Prioritizing TCR Repertoire Network Properties - A Novel Approach to Select Network Signatures

in collaboration with University of California San Francisco

Thesis Defense by: Shilpika Banerjee

San Francisco State University

MS in Statistical Data Science, 2021-22

Supervised by Dr. Tao He (PhD., Associate Professor)

Dec 15th, 2022.

01 Introduction

02 Methods

03 Implementation

04 Simulation Study

05 Results & Discussion

1 Introduction

Background

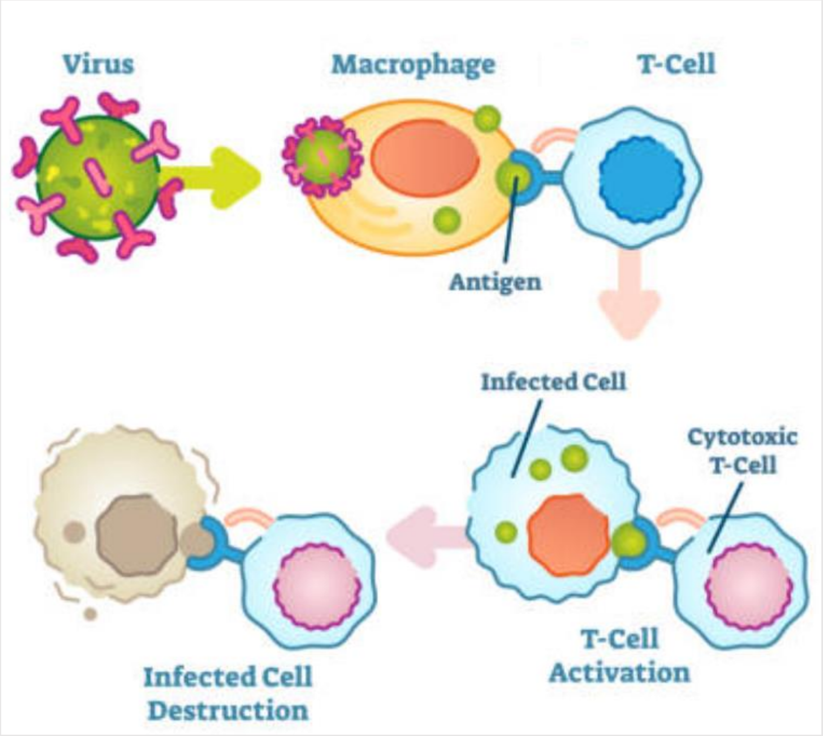
The T cells are one of the major types of white blood cells in the immune system. They play a key role in the coordination of the immune response. T cells are central to adaptive immunity and are involved in almost all adaptive immune responses.

T cell receptors (TCRs) are protein complexes on the T cell surface. TCRs mount a response to harmful foreign invaders and act as the arms of the T cells with memory. They can remember harmful pathogens - provide a life-long protection against the known pathogens in the future.

TCR repertoires continually shape throughout the lifetime of an individual in response to pathogenic exposures. TCRs can serve as a fingerprint of an individual's current immunological profile.

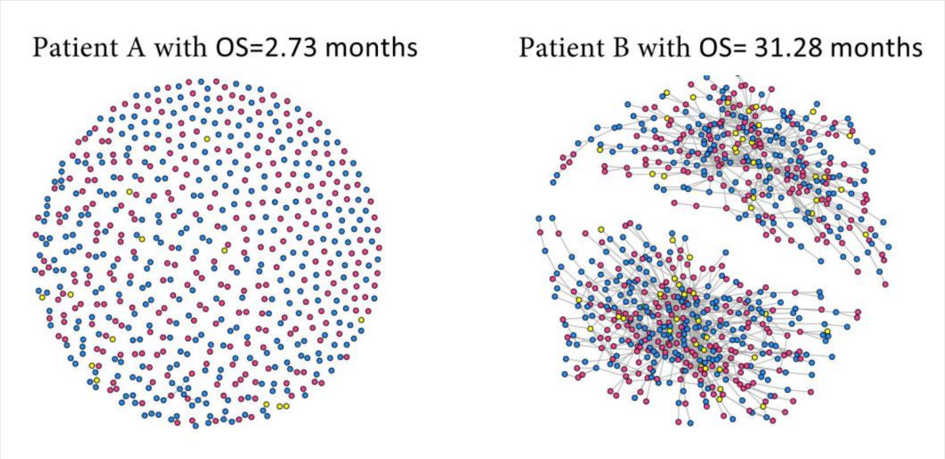
Image sourced from: Stock Photo

Overview of T cell activation



Stage III Lung Cancer Patient data from Phase I trial of Durvalumab

Durvalumab, an immune checkpoint inhibitor, is designed to activate exhausted tumor-reactive T cells in patients with stage III, non-small cell lung cancer (NSCLC). 65 Patients from the Phase I trial (NCT01693562, 14 September 2012) and their immunophenotypic responses were observed. Patients exhibiting increased TCR repertoire diversity on day 15 attained significantly longer overall survival (OS) than those with decreased diversity. Patients with larger and denser TCR clusters showed improved Overall Survival (OS) than patients with smaller TCR clusters.


Patient A with OS=2.73 months
Patient B with OS= 31.28 months

TCR network diversity of two patients from the drug trial.

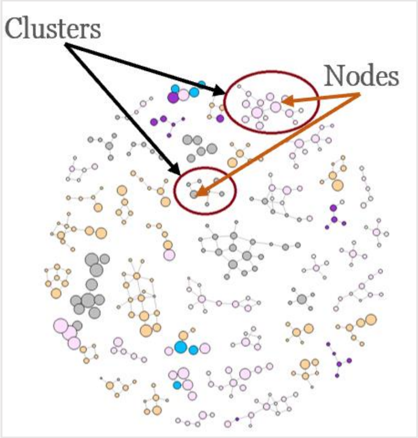
Image sourced from: https://doi.org/10.1007/s00262-020-02833-z

Motivation and Objective

- Similarity among TCRs sequence directly influences the antigen recognition breadth. Using network analysis, which allows interrogation of sequence similarity, can add an important layer of information.
- Draw quantitative analysis of TCR repertoire diversity in the 'longer overall survival' and the 'shorter overall survival' cohorts can provide a better understanding of the immune landscape involving the T cell response.
- Help to develop tools to improve patient stratification, prediction of disease outcome, and patient response to treatments.
- Develop statistical methods to prioritize the important network properties that are associated with the clinical outcome of increased overall survival.

Data Preview

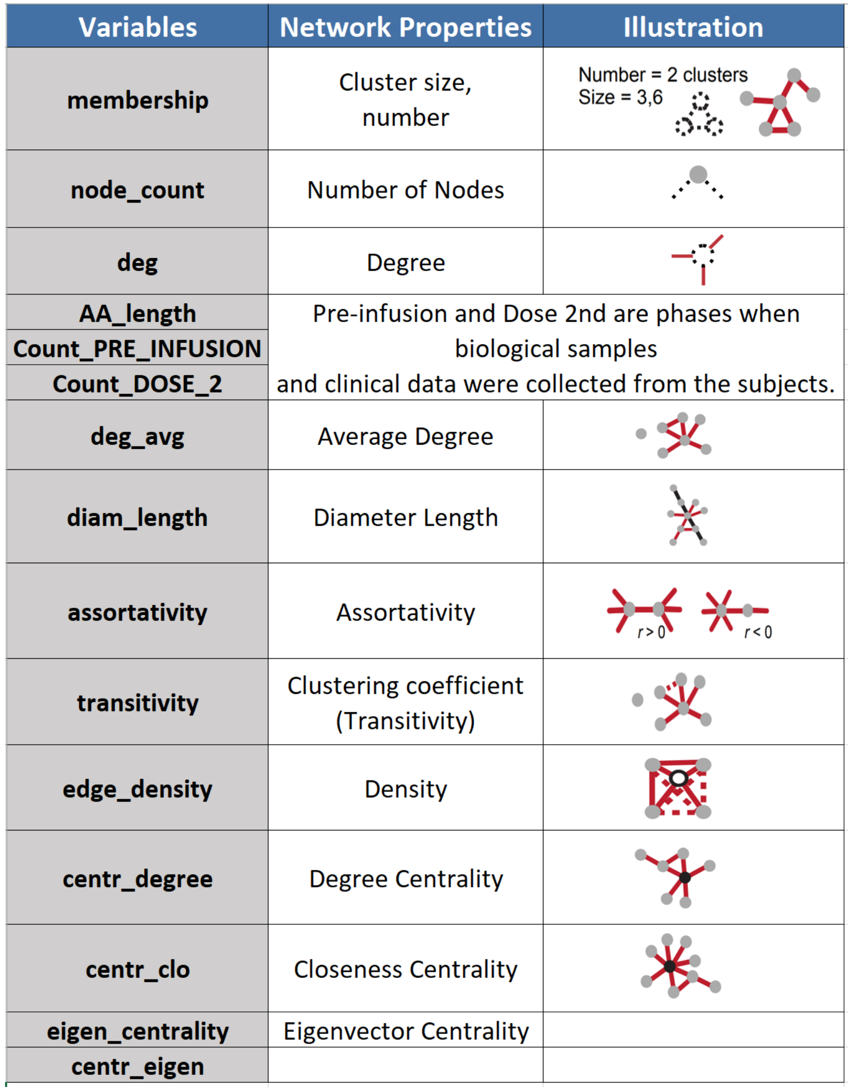
The data set comprises a total of 15 network properties at cluster level. For each patient, the overall survival month data is also provided. The median overall survival months (*OS_mon*) is 20.3 and patients with '*OS_mon* ≥ 20.3' have a higher survival chance than the other patients.



Elements of TCR Network Repertoire

Challenges

- TCR repertoire is continually shaping since it adapts to the health and the environmental factors of the patient.
- Heterogeneous nature of the TCR repertoire makes it difficult to perform statistical inference or machine learning directly between subjects.
  - TCR repertoire is never the same for any two patients. Less than 20% overlap is observed in the TCR repertoires for the same subject.
  - TCR network data is a mix of global and local variables.
  - Global variables are described by a single set of values, while the local variables have varying lengths.
- Heterogeneity complicates the data simulation process essential for simulation study.



Variables	Network Properties	Illustration

membership	Cluster size, number	Number = 2 clusters
		Size = 3,6

node_count	Number of Nodes	
deg	Degree	
AA_length	Pre-infusion and Dose 2nd are phases when biological samples and clinical data were collected from the subjects.	
Count_PRE_INFUSION		
Count_DOSE_2		
deg_avg	Average Degree	
diam_length	Diameter Length	
assortativity	Assortativity	$r > 0$ $r < 0$
transitivity	Clustering coefficient (Transitivity)	
edge_density	Density	
centr_degree	Degree Centrality	
centr_clo	Closeness Centrality	
eigen_centrality	Eigenvector Centrality	
centr_eigen		

Image sourced from: https://doi.org/10.1038/s41467-019-09278-8

Thesis Outline

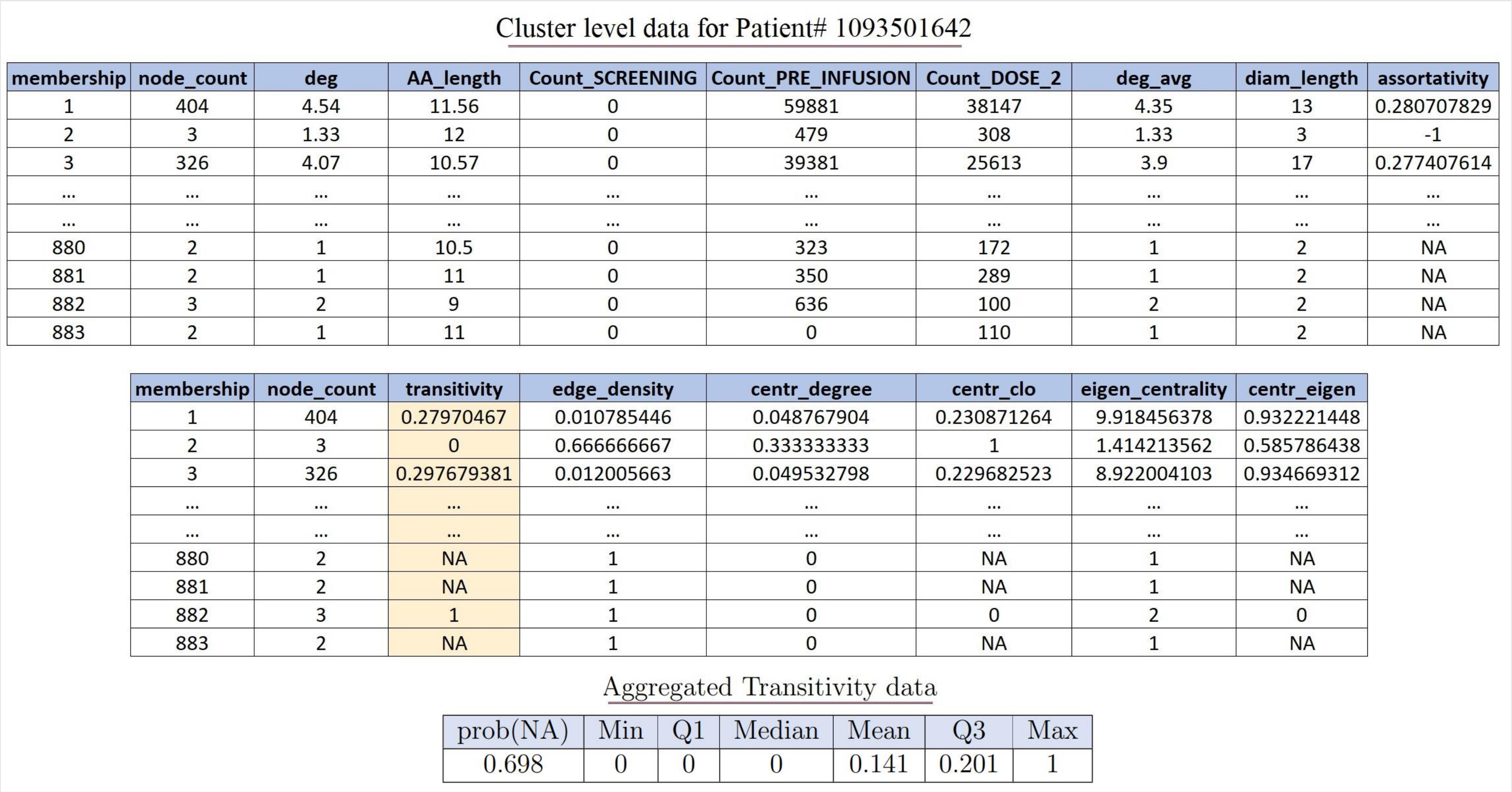
- Formulate a technique to handle the heterogeneity in the TCR network data.
- Use group variable selection models to identify the significant network properties – Group Lasso (Yuan and Lin, 2005)
- Identify the top network features from the network properties – Lasso (Tibshirani, 1996), and Exclusive Lasso (Zhou, Jin and Hoi, 2010)
- Hyperparameter Tuning – Use Cross-Validation (CV) and Permutation Assisted Tuning (PAT) for deducing an optimal tuning parameter.
- Use simulation study to compare and evaluate model performances — Sensitivity, False Discovery Rate, F1 score, Power, and Stability.

Contribution

- Proposed a strategy to extract features from heterogeneous global/local TCR network properties.
- Developed a novel statistical method to prioritize the network properties that are associated with the outcome of interest, based on the extracted network features.
- Proposed a procedure to simulate network properties using the real data, to mimic real property distributions and correlation structure.
- Demonstrated proposed method and schemes via both real data analysis and simulation study.

2 Methods

Feature Extraction and Aggregation



Cluster level data for Patient# 1093501642

membership	node_count	deg	AA_length	Count_SCREENING	Count_PRE_INFUSION	Count_DOSE_2	deg_avg	diam_length	assortativity
1	404	4.54	11.56	0	59881	38147	4.35	13	0.280707829
2	3	1.33	12	0	479	308	1.33	3	-1
3	326	4.07	10.57	0	39381	25613	3.9	17	0.277407614
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
880	2	1	10.5	0	323	172	1	2	NA
881	2	1	11	0	350	289	1	2	NA
882	3	2	9	0	636	100	2	2	NA
883	2	1	11	0	0	110	1	2	NA

membership	node_count	transitivity	edge_density	centr_degree	centr_clo	eigen_centrality	centr_eigen
1	404	0.27970467	0.010785446	0.048767904	0.230871264	9.918456378	0.932221448
2	3	0	0.666666667	0.333333333	1	1.414213562	0.585786438
3	326	0.297679381	0.012005663	0.049532798	0.229682523	8.922004103	0.934669312
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
880	2	NA	1	0	NA	1	NA
881	2	NA	1	0	NA	1	NA
882	3	1	1	0	0	2	0
883	2	NA	1	0	NA	1	NA

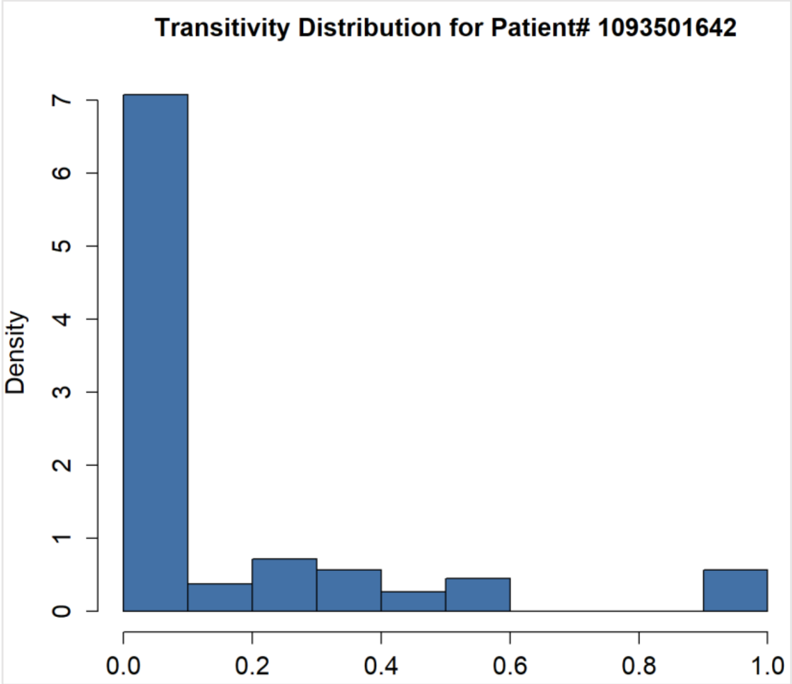
Aggregated Transitivity data

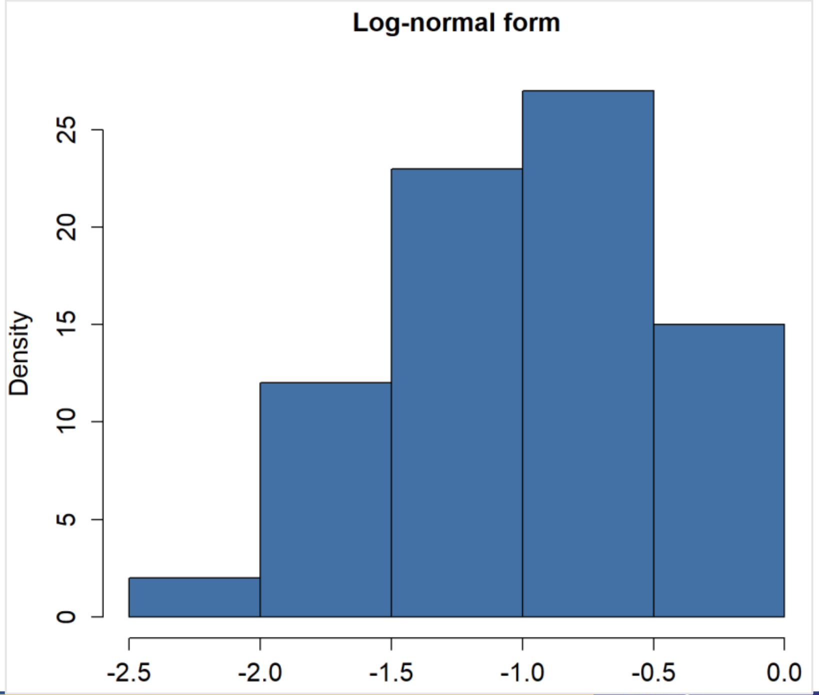
prob(NA)	Min	Q1	Median	Mean	Q3	Max
0.698	0	0	0	0.141	0.201	1

Aggregated Network Data Layout for a Patient

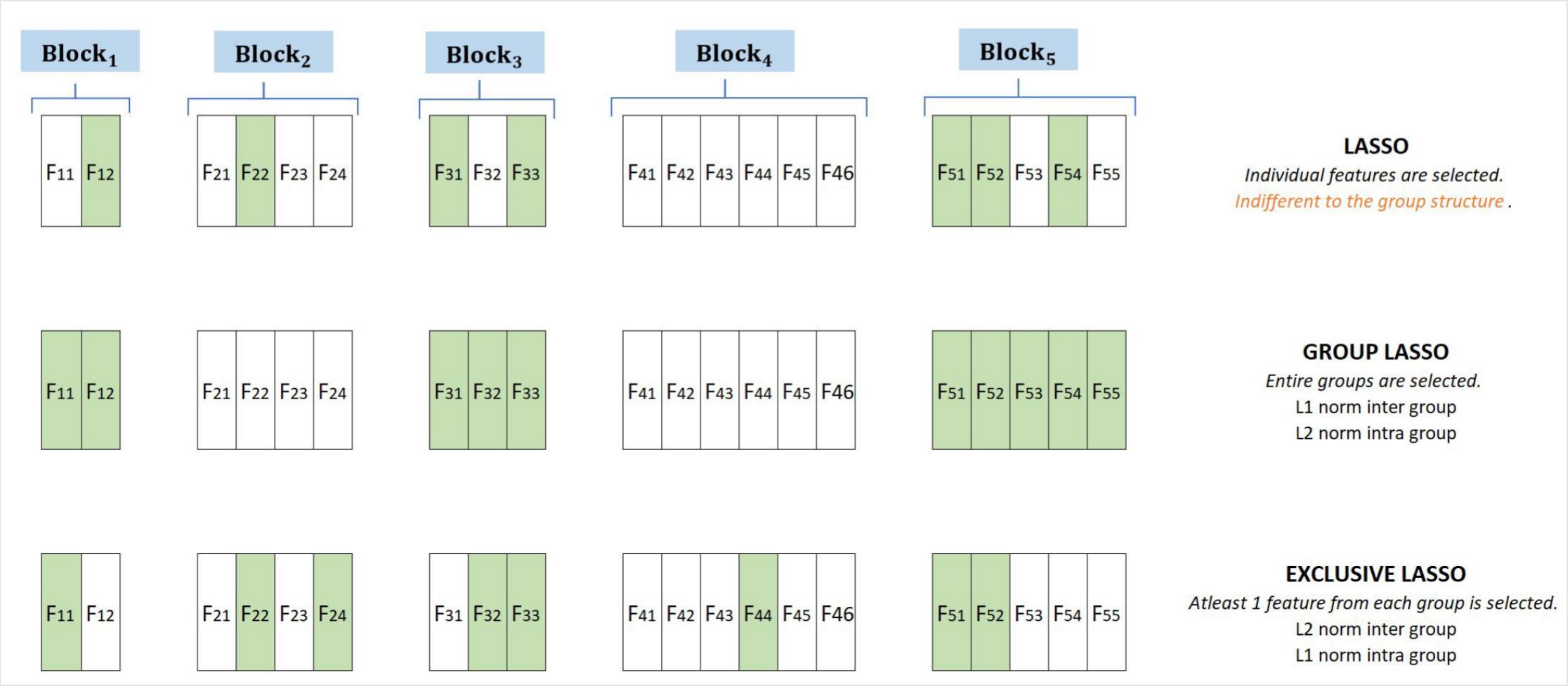
	Membership	Node_Count	. . .	Centr_Eigen
Patient-1	# of clusters	(Min, Q1, Median, Mean, Q3, Max)	(..),...,(..)	(prob(NA), Min, Q1, Median, Mean, Q3, Max)

By deducing the summary statistics of the network properties, we try to approximate the likely distributions of the network properties. In the below density plots, Transitivity data has the distribution shown on the left. The log-normal form on the right is approximately 'normal'. We will utilize this information when simulating data for these network properties.





Variable Selection Techniques



LASSO

Individual features are selected.

Indifferent to the group structure .

GROUP LASSO

Entire groups are selected.

L1 norm inter group

L2 norm intra group

EXCLUSIVE LASSO

Atleast 1 feature from each group is selected.

L2 norm inter group

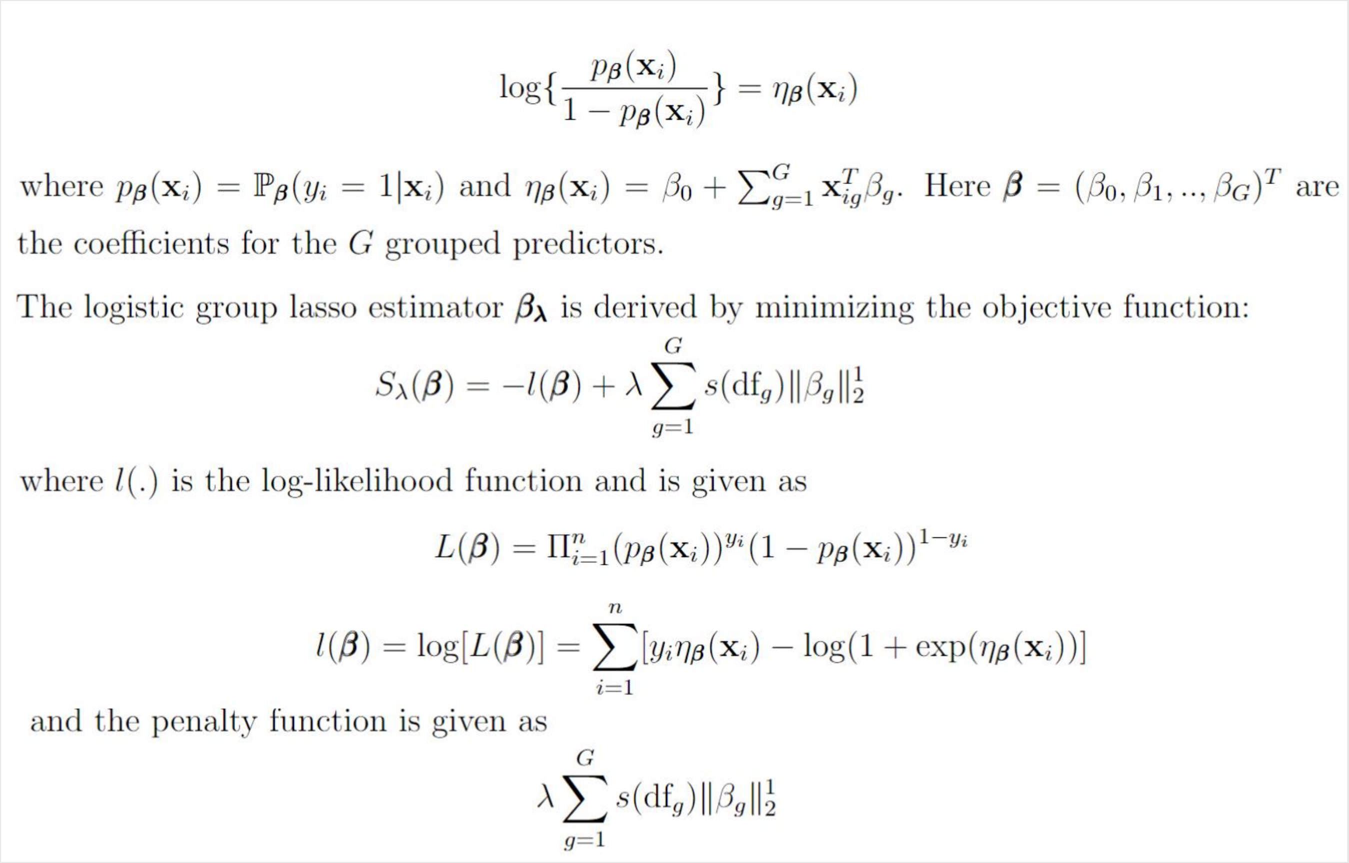
L1 norm intra group

GROUP LASSO: Performs variable selection on grouped variables (feature blocks)

Assume we have iid observations ($\mathbf{x_i}, \mathrm{y}_i$) where i = 1, …, n of a p-dimensional vector $\mathbf{x_i} \in \mathrm{R}^p$, binary response **y** = $(\mathrm{y}_1, .., \mathrm{y}_n\,)^T$ where ($\mathrm{y}_i \in 0, 1$) for all the n observations.

The predictors $\mathbf{x_i}$ = $(\mathrm{x}_{i1}, \ldots, \mathrm{x}_{ip})^T$ when grouped (based on factors) to form a total of G groups of predictors can be written as $\mathbf{x_i}$ = $(\mathrm{x}_{i1}, \ldots, \mathrm{x}_{iG})^T$ using the grouped variables $\mathrm{x}_{ig} \in \mathrm{R}^{\mathrm{df}_g}$ where g = 1, …, G and $\mathbf{df}_g$ is the degrees of freedom for the g$^{\mathrm{th}}$ predictor group.

The log of odds (logit) for a Logistic Regression model can be written as:



$$\log\{\frac{p_{\boldsymbol{\beta}}(\mathbf{x}_i)}{1-p_{\boldsymbol{\beta}}(\mathbf{x}_i)}\} = \eta_{\boldsymbol{\beta}}(\mathbf{x}_i)$$

where $p_{\boldsymbol{\beta}}(\mathbf{x}_i) = \mathbb{P}_{\boldsymbol{\beta}}(y_i = 1|\mathbf{x}_i)$ and $\eta_{\boldsymbol{\beta}}(\mathbf{x}_i) = \beta_0 + \sum_{g=1}^{G} \mathbf{x}_{ig}^T \beta_g$. Here $\boldsymbol{\beta} = (\beta_0, \beta_1, .., \beta_G)^T$ are the coefficients for the G grouped predictors.

The logistic group lasso estimator $\boldsymbol{\beta_\lambda}$ is derived by minimizing the objective function:

$$S_\lambda(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \lambda \sum_{g=1}^{G} s(\mathrm{df}_g)\|\beta_g\|_2^1$$

where $l(.)$ is the log-likelihood function and is given as

$$L(\boldsymbol{\beta}) = \Pi_{i=1}^{n}(p_{\boldsymbol{\beta}}(\mathbf{x}_i))^{y_i}(1-p_{\boldsymbol{\beta}}(\mathbf{x}_i))^{1-y_i}$$

$$l(\boldsymbol{\beta}) = \log[L(\boldsymbol{\beta})] = \sum_{i=1}^{n}[y_i\eta_{\boldsymbol{\beta}}(\mathbf{x}_i) - \log(1+\exp(\eta_{\boldsymbol{\beta}}(\mathbf{x}_i))]$$

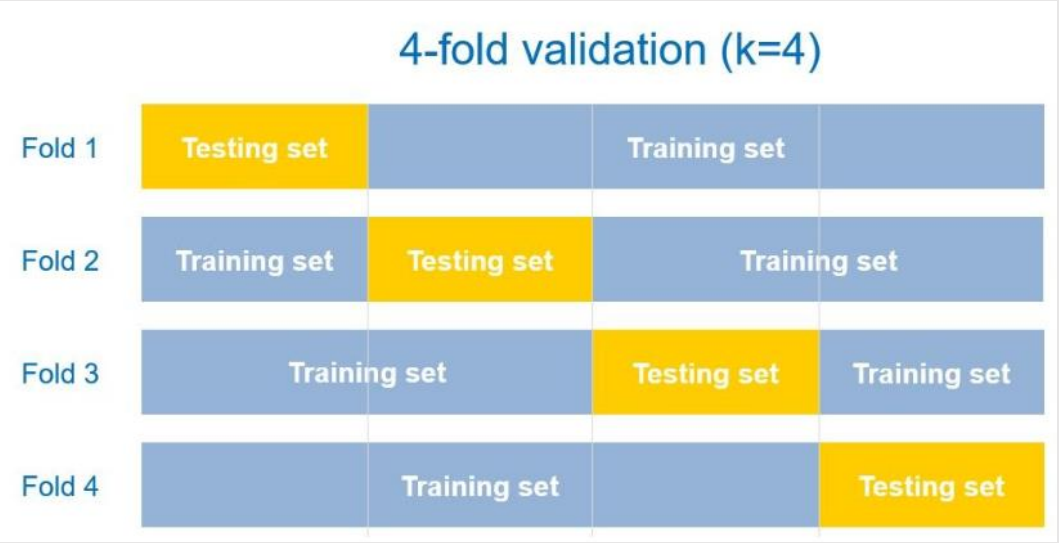
and the penalty function is given as

$$\lambda \sum_{g=1}^{G} s(\mathrm{df}_g)\|\beta_g\|_2^1$$

The tuning parameter $\lambda \geq 0$, controls the amount of penalization. The function $s(.)$ is used to rescale the penalty with respect to the dimensionality of the parameter vector βg. In the penalty function $\|\beta g\|_2^1$ implies L_1 –norm inter-group and L_2 –norm intra-group.

Hyperparameter Tuning: For finding an optimal value for the tuning parameter λ

Cross-validation is a statistical method for estimating machine learning model performance (or accuracy) through resampling. Prevents overfitting in a predictive model, especially when the amount of data available is limited. For hyperparameter tuning use different values of λ for each 'sub-problem'. The λ value which gives the lowest test error is chosen.



4-fold validation (k=4)

Fold 1: Testing set | Training set

Fold 2: Training set | Testing set | Training set

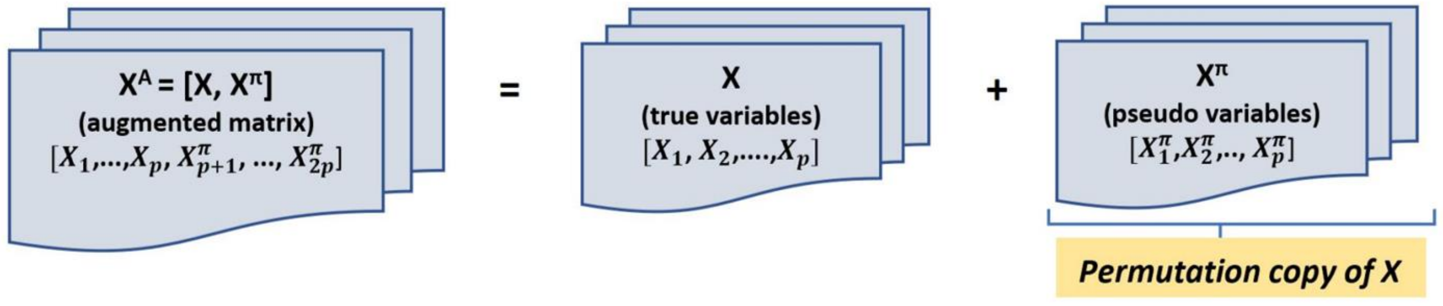
Fold 3: Training set | Testing set | Training set

Fold 4: Training set | Testing set

Cross Validation with k=4 folds

Image sourced from: https://www.mathworks.com/discovery/cross-validation.html

Permutation Assisted Tuning creates permutation copy (set of pseudo-variables) of the original set of predictor variables. Augmenting the pseudo-variables to the original set to increases the dimension of the predictor set while keeping the sample size constant.



Augmented matrix in Permutation Assisted Tuning

Objective is to disrupt the correlation structure of the explanatory variable with the response variable by creating the permutation copy. Using the augmented matrix, the model is enforced to tune the λ values such that only the true variables are picked. The variable selection is stabilized by using different permutations iteratively and evaluating the frequency of the selected variables across those iterations.

So far, Permutation Assisted Tuning has been used only on Lasso models and is know to have lower false positives than cross-validation. We add novelty by extending the application of permutation tuning to Group Lasso model and formulating the required setup.

Group Lasso with Permutation Assisted Tuning

The augmented design matrix can be rewritten as x_i^A = (x_{i1}, ..., x_{iG},x_{i1}^{π}, ..., x_{iG}^{π})^T using the grouped variable $x_{ig} \in R^{df_g}$ where g = 1, ..., 2G and the grouping of [1, 2, .., G] groups = [G+1, G+2, ..., 2G] groups. Note that the grouping strategy used for the original variables is retained and is also applied on to the pseudo-variables. The objective function $S_\lambda(\beta^A)$ is as below.

$$S_\lambda(\boldsymbol{\beta}^A) = -\sum_{i=1}^{n}[y_i\eta_{\boldsymbol{\beta}^A}(\mathbf{x}_i) - \log(1+\exp(\eta_{\boldsymbol{\beta}^A}(\mathbf{x}_i)))] + \lambda \boxed{\sum_{g=1}^{2G} s(\mathrm{df}_g)\|\beta_g^A\|_2^1}$$

Let W_g be an importance metric for the gth variable group. Let C_π be the benchmark to separate active variables from inactive ones. $\hat{S}_\pi$ is the estimator of true variables under the permutation π. In terms of increasing λ values, those original variables will be selected which have an importance metric $W_g > C_\pi$.

$$W_g = \sup\{\lambda : \hat{\beta}_g^A(\lambda) \neq 0\}; g = 1, ..., 2G$$

$$C_\pi = \max_{(G+1)\leq g\leq 2G}(W_g)$$

$$\hat{S}_\pi = \{g : W_g > C_\pi, g = 1, ..., G\}$$

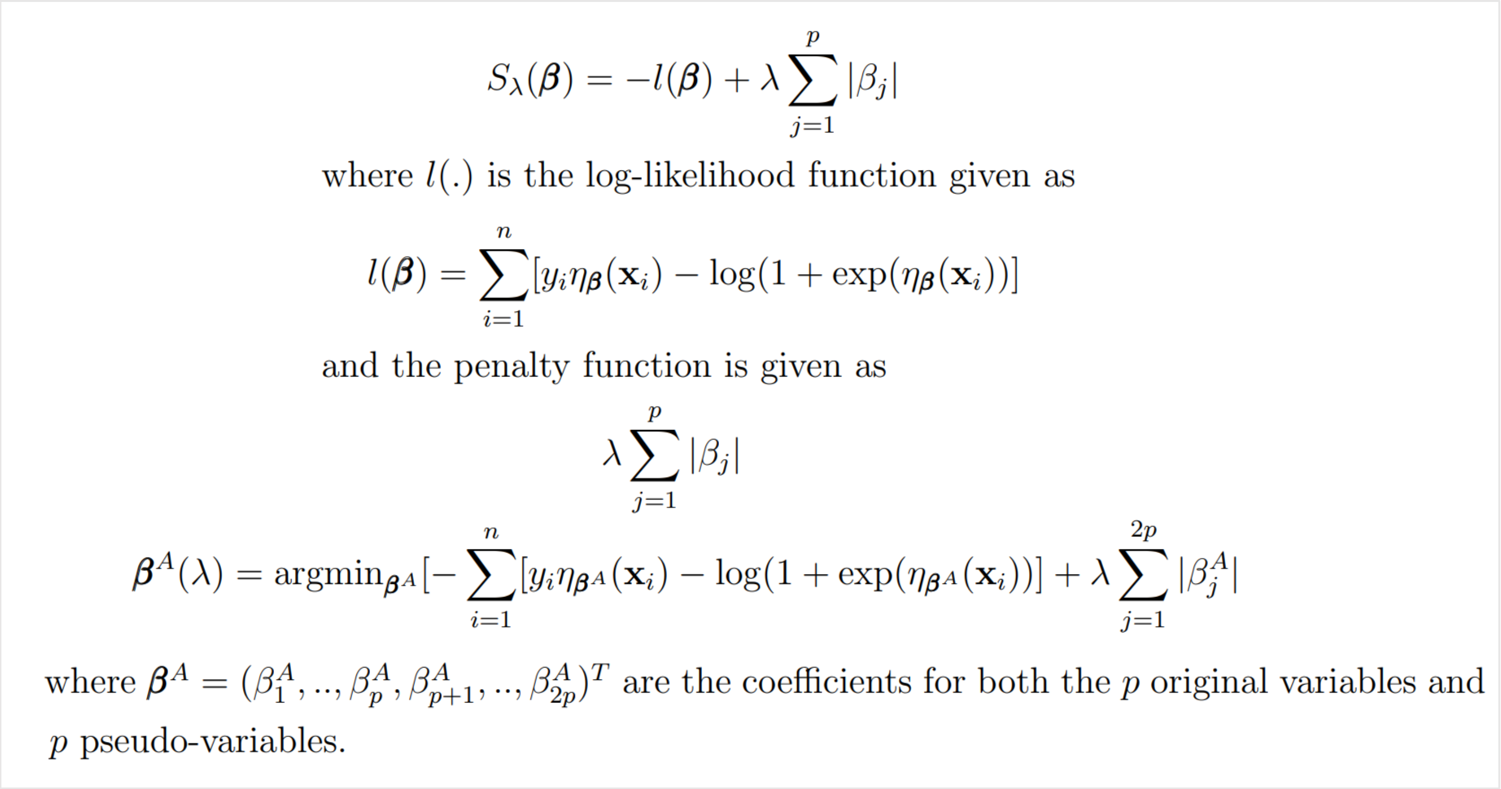
In terms of decreasing λ values, iterate through all λ values till the point where all pseudo-variables have W_g = 0 where g ∈ G+1, ..., 2G.

LASSO: Least Absolute Shrinkage and Selection Operator

Lasso is indifferent to group structure.

The penalty term |βj| implies $\mathbf{L_1}$ −norm among all the variables.

We use both Cross-validation and Permutation assisted tuning to find the optimal tuning parameter, λ.



$$S_\lambda(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|$$

where $l(.)$ is the log-likelihood function given as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \eta_{\boldsymbol{\beta}}(\mathbf{x}_i) - \log(1 + \exp(\eta_{\boldsymbol{\beta}}(\mathbf{x}_i))]$$

and the penalty function is given as

$$\lambda \sum_{j=1}^{p} |\beta_j|$$

$$\boldsymbol{\beta}^A(\lambda) = \text{argmin}_{\boldsymbol{\beta}^A}[-\sum_{i=1}^{n}[y_i \eta_{\boldsymbol{\beta}^A}(\mathbf{x}_i) - \log(1 + \exp(\eta_{\boldsymbol{\beta}^A}(\mathbf{x}_i))] + \lambda \sum_{j=1}^{2p} |\beta_j^A|$$

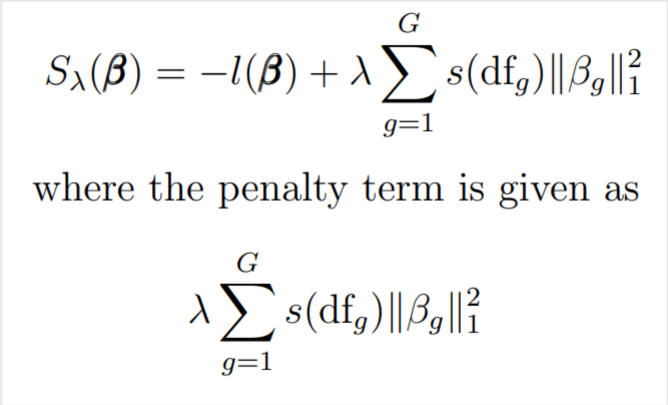
where $\boldsymbol{\beta}^A = (\beta_1^A, .., \beta_p^A, \beta_{p+1}^A, .., \beta_{2p}^A)^T$ are the coefficients for both the p original variables and p pseudo-variables.

EXCLUSIVE LASSO: Performs variable selection from all grouped variables (feature blocks)

The penalty function of Exclusive Lasso is $\|\beta g\|_1^2$ which implies $\mathbf{L_2}$ –norm inter-group and $\mathbf{L_1}$ –norm intra-group. Therefore, it selects at least one variable from each group.

The function *s*(.) is used to rescale the penalty with respect to the dimensionality of the parameter vector βg. The tuning parameter λ ≥ 0, controls the amount of penalization. An optimal value of λ can be derived using cross-validation.

However, permutation assisted tuning cannot be used on Exclusive Lasso model when using decreasing λ values.



$$S_\lambda(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \lambda \sum_{g=1}^{G} s(\mathrm{df}_g)||\beta_g||_1^2$$

where the penalty term is given as

$$\lambda \sum_{g=1}^{G} s(\mathrm{df}_g)||\beta_g||_1^2$$

Some Nomenclature

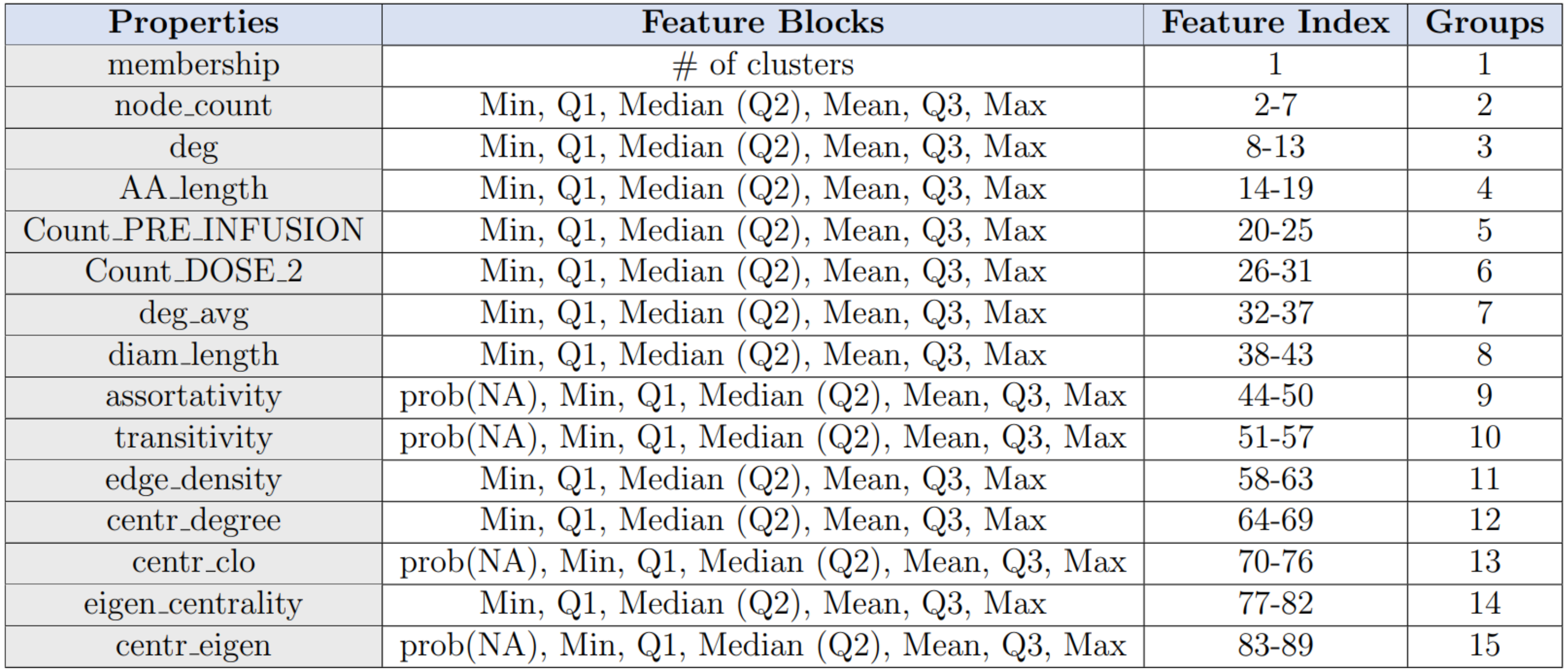
- *Plasso*: Lasso with Permutation assisted tuning (Yang *et al.,* 2020)
- *Lasso_CV*: Lasso with Cross-validation.
- *Group Plasso*: Group Lasso model using Permutation assisted tuning for finding the tuning parameter.
- *Group Lasso_CV*: Group Lasso model using Cross-validation technique for tuning parameter.
- *Feature Block/Group*: Aggregated summary statistics for a TCR network property.
- *Network Features*: Individual summary statistics.

3 Implementation

- **Prioritize significant network properties:** Use Group Lasso_CV and Group Plasso models to identify the significant feature blocks.

- **Select top network features:** Use Lasso_CV and Plasso models to identify the top features. Apply Exclusive Lasso to confirm the variables selected using the Lasso models.

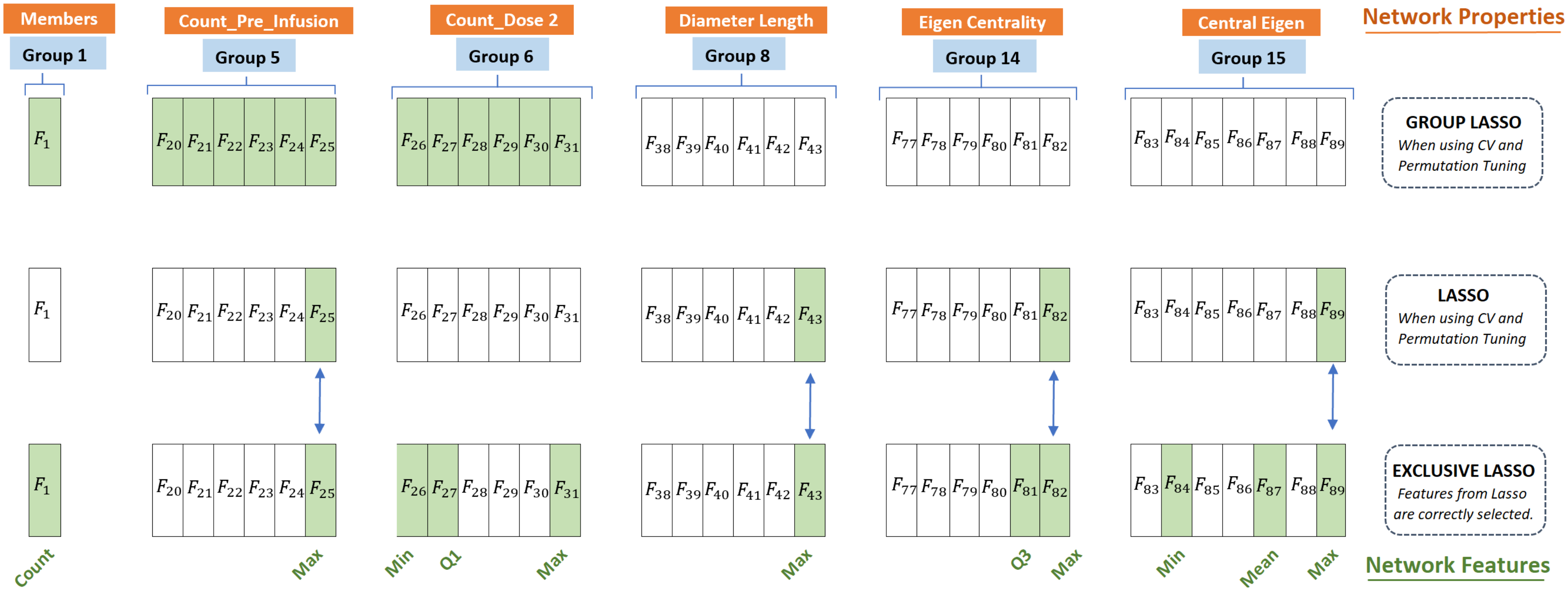
TCR Network Properties, Features Indexes, and Group Indexes



Properties	Feature Blocks	Feature Index	Groups
membership	# of clusters	1	1
node_count	Min, Q1, Median (Q2), Mean, Q3, Max	2-7	2
deg	Min, Q1, Median (Q2), Mean, Q3, Max	8-13	3
AA_length	Min, Q1, Median (Q2), Mean, Q3, Max	14-19	4
Count_PRE_INFUSION	Min, Q1, Median (Q2), Mean, Q3, Max	20-25	5
Count_DOSE_2	Min, Q1, Median (Q2), Mean, Q3, Max	26-31	6
deg_avg	Min, Q1, Median (Q2), Mean, Q3, Max	32-37	7
diam_length	Min, Q1, Median (Q2), Mean, Q3, Max	38-43	8
assortativity	prob(NA), Min, Q1, Median (Q2), Mean, Q3, Max	44-50	9
transitivity	prob(NA), Min, Q1, Median (Q2), Mean, Q3, Max	51-57	10
edge_density	Min, Q1, Median (Q2), Mean, Q3, Max	58-63	11
centr_degree	Min, Q1, Median (Q2), Mean, Q3, Max	64-69	12
centr_clo	prob(NA), Min, Q1, Median (Q2), Mean, Q3, Max	70-76	13
eigen_centrality	Min, Q1, Median (Q2), Mean, Q3, Max	77-82	14
centr_eigen	prob(NA), Min, Q1, Median (Q2), Mean, Q3, Max	83-89	15

Implementation Results:

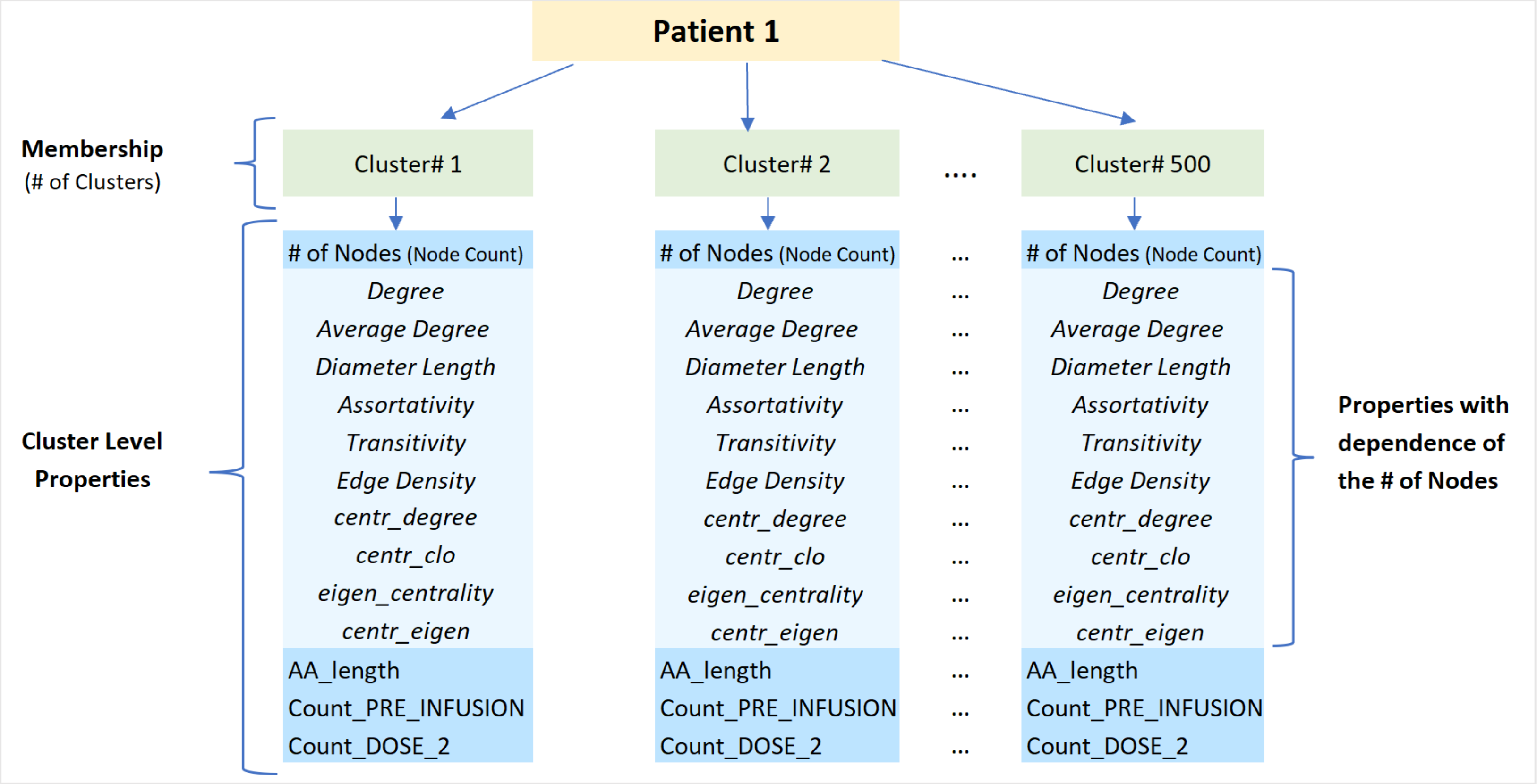
- Group Lasso Models selected: Groups 1, 5, 6
- Lasso and Exclusive Lasso Models selected: 25, 43, 82, 89 (part of Groups: 5, 8, 14, 15)


Members
Group 1
Count_Pre_Infusion
Group 5
Count_Dose 2
Group 6
Diameter Length
Group 8
Eigen Centrality
Group 14
Central Eigen
Group 15
Network Properties
F_1
F_{20} F_{21} F_{22} F_{23} F_{24} F_{25}
F_{26} F_{27} F_{28} F_{29} F_{30} F_{31}
F_{38} F_{39} F_{40} F_{41} F_{42} F_{43}
F_{77} F_{78} F_{79} F_{80} F_{81} F_{82}
F_{83} F_{84} F_{85} F_{86} F_{87} F_{88} F_{89}
GROUP LASSO
When using CV and Permutation Tuning
LASSO
When using CV and Permutation Tuning
EXCLUSIVE LASSO
Features from Lasso are correctly selected.
Count
Max
Min
Q1
Max
Max
Q3
Max
Min
Mean
Max
Network Features

4 Simulation Study

Data Simulation Scheme (for Explanatory Variables)

- Unravel dependencies and correlations between the explanatory variables.
- Approximate the density functions of the variables.
- Simulate the heterogeneous network data and then aggregate the summary statistics for 1000 dummy patients.



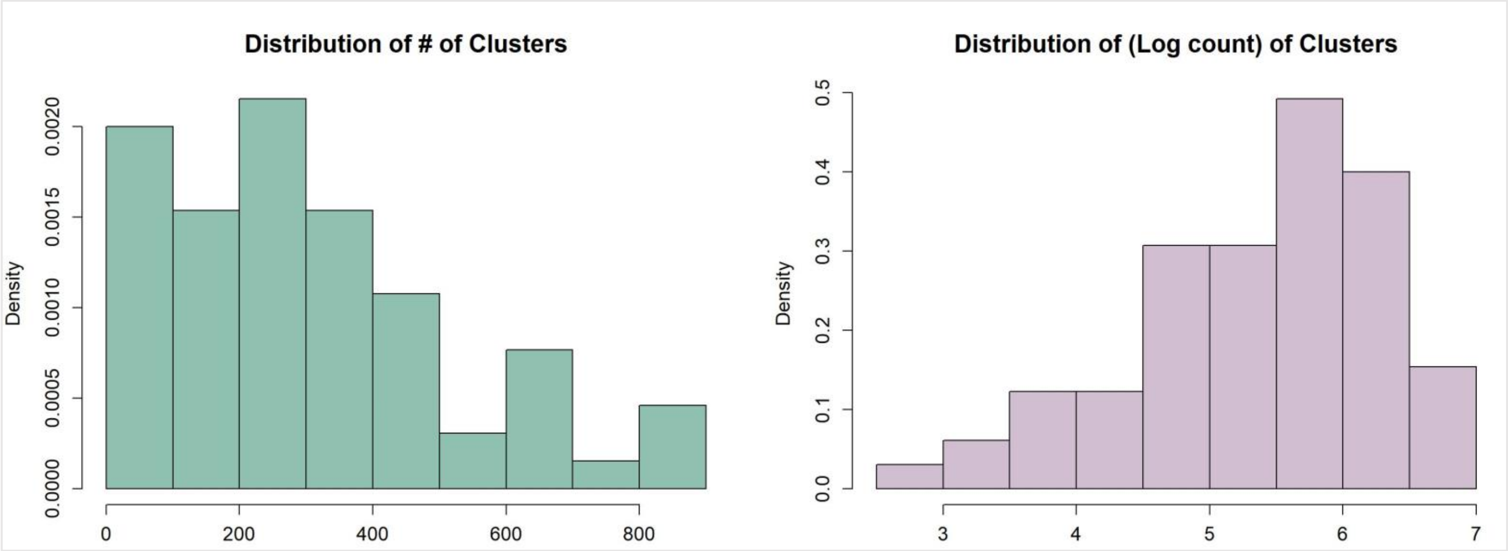
Patient 1

Membership
(# of Clusters)

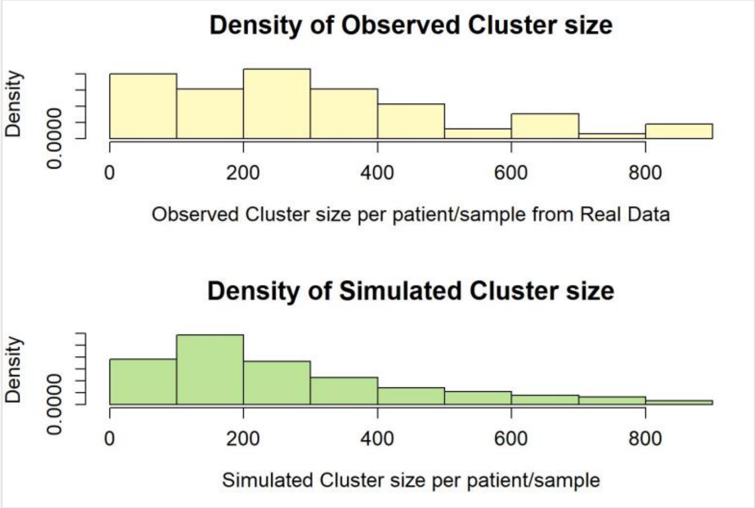
Cluster Level Properties

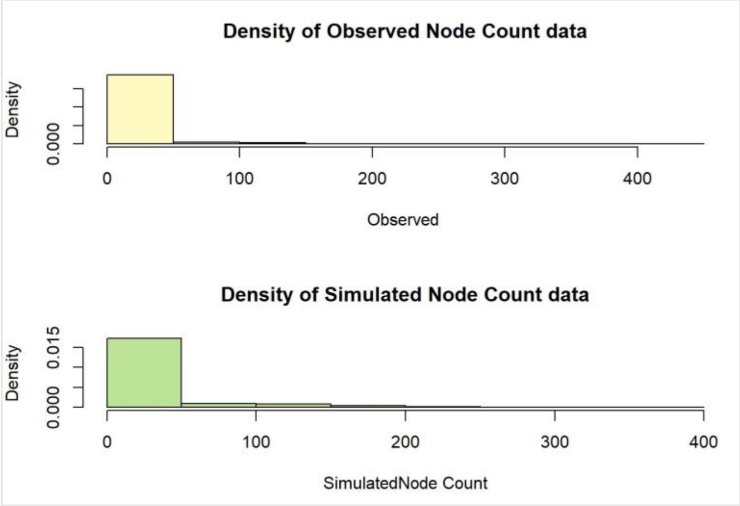
Cluster# 1	Cluster# 2	....	Cluster# 500
# of Nodes (Node Count)	# of Nodes (Node Count)	...	# of Nodes (Node Count)
Degree	*Degree*	...	*Degree*
Average Degree	*Average Degree*	...	*Average Degree*
Diameter Length	*Diameter Length*	...	*Diameter Length*
Assortativity	*Assortativity*	...	*Assortativity*
Transitivity	*Transitivity*	...	*Transitivity*
Edge Density	*Edge Density*	...	*Edge Density*
centr_degree	*centr_degree*	...	*centr_degree*
centr_clo	*centr_clo*	...	*centr_clo*
eigen_centrality	*eigen_centrality*	...	*eigen_centrality*
centr_eigen	*centr_eigen*	...	*centr_eigen*
AA_length	AA_length	...	AA_length
Count_PRE_INFUSION	Count_PRE_INFUSION	...	Count_PRE_INFUSION
Count_DOSE_2	Count_DOSE_2	...	Count_DOSE_2

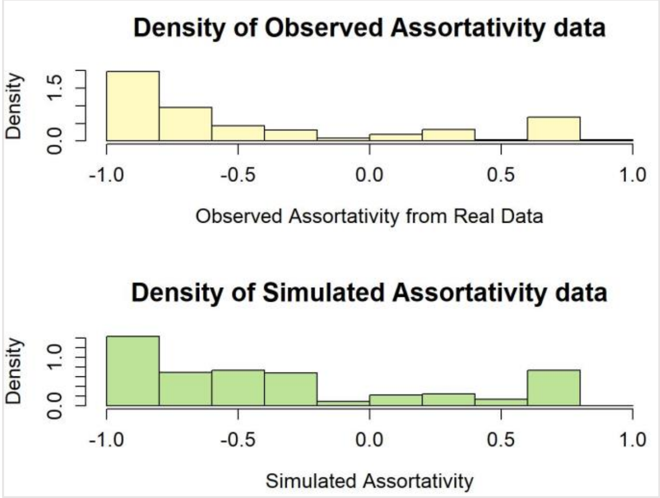
Properties with dependence of the # of Nodes



Properties	Value	Value
Node_Count	2	3
Assortativity	NA	-1 (99.7%)
Transitivity	NA	
Degree	1	
diam_length	2	
deg_avg	1	
edge_density	1	
centr_degree	0	
centr_clo	NA	
eigen_centrality	1	
centr_eigen	NA	



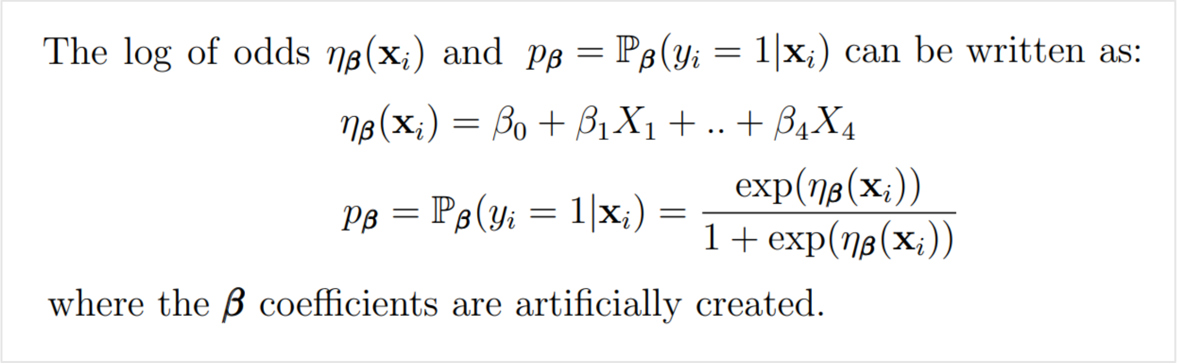




Data Simulation for the Response Variable

- Use the causal variables identified from the Lasso models (feature indexes: 25, 43, 82, 89).

- Fit the causal variables to a logistic regression model and derive the probability of success p_β using artificial beta coefficients.



The log of odds $\eta_{\boldsymbol{\beta}}(\mathbf{x}_i)$ and $p_{\boldsymbol{\beta}} = \mathbb{P}_{\boldsymbol{\beta}}(y_i = 1|\mathbf{x}_i)$ can be written as:

$$\eta_{\boldsymbol{\beta}}(\mathbf{x}_i) = \beta_0 + \beta_1 X_1 + .. + \beta_4 X_4$$

$$p_{\boldsymbol{\beta}} = \mathbb{P}_{\boldsymbol{\beta}}(y_i = 1|\mathbf{x}_i) = \frac{\exp(\eta_{\boldsymbol{\beta}}(\mathbf{x}_i))}{1 + \exp(\eta_{\boldsymbol{\beta}}(\mathbf{x}_i))}$$

where the $\boldsymbol{\beta}$ coefficients are artificially created.

- Sample randomly from a Bernoulli distribution using the function *rbern()* such that 1000 random samples are generated with probability = p_β. Use these values as the simulated response values.

- Down sample the simulated data set to accommodate for the outcome of increased overall survival (y_i=1).

Model Evaluation Criteria: Performance Measures

- *Sensitivity*: Measures the proportion of correctly identifying the true causal variables among the selected variables from a single iteration.

- *False Discovery Rate (FDR):* Measures the frequency of false-positive findings among all variables selected in each iteration.

- *F1 Score*: Harmonic mean of the Sensitivity and (1-FDR). Measures the performance accuracy of the model.

- *Power*: For each causal variable, the power is calculated as the frequency of correct selections among all iterations.

- *Stability*: Computes the consistency of a model. All pairwise combinations of the selected variables list from each iteration are considered. The stability of the two lists of selected variables is determined using the Jaccard's index. The average of all pairs is used as the stability value for that method.

$$J(A_i, A_j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}. \quad \begin{aligned} &= 0 \text{ if the two lists do not overlap} \\ &= 1 \text{ if the two lists contain the same variables.} \end{aligned}$$

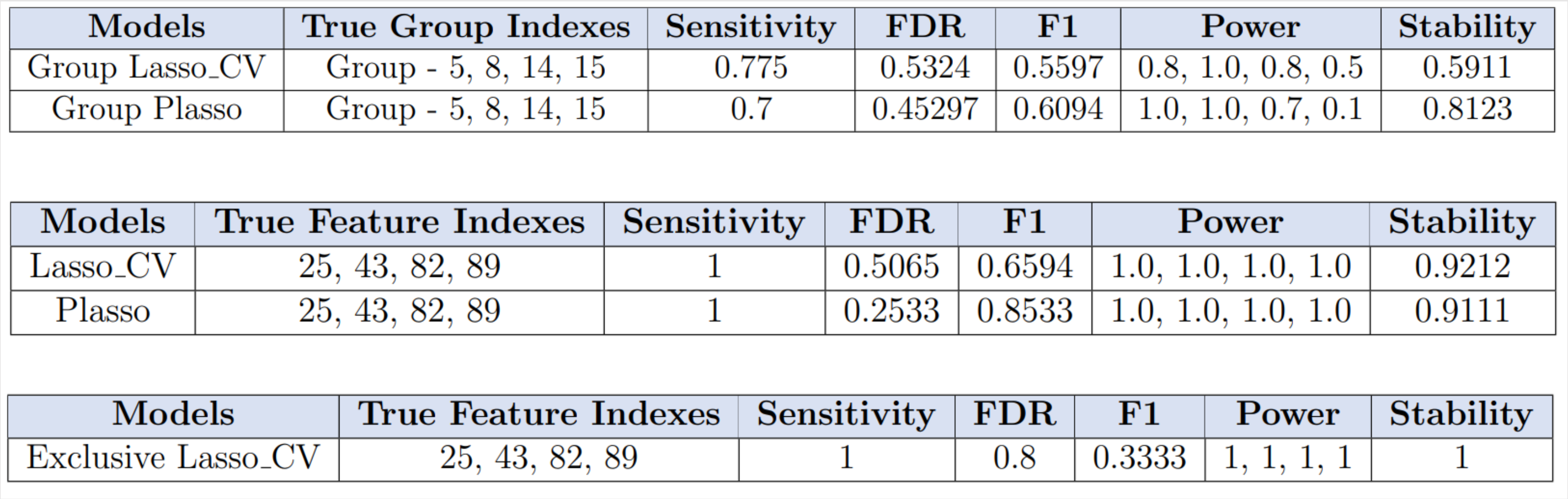
.

5 Results & Discussion

Real Data Analysis: Identifying the significant network variables and top network features

- Group Lasso Models selected: Groups 1, 5, 6 {‘Membership (# of Clusters)’, ‘Count_Pre_Infusion’, and ‘Count_Dose 2’ network properties}
- Lasso and Exclusive Lasso Models selected: 25, 43, 82, 89 (part of Groups: 5, 8, 14, 15){Max. summary statistics for the TCR network properties ‘Count_Pre_Infusion’, ‘diam length’, ‘eigen centrality’, ‘centr eigen’}

Simulation Study: Model Comparison



Models	True Group Indexes	Sensitivity	FDR	F1	Power	Stability
Group Lasso_CV	Group - 5, 8, 14, 15	0.775	0.5324	0.5597	0.8, 1.0, 0.8, 0.5	0.5911
Group Plasso	Group - 5, 8, 14, 15	0.7	0.45297	0.6094	1.0, 1.0, 0.7, 0.1	0.8123

Models	True Feature Indexes	Sensitivity	FDR	F1	Power	Stability
Lasso_CV	25, 43, 82, 89	1	0.5065	0.6594	1.0, 1.0, 1.0, 1.0	0.9212
Plasso	25, 43, 82, 89	1	0.2533	0.8533	1.0, 1.0, 1.0, 1.0	0.9111

Models	True Feature Indexes	Sensitivity	FDR	F1	Power	Stability
Exclusive Lasso_CV	25, 43, 82, 89	1	0.8	0.3333	1, 1, 1, 1	1

Simulation Study: Model Evaluation

- F1 scores: Group Plasso > Group Lasso_CV.

- Group Plasso model has a stability of ~ 81% while Group Lasso CV has only ~ 59% stability. This shows that permutation assisted tuning improved the model stability signifcantly over cross-validation technique.

- F1 scores: Plasso >> Lasso_CV.

- Permutation assisted tuning not only improves the FDR and F1 values but significantly improves the stability of the group variable selection models over cross-validation.

- Using permutation assisted tuning for hyperparameter tuning could be a preferred method for variable selection than cross-validation.

References

- Elliot Naidus, Jerome Bouquet, David Y. Oh, Timothy J. Looney, Hai Yang, Lawrence Fong, Nathan E. Standifer, Li Zhang. "Early Changes In the Circulating T Cells are Associated with Clinical Outcomes after PD-L1 Blockade by Durvalumab in Advanced NSCLC Patients". In: *Cancer Immunology, Immunotherapy* 70:2095-2102 (2021).

- Enkelejda Miho, Rok Roskar, Victor Greiff And Sai T.Reddy. "Large-scale Network Analysis Reveals the Sequence Space Architecture of Antibody Repertoires". In: *Nature Communications* 10:1321 (2019).

- Ming Yuan And Yi Lin. "Model Selection and Estimation in Regression with Grouped Variables". In: *Journal Of The Royal Statistical Society. Series B* 68.Part 1 (2006), Pp. 49-67.

- Songshan Yang, Jiawei Wen, Scott T. Eckert, Yaqun Wang, Dajiang J. Liu, Rongling Wu, Runze Li1 And Xiang Zhan. "Prioritizing Genetic Variants in GWAS with Lasso using Permutation-assisted Tuning". In: *Bioinformatics* 36:3811-7 (2020).

- Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of The Royal Statistical Society. Series B* (Methodological) 58 (1996), Pp. 267-288.

- Yang Zhou, Rong Jin, Steven C. H. Hoi. "Exclusive Lasso for Multi-task Feature Selection". In: *JMLR Workshop And Conference Proceedings: 13th International Conference On Artificial Intelligence And Statistics* (AISTATS) 9 (2010), Pp. 988-995.