Cross-Validation for training and testing co-occurrence network inference algorithms

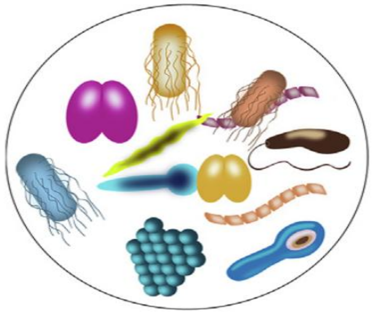
Daniel Agyapong
da2343@nau.edu
PhD student

Dr. Toby Hocking
Toby.Hocking@nau.edu
Machine Learning Director

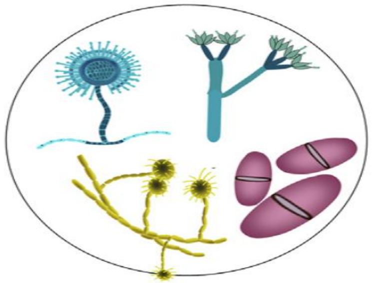
NAU
NORTHERN ARIZONA UNIVERSITY
School of Informatics, Computing, and Cyber Systems

INTRODUCTION

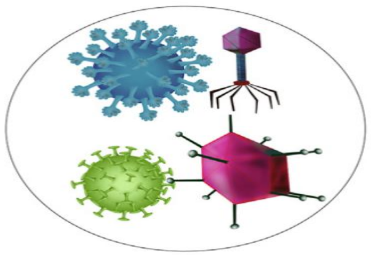
Bacteria



Fungi



Virome



- Microbial communities consist of micro-organisms such as bacteria, virus and fungi.
- Micro-organisms have built robust ecosystems in various environments such as soil, sea water and various human organs.
- Microbiome has been associated with conditions such as obesity, colorectal cancer and inflammatory bowel disease.
- Understanding microbial interactions and relationships may provide great insights in restoring a healthy microbial community.

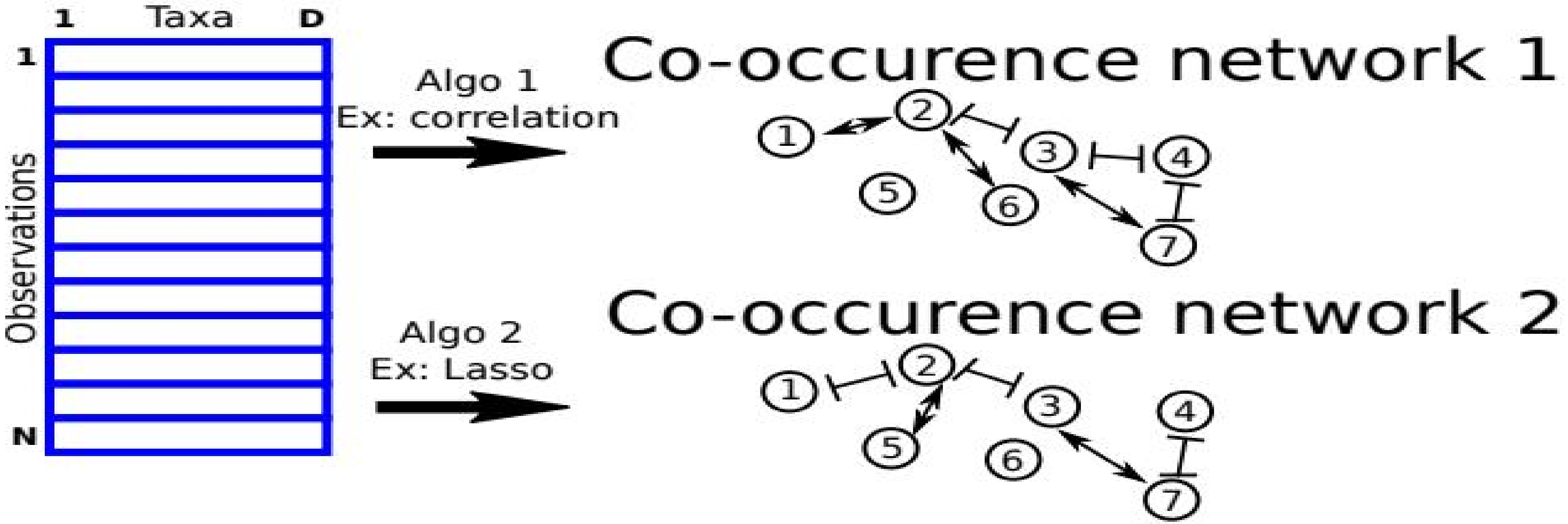
Real Microbiome Abundance Data

Data	Citation	Samples	Taxa
amgut1	https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004226	289	127
amgut2		296	138
hmp216S	https://ibdmdb.org/tunnel/public/summary.html	47	45
hmp2prot		47	43
enterotype	https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061217	280	553
esophagus		3	58
crohns	https://www.mcgill.ca/statisticalgenetics/software	100	5
Baxter_CRC	http://www.raeslab.org/companion/ocean-interactome.html	490	117
glne007		490	338
iOraldat	https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03911-w	86	63

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

$$\begin{array}{c|ccc} & & \textbf{Taxa} & \\ \hline \textbf{Samples} & 0 & 15 & 761 \\ & 4 & 0 & 98 \\ & 53 & 74 & 0 \\ & 0 & 32 & 0 \\ & 11 & 0 & 0 \\ & 0 & 24 & 65 \end{array}$$

Different algorithms infer different co-occurence networks



Algo 1
Ex: correlation

Co-occurence network 1

Algo 2
Ex: Lasso

Co-occurence network 2

Which is a more accurate interpretation for these data?

Associations:
Positive ↔
Negative ⊢⊣

What are some of the Existing Algorithms?

- There are many existing algorithms, each with various hyper-parameters which determine the sparsity (number of edges) in network.

➢ Linear (Pearson/Spearman) Correlation: threshold on correlation constant.

➢ Least Absolute Shrinkage and Selection Operator (LASSO): degree of L1 regularization.

➢ Gaussian Graphical Model : Inverse Covariance (Precision) Matrix.

Research Questions

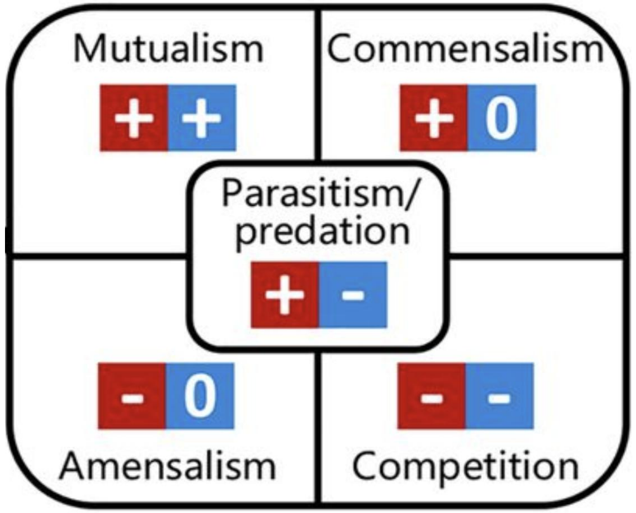
For a particular real data set, like the ones we will be gathering in this project :

- 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 error ?

- How many samples are needed for Cross Validation to be useful.

BACKGROUND



Microbial relationships

- Micro-organisms form complex ecological interactions :
  - **Mutualism**: Both parties benefit. Mutual cross-feeding.
  - **Parasitism/Predation**: One side benefits whilst the other side loses. Relationships such as predator-prey and host-parasite interactions.
  - **Competition**: Both parties lose. When there is insufficient resources for both organisms, they compete for the limited resources.
  - **Commensalism**: One organism benefits without harming the other.
  - **Amensalism**: One organism is harmed but the other is unaffected.
- Reconstructing microbial ecological networks to represent these interactions would help to understand the complex behaviors in microbial communities.

Existing Correlation Based Methods

Method	SparCC (2012)	REBACCA(2015)
Link	https://rdrr.io/github/zdk123/SpiecEasi/man/sparcc.html	https://faculty.wcas.northwestern.edu/hji403/REBACCA.htm
Algorithms Compared	SparCC, Pearson	REBACCA, SparCC, BP, ReBoot
How they compare	Computing the number of true-positives (TP), false-positives (FP), true-negatives (TN) and false-negatives (FN) detected in the Pearson network by treating the SparCC network as the true one.	Consistency of correlated pairs identified independently from the three datasets (A correlated pair of OTUs is consistent between two datasets if the pair has the same signs of correlations in both datasets).
Category of Evaluation Type	External data	External data

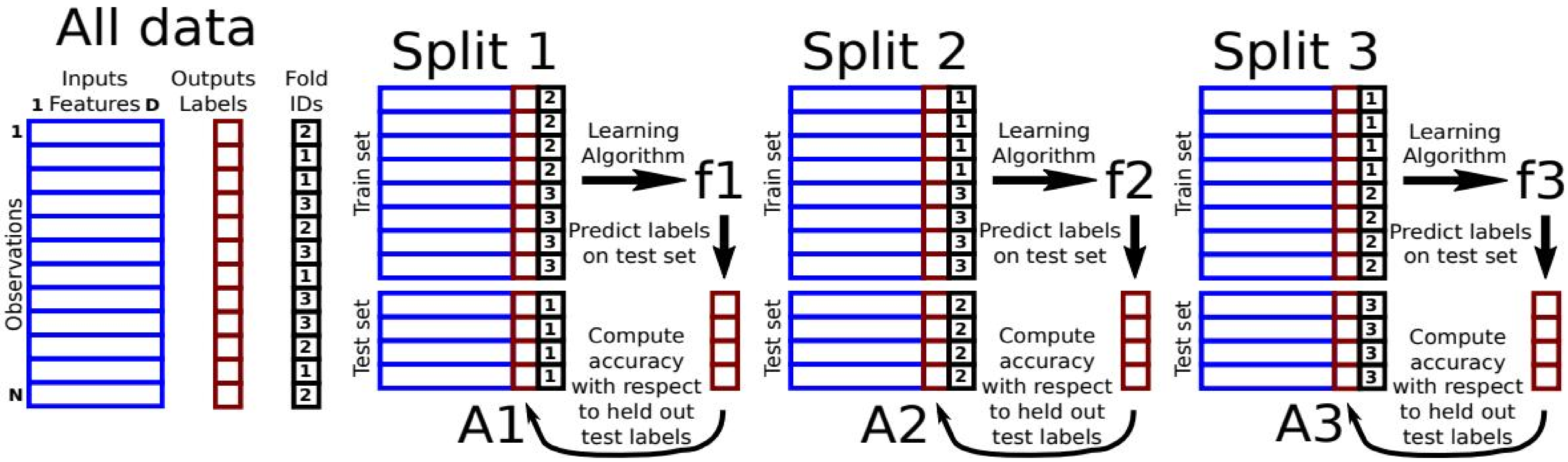
Existing LASSO Based Methods

Method	SPIEC-EASI (2015)	CCLasso (2015)
Link	https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004226#pcbi-1004226-g006	https://github.com/huayingfang/CCLasso
Algorithms Compared	SPIEC-EASIE, SparCC, CCREPE	CCLasso, SparCC
How they compare	Consistency between two models by computing Hamming Distance (the difference between the upper triangular part of the two adjacency matrices) between reference and new models.	Frobenius Accuracy with respect to estimating correlation matrix from data using half samples (measured by the Frobenius norm distance between the estimated correlation matrices and a reference correlation matrix). Reproducibility (measured by the fraction of the same edges shared for the two steps in the first reference network which only the top 1/4 edges is used)
Category of Evaluation Type	External Data (Amgut Dataset)	Sub-sample analysis

Existing Gaussian Graphical Based Methods

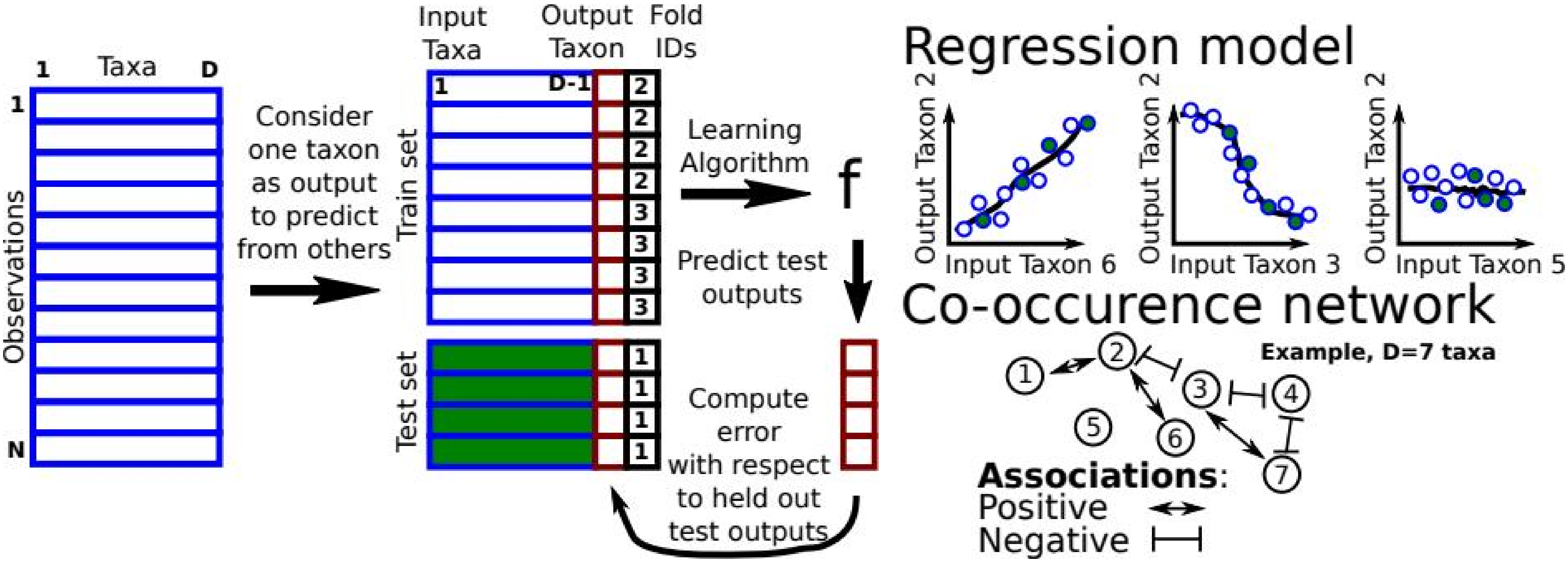
Method	gCoda(2017)	mLDM(2020)
Link	https://doi.org/10.1089/cmb.2017.0054	https://www.science.org/doi/abs/10.1126/science.1262073
Algorithms Compared	gCoda, SPIEC-EASIE	mLDM, SparCC, CClasso
How they compare	False-Positive count and the running when methods are tested on shuffled OTU data.	Power of association inference when compared to a reference association inference data from the research paper.
Category of Evaluation Type	External Data (Mouse Skin Microbiome Data)	External data (Tara Oceans Eukaryotic Data)

Cross-validation algorithm 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

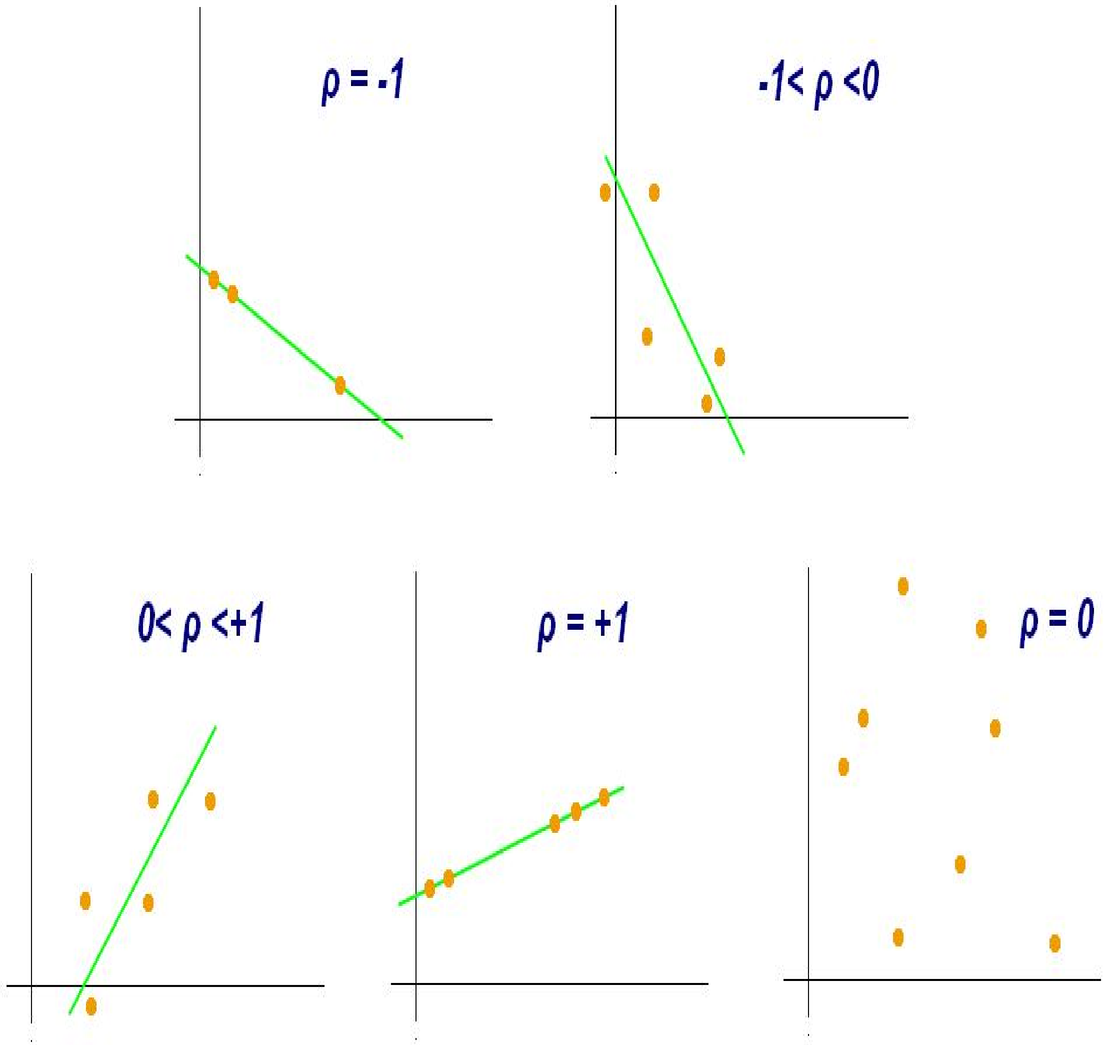


Repeat for each output taxon and Fold ID

PEARSON CORRELATION

- Pearson’s correlation coefficient is the standard tool to infer a network through correlation analysis among all pairs of OTU (Operational Taxonomic Unit) samples.
- It is a number that ranges from –1 to 1 and measures the strength and direction of the relationship between two variables.
- Where x and y are the two taxa being compared, ρ is the correlation constant, μ is the mean, σ is the standard deviation, the expected (predicted) value is given by:

$$\mathrm{E}(Y \mid X) = \mu_Y + \rho_{X,Y} \frac{\sigma_Y}{\sigma_X}(X - \mu_X)$$



SPEARMAN’S RANK CORRELATION

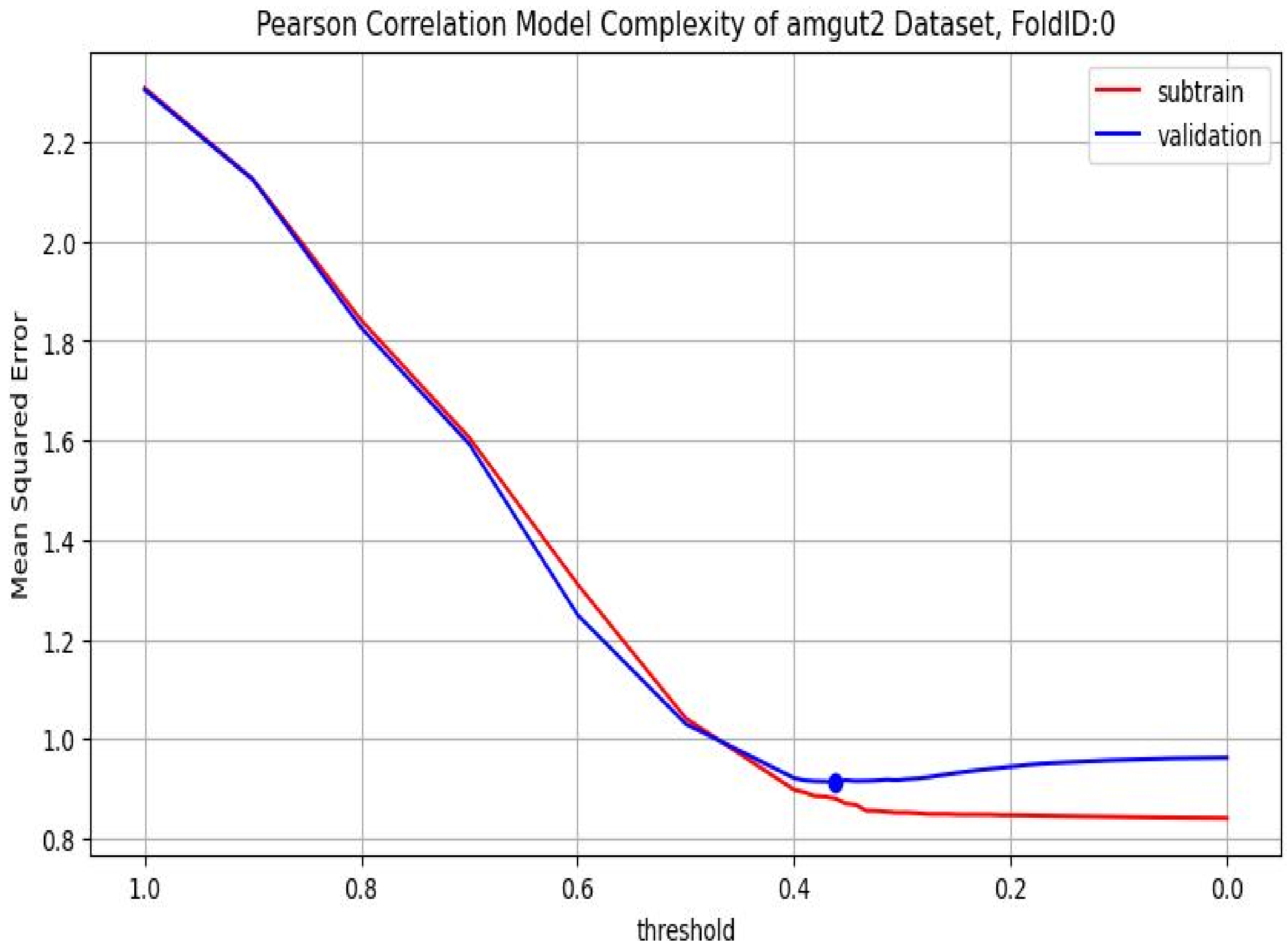
- Spearman's Rank Correlation coefficient is another popular correlation method for microbial network inference.

- It is often adopted as an alternative to the Pearson Correlation Coefficient especially when dealing with non-linear relationships between taxa.

- The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables.

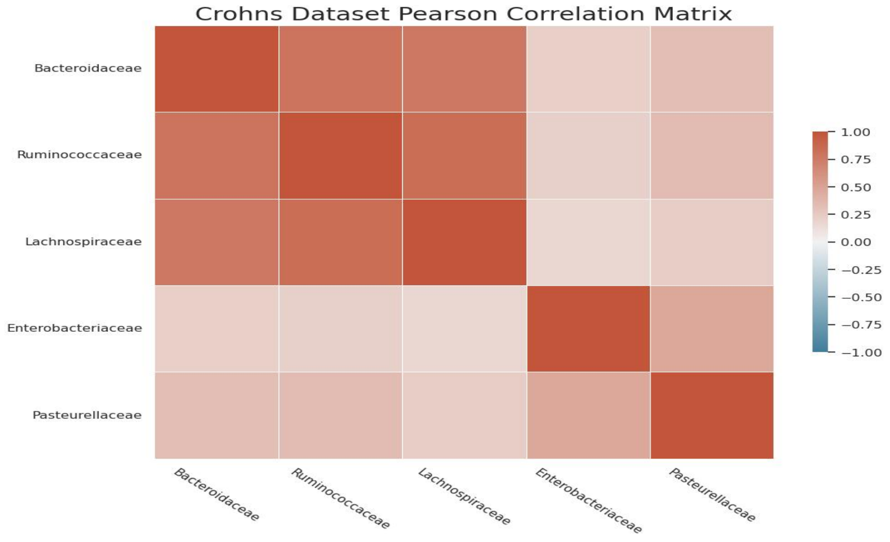
- Just like Pearson, Spearman’s rank correlation coefficient ranges from –1 to +1, with -1 indicating a perfect negative monotonic relationship, 0 indicating no monotonic relationship, and +1 indicating a perfect positive monotonic relationship.

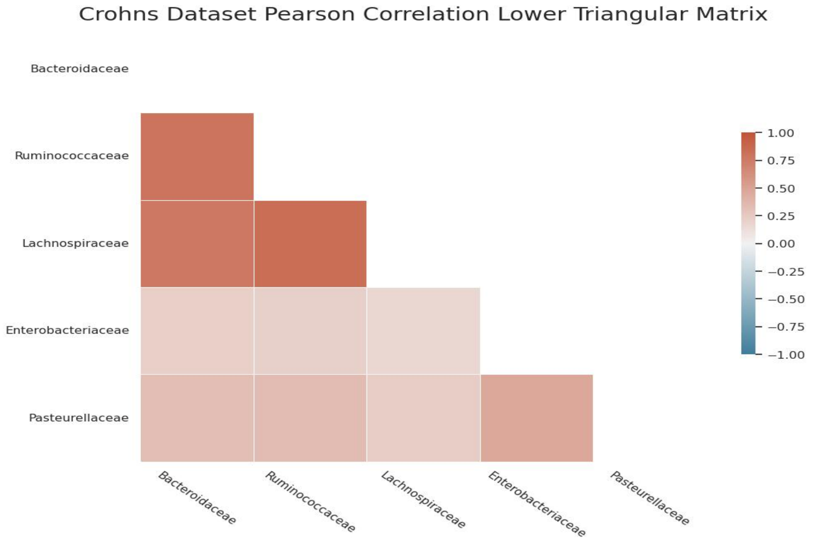
Results: Training the Pearson correlation threshold using cross-validation



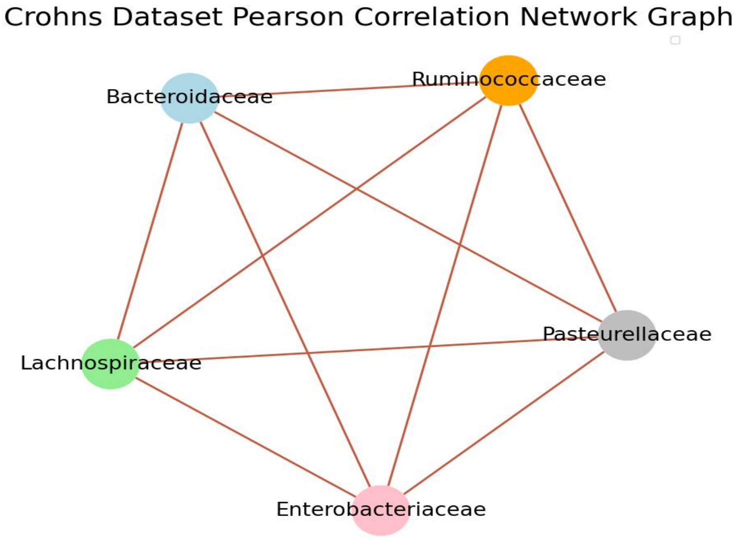
- Subtrain error decreases as the model complexity increases whilst the validation error shows a U shape.
- We select the threshold which gives the minimum validation error, in this example r2=0.35 (any smaller r2 values will have no edge in the co-occurrence network).

PEARSON CORRELATION MATRIX

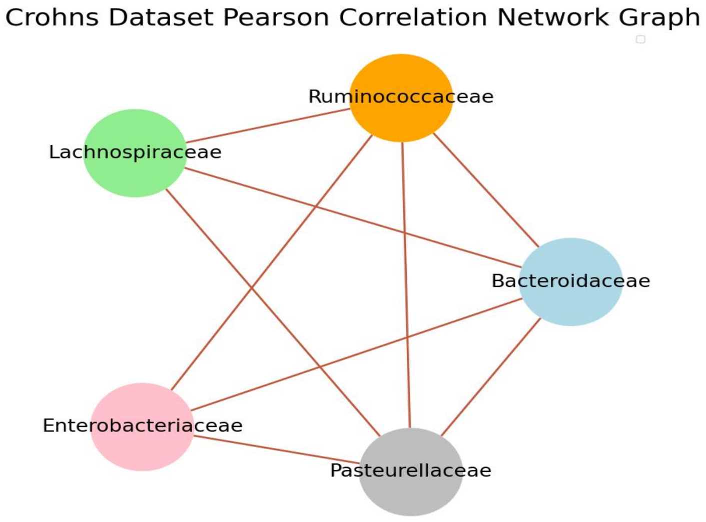




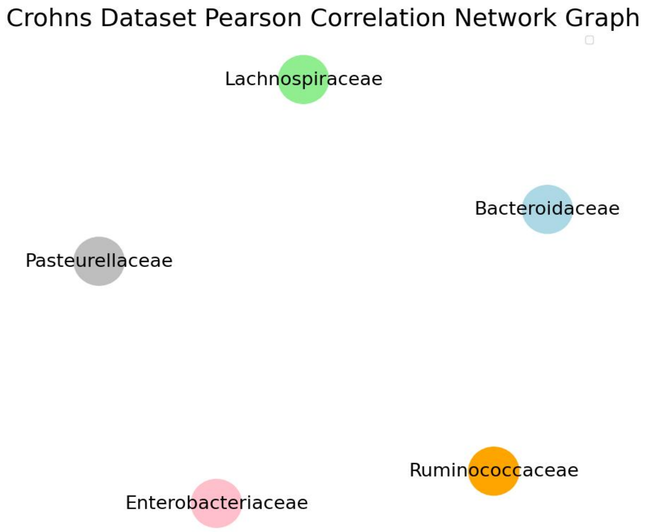
NETWORK GRAPH



Threshold = 0



Threshold = 0.2



Threshold = 1

LASSO REGRESSION MODEL

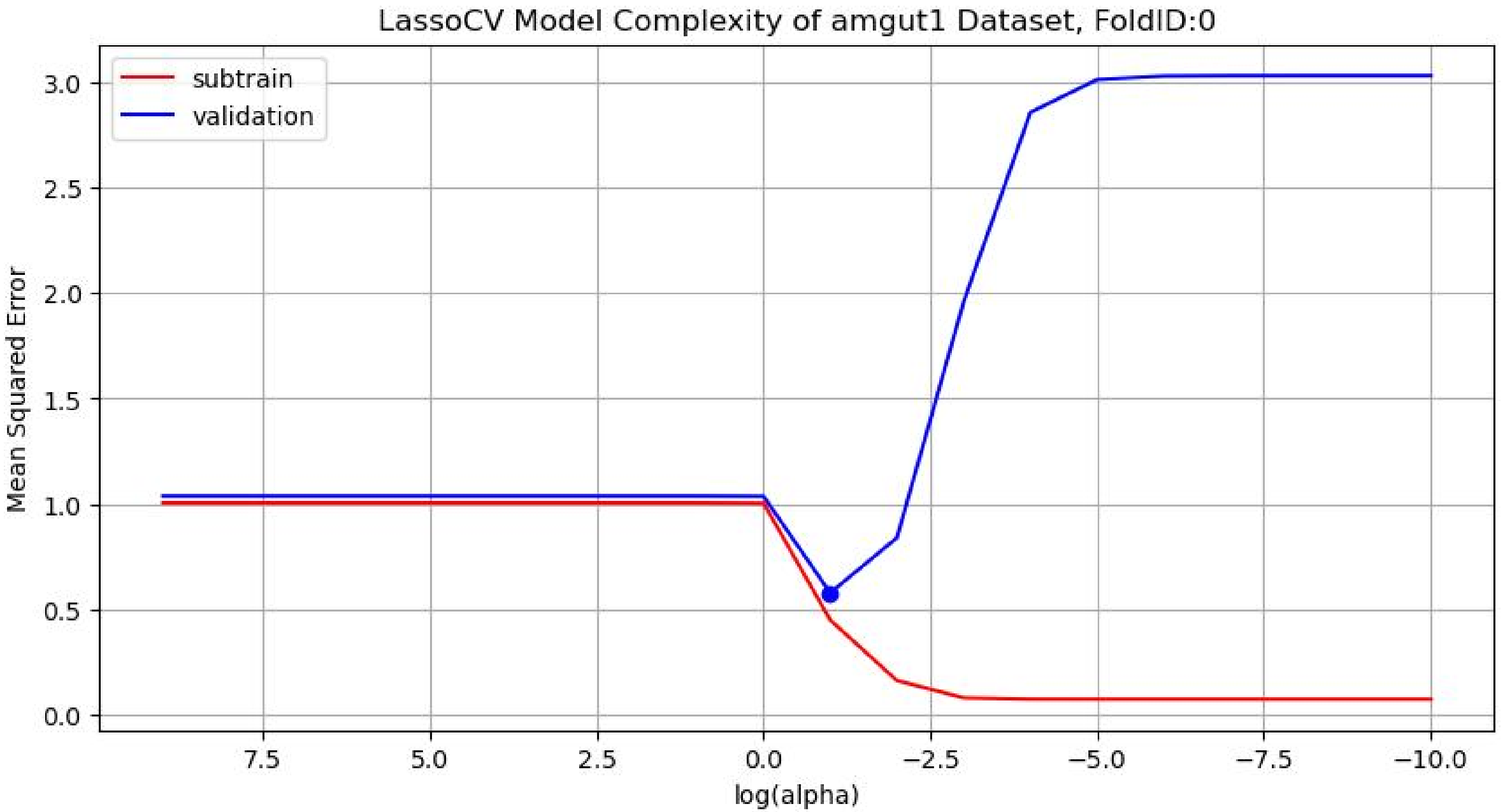
- The LASSO is also known as Least Absolute Shrinkage and Selection Operator. It is a form of linear regression which uses L1 regularization technique and variable/taxa selection to increase the accuracy of prediction.

- L1 regularization adds a penalty which causes the regression coefficient of the less contributing variable to shrink to zero or near zero.

- Loss function: ß values are the coefficients to be learned. λ (lamda or alpha) is a tuning parameter (amount of shrinkage). When λ = 0, no parameters(taxa) are eliminated.

$$L(\beta_0, \beta) = \frac{1}{2n}||y - \beta_0 - X\beta||_2^2 + \lambda||\beta||_1$$

Results: Training the Lasso algorithm with cross-validation



- Train set split into subtrain set (used to learn regression coefficients) and validation set (used to learn model complexity, degree of L1 regularization, which controls sparsity / number of edges in co-occurence network).
- Subtrain error decreases, while the validation error shows expected U shape.
- We select the alpha value (degree of L1 regularization) which has the minimimum the validation error.

GAUSSIAN GRAPHICAL MODEL

- The Gaussian distribution is a continuous and symmetrical probability distribution that explains how the outcomes of a random variable are distributed.

- Most observations are clustered around the mean so there is a less chance of occurrence as the observations move further away of mean.

The Probability Density Function (PDF) of the distribution is :

x is a k dimensional vector variable.

Σ is the k × k covariance matrix

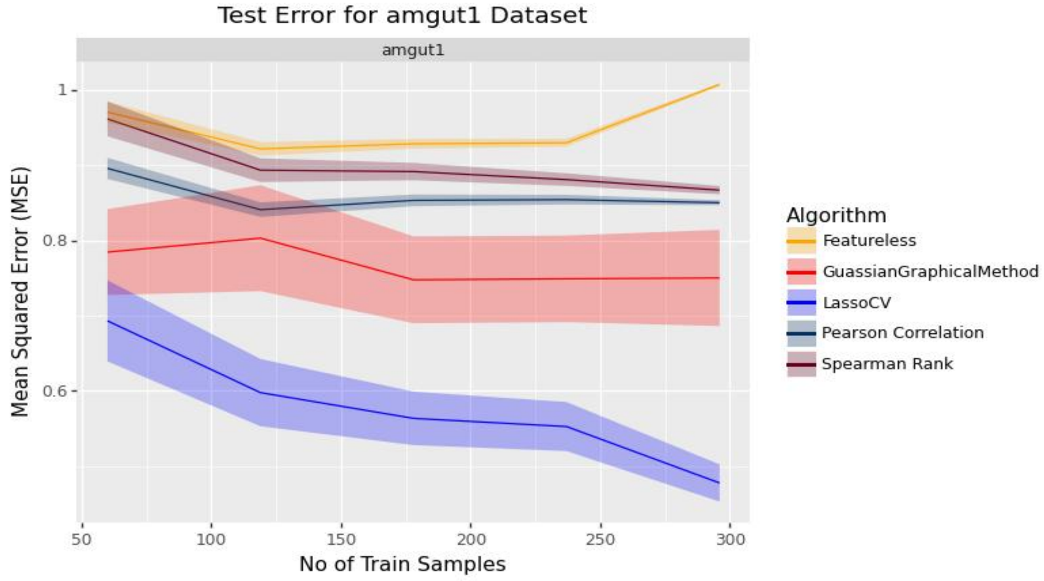
ω is a value in the k x k precision matrix

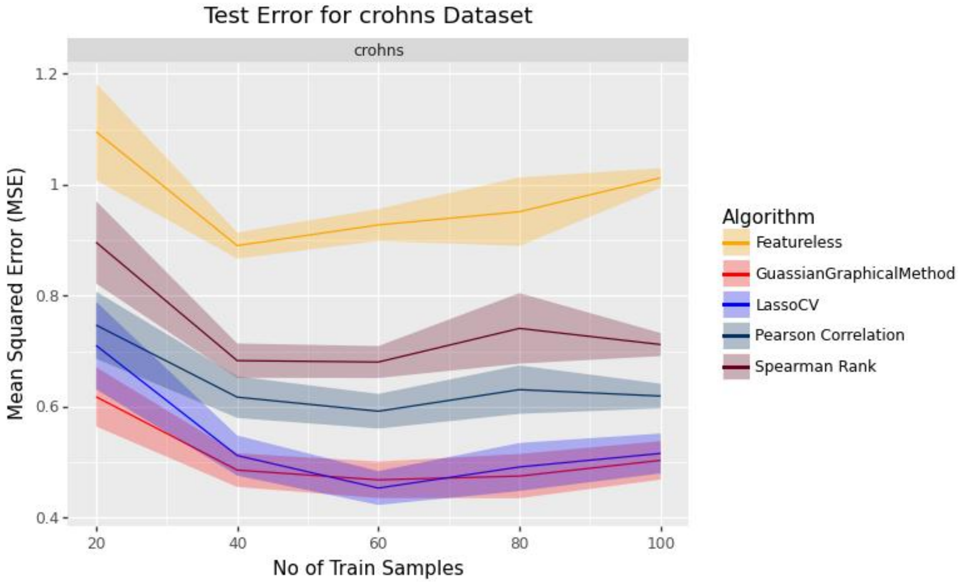
$$f(x) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}) \right)$$

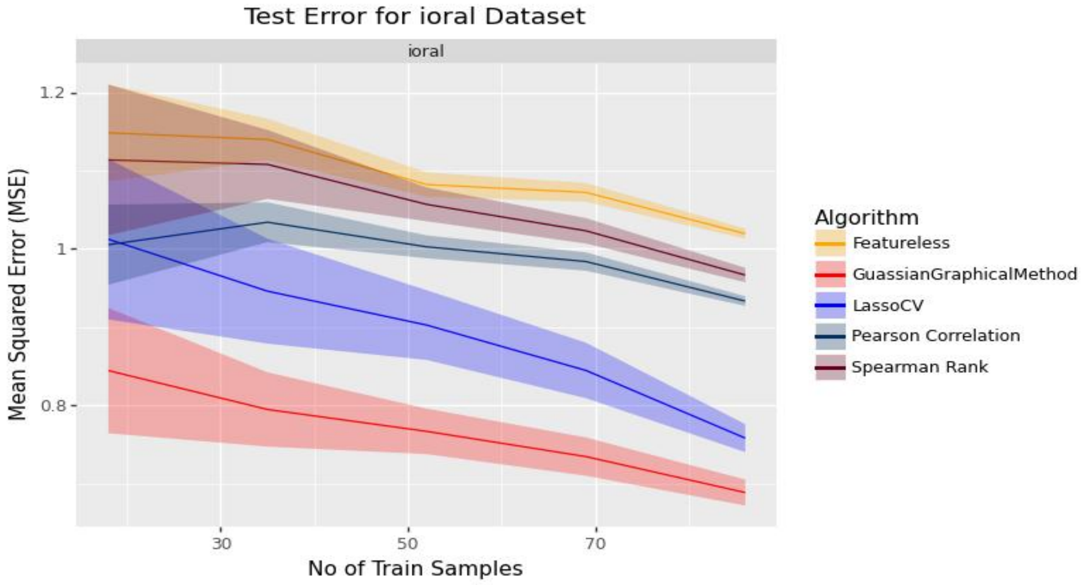
Conditional mean (predicted value) is :

$$x_1 = \frac{-1}{2\omega_{11}} \left(\sum_{i=2}^{n} \omega_{i1} x_i + \sum_{j=2}^{n} \omega_{1j} x_j \right)$$

Results: Algorithms can be compared using test error


Test Error for amgut1 Dataset

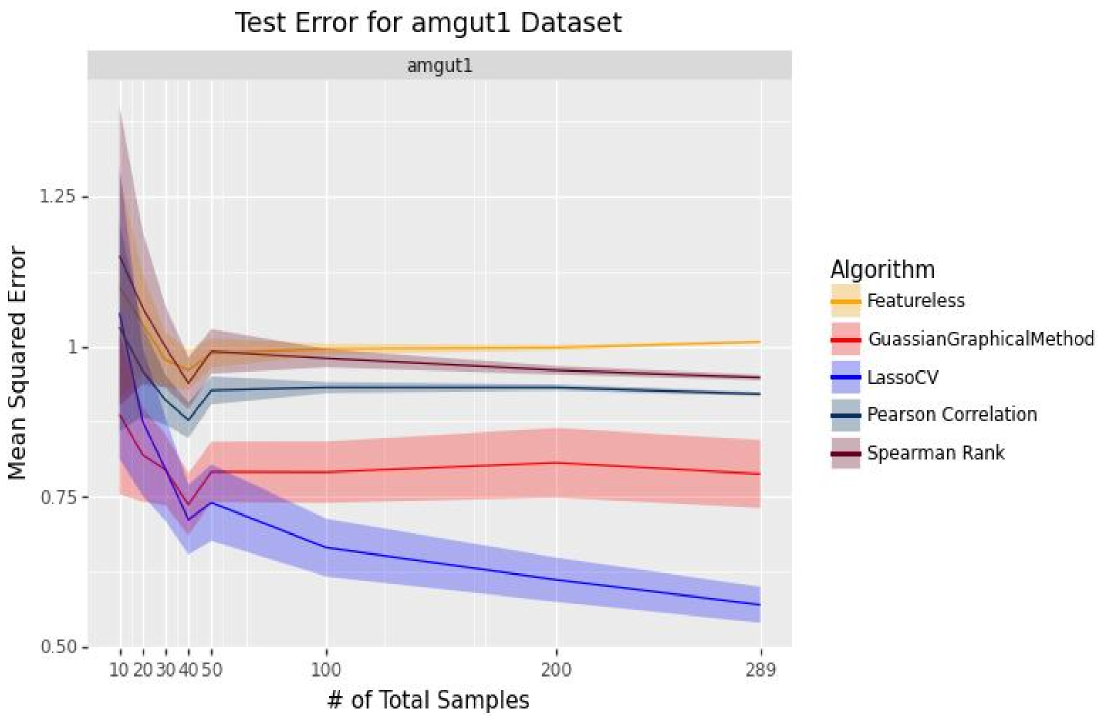

Test Error for crohns Dataset

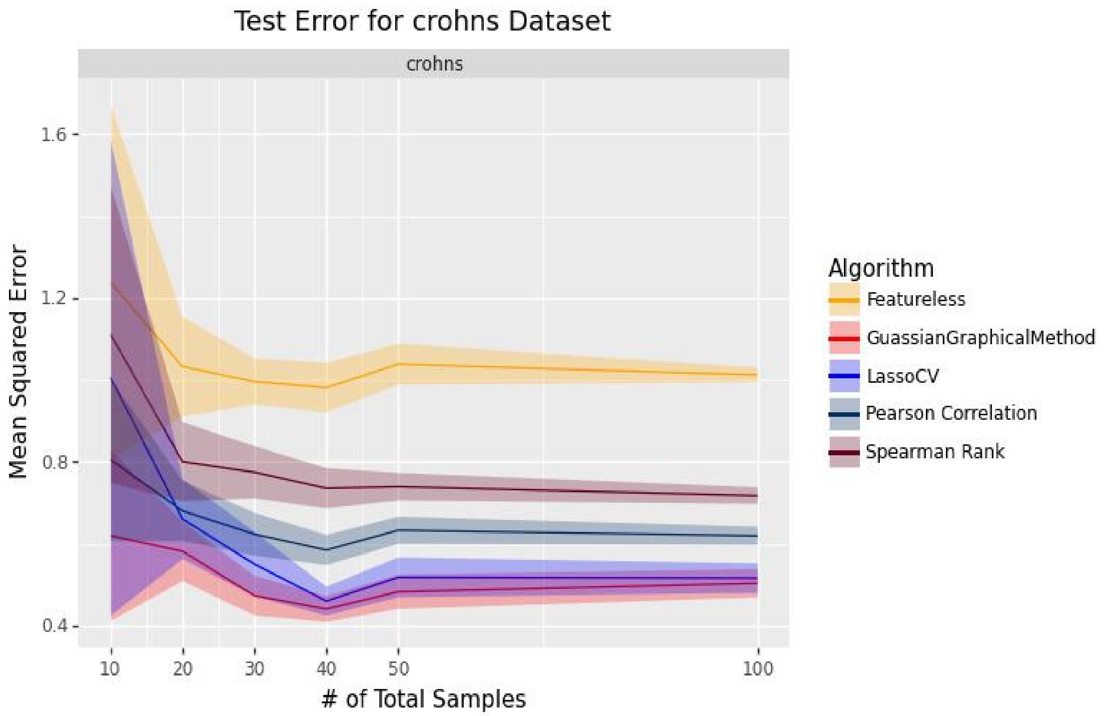

Test Error for ioral Dataset

- LassoCV and Gaussian Graphical Model perform better than the other algorithms.

- The test error reduces as the number of train samples increases.

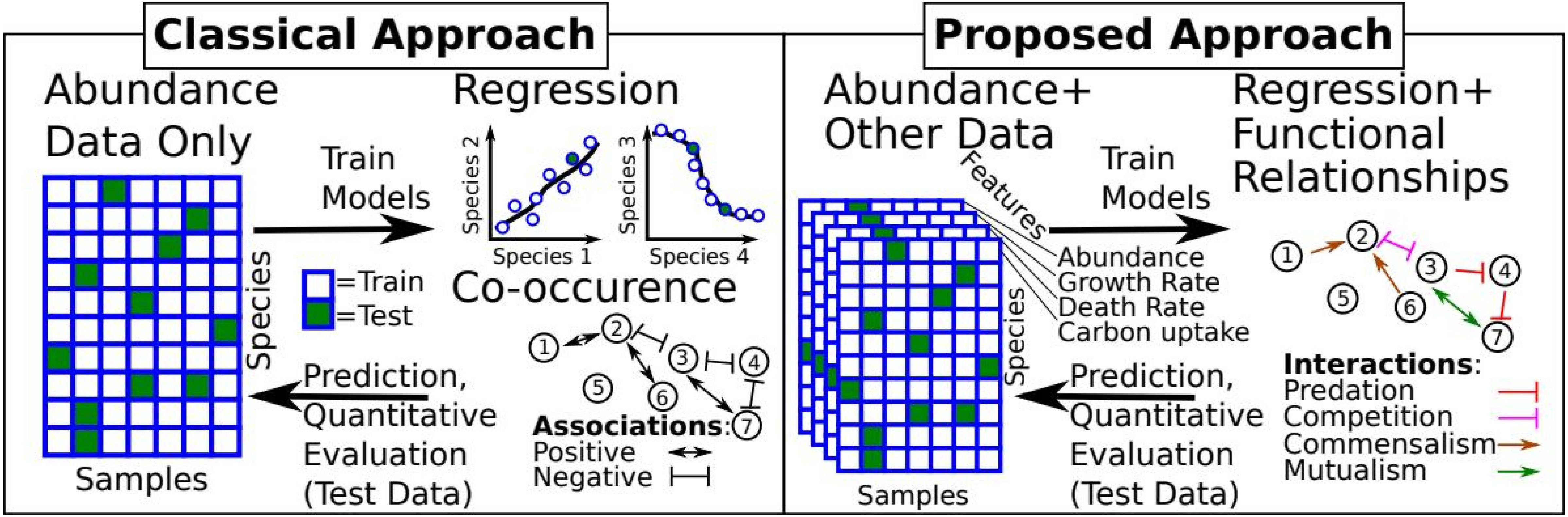
Results: How many samples are needed for CV to be useful





- At every iteration (# of Total Samples), we use the full dataset by dividing the total samples into small sub-samples and we find the average test error of them.
- We are interested in the minimum # of total samples where we see a clear change in test error between the various algorithms.

Future work: Cross-validation for training and testing interaction network inference algorithms, using several qSIP data features



REFERENCES

- https://www.liebertpub.com/doi/10.1089/cmb.2021.0406
- https://smnh.tau.ac.il/en/interactions-among-living-organisms/
- https://scikit-learn.org/stable/modules/cross_validation.html
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7768662/
- https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004226
- https://doi.org/10.1128/mSystems.00124-19
- https://www.thoughtco.com/commensalism-definition-and-examples-4114713
- https://www.sciencedaily.com/releases/2018/05/180515092931.htm

Contact: toby.hocking@nau.edu

Reproducibility: https://github.com/EngineerDanny/CS685-Microbe-Network-Research

THANK YOU

ANY QUESTIONS?