Results

Timing:

| dataset | distrib. comp. | re-run | serial |
|---|---|---|---|
| African soil | 31.32 | 4.02 | 33.02 |
| Million Song | 138.80 | 75.61 | 333.11 |

Accuracy:

| dataset | distrib. comp. | re-run | serial |
|---|---|---|---|
| African Soil. qeLin | 0.378 | 0.415 | 0.365 |
| African Soil, qeGBoost | 0.49 | 0.49 | 0.49 |
| Million Song, qeLin | 6.94 | 7.08 | 6.95 |
| Million Song, qeGBoost | 7.30 | 7.20 | 7.23 |

Norm Matloff University of California at Davis

# Trends, Here and Other Datasets

# Trends, Here and Other Datasets

- The proposed grouping approach does improve (in some cases not shown, dramatically).
- Accuracy is generally maintained.