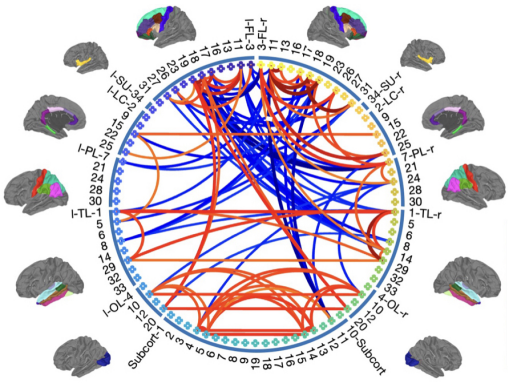
# Uncertainty Quantification for Structure Learning and Interpretable Machine Learning

Lili Zheng
Department of Electrical and Computer Engineering, Rice University

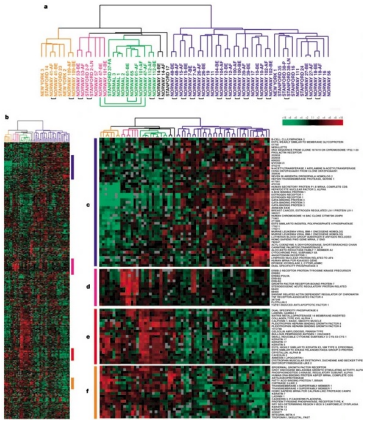
# Table of contents

# Background

# Data Science Propels Discoveries




Functional connectivity from fMRI data

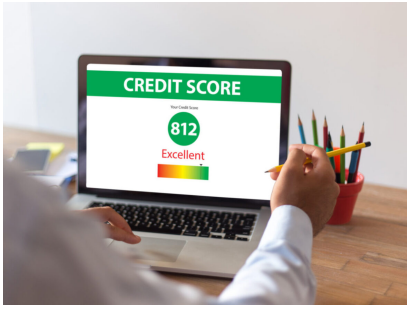


Hierarchical clustering for tumor data

(Perou et al., 2000)

# Data Science Propels Decision-making


Healthcare




Loan approval

Picture source: https://www.aamc.org/news/electronic-health-records-what-will-it-take-make-them-work;

https://auto.economictimes.indiatimes.com/news/auto-technology/us-lawmakers-raise-concerns-over-chinese-self-driving-testing-data-collection/105283633

# Trust in Data-driven Discoveries and Decision-making?

- For discovery: replicability in science
- For decision-making: reduce risk in critical applications

# Trust in Data-driven Discoveries and Decision-making?

- For discovery: replicability in science
- For decision-making: reduce risk in critical applications

**One Potential Solution**

Provide uncertainty quantification (UQ) associated with any data-driven discoveries and for machine learning interpretations!

# Uncertainty Quantification: Challenges in the Modern Era

- Great tools in statistics & machine learning: selective inference, conformal inference, Bayesian inference...
- Numerous challenges from large-scale, complex data and models!

**Rigorous uncertainty quantification in practical scenarios?**

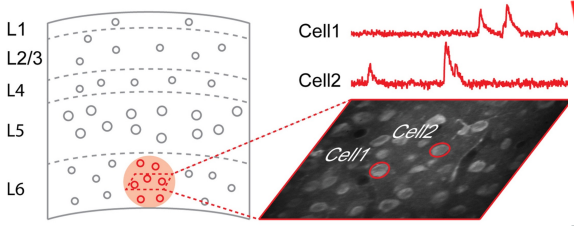
# Uncertainty Quantification: Challenges in the Modern Era

- First part: UQ for graph learning (scientific discovery)
  - Real application challenges: **unconventional data structure**
- Second part: UQ for model-agnostic machine learning interpretations (decision-making)
  - Algorithmic challenges: **large-scale data and black-box models**

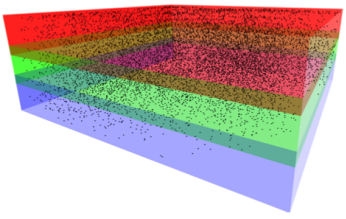
# Uncertainty Quantification for Statistical Structure (Graph) Learning

# Challenges from Data: Erose Measurements

**Erose measurements**: irregular, highly uneven measurements over a large system



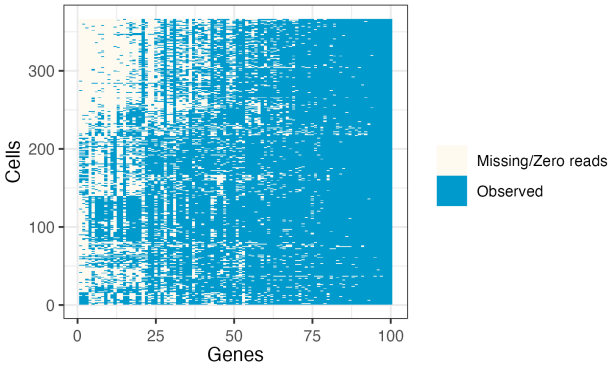

Calcium imaging data in neuroscience (Birkner et al., 2017)



Measurements in semi-overlapping cubes; also called graph quilting (Vinci et al., 2019)

# Challenges from Data: Erose Measurements

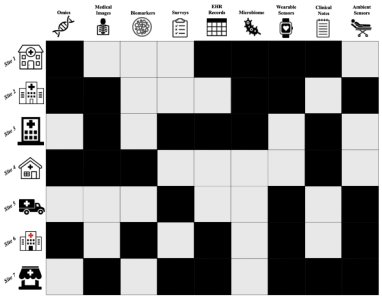
**Erose measurements**: irregular, highly uneven measurements over a large system




Single-cell RNA sequencing (Darmanis et al., 2015)

# Challenges from Data: Erose Measurements

**Erose measurements**: irregular, highly uneven measurements over a large system




**patchwork learning in healthcare** (Rajendran et al., 2023)

**Table 1.** Some examples of unequally spaced surveys.

| Panel A: household surveys for monitoring poverty in developing countries[a] | | |
|---|---|---|
| *Country* | *Survey* | *Survey periods* |
| Bolivia | Encuesta Integrada de Hogares (EIH) | Mar 89, Nov 89, Sept 90, Nov 91, Nov 92, July–Dec 93, July–Dec 94, June 95 |
| Brazil | Pesquisa Nacional por Amostra de Domicillios (PNAD) | Annual surveys since 1971, but surveys not taken in census years 1980 and 1991 |
| Chile | Caracterización Socioeconómica Nacional (CASEN) | 1985, 87, 90, 92, 94, 96 |
| Ethiopia | Welfare monitoring survey | 1995, 97, 98 |
| Ghana | Ghana living standards survey | 1987, 88, 91, 98 |
| Kenya | Welfare monitoring survey | 1992, 94, 97 |
| Kyrgyz Republic | Poverty monitoring survey | 1993, 96, 96, 97, 98 |
| Mexico | Encuesta nacional de Ingreso-Gasto de los hogares (ENIGH) | 1984, 89, 92, 94, 96 |
| Nigeria | National consumer survey | 1980, 85, 92, 96 |
| Panama | Encuesta de Hogares–Mano de Obra (EMO) | 1979, 89, 91, 95, 96 |
| Peru | Encuesta Nacional de Hogares Sobre Medición de Niveles de Vida (ENNIV) | 1985, 90, 91, 94 |
| Senegal | Enquête Démographique et de Santé | 1986, 92, 97 |
| Thailand | Thailand Socio-Economic Survey (SES) | 1975, 81, 86, 88, 90, 92, 94, 96, 98 |

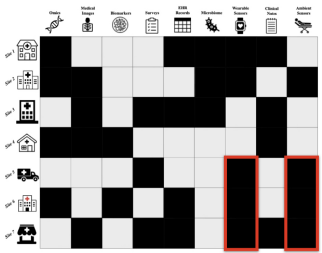
1985, 87, 90, 92, 94, 96
1995, 97, 98
1987, 88, 91, 98

**unevenly spaced time series in econometrics** (Millimet and McDonough, 2017)

# Structure Learning from Erose Measurements?

Common practices

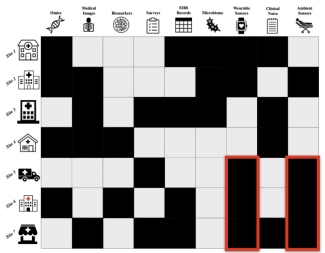
- Downsampling: focus on complete block;
  - throw too much data away!



# Structure Learning from Erose Measurements?

Common practices

- Downsampling: focus on complete block;
  – throw too much data away!
- Ad-hoc imputation + downstream analysis on the imputed;
  – low-rank completion methods?
    - provable mainly for random missingness
    - not low-rank?
    - extra uncertainty from imputation



Focus on graph learning from erose measurements in this talk

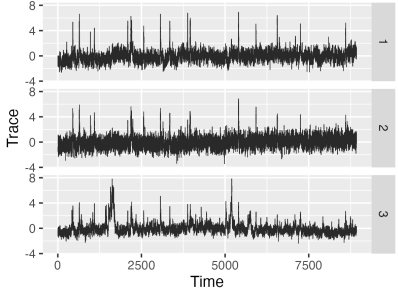
# Why Graph Learning?

## Graphical Model Structural Learning

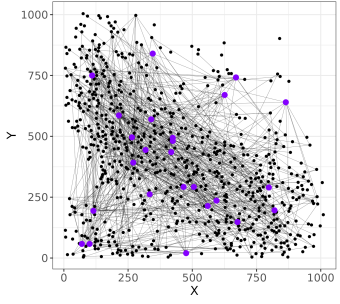
Extract **conditional dependency** relationships:

Edge $(i, j) \iff X_i$ and $X_j$ are conditional dependent given all other nodes.

**Functional Connectivity**: a graph between neurons that reflect their co-firing patterns




$\Rightarrow$




Many applications: gene co-expression networks, sensor networks, statistical physics, ...

# Gaussian Graphical Model Learning from Erose Measurements

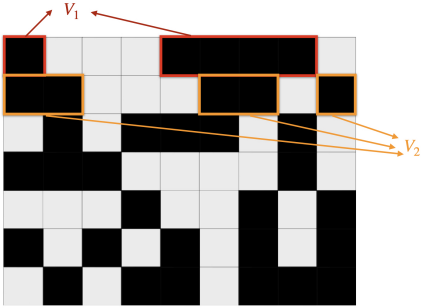
- Focus on Gaussian graphical models in this talk
  - $p$-dimensional $X_1, \dots X_n \sim \mathcal{N}(0, \Theta^{*-1})$;
  - Nodes: $V = [p]$;
  - Edges: $E = \{(i,j) : 1 \leq i, j \leq p, \Theta^*_{i,j} \neq 0\}$;
  - **Goal: identify non-zero entries in $\Theta^*$**

- Erose measurements
  - $X_{i,V_i}$, $1 \leq i \leq n$; $V_i \subset [p]$ are irregular feature subsets
  - Pairwise joint sample size $\{n_{j,k} : 1 \leq j, k \leq p\}$ are highly different
    $n_{j,k} = \sum_{i=1}^{n} \mathbb{1}_{\{j,k \in V_i\}}$

# Gaussian Graphical Model Learning from Erose Measurements

- Erose measurements
  - $X_{i,V_i}$, $1 \leq i \leq n$; $V_i \subset [p]$ are irregular feature subsets
  - Pairwise joint sample size $\{n_{j,k} : 1 \leq j, k \leq p\}$ are highly different

    $n_{j,k} = \sum_{i=1}^{n} \mathbb{1}_{\{j,k \in V_i\}}$



# Prior Works on Gaussian Graphical Models from Partial Observations

Estimation

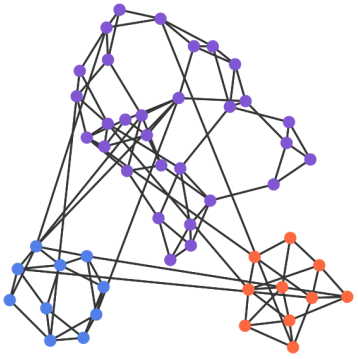
- If all node pairs jointly observed at least twice, plug in covariance estimates into graphical Lasso (Kolar and Xing, 2012; Park et al., 2021)
- Theory assumes missing with the same / similar probability!
- Existing characterization in minimum pairwise sample size
- **Limited insight for our setting**

Inference

- Fully observed data
- Missing independently with same probability
- **Not applicable for our setting**

# Toy Example: UQ Promotes Reliable Graph Learning

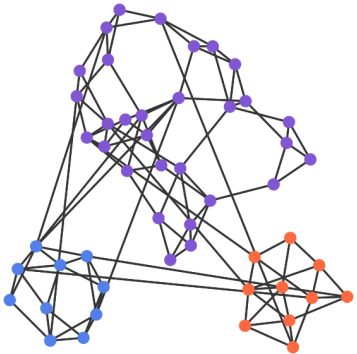
- Toy example: irregular patchwise observations
- $p = 40 + 20 + 20 = 80$ nodes in total



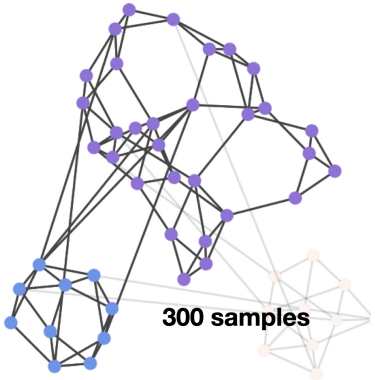
True graph

# Toy Example: UQ Promotes Reliable Graph Learning

- Toy example: irregular patchwise observations
- $p = 40 + 20 + 20 = 80$ nodes in total



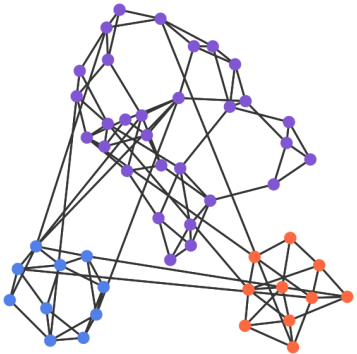
True graph



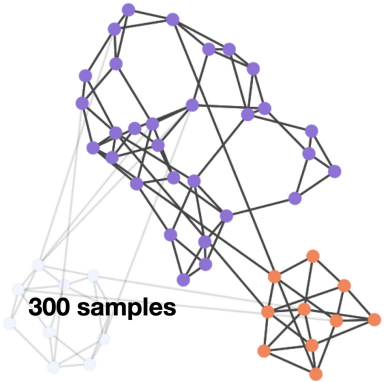

Measurement 1

# Toy Example: UQ Promotes Reliable Graph Learning

- Toy example: irregular patchwise observations
- $p = 40 + 20 + 20 = 80$ nodes in total



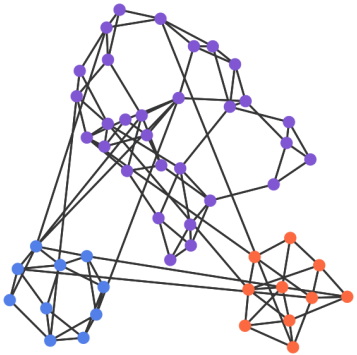
True graph



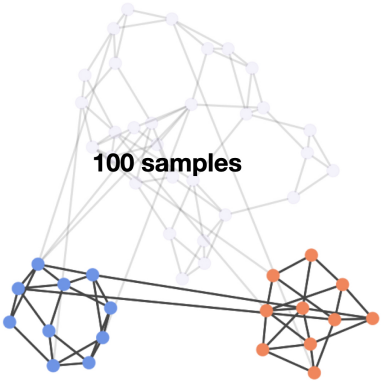

Measurement 2

# Toy Example: UQ Promotes Reliable Graph Learning

- Toy example: irregular patchwise observations
- $p = 40 + 20 + 20 = 80$ nodes in total



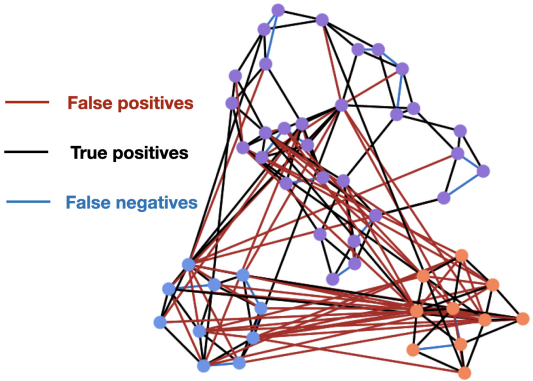
True graph




Measurement 3

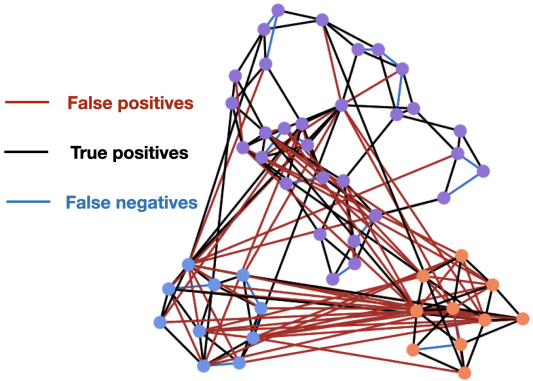
# Toy Example: UQ Promotes Reliable Graph Learning

- Plug-in estimate using graphical lasso

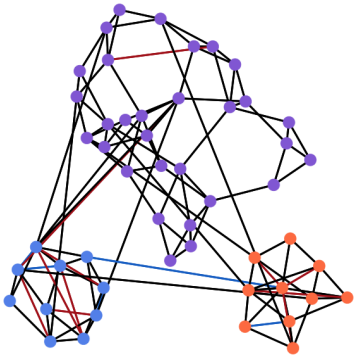


# Toy Example: UQ Promotes Reliable Graph Learning

- Plug-in estimate using graphical lasso




- We develop GI-JOE (**G**raph **I**nference when **J**oint **O**bservations are **E**rose) with FDR control



# Problem Setup and Proposed Method

# Recall: Model Setup

Gaussian graphical model:

- $p$-dimensional $X_1, \dots X_n \sim \mathcal{N}(0, \Theta^{*-1})$;
- Nodes: $V = [p]$;
- Edges: $E = \{(i,j) : 1 \leq i,j \leq p, \Theta_{i,j} \neq 0\}$;

Observations

- $X_{i,V_i}$, $1 \leq i \leq n$; $V_i \subset [p]$ are irregular feature subsets independent from $X_i$
- Pairwise joint sample size $\{n_{j,k} : 1 \leq j,k \leq p\}$ are highly different

# Recall: Model Setup

Observations

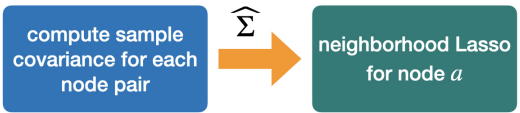
- $X_{i,V_i}$, $1 \leq i \leq n$; $V_i \subset [p]$ are irregular feature subsets independent from $X_i$
- Pairwise joint sample size $\{n_{j,k} : 1 \leq j, k \leq p\}$ are highly different

**Edgewise-testing:** $\mathcal{H}_0 : (a, b) \notin E$ **for** $a, b \in [p]$ **(FDR control later)**

# Edgewise Inference: Debiased Neighborhood Lasso

- Many existing methods are covariance-based (Meinshausen and Bühlmann, 2006; Van de Geer et al., 2014).

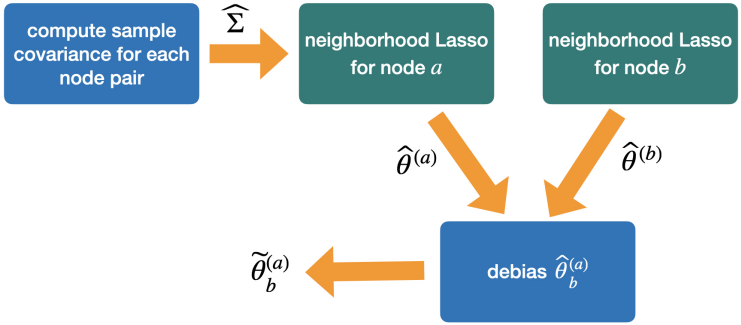
- **Step 1**: Plug in pairwise sample covariance into neighborhood Lasso and perform debiasing :




$$\widehat{\theta}^{(a)} = \arg \min_{\theta \in \mathbb{R}^p, \theta_a = 0} \frac{1}{2} \theta^\top \widehat{\Sigma} \theta - \widehat{\Sigma}_{a,:} \theta + \sum_{j=1}^{p} \lambda_j |\theta_j|,$$

# Edgewise Inference: Debiased Neighborhood Lasso

- **Step 1**: Plug in pairwise sample covariance into neighborhood Lasso and perform debiasing :



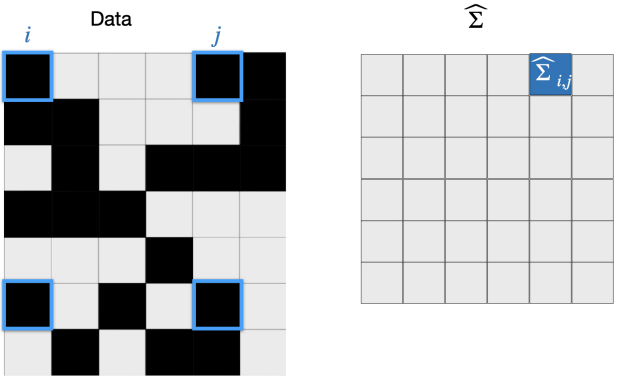

$\widehat{\theta}_b^{(a)}$, $\widetilde{\theta}_b^{(a)}$ are estimates of $\frac{\Theta^*_{a,b}}{\Theta^*_{a,a}}$;

Larger $|\widetilde{\theta}_b^{(a)}|$ indicates strong dependence between $(a, b)$

- **Step 2**: Normal approximation for $\widetilde{\theta}_b^{(a)}$ and variance estimation
  **Challenge: $\widehat{\Sigma}$ computed from irregular data patches**

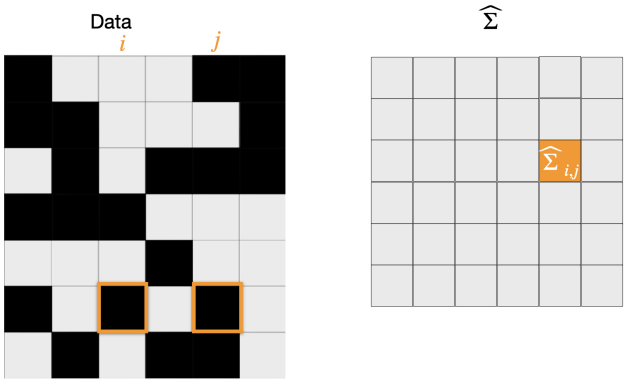
# Edgewise Inference: Debiased Neighborhood Lasso

- **Step 2**: Normal approximation for $\widetilde{\theta}_b^{(a)}$ and variance estimation
  **Challenge: $\widehat{\Sigma}$ computed from irregular data patches**



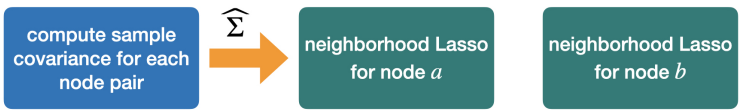
# Edgewise Inference: Debiased Neighborhood Lasso

- **Step 2**: Normal approximation for $\widetilde{\theta}_b^{(a)}$ and variance estimation
  **Challenge: $\widehat{\Sigma}$ computed from irregular data patches**



- **Step 2**: Normal approximation for $\widetilde{\theta}_b^{(a)}$ and variance estimation
  **Challenge: $\widehat{\Sigma}$ computed from irregular data patches**

  All entries of $\widehat{\Sigma}$ play a role!



# Characterization of Debiased Neighborhood Lasso

## A Closer Look into $\widetilde{\theta}_b^{(a)}$

With appropriately chosen tuning parameters in the neighborhood Lasso,

$$\widetilde{\theta}_b^{(a)} = -\frac{\Theta^*_{a,b}}{\Theta^*_{a,a}} + \text{mean-zero first order term} + \text{high-order residuals}$$

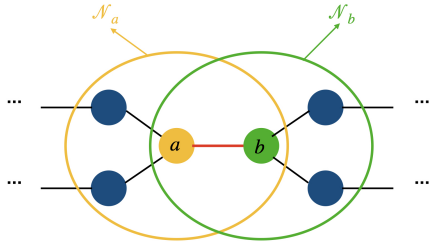
# Characterization of Debiased Neighborhood Lasso

## A Closer Look into $\widetilde{\theta}_b^{(a)}$

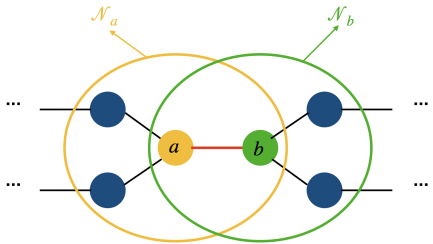
With appropriately chosen tuning parameters in the neighborhood Lasso,

$$\widetilde{\theta}_b^{(a)} = -\frac{\Theta^*_{a,b}}{\Theta^*_{a,a}} + \text{mean-zero first order term} + \text{high-order residuals}$$

- mean-zero first-order term
  $\propto \sum_{j,k}(\widehat{\Sigma}_{j,k} - \Sigma^*_{j,k})\Theta^*_{a,j}\Theta^*_{b,k}$
- only involve neighbors of $a$ and $b$!



# GI-JOE: Edge-wise Uncertainty Quantification



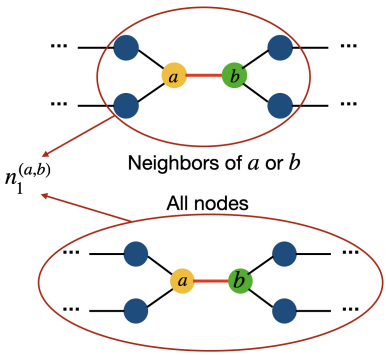

- **Step 2**: Estimate variance of first-order term
  - variance from each node pair $(j, k)$: $\widehat{\theta}_j^{(a)}$, $\widehat{\theta}_k^{(b)}$, $1/n_{j,k}$
  - plus some edge-edge correlations
  - Obtain $\widehat{\sigma}_n^2(a, b)$

- **Output**: Reject $\mathcal{H}_0 : (a, b) \notin E$ if $\frac{|\widetilde{\theta}_b^{(a)}|}{\widehat{\sigma}_n(a,b)} > z_{\alpha/2}$.

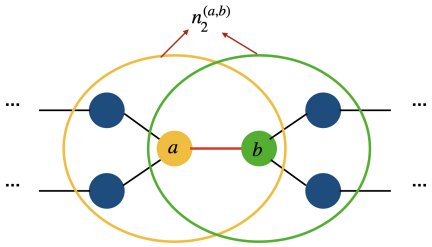
# Edgewise Testing: Theoretical Guarantees

# Assumption for Validity: Sufficient Local Sample Sizes

$$n_1^{(a,b)} = \min_{j \in \overline{\mathcal{N}}_a \cup \overline{\mathcal{N}}_b, k \in [p]} n_{j,k}, \quad n_2^{(a,b)} = \min_{j \in \overline{\mathcal{N}}_a, k \in \overline{\mathcal{N}}_b} n_{j,k}$$




For good performance of neighborhood Lasso




First-order dependency

# Assumption for Validity: Sufficient Local Sample Sizes

## Main Assumption

The local sample sizes $n_1^{(a,b)}$, $n_2^{(a,b)}$, degrees of node $a$, $b$ ($d_a$, $d_b$), graph size $p$ satisfy

$$n_1^{(a,b)} \gg (d_a + d_b)^2 (\log p)^2 \frac{n_2^{(a,b)}}{n_1^{(a,b)}}.$$

# Assumption for Validity: Sufficient Local Sample Sizes

**Main Assumption**

The local sample sizes $n_1^{(a,b)}$, $n_2^{(a,b)}$, degrees of node $a$, $b$ ($d_a$, $d_b$), graph size $p$ satisfy

$$n_1^{(a,b)} \gg (d_a + d_b)^2 (\log p)^2 \frac{n_2^{(a,b)}}{n_1^{(a,b)}}.$$

- Arbitrary data-independent missing pattern!
- Reduces to prior requirements when sample sizes are the same: $n \gg d^2 \log^2 p$
- Localized sample sizes: careful analysis and disentangling dependencies over the graph

# Statistical Validity of GI-JOE (Edge-wise Testing)

**Main Theorem: Type I error and power**

Suppose Assumption A1 hold. For testing $\mathcal{H}_0 : (a, b) \notin E$:

1. GI-JOE (edgewise testing) has asymptotically valid type I error control;
2. The asymptotic power is an increasing function of $|\Theta^*_{a,b}|\sqrt{n_2^{(a,b)}}$.

- Valid confidence intervals for the entries of precision $\Theta^*_{a,b}$ also available.

# GI-JOE: FDR control

# GI-JOE: FDR Control

Whole graph testing with FDR control?

- Inspired by FDR control for debiased Lasso (Javanmard and Javadi, 2019)
- Take edgewise $p$-values, apply Benjamini-Hochberg's procedure but with a truncation step
- Key idea: Weak/sparse edge-edge correlation $\Rightarrow$ valid FDR control

# Theoretical Gaurantees

**Theorem: Valid FDR control**

Assume

1. $n_1^{(a,b)}$ is sufficiently large for all $(a, b)$;
2. Most edge pairs $(a, b)$, $(a', b')$ are only weakly correlated.

The edge set selected by GI-JOE (FDR) has asymptotically valid FDR control.

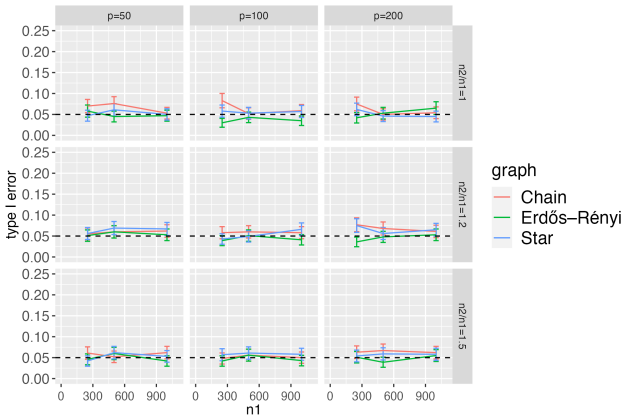
- Still allow highly unenven pairwise sample sizes;
- Weaker condition than literature with full observations
- Correlation condition empirically supported

# Empirical Studies

# Simulation: Edge-wise Testing

Type I Error

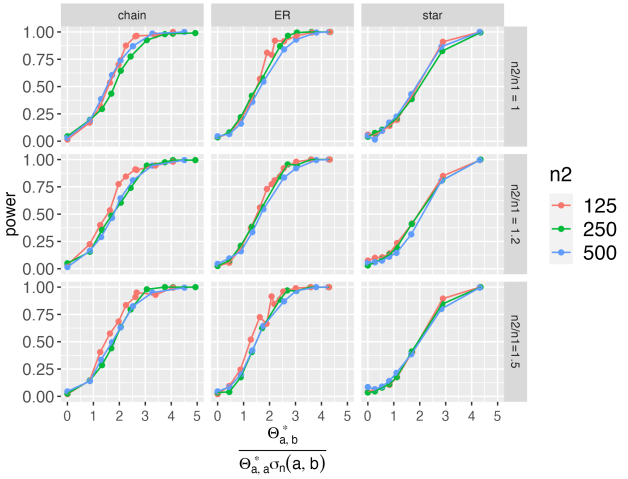
- Pairwise measurements mostly 50
- Change $n_1^{(a,b)}$ and $n_2^{(a,b)}$
- Type I error fluctuates around nominal level 0.05



# Simulation: Edge-wise Testing

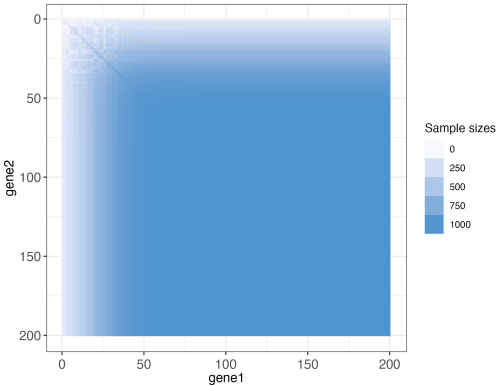
Power

- Changing signal strength and $n_1^{(a,b)}$, $n_2^{(a,b)}$
- X-axis: signal strength / standard deviation
- Lines within each column aligns well with each other



# Simulation: Graph Selection Comparison

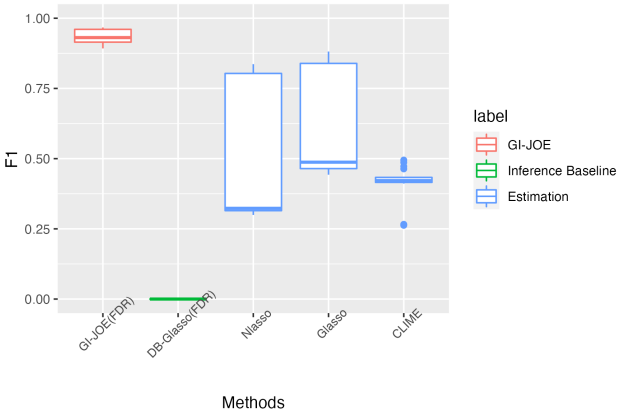
- Simulate data from a scale-free graph with 200 nodes
- The measurement pattern is the same as a real single-cell RNA sequencing data set (Chu et al., 2016)



# Simulation: Graph Selection Comparison

F1-score comparison:
$2/(\text{TPR}^{-1}+\text{TDR}^{-1})$

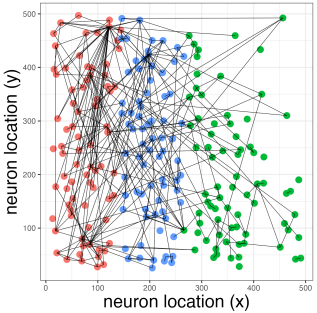
- Our inference methods with FDR control
- Baseline inference methods: Plug-in method with debiased graphical lasso, minimum sample size
- Estimation methods: graphical lasso, neighborhood lasso, CLIME



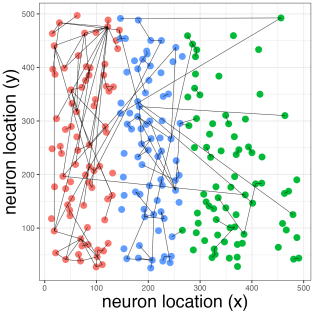
# Application to Neuronal Functional Data

Allen Brain Atlas data; $p = 227$ neurons, $n = 8931$ time points; spontaneous period

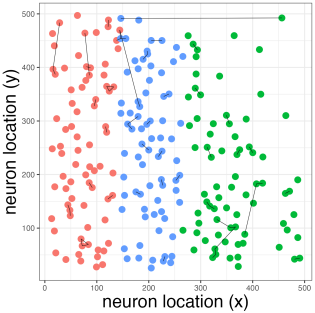
Manually mask functional data; three sets of neurons observed with high, median, low probabilities;




FDR-selected graph with full data




GI-JOE (FDR), applied to erose data




DB-Glasso with minimum sample size, applied to erose data

# Conclusion

- Erose measurements: challenge for reliable graph learning

# Conclusion

- Erose measurements: challenge for reliable graph learning
- Edge-wise uncertainty hinges on neighbors; can be estimated by GI-JOE;

# Conclusion

- Erose measurements: challenge for reliable graph learning
- Edge-wise uncertainty hinges on neighbors; can be estimated by GI-JOE;
- Quantify different uncertainty levels over the graph with FDR control $\Rightarrow$ Better graph selection with erose data!
- Open questions

# Conclusion

- Erose measurements: challenge for reliable graph learning
- Edge-wise uncertainty hinges on neighbors; can be estimated by GI-JOE;
- Quantify different uncertainty levels over the graph with FDR control $\Rightarrow$ Better graph selection with erose data!
- Open questions
  - Non-Gaussian data
  - Latent variables?
  - Zero sample sizes for certain edges?
- **L. Zheng**, G. I. Allen, "Graphical Model Inference with Erosely Measured Data", *Journal of the American Statistical Association, Theory and Methods*, 2023.

# From Complex Data Collection to Complex Machine Learning Systems

- First part concerned with making discoveries from realistic data
- Based on a generative model and a model-specific method
- What about UQ for blackbox ML models?

# Uncertainty Quantification for Interpretable Machine Learning

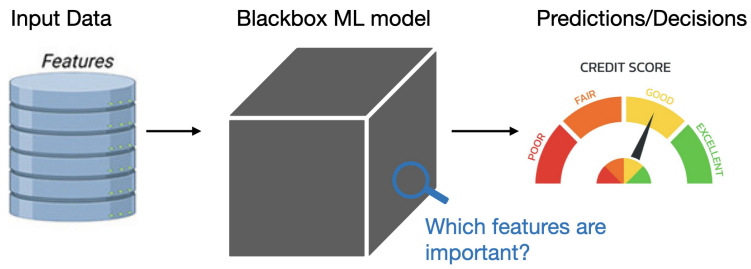
# Interpretable Machine Learning (IML)

Machine learning widely applied in **high-stake applications**:

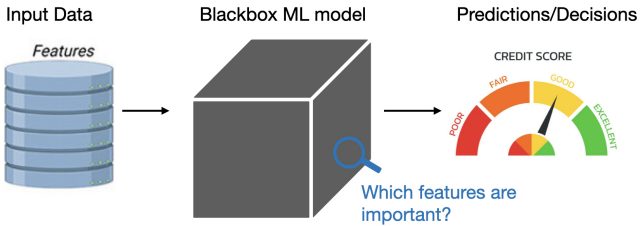


Can we trust machine learning? Make it interpretable!

**Feature importance**: How does my model's prediction rely on each feature?



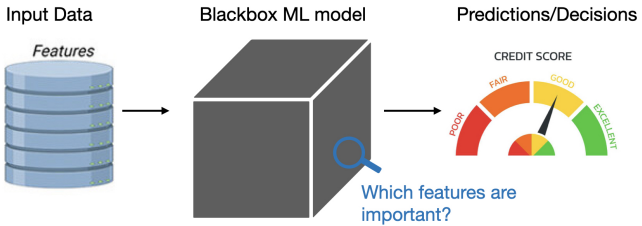
# Example Usage of Feature Importance as ML Interpretation




## Model diagnostics

- Example: ML model decides who gets home mortgage
- Unfair w.r.t. races
- Why? Model is race-blind
- Check feature importance: model depends heavily on zip code (proxy of race)?

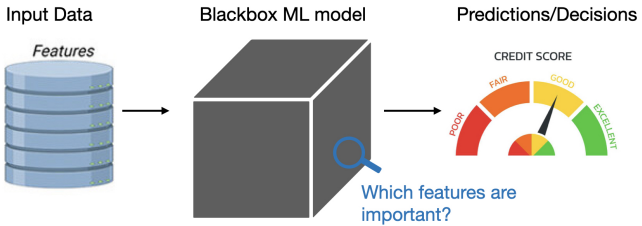
# Example Usage of Feature Importance as ML Interpretation




## Model auditing

- Regular checks & human evaluation
- Does the model align well with expectation / common sense / ethics?

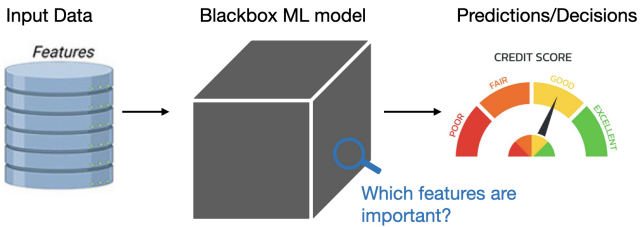
# Example Usage of Feature Importance as ML Interpretation




## Model deployment

- Which features lead to good prediction?
- Improve data quality for important features

# Example Usage of Feature Importance as ML Interpretation




**Model deployment**

- Which features lead to good prediction?
- Improve data quality for important features

Can we trust feature importance for decision-making?

# UQ for Feature Importance

Focus on global, model-agnostic feature importance

- **Occlusion-based** (Lei et al., 2018; Rinaldo et al., 2019)
- Shapley-value-based (Williamson and Feng, 2020; Lundberg et al., 2018)
- Permutation-based (Fisher et al., 2019)
- Knockoff-based (Watson and Wright, 2021)
- Related but different: feature importance as conditional dependency; linear model coefficients; Floodgate (Zhang and Janson, 2020), GCM (Shah and Peters, 2020), VIMP (Williamson et al., 2022)

# Prior Work: LOCO Inference

**Leave-One-Covariate-Out (LOCO) Inference**:

- (Lei et al., 2018; Rinaldo et al., 2019)

# Prior Work: LOCO Inference

**Leave-One-Covariate-Out (LOCO) Inference**:

- (Lei et al., 2018; Rinaldo et al., 2019)


**Inference target**: Predictive power without feature $j$ vs. with feature $j$.

$$\Delta_j^*(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left[\text{Error}\left(Y^{\text{test}}, \mu_{-j}(X_{-j}^{\text{test}}; \boldsymbol{X}_{-j}, \boldsymbol{Y})\right) - \text{Error}\left(Y^{\text{test}}, \mu(X^{\text{test}}; \boldsymbol{X}, \boldsymbol{Y})\right) | \boldsymbol{X}, \boldsymbol{Y}\right]$$

# Prior Work: LOCO Inference

**Leave-One-Covariate-Out (LOCO) Inference**:

- (Lei et al., 2018; Rinaldo et al., 2019)


**Inference target**: Predictive power without feature $j$ vs. with feature $j$.

$$\Delta_j^*(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left[\text{Error}\left(Y^{\text{test}}, \mu_{-j}(X_{-j}^{\text{test}}; \boldsymbol{X}_{-j}, \boldsymbol{Y})\right) - \text{Error}\left(Y^{\text{test}}, \mu(X^{\text{test}}; \boldsymbol{X}, \boldsymbol{Y})\right) | \boldsymbol{X}, \boldsymbol{Y}\right]$$

- Property of the current models
- Q1: Which feature does my model depend on?

# Prior Work: LOCO Inference

**Leave-One-Covariate-Out (LOCO) Inference**:

- (Lei et al., 2018; Rinaldo et al., 2019)

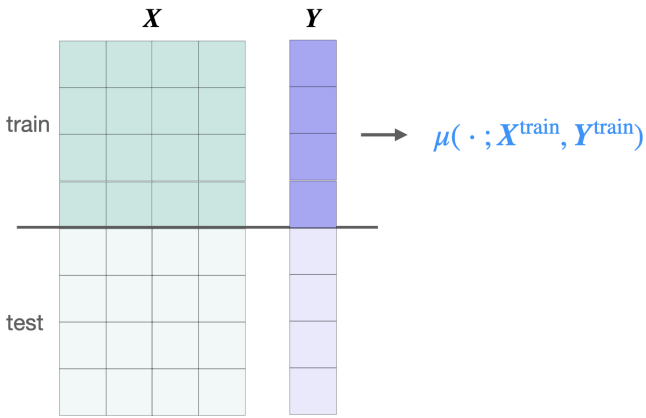

**Inference target**: Predictive power without feature $j$ vs. with feature $j$.

$$\Delta_j^*(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left[\text{Error}\left(Y^{\text{test}}, \mu_{-j}(X_{-j}^{\text{test}}; \boldsymbol{X}_{-j}, \boldsymbol{Y})\right) - \text{Error}\left(Y^{\text{test}}, \mu(X^{\text{test}}; \boldsymbol{X}, \boldsymbol{Y})\right) | \boldsymbol{X}, \boldsymbol{Y}\right]$$

- Property of the current models
- Q1: Which feature does my model depend on?
- Q2: Which feature helps my model's performance?

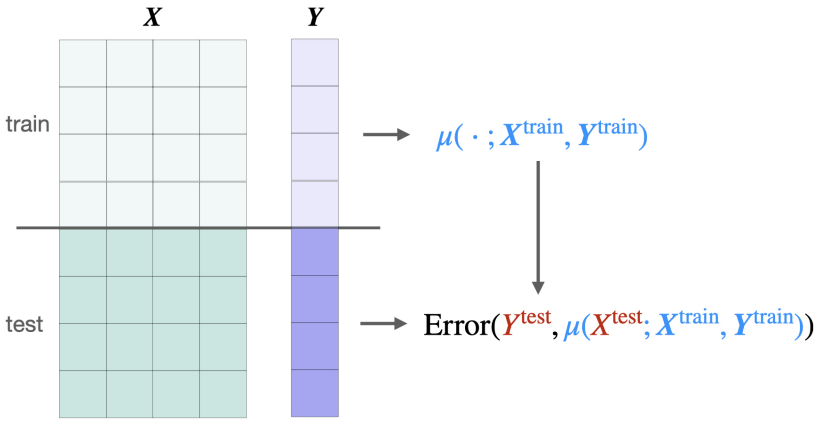
# Prior Work: LOCO Inference

**LOCO inference approach**: Data-splitting and model-refitting



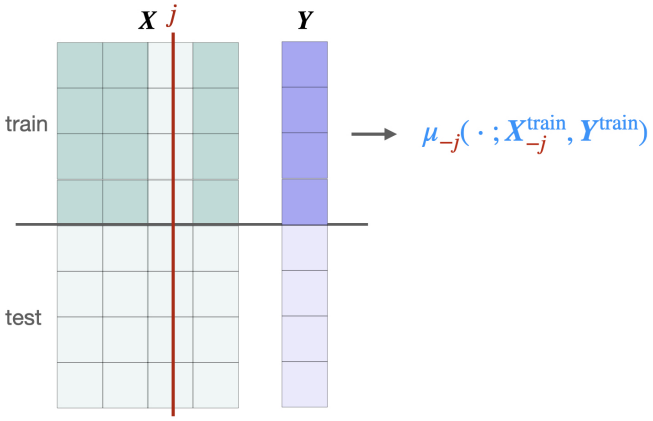
# Prior Work: LOCO Inference

**LOCO inference approach**: Data-splitting and model-refitting



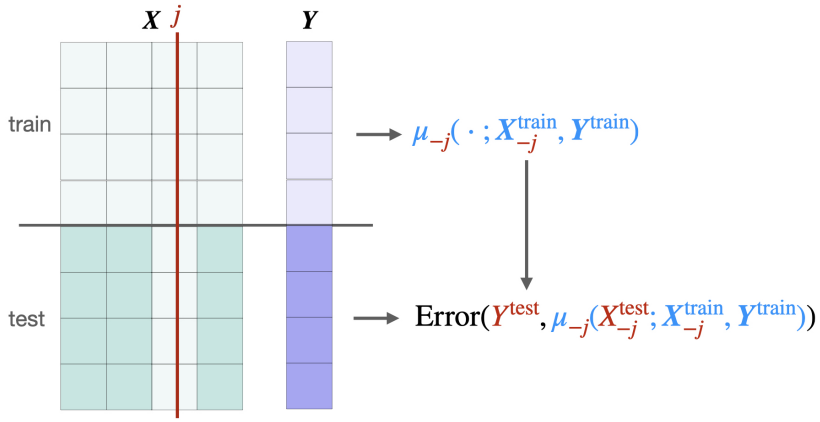
# Prior Work: LOCO Inference

**LOCO inference approach**: Data-splitting and model-refitting



# Prior Work: LOCO Inference

**LOCO inference approach**: Data-splitting and model-refitting



# Prior Work: LOCO Inference

**LOCO inference approach**: Data-splitting and model-refitting

Construct asymptotic normal confidence intervals from

$$\text{Error}\left(Y^{\text{test}}, \mu_{-j}(X^{\text{test}}_{-j}; \boldsymbol{X}^{\text{train}}_{-j}, \boldsymbol{Y}^{\text{train}})\right) - \text{Error}\left(Y^{\text{test}}, \mu(X^{\text{test}}; \boldsymbol{X}^{\text{train}}, \boldsymbol{Y}^{\text{train}})\right)$$

**Advantages**:

- Model-agnostic (applicability).
- Valid without assuming data distribution/model choice.

# Prior Work: LOCO Inference

**LOCO inference approach**: Data-splitting and model-refitting

Construct asymptotic normal confidence intervals from

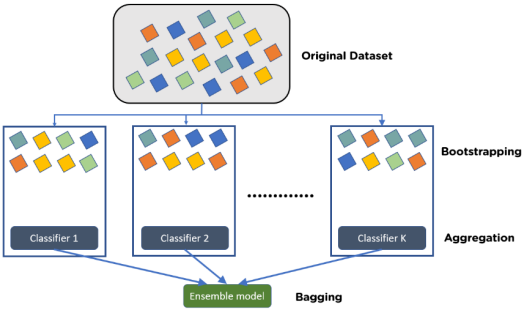
$$\text{Error}\left(Y^{\text{test}}, \mu_{-j}(X^{\text{test}}_{-j}; \boldsymbol{X}^{\text{train}}_{-j}, \boldsymbol{Y}^{\text{train}})\right) - \text{Error}\left(Y^{\text{test}}, \mu(X^{\text{test}}; \boldsymbol{X}^{\text{train}}, \boldsymbol{Y}^{\text{train}})\right)$$

**Challenges**:

- Data splitting loses statistical power;
- Interpretation is not for the full model & depends on random data splitting
- Model refitting for each feature: prohibitive computation after model training

# Proposed Method: LOCO Inference for an Ensemble Framework

**Feature importance inference for ensemble methods?**




Bagging

Picture source: https://www.simplilearn.com/tutorials/machine-learning-tutorial/bagging-in-machine-learning

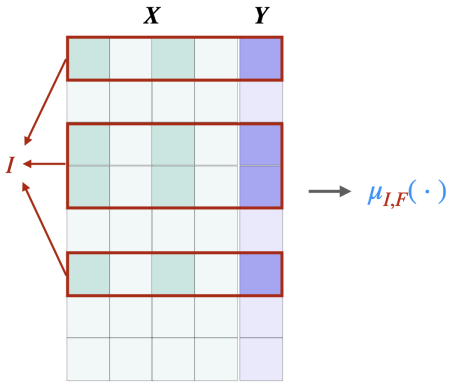
**Inspiration**: Jacknife+ After Bootstrap (Kim et al., 2020).

- Ensemble methods give good predictions
- Conformal inference (Jacknife+) is computationally free for bagging/subbagging!

**Idea: Minipatch Ensembles**.

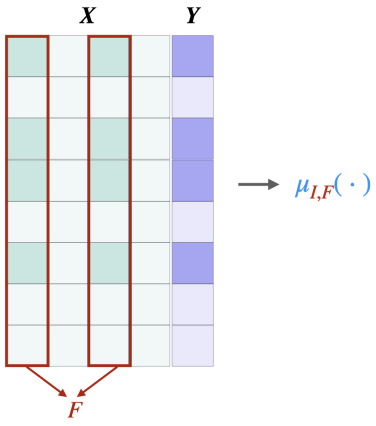
# Minipatch Ensemble Learning

**Minipatch ensembles**: like subbagging, but subsample both observations and features. (Yao et al., 2020)



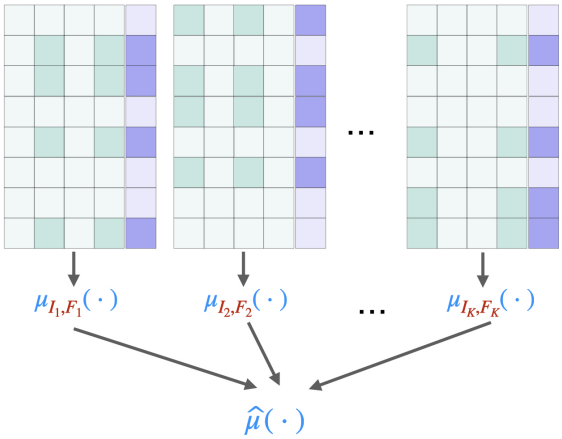
# Minipatch Ensemble Learning

**Minipatch ensembles**: like subbagging, but subsample both observations and features. (Yao et al., 2020)



# Minipatch Ensemble Learning

**Minipatch ensembles**: like subbagging, but subsample both observations and features. (Yao et al., 2020)

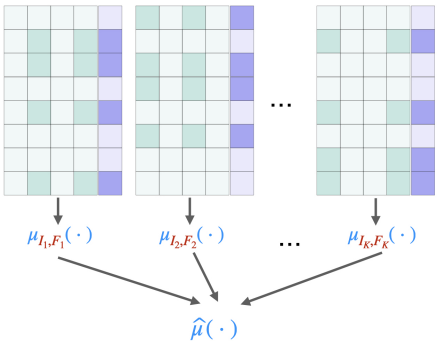


# Minipatch Ensemble Learning

**Inspiration**: Bagging; Random Forests (Louppe and Geurts, 2012); Stochastic Optimization & Dropout.

**Advantages**:

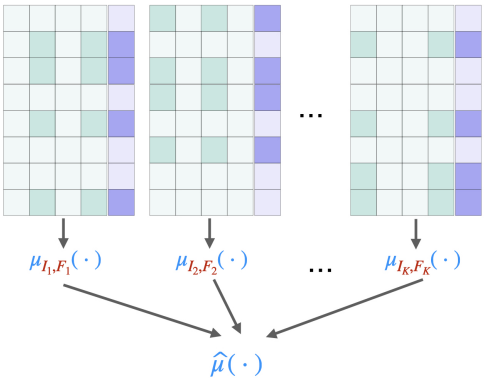
- Fast and easily parallelizable
- Ensemble diversity; **implicit regularization** (LeJeune et al., 2020; Yao et al., 2021)




**LOCO Inference for Minipatch Ensembles?**

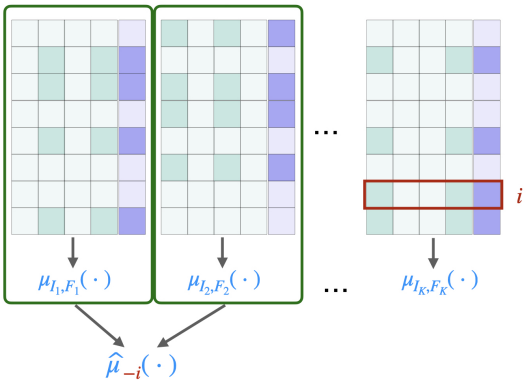
# Algorithm: LOCO for Minipatch

1. Fit minipatch learning predictor: $\hat{\mu}$.



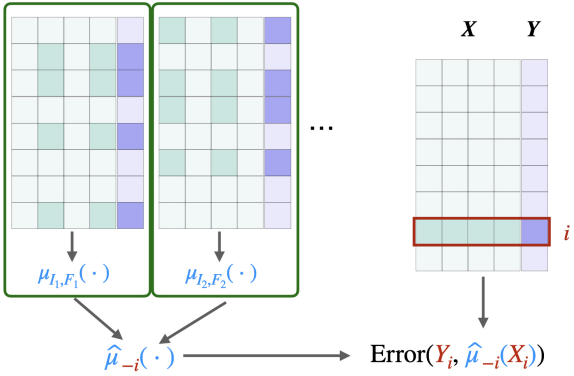
# Algorithm: LOCO for Minipatch

2. **LOO** (leave-one-observation-out) predictor: $\hat{\mu}_{-i}(X_i)$.
   - Ensemble minipatches without observation $i$.
   - Compute test error on $i$



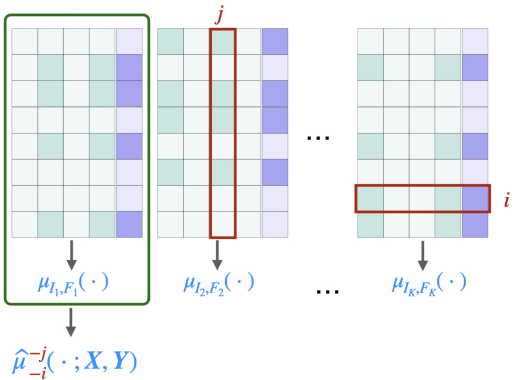
# Algorithm: LOCO for Minipatch

2. **LOO** (leave-one-observation-out) predictor: $\hat{\mu}_{-i}(X_i)$.
   - Ensemble minipatches without observation $i$.
   - Compute test error on $i$



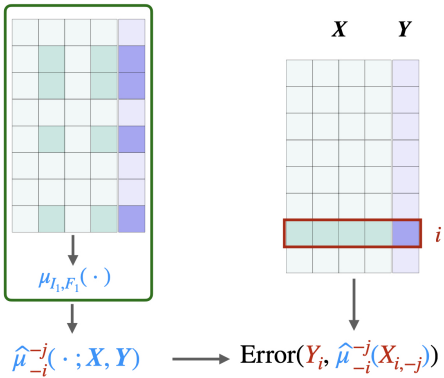
# Algorithm: LOCO for Minipatch

3. **LOCO-LOO** predictor: $\hat{\mu}_{-i}^{-j}(X_i)$.
   - Ensemble minipatches without observation $i$ and without feature $j$.
   - Compute test error on $i$



# Algorithm: LOCO for Minipatch

3. **LOCO-LOO** predictor: $\hat{\mu}_{-i}^{-j}(X_i)$.
   - Ensemble minipatches without observation $i$ and without feature $j$.
   - Compute test error on $i$



# Algorithm: LOCO for Minipatch

4. Compute feature occlusion score for observation $1 \leq i \leq N$

$$\widehat{\Delta}_j(X_i, Y_i) = \text{Error}(Y_i, \hat{\mu}_{-i}^{-j}(X_i)) - \text{Error}(Y_i, \hat{\mu}_{-i}(X_i)).$$

**Importance of feature $j$ for predicting sample $i$**
Recall our inference target:

$$\Delta_j^*(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left[\text{Error}\left(Y^{\text{test}}, \mu_{-j}(X_{-j}^{\text{test}}; \boldsymbol{X}_{-j}, \boldsymbol{Y})\right) - \text{Error}\left(Y^{\text{test}}, \mu(X^{\text{test}}; \boldsymbol{X}, \boldsymbol{Y})\right) | \boldsymbol{X}, \boldsymbol{Y}\right]$$

# Algorithm: LOCO for Minipatch

5. Construct asymptotically normal interval from $\{\widehat{\Delta}_j(X_i, Y_i)\}_{i=1}^N$:

$$\hat{\mathbb{C}}_j = \left[\bar{\Delta}_j - \frac{z_{\alpha/2}\hat{\sigma}_j}{\sqrt{N}}, \bar{\Delta}_j + \frac{z_{\alpha/2}\hat{\sigma}_j}{\sqrt{N}}\right],$$

$\bar{\Delta}_j$: mean occlusion score, $\hat{\sigma}_j$: standard deviation.

**Algorithmic Advantages**

- **No data-splitting** $\Rightarrow$ powerful; feature importance inference for the current model at hand!
- **No model-refitting** $\Rightarrow$ once predictive model is trained, confidence intervals are computationally free!

# Theoretical Guarantees

Does LOCO-MP confidence interval have valid coverage?

- **Leave-one-observation-out instead of data-splitting** $\Rightarrow$ dependency amongst $\{\hat{\Delta}_j(X_i, Y_i)\}_{i=1}^N$!
- $\Delta_j(X_i, Y_i)$ and $\Delta_j(X_{i'}, Y_{i'})$ share $N-2$ training samples!
- Asymptotic normality?

# Theoretical Guarantees

A1. $\mathrm{Error}()$ is Lipschitz-$L$.

A2. Bounded difference in MP predictions $||\hat{\mu}_{I,F}(X) - \hat{\mu}_{I',F'}(X)|| \leq B$.

# Theoretical Guarantees

A1. $\mathrm{Error}()$ is Lipschitz-$L$.

A2. Bounded difference in MP predictions $||\hat{\mu}_{I,F}(X) - \hat{\mu}_{I',F'}(X)|| \leq B$.
(automatically hold for classification)

# Theoretical Guarantees

A1. $\mathrm{Error}()$ is Lipschitz-$L$.

A2. Bounded difference in MP predictions $||\hat{\mu}_{I,F}(X) - \hat{\mu}_{I',F'}(X)|| \leq B$.
(automatically hold for classification)

A3. MP size: $n = o(\frac{\sigma_j}{LB}\sqrt{N})$.

# Theoretical Guarantees

A1. $\mathrm{Error}()$ is Lipschitz-$L$.

A2. Bounded difference in MP predictions $||\hat{\mu}_{I,F}(X) - \hat{\mu}_{I',F'}(X)|| \leq B$.
(automatically hold for classification)

A3. MP size: $n = o(\frac{\sigma_j}{LB}\sqrt{N})$.

A4. MP number: $K \gg (\frac{L^2B^2N}{\sigma_j^2} + \frac{LB\sqrt{N}}{\sigma_j} + 1)\log(N)$.

# Theoretical Guarantees

A1. $\mathrm{Error}()$ is Lipschitz-$L$.

A2. Bounded difference in MP predictions $||\hat{\mu}_{I,F}(X) - \hat{\mu}_{I',F'}(X)|| \leq B$.
(automatically hold for classification)

A3. MP size: $n = o(\frac{\sigma_j}{LB}\sqrt{N})$.

A4. MP number: $K \gg (\frac{L^2B^2N}{\sigma_j^2} + \frac{LB\sqrt{N}}{\sigma_j} + 1)\log(N)$.

**Theorem**
Suppose samples $(\boldsymbol{X}_i, Y_i)$ are i.i.d., and assumptions A1-A4 hold. Then

$$\sqrt{N}\hat{\sigma}_j^{-1}(\bar{\Delta}_j - \Delta_j^*) \xrightarrow{d} \mathcal{N}(0,1).$$

**Theorem**
Suppose samples $(\boldsymbol{X}_i, Y_i)$ are i.i.d., and assumptions A1-A4 hold. Then

$$\sqrt{N}\hat{\sigma}_j^{-1}(\bar{\Delta}_j - \Delta_j^*) \xrightarrow{d} \mathcal{N}(0,1).$$

**Corollary**

$$\lim_{N\to\infty} \mathbb{P}(\Delta_j^* \in \hat{\mathbb{C}}_j) = 1 - \alpha.$$

**Valid asymptotic coverage**!

**Proof Idea**:

- Bayle et al., 2020 show CLT for cross-validation error, if training algorithm is stable.
- Minipatch ensembles are stable with any base model and any data distribution!
- Characterize conditions in MP differences, number, size

**Predictive inference is also free after training!**

- Similar to Jacknife+ after bootstrap (Barber et al., 2021)
- Valid $1 - 2\alpha$ coverage under exchangeability assumptions

# Empirical Studies

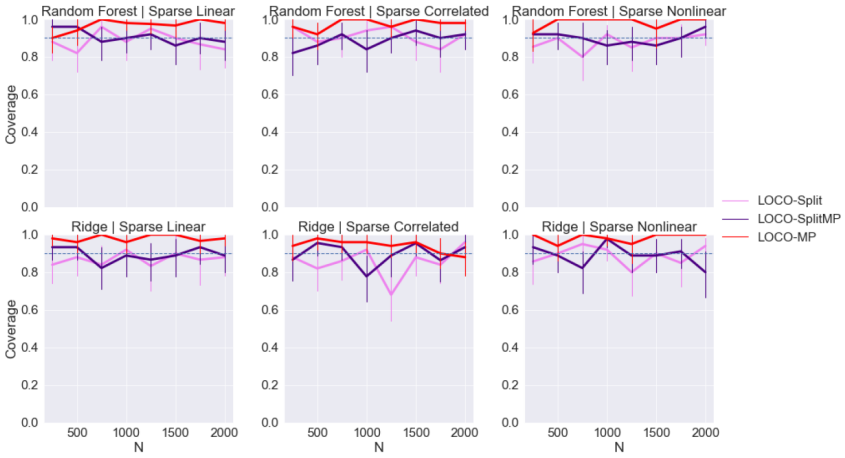
# Simulations: Validate Coverage

Simulation Set-up:

- $M = 200$, varying $N$, 10 signal features
- Sparse linear & nonlinear (logistic) regression; i.i.d. or correlated features
- LOCO-MP with $m = \sqrt{M}$, $n = \sqrt{N}$, $K = 10,000$

# Simulations: Comparative Results

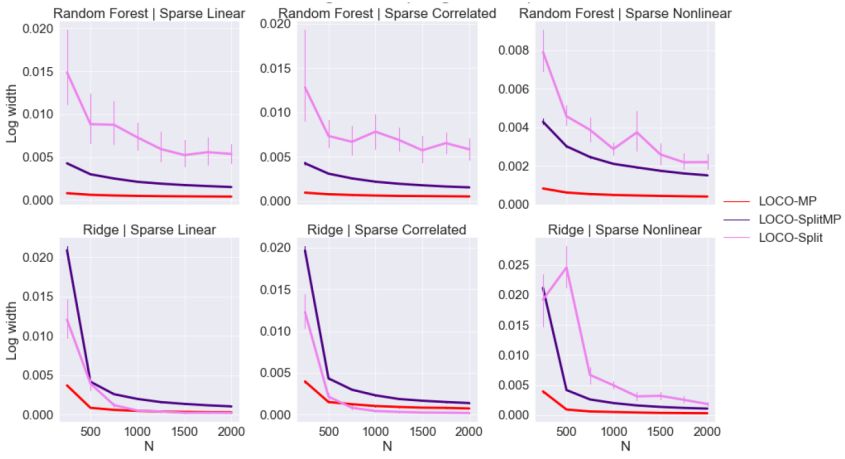
Theory Validation: Coverage.




Coverage for regression simulations for a null feature.

# Simulations: Comparative Results

Interval Width:

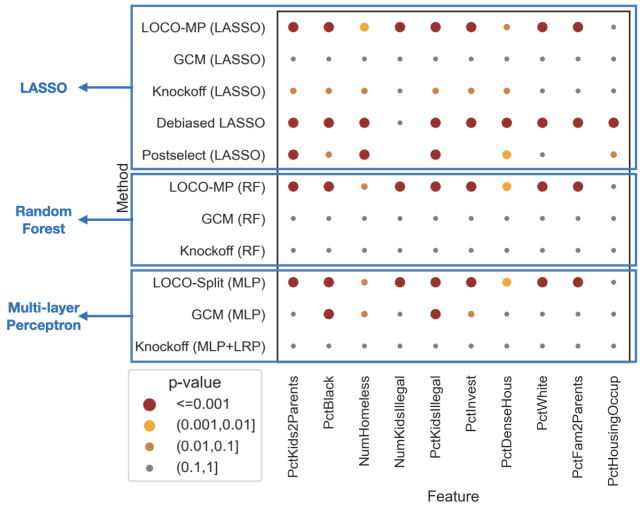



Log interval width for regression simulations for a null feature.

# Real Data Example

- Communities and Crimes data (Redmond M. 2009)
- $n = 1994$ observations, $p = 122$ features
- Predict the per capita violent crime rate based on community features

# Real Data Example

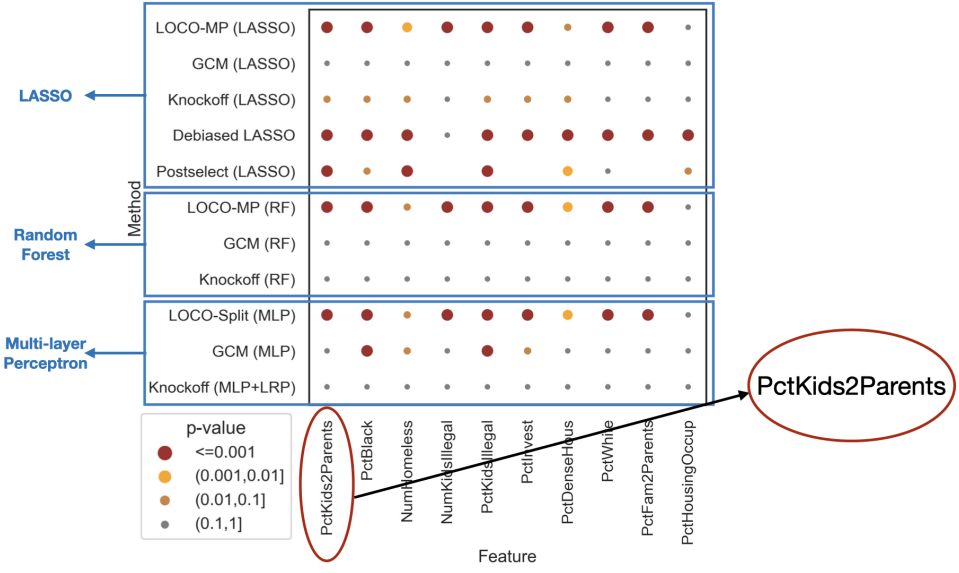



Feature importance inference

- Model-based: debiased Lasso, post-selective inference for Lasso
- Model-agnostic: GCM, Knockoff, LOCO-MP, LOCO-split

LOCO methods identify the most consistent features across methods

# Real Data Example



# Real Data Example

**Going beyond model diagnostics: LOCO-MP for discovery?**

- Revealing important features associated with certain response?
- Relationship to parametric feature importance?

# Conclusion

- **Uncertainty quantification for feature importance for minipatch ensembles**

# Conclusion

- **Uncertainty quantification for feature importance for minipatch ensembles**
  - Free computationally (after minipatch learning).
  - Also (free) predictive intervals.
  - Statistically powerful; assumption-light

# Conclusion

- **Uncertainty quantification for feature importance for minipatch ensembles**
- Open Questions:
  - Feature importance inference for adaptive sampled minipatches?
  - Shapley value-based feature importance.
  - Beyond feature importance (e.g. feature interactions).

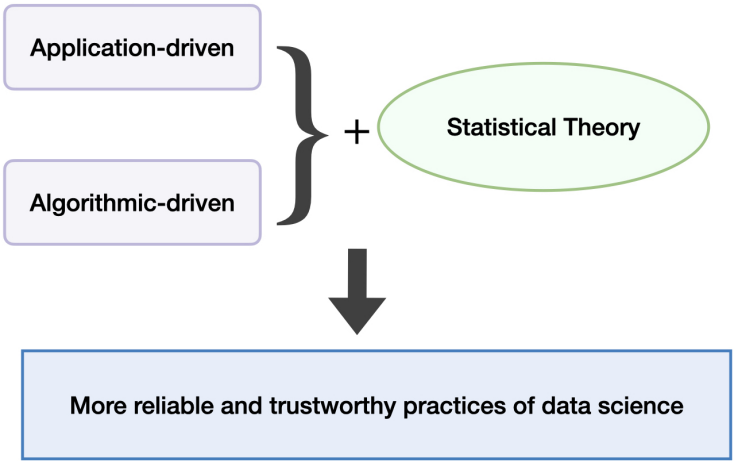
# Conclusion

- **Uncertainty quantification for feature importance for minipatch ensembles**

- L. Gan*, **L. Zheng***, G. I. Allen (*: equal contribution), "Model-Agnostic Confidence Intervals for Feature Importance: A Fast and Powerful Approach Using Minipatch Ensembles", https://arxiv.org/abs/2206.02088.

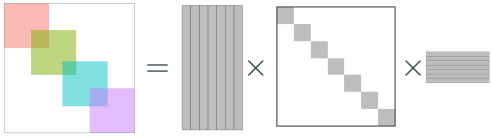
# Other Works and Future Directions

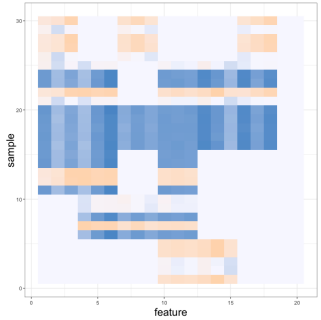
# Research Theme



## Structure Learning with Erose Measuremets

- Low-rank graph quilting
- Applications to neuroscience

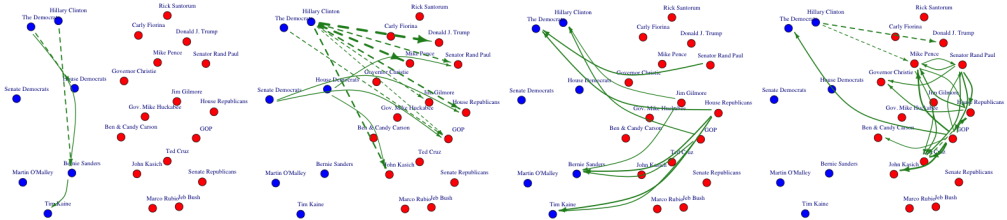




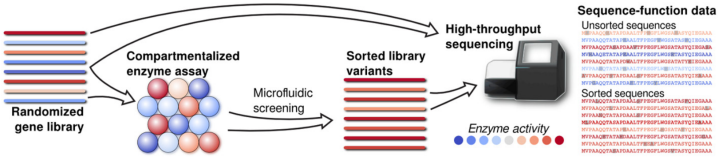

- Spectral clustering for patchwork learning
- Applications to healthcare

## Granger Causal Network Learning and Inference

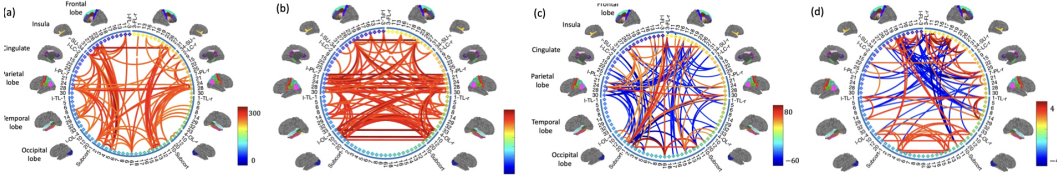
- Applications to social network, neuroscience, finance
- Hypothesis testing for Granger causal edges in linear AR(p) models (Electronic Journal of Statistics, 2019)
- Context-dependent Granger causal network learning for mixed data types (Journal of Machine Learning Research, 2020)



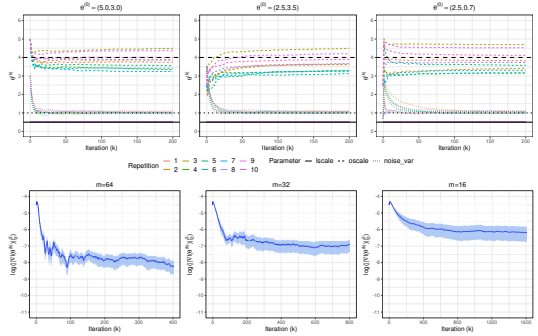
## Presence-only Data Classification with Applications to Protein Engineering




## Joint Tensor PCA for Multi-modal Populations of Networks in Neuroimaging



**Provable Convergence: Stochastic Gradient Descent can Speed up Gaussian Processes!** (Journal of Machine Learning Research, 2022)



# Future Directions

**Can we trust statistical structure learning with latent variables?**

- Ignore latent variable in UQ $\Rightarrow$ inflated type I error & FDR
- Leverage GI-JOE, latent variable graph learning, topological assumptions

# Future Directions

**Adaptive sampling strategies under measurements/computational constraints?**

- GI-JOE $\Rightarrow$ with limited measurements, should focus on neighbors $\Rightarrow$ adaptive experimental design
- Limited memory/computation $\Rightarrow$ subsampling-based ensembles? Adaptive subsampling $\Rightarrow$ statistical & computational advantage
- Leverage graph theory & ensemble theory

**Fairness in machine learning interpretations?**

ML interpretations can

- inherit bias from data
- only depicts majority group

Predictive fairness is much more studied than interpretational fairness. Can leverage predictive fairness literature & ML interpretation methods & graph modeling

# Publication List

## Peer-reviewed Journal Publications

1. **Lili Zheng**, Genevera I. Allen, "Graphical Model Inference with Erosely Measured Data", **Journal of the American Statistical Association**, 2023.
2. Andersen Chang*, **Lili Zheng***, Gautam Dasarathy, Genevera I. Allen, "Nonparanormal Graph Quilting with Applications to Calcium Imaging", **STAT**, 2023.
3. Genevera I. Allen, Luqin Gan, **Lili Zheng**, "Interpretable Machine Learning for Discovery: Statistical Challenges & Opportunities", **Annual Review of Statistics and Its Application**, 2023.
4. Hao Chen*, **Lili Zheng***, Raed Al Kontar, Garvesh Raskutti ($*$: equal contribution), "Gaussian Process Parameter Estimation Using Mini-batch Stochastic Gradient Descent: Convergence Guarantees and Empirical Benefits", **Journal of Machine Learning Research**, 23.1 (2022): 10298-10356.

# Publication List

## Peer-reviewed Journal Publications

5. Yuchen Zhou, Anru R. Zhang, **Lili Zheng**, Yazhen Wang, "Optimal High-order Tensor SVD via Tensor-train Orthogonal Iteration", **IEEE Transactions on Information Theory**, 68.6 (2022): 3991-4019.
6. **Lili Zheng**, Garvesh Raskutti, Rebecca Willett, Benjamin Mark, "Context-dependent Networks in Multivariate Time Series: Models, Methods, and Risk Bounds in High Dimensions", **Journal of Machine Learning Research**, 22.1 (2021): 9771-9858.
7. **Lili Zheng**, Garvesh Raskutti, "Testing for High-dimensional Network Parameters in Auto-regressive Models", **Electronic Journal of Statistics**, (2019): 4977-5043.

# Publication List

## Peer-reviewed Conference Publications

8. **Lili Zheng**, Zach T. Rewolinski, Genevera I. Allen, "A Low-Rank Tensor Completion Approach for Imputing Functional Neuronal Data from Multiple Recordings", **IEEE Data Science and Learning Workshop (DSLW)**. IEEE, 2022.
9. **Lili Zheng**, Genevera I. Allen, "Learning Gaussian Graphical Models with Differing Pairwise Sample Sizes", **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. IEEE 2022.
10. Hao Chen*, **Lili Zheng***, Raed Al Kontar, Garvesh Raskutti (*: equal contribution), "Stochastic Gradient Descent in Correlated Settings: A Study on Gaussian Processes", **Neural Information Processing Systems (NeurIPS)**, 33 (2020): 2722-2733.

# Publication List

## Preprints

11. **Lili Zheng**, Garvesh Raskutti, "High-dimensional Multi-class Classification with Presence-only Data", *under revision at Electronic Journal of Statistics*.
12. Luqin Gan*, **Lili Zheng***, Genevera I. Allen (*: equal contribution), "Model-Agnostic Confidence Intervals for Feature Importance: A Fast and Powerful Approach Using Minipatch Ensembles", arXiv preprint arXiv: 2206.02088, 2023.
13. Andersen Chang, **Lili Zheng**, Genevera I. Allen, "Low-Rank Covariance Completion for Graph Quilting with Applications to Functional Connectivity". *under revision at Journal of the American Statistical Association, Applications and Case Studies*, arXiv preprint arXiv: 2209.08273, 2023.

# Acknowledgments

Coauthors

Luqin Gan

Genevera I. Allen

# Thank you!

**Assumption (Sample Size Condition)**
*For all node pairs* $(a, b) \in [p] \times [p]$,

$$n_1^{(a,b)} \gg C(d+1)^2(\log p)^5 \log\log p \left(\frac{n_2^{(a,b)}}{n_1^{(a,b)}}\right)^2, \quad n_2^{(a,b)} \geq C(d+1)^6(\log p)^6.$$

- Stronger than edge-wise testing for uniform Gaussian approximation results;
- Weaker assumption than $p < n^C$ for $C > 0$ in prior literature (Liu, 2013);
- Let $g(d, p) = C(d+1)^2(\log p)^5 \log\log p$, then this is implied by

$$n_{\min} \gg g(d, p) \left(\frac{n_{\max}}{g(d, p)}\right)^{2/3}$$

**Assumption (Edge-edge correlations)**
*Total number of edge pairs:* $p^4$.

- $\mathcal{A}_1$*: set of strongly correlated edge pairs;* $|\mathcal{A}_1| \leq Cp^2$
- $\mathcal{A}_2$*: set of moderately correlated edge pairs;* $|\mathcal{A}_2| \ll p^{4-\varepsilon}$ *for a small constant* $\varepsilon > 0$.

- In full observational setting, this is implied by (i) each node only has constant number of strongly connected neighbors; (ii) $d \ll p^{1-c}$;
- Empirical evidence supports this assumption for general graph and measurement patterns.

# Proof Sketch for GI-JOE: Edgewise Testing

$$\widetilde{\theta}_b^{(a)} = -\frac{\Theta^*_{a,b}}{\Theta^*_{a,a}} + \text{mean-zero first-order term} + \text{high-order residuals}$$

- Mean-zero first-order term $\asymp \frac{1}{\sqrt{n_2^{(a,b)}}}$
- High-order residuals carefully controlled: $\lesssim \frac{\log p}{n_1^{(a,b)}}$ (collects errors from neighborhood Lasso)
- Variance estimates depend on (i) neighborhood Lasso; (ii) $\widehat{\Sigma}_{j,k} - \Sigma^*_{j,k}$ mainly for the $j \in \mathcal{N}_a$, $k \in \mathcal{N}_b$.

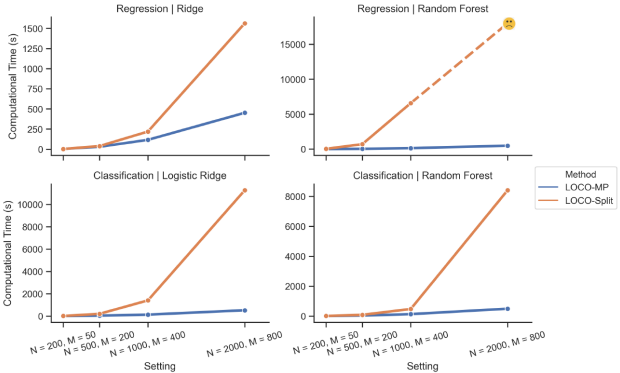
# LOCO-MP Simulations: Validate Coverage

Simulation Set-up:

- Vary $N$, $M = 200$ (unless otherwise specified) & 10 true features.
- 3 Scenarios:
  1. Sparse Linear Regression (or Logistic Regression); iid features.
  2. Sparse Linear Regression (or Logistic Regression); correlated features.
     - Adjacent features have correlation 0.5.
  3. Sparse Non-linear Regression (or Logistic Regression); iid features.
     - Polynomial and MARS spline non-linearity.
- Minipatch LOCO (LOCO-MP) run with $m = \sqrt{M}$ and $n = \sqrt{N}$ and $K = 10,000$.

# LOCO-MP Simulations: Comparative Results

Computational Time:




Computational time for inference on all features in sparse linear regression and classification simulations.

## References

Birkner, A., Tischbirek, C. H., and Konnerth, A. (2017). Improved deep two-photon calcium imaging in vivo. *Cell calcium*, 64:29–35.

Chu, L.-F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D. T., Choi, J., Kendziorski, C., Stewart, R., and Thomson, J. A. (2016). Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, 17(1):1–20.

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Hayden Gephart, M. G., Barres, B. A., and Quake, S. R. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290.

Javanmard, A. and Javadi, H. (2019). False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1):1212–1253.

Kolar, M. and Xing, E. P. (2012). Estimating sparse precision matrices from data with missing values.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462.

Millimet, D. L. and McDonough, I. K. (2017). Dynamic panel data models with irregular spacing: With an application to early childhood development. *Journal of Applied Econometrics*, 32(4):725–743.

Park, S., Wang, X., and Lim, J. (2021). Estimating high-dimensional covariance and precision matrices under general missing dependence. *Electronic Journal of Statistics*, 15(2):4868–4915.

Rajendran, S., Pan, W., Sabuncu, M. R., Zhou, J., and Wang, F. (2023). Patchwork learning: A paradigm towards integrative analysis across diverse biomedical data sources. *arXiv preprint arXiv:2305.06217*.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Vinci, G., Dasarathy, G., and Allen, G. I. (2019). Graph quilting: graphical model selection from partially observed covariances. *arXiv preprint arXiv:1912.05573*.