Cross-Validation for Training and Testing Co-occurrence Network Inference Algorithms

Daniel Agyapong

Dr. Toby Hocking

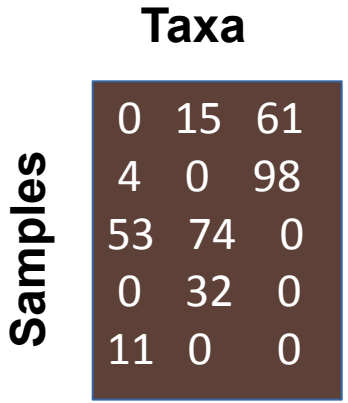
NAU
NORTHERN ARIZONA UNIVERSITY

School of Informatics, Computing, and Cyber Systems

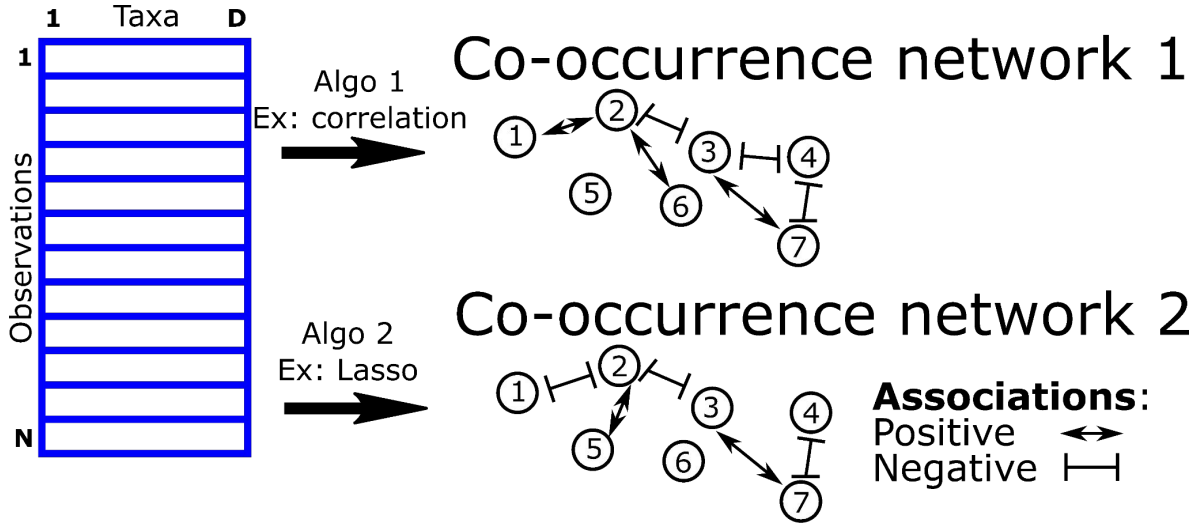
Real Microbiome Abundance Data

Data	Citation	Samples	Taxa
amgut1	https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004226	289	127
amgut2		296	138
crohns	https://www.mcgill.ca/statisticalgenetics/software	100	5
iOraldat	https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03911-w	86	63

Each data set is a matrix of counts, for example:



Different Algorithms Infer Different Co-occurrence Networks



Previous algorithms

SparCC (Jonathan et al 2012)
SPIEC-EASI (Zachary et al 2015)
COZINE (Min et al 2020)

Which is a more accurate interpretation for these data?

Categories of Some of the Existing Algorithms?

- There are a lot of existing algorithms, each with various hyper-parameters which determine the sparsity (number of edges) in network.
- We grouped them into three different categories.

Pearson/Spearman Correlation

- Threshold on correlation constant
- **CoNet(Karoline et al 2016)**

Least Absolute Shrinkage and Selection Operator (LASSO)

- Degree of L1 regularization
- **CCLasso (Huaying et al 2015)**

Gaussian Graphical Model (GGM)

- Inverse Covariance (Precision) Matrix
- **COZINE(Min et al 2020)**

Research Questions

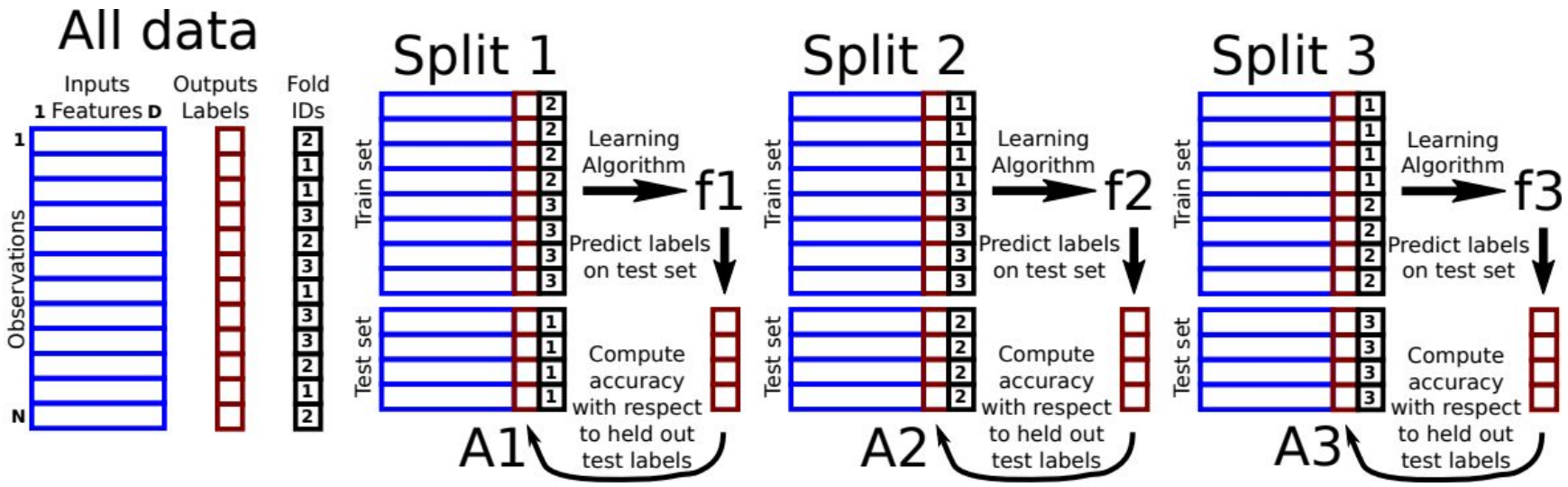
For some particular real data sets:

- How can we automatically learn hyper-parameters? (let the data tell us the “**best**” threshold, rather than choosing arbitrarily)
- Which of the available microbial network analysis algorithms is most accurate and gives least test error ?
- How many samples are needed for cross validation to be useful.

Existing Evaluation Types

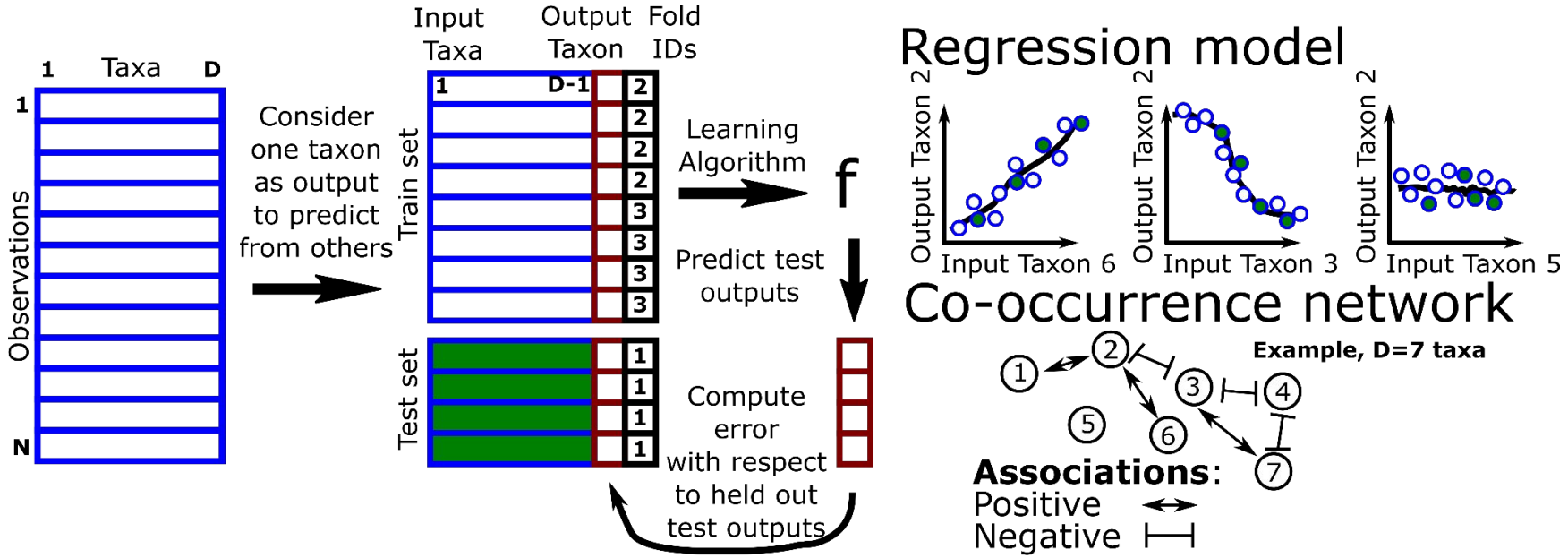
Category of Evaluation Type	External Data	Network consistency between sub-samples
Papers	**SparCC (Jonathan et al 2012), SPIEC-EASI (Zachary et al 2015)**	**CCLasso (Huaying et al 2015)**
Issues	• Lack of ground truth (external data are not always available) • Biases in external data	• Trivial model has perfect consistency (featureless model with no edges has 100% accuracy)

Cross-Validation for Supervised Learning



- K-Fold cross-validation: each observation assigned a fold ID, K=3 means fold IDs from 1 to 3.
- For each fold ID, use corresponding observations as a test set to evaluate generalization ability of learning algorithm (trained on all other observations).

Proposal: Cross-Validation for Training and Testing Co-occurrence Network Inference Algorithms



Repeat for each output taxon and Fold ID

Parallelizing Test Error Script on Monsoon

N samples
x
D taxa

1	2	3
5	0	7
0	0	0
2	1	3

Dataset	# of Total Samples	Index of prediction column
amgut2	10	0
⋮	⋮	⋮
amgut2	296	0

MONSOON HPC

Case 1

n_subsamples = 296//10 = 29
total_samples = 29*10 = 290

Split dataset into **29** different subsamples each with **10** observations randomly.

Case 2

n_subsamples = 296//296 = 1
total_samples = 1*296 = 296

Use full dataset as **1** big subsample with **296** observations.

```
For each ss_index, ss of the subsamples:
  For each fold in the 3-fold CV:
    For each Algo:
      algo.fit(train)
      pred_y = algo.predict(testX)
      test_error = mse(test_y, pred_y)
      save(test_error, ss_index,fold_id,...)
```

Concat the saved df from all the results

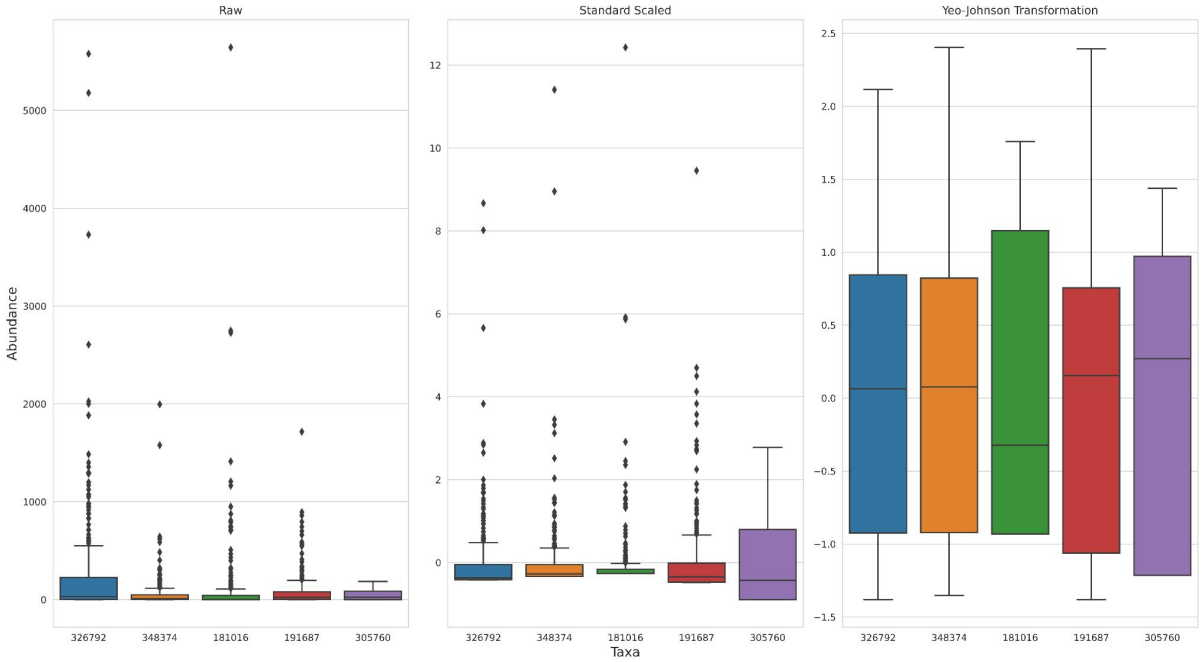
Dataset	# of Total Samples	Index of prediction column	Index of Subsample	FoldID	Algorithm	MSE on Test
amgut2	10	0	0	1	LASSO	0.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Group by Dataset, # of Total Samples and Algo

Find the average and the std of the MSE on test set

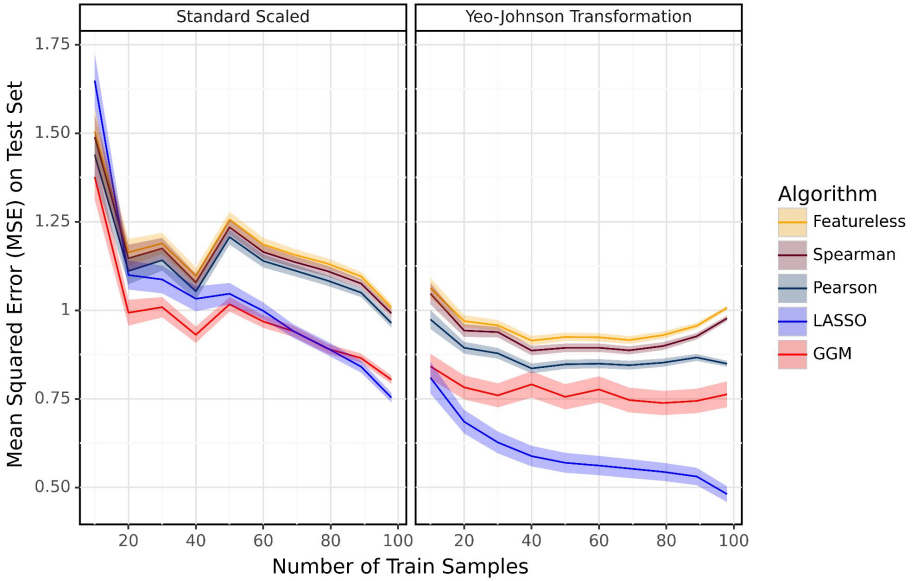
Dataset	# of Total Samples	Algorithm	Average of MSE on Test	std of MSE on Test
amgut2	10	LASSO	0.5	0.001
⋮	⋮	⋮	⋮	⋮

Boxplot of Different Data Transformation Techniques on first 5 taxa of the Amgut2 data set



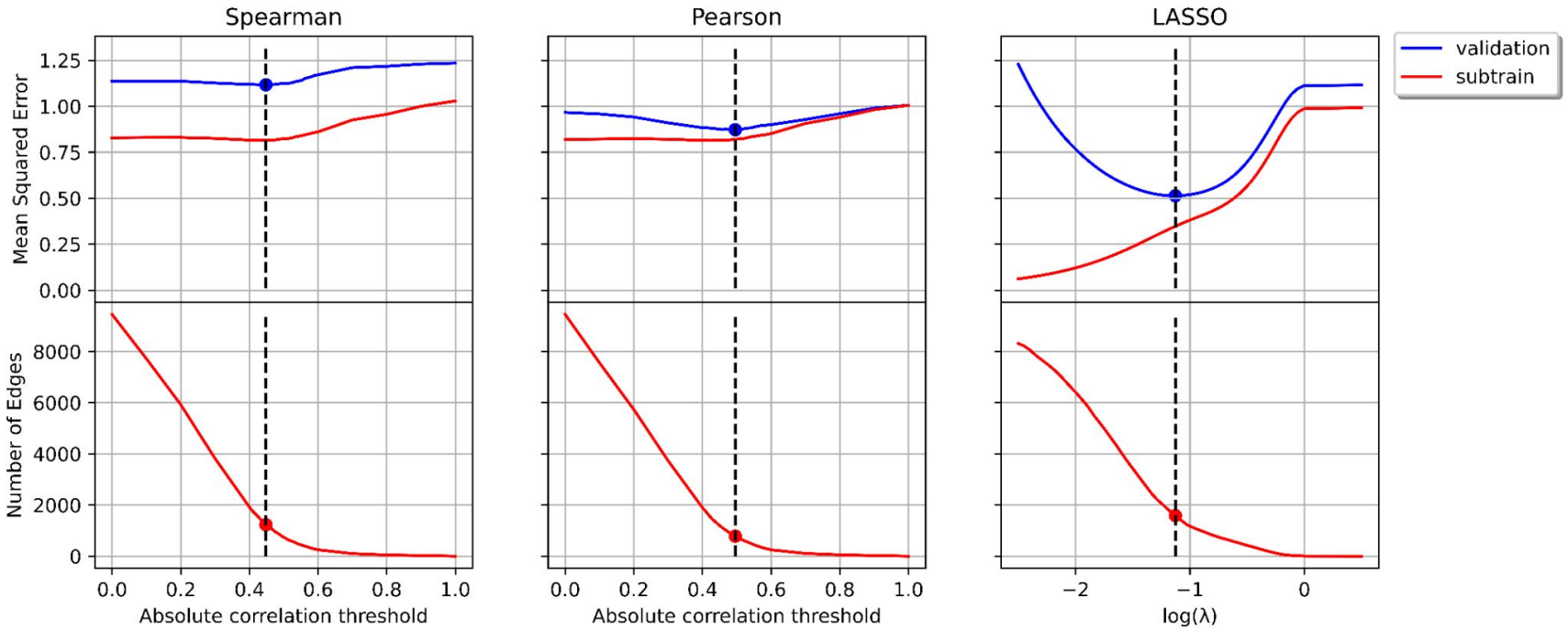
- Both raw and standard scale contain sparse data. Hence a lot of outliers
- After Yeo-Johnson transformation (and Standard Scaling), the data becomes more normally distributed and has less outliers.

How Different Data Transformation Techniques affect the Test Error



- This figure evaluates the performance of the various algorithms on the **Amgut2** real dataset.
- The results imply that just standard scaling alone yields lower accuracy than the combination of Yeo-Johnson and standard scaling for the algorithms compared.

Hyperparameter training on Amgut2 data set

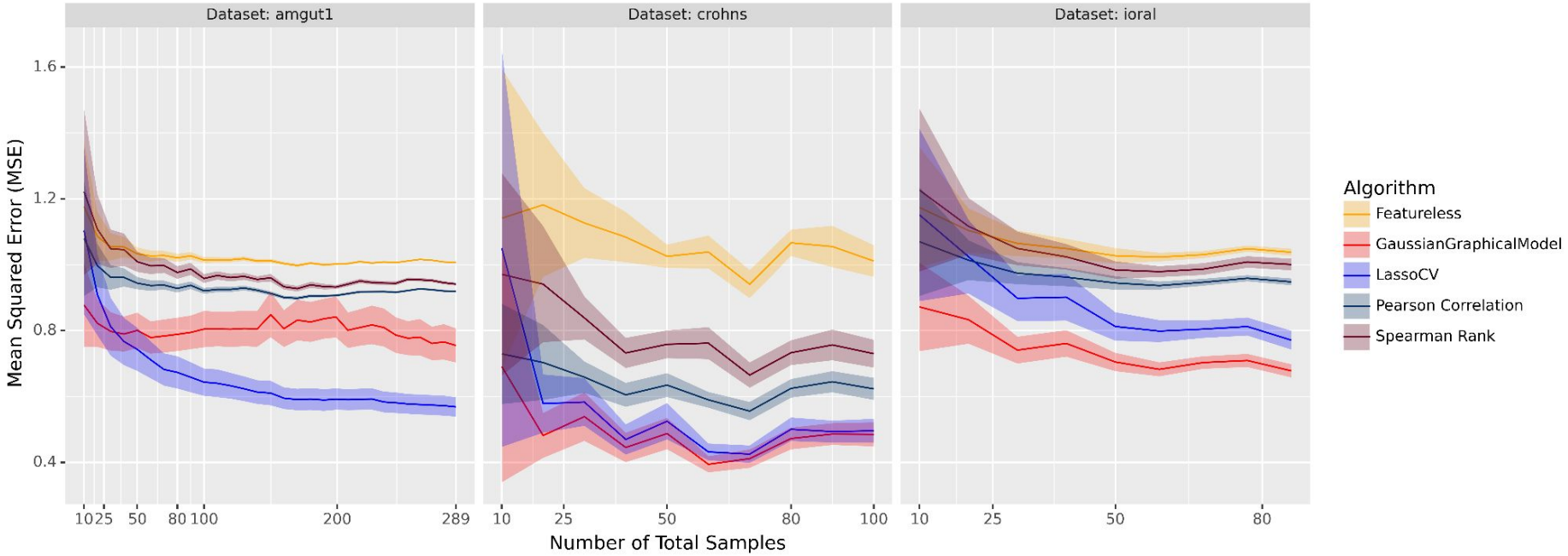


Threshold = 0.448
Edges = 1231

Threshold = 0.495
Edges = 785

log(λ) = -1.12
Edges = 1585

Proposed Cross-Validation Method requires only 10-20 Samples to see Differences in Test Error between Algorithms



LASSO best

LASSO and GGM similar

GGM best

(A) Model Comparison using Inferred Positive and Negative Association
(B) Microbial Network Graph of crohns Data Set