

Uncertainty Quantification for Interpretable Machine Learning

- For Trustworthy Discoveries & Decision-making**
-

Lili Zheng

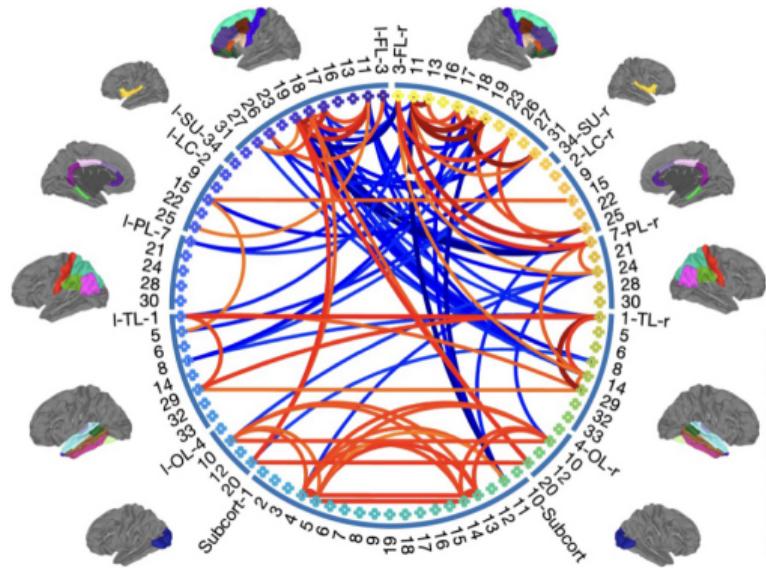
Department of Electrical and Computer Engineering, Rice University

Table of contents

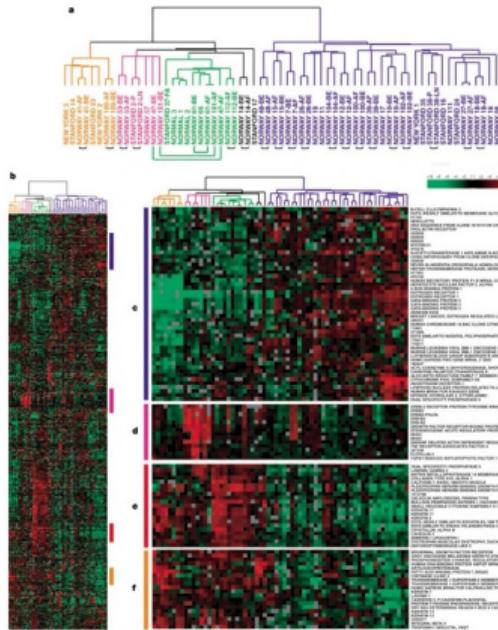
1. Background
2. Uncertainty Quantification for Statistical Structure (Graph) Learning
3. Uncertainty Quantification for Model-agnostic Machine Learning Interpretations
4. Other Works and Future Directions

Background

Machine Learning Propels Discoveries



Association between brain regions from
fMRI data



Hierarchical clustering for tumor data
(Perou et al., 2000)

Machine Learning Propels Decision-making



Treatment in healthcare



Loan approval

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

Interpretable Machine Learning (IML)

Interpretable Machine Learning

Generate human-understandable insights into **the data, the ML model, or the model output**

Interpretable Machine Learning (IML)

Interpretable Machine Learning

Generate human-understandable insights into **the data, the ML model, or the model output**

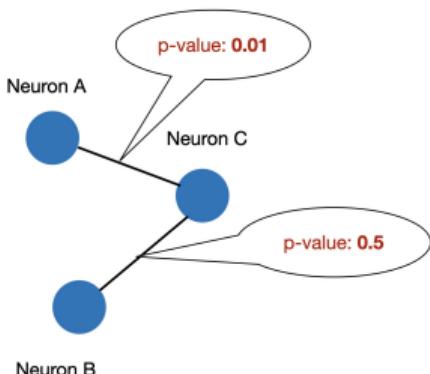
- **Insights into the data:** functional association between brain regions; which treatment is more effective?
 - **Insights into the model:** model-diagnostics; safety check
- ⇒ **scientific discoveries, decision-making**

Can we trust interpretable machine learning for discoveries and decision-making?

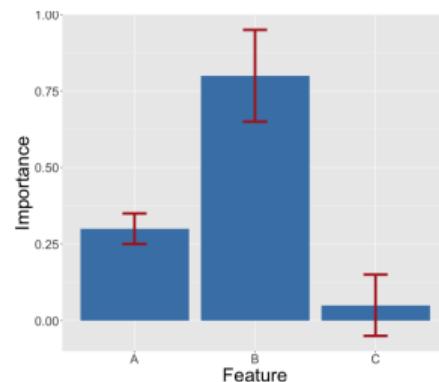
Trust in IML?

One Potential Solution

Provide **uncertainty quantification** (UQ) associated with machine learning interpretations!



p-values for
detected
association



Confidence
intervals for
feature
importance

– draw conclusions/make decisions only based on *significant signals*.

Uncertainty Quantification: Challenges in the Modern Era

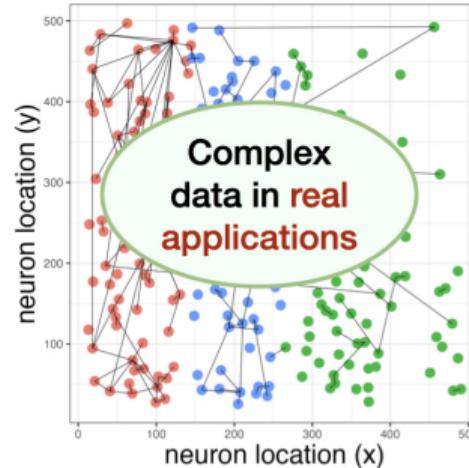
- Great tools in statistics & machine learning: selective inference, conformal inference, Bayesian inference...
- Numerous challenges from **large-scale, complex data and models!**

Rigorous uncertainty quantification in practical and complex scenarios?

Uncertainty Quantification: Challenges in the Modern Era

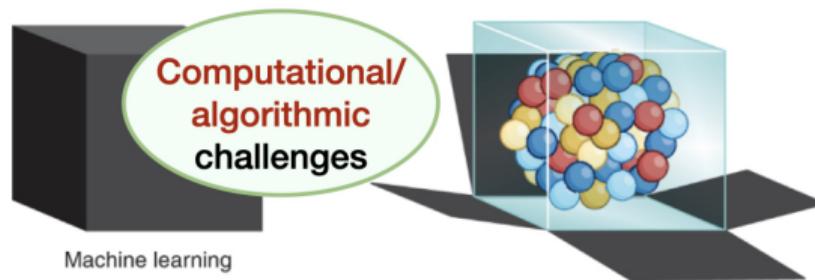
First part:

ML in science
UQ for graph learning



Second part:

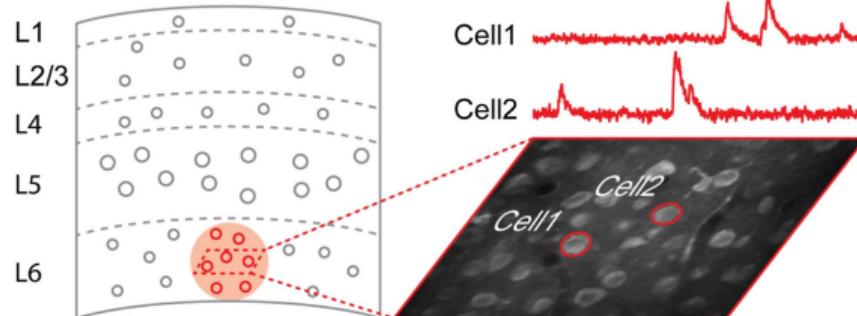
ML in the society
UQ for model-agnostic ML interpretations



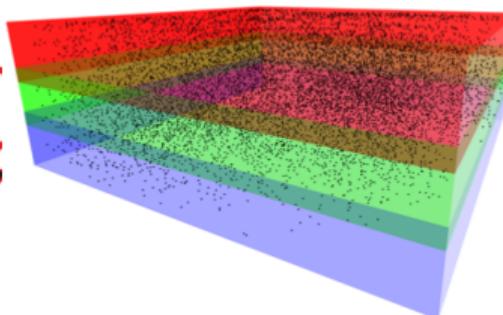
Uncertainty Quantification for Statistical Structure (Graph) Learning

Challenges from Data: Erose Measurements

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



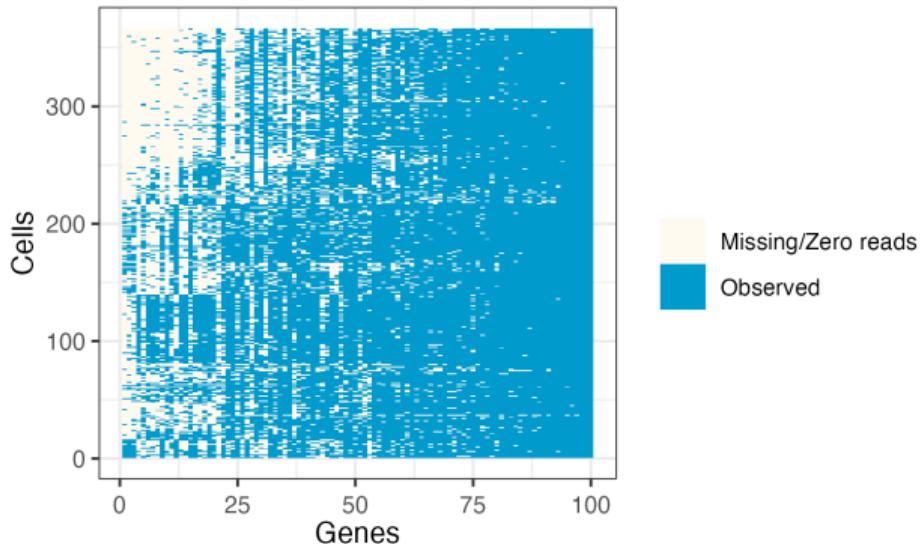
Calcium imaging data in neuroscience
(Birkner et al., 2017)



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

Challenges from Data: Erose Measurements

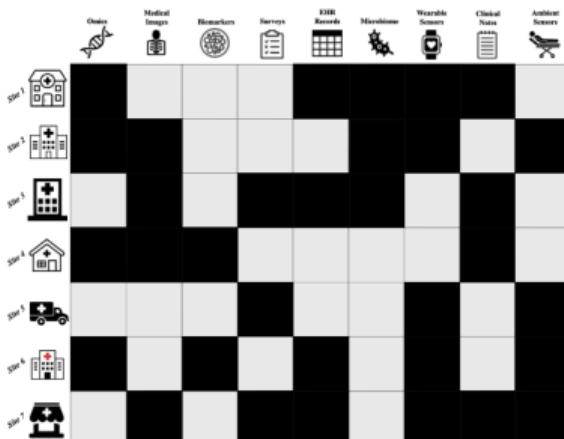
Erose measurements: irregular, highly uneven measurements over a large system



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)

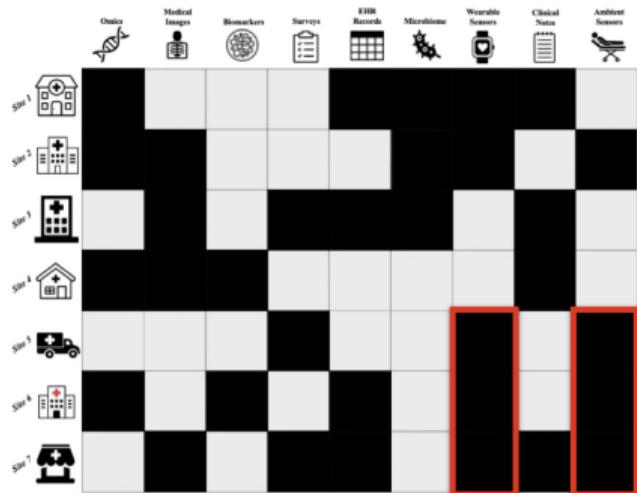
Table 1. Some examples of unequally spaced surveys.		
Country	Panel A: household surveys for monitoring poverty in developing countries ^a	
	Survey	Survey periods
Bolivia	Encuesta Integrada de Hogares (EIH)	Mar 89, Nov 89, Sept 90, Nov 91, Nov 92, July-Dec 93, July-Dec 94, June 95
Brazil	Pesquisa Nacional por Amostra de Domicílios (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

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

Structure Learning from Erose Measurements?

Common practices

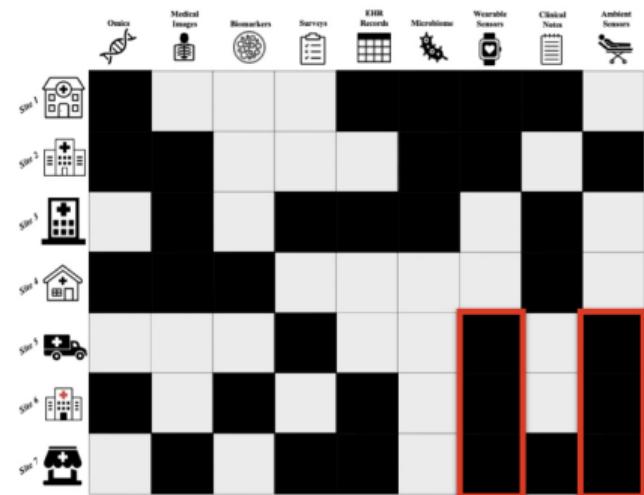
- Downsampling: focus on a complete block;
 - throw too much data away!



Structure Learning from Erode Measurements?

Common practices

- Downsampling: focus on a 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 erode measurements in this talk

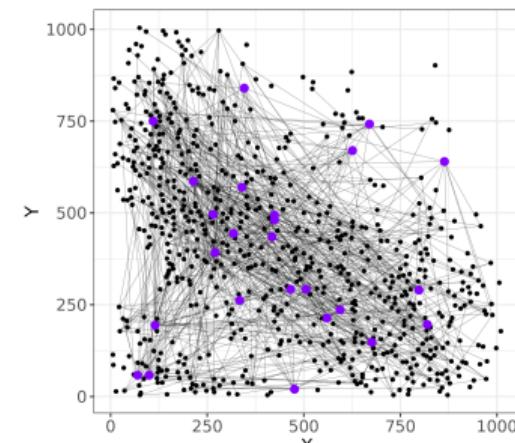
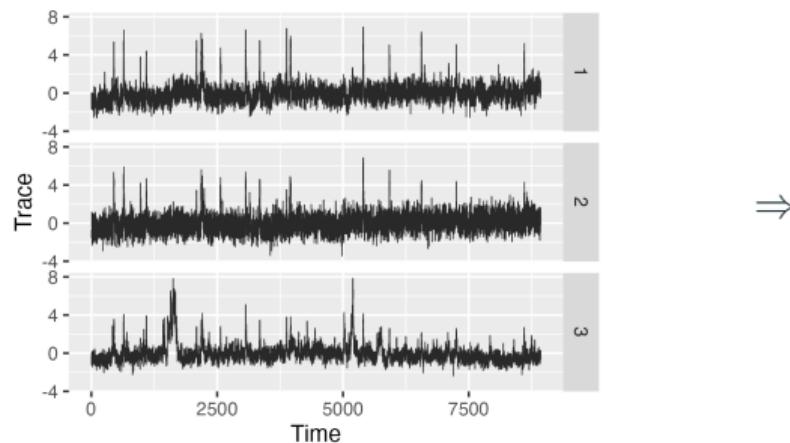
Why Graph Learning?

Graphical Model Structural Learning

Extract **conditional dependency** relationships:

An edge between node j and $k \iff$ Observations for j and k are conditionally dependent given all other nodes.

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



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
 - Nodes: $V = [p]$;
 - n samples of p -dimensional r.v.s:
 $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}(0, \Theta^{*-1})$;
 - $\Theta_{j,k}^* \neq 0 \Leftrightarrow j \not\perp\!\!\!\perp k | \text{all other nodes}$
 - Edges:
 $E = \{(j, k) : 1 \leq j, k \leq p, \Theta_{j,k}^* \neq 0\}$;
 - **Goal:** identify non-zero entries in Θ^*

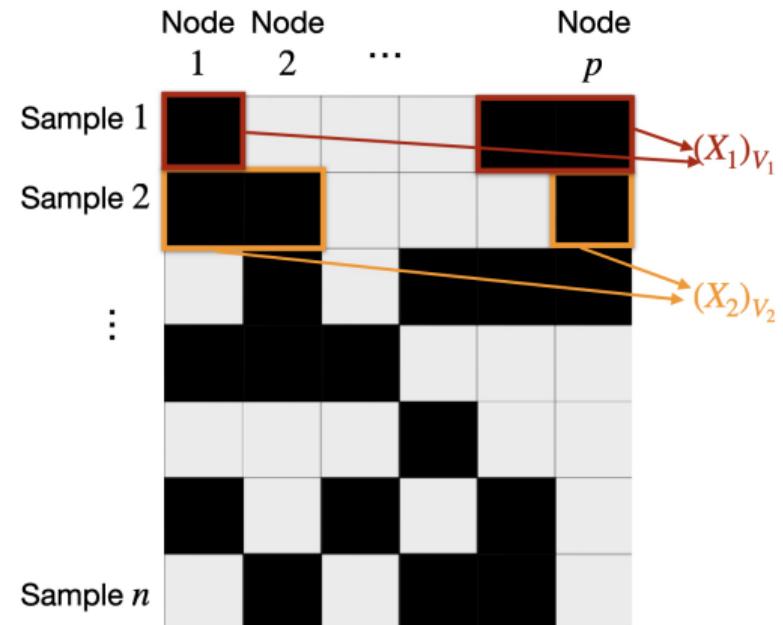
Gaussian Graphical Model Learning from Erose Measurements

- Focus on Gaussian graphical models in this talk

- Nodes: $V = [p]$;
- n samples of p -dimensional r.v.s: $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}(0, \Theta^{*-1})$;
- $\Theta_{j,k}^* \neq 0 \Leftrightarrow j \not\perp\!\!\!\perp k \mid \text{all other nodes}$
- Edges:
- $E = \{(j, k) : 1 \leq j, k \leq p, \Theta_{j,k}^* \neq 0\}$;
- Goal: identify non-zero entries in Θ^***

- Erose measurements**

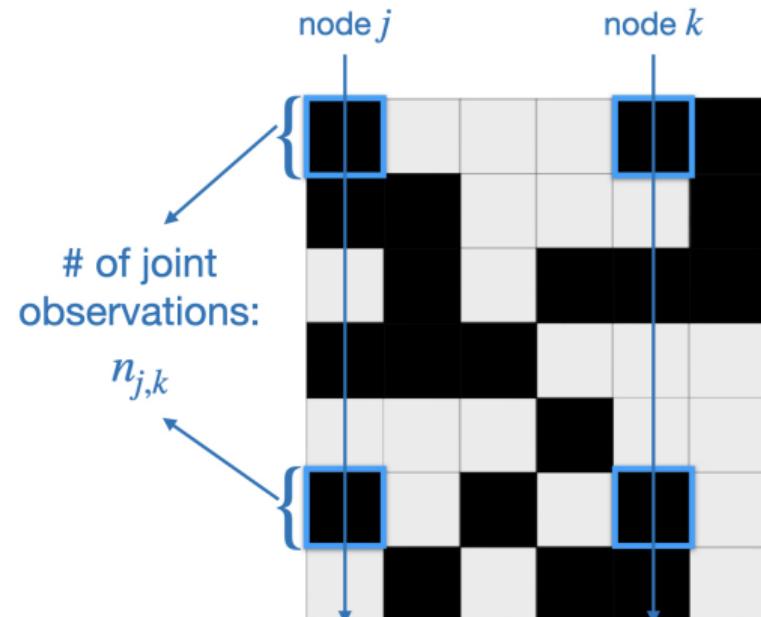
Observe $(\mathbf{X}_i)_{V_i}, 1 \leq i \leq n; V_i \subset [p]$
are irregular feature subsets,
independent from \mathbf{X}_i .



Gaussian Graphical Model Learning from Erose Measurements

- **Erose measurements**

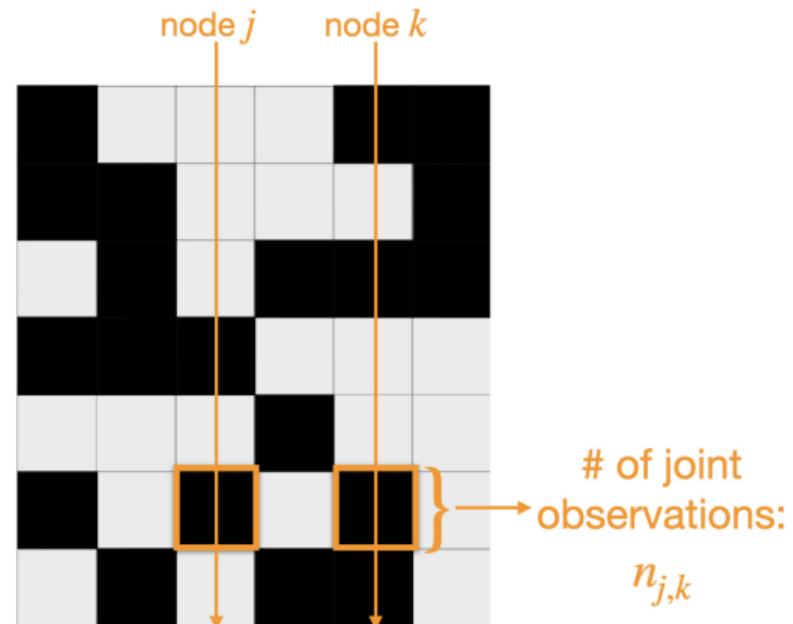
Even for assessing marginal dependency:
joint sample sizes for node pairs
 $\{n_{j,k} : 1 \leq j, k \leq p\}$ are **highly**
different



Gaussian Graphical Model Learning from Erose Measurements

- **Erose measurements**

Even for assessing marginal dependency:
joint sample sizes for node pairs
 $\{n_{j,k} : 1 \leq j, k \leq p\}$ are **highly** different



Prior Works on Gaussian Graphical Models from Partial Observations

Estimation

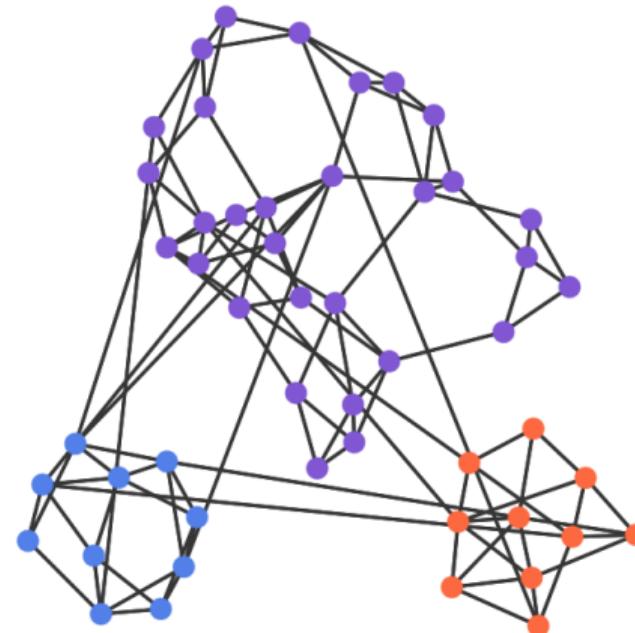
- Plug in covariance estimates into graphical Lasso (Kolar and Xing, 2012; Park et al., 2021)
- Most theory assumes missing with the **same/similar probability!**
- Existing characterization in minimum pairwise sample size
- **Limited insights for our setting**

Inference

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

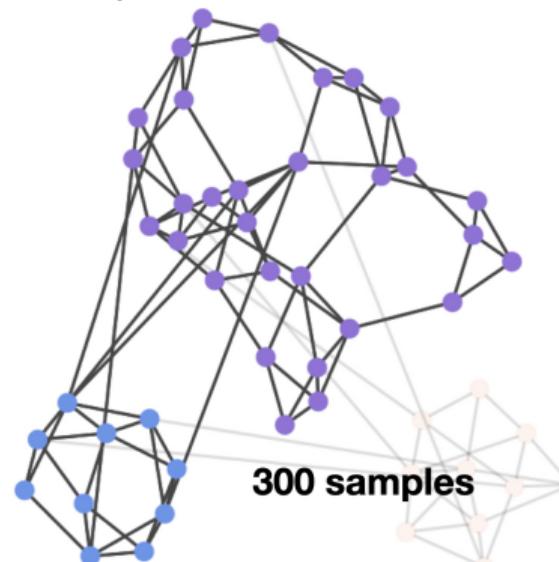
Toy Example: UQ Promotes Reliable Graph Learning

- Toy example: irregular patchwise observations
- $p = 30 + 10 + 10 = 50$ nodes in total

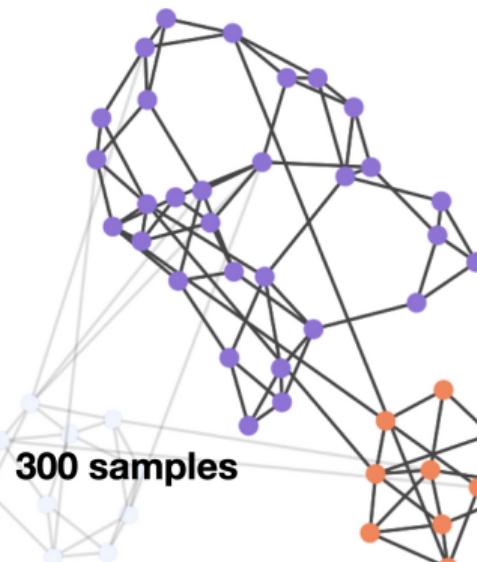


Toy Example: UQ Promotes Reliable Graph Learning

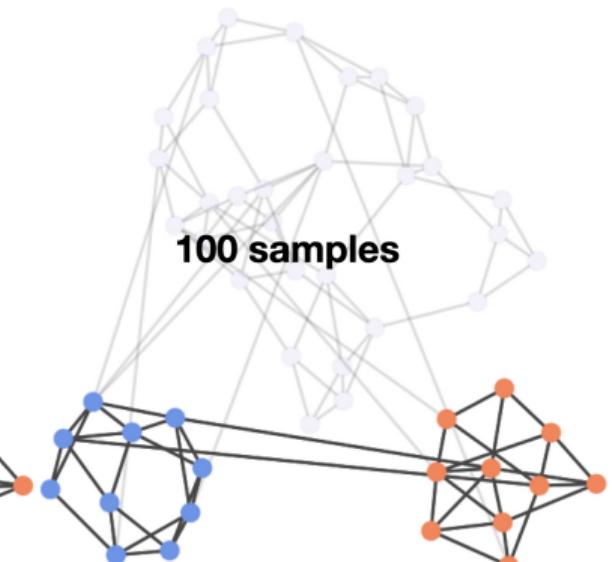
- Toy example: irregular patchwise observations
- $p = 30 + 10 + 10 = 50$ nodes in total



Measurement 1



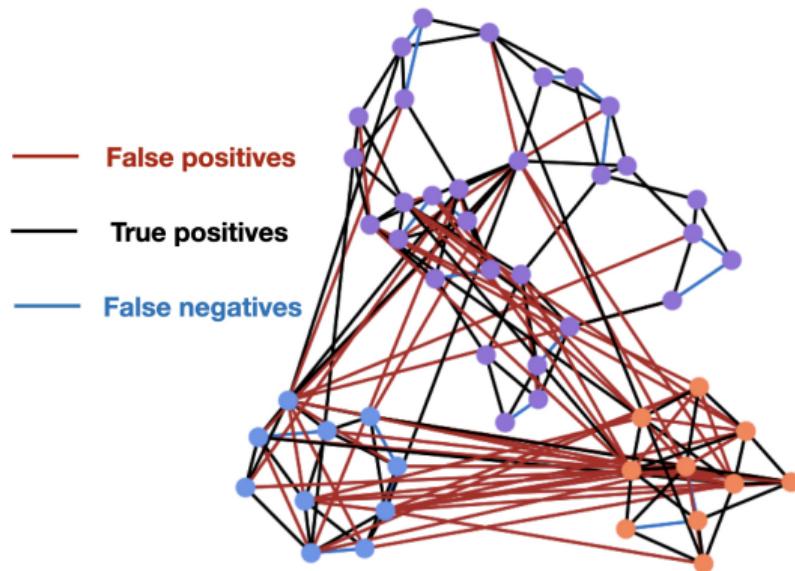
Measurement 2



Measurement 3

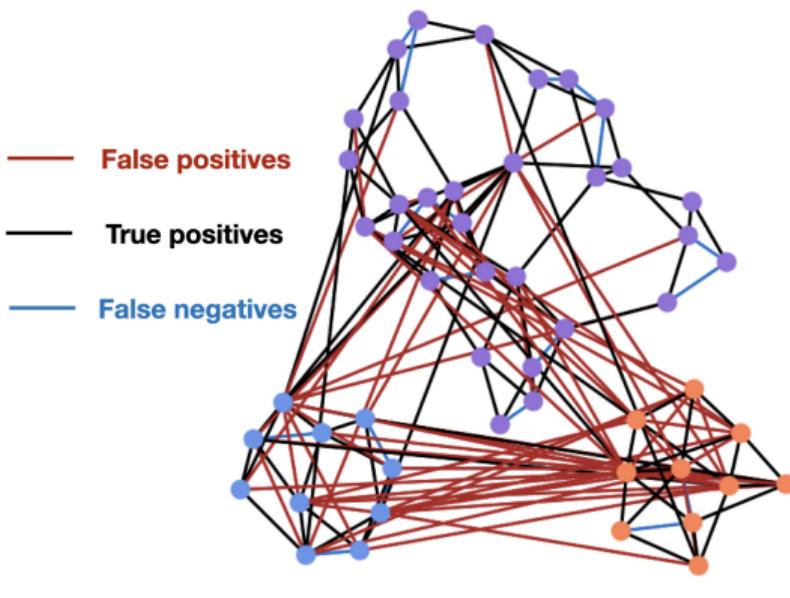
Toy Example: UQ Promotes Reliable Graph Learning

- Plug-in estimate using graphical lasso

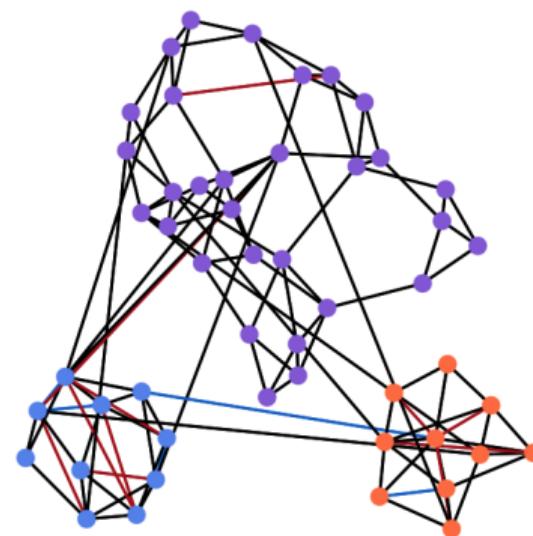


Toy Example: UQ Promotes Reliable Graph Learning

- Plug-in estimate using graphical lasso



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



Problem Setup and Proposed Method

Recall: Model Setup

Gaussian graphical model:

- p -dimensional $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}(0, \Theta^{*-1})$;
- Nodes: $V = [p]$;
- Edges: $E = \{(j, k) : 1 \leq j, k \leq p, \Theta_{j,k}^* \neq 0\}$;

Observations

- $(\mathbf{X}_i)_{V_i}, 1 \leq i \leq n; V_i \subset [p]$ are irregular feature subsets independent from \mathbf{X}_i
- Pairwise joint sample sizes $\{n_{j,k} : 1 \leq j, k \leq p\}$ are highly different

Recall: Model Setup

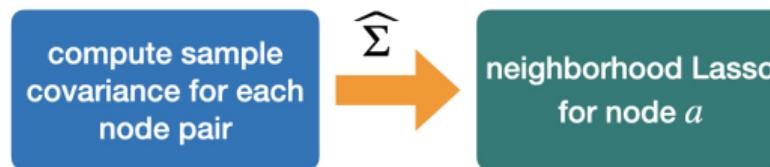
Observations

- $(\mathbf{X}_i)_{V_i}, 1 \leq i \leq n; V_i \subset [p]$ are irregular feature subsets independent from \mathbf{X}_i
- Pairwise joint sample sizes $\{n_{j,k} : 1 \leq j, k \leq p\}$ are highly different

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

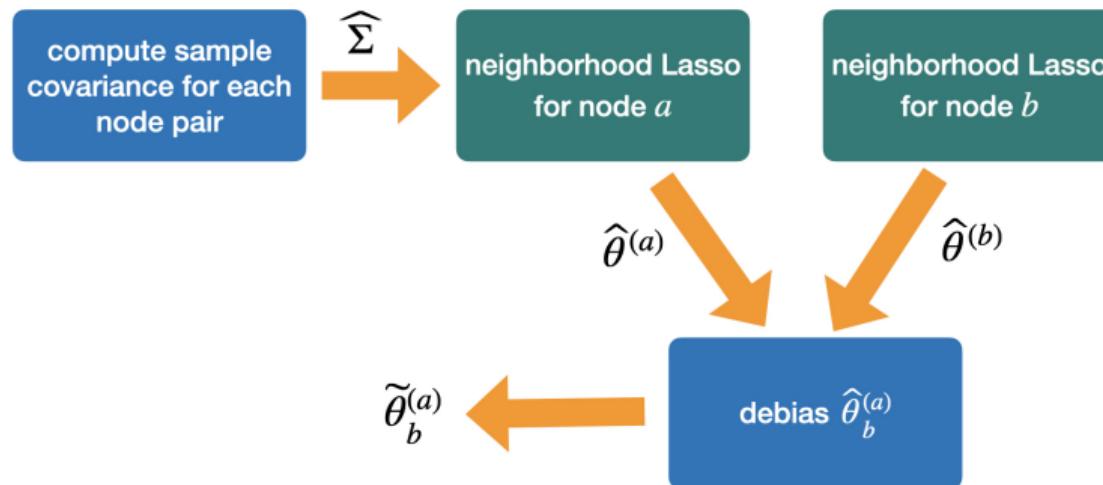


$$\hat{\theta}^{(a)} = \arg \min_{\theta \in \mathbb{R}^p, \theta_a=0} \frac{1}{2} \theta^\top \hat{\Sigma} \theta - \hat{\Sigma}_{a,:} \theta + \sum_{j=1}^p \lambda_j |\theta_j|,$$

$|\hat{\theta}_b^{(a)}|$ indicates edge strength of (a, b)

Edgewise Inference: Debiased Neighborhood Lasso

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



$|\widetilde{\theta}_b^{(a)}|$ also indicates edge strength of (a, b)

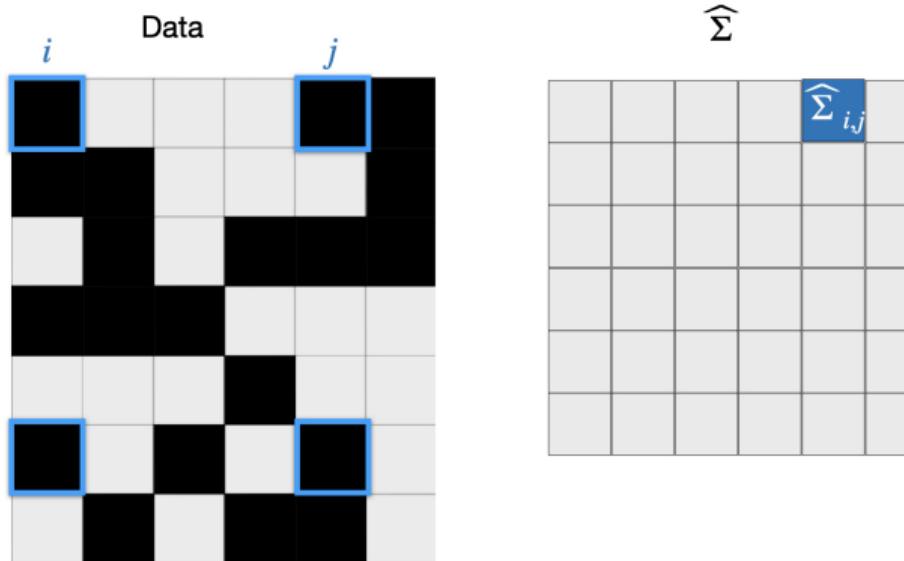
Edgewise Inference: Debiased Neighborhood Lasso

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

Edgewise Inference: Debiased Neighborhood Lasso

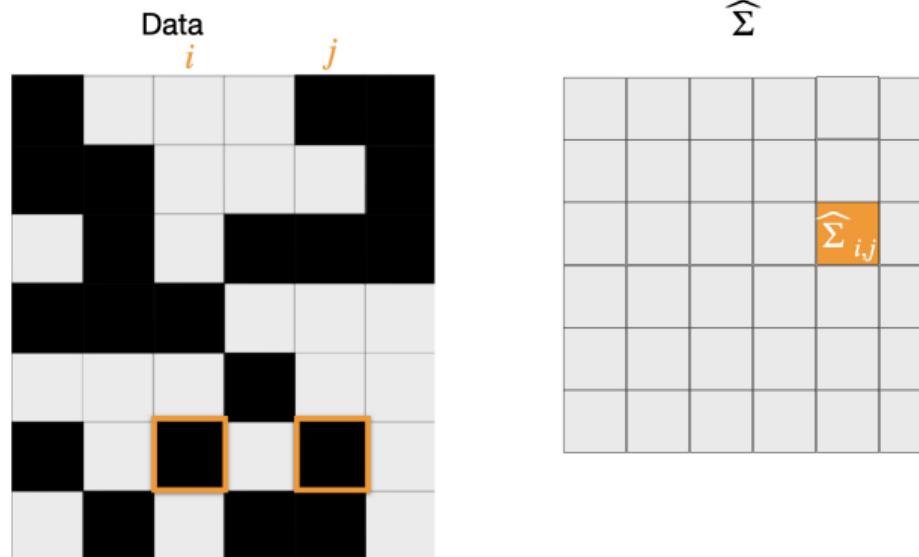
- Step 2: Normal approximation for $\tilde{\theta}_b^{(a)}$ and variance estimation

Challenge: $\widehat{\Sigma}$ computed from irregular data patches



Edgewise Inference: Debiased Neighborhood Lasso

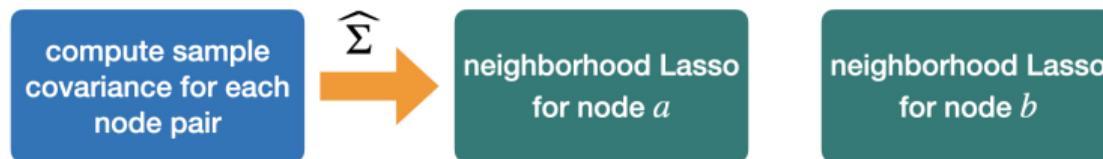
- Step 2: Normal approximation for $\tilde{\theta}_b^{(a)}$ and variance estimation
Challenge: $\widehat{\Sigma}$ computed from irregular data patches



Edgewise Inference: Debiased Neighborhood Lasso

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

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



Prior estimation theory hinges on the **minimum pairwise sample size**...

Characterization of Debiased Neighborhood Lasso

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

With appropriately chosen tuning parameters in the neighborhood Lasso,

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

Characterization of Debiased Neighborhood Lasso

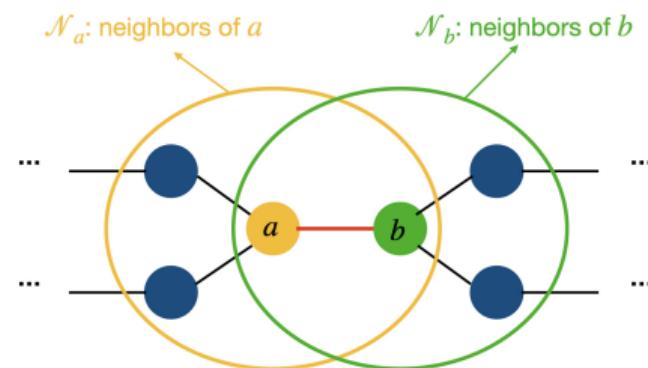
A Closer Look into $\tilde{\theta}_b^{(a)}$

With appropriately chosen tuning parameters in the neighborhood Lasso,

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

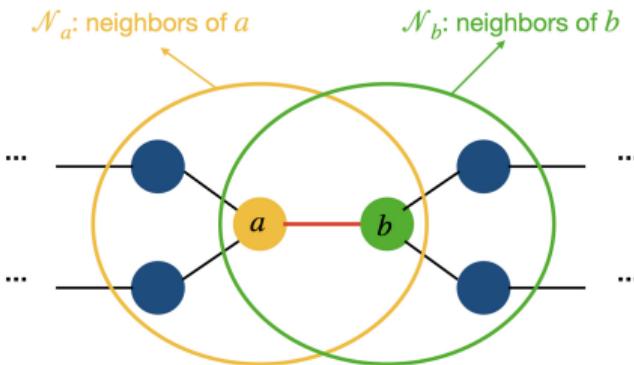
- mean-zero first-order term

$$\propto \sum_{j,k} (\hat{\Sigma}_{j,k} - \Sigma_{j,k}^*) \underbrace{\Theta_{a,j}^* \Theta_{b,k}^*}_{\text{weight of node pair } (j, k)}$$



- only involve neighbors of a and b !

GI-JOE: Edge-wise Uncertainty Quantification

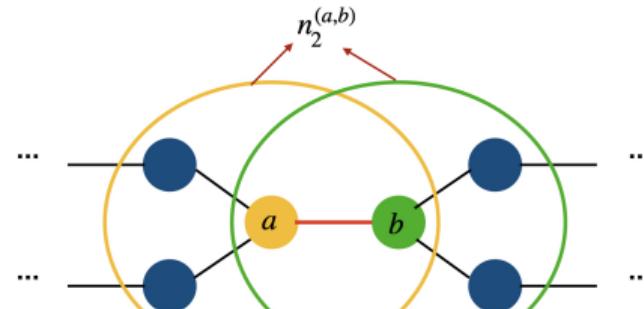
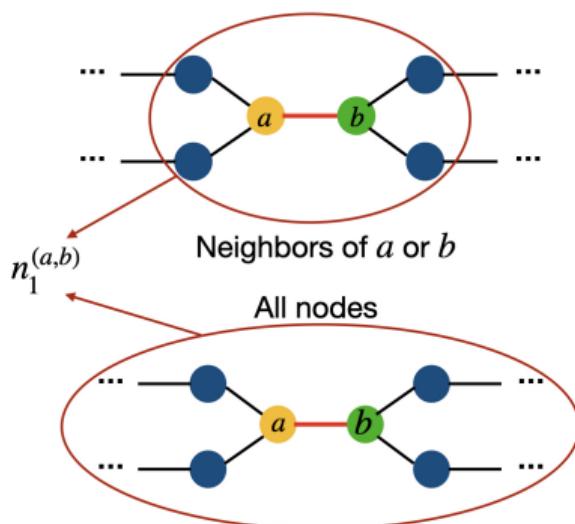


- **Step 2:** Estimate variance of first-order term
 - Variance contribution from each node pair (j, k) :
 $\widehat{\theta}_j^{(a)}, \widehat{\theta}_k^{(b)}, 1/n_{j,k}$
 - Plus some edge-edge correlations
 - Obtain $\widehat{\sigma}_n^2(a, b)$
- **Output:** Reject $\mathcal{H}_0 : (a, b) \notin E$ if
$$\frac{|\widehat{\theta}_b^{(a)}|}{\widehat{\sigma}_n(a, b)} > z_{\alpha}/2.$$

Edgewise Testing: Theoretical Guarantees

Assumption for Validity: Sufficient Local Sample Sizes

Minimum pairwise sample size between $\begin{cases} \text{neighbors of } a \text{ or } b \text{ and all nodes: } n_1^{(a,b)} \\ \text{neighbors of } a \text{ and neighbors of } b: n_2^{(a,b)} \end{cases}$



Assumption for Validity: Sufficient Local Sample Sizes

Main Assumption

A1. The local sample size $n_1^{(a,b)}$ is sufficiently large, as a function of node degrees and dimensions.

Assumption for Validity: Sufficient Local Sample Sizes

Main Assumption

A1. The local sample size $n_1^{(a,b)}$ is sufficiently large, as a function of node degrees and dimensions.

- Same sample size $n_{j,k} = n$ setting: reduces to prior requirements $n \gg d^2 \log^2 p$
- **First theory that allows for general erose measurements:**
 - arbitrary data-independent missing pattern!
 - localized sample size requirement: not the minimum pairwise sample size as in existing theory!

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 $\Theta_{a,b}^*$ are also available.

GI-JOE: FDR control

Whole graph testing with FDR control?

- Want: 95% of the selected edges are true positives
- Take edgewise p -values, apply Benjamini-Hochberg's procedure but with a truncation step
- Inspired by FDR control for the debiased Lasso (Javanmard and Javadi, 2019)

Theoretical Guarantees

Theorem: Valid FDR control

Assume

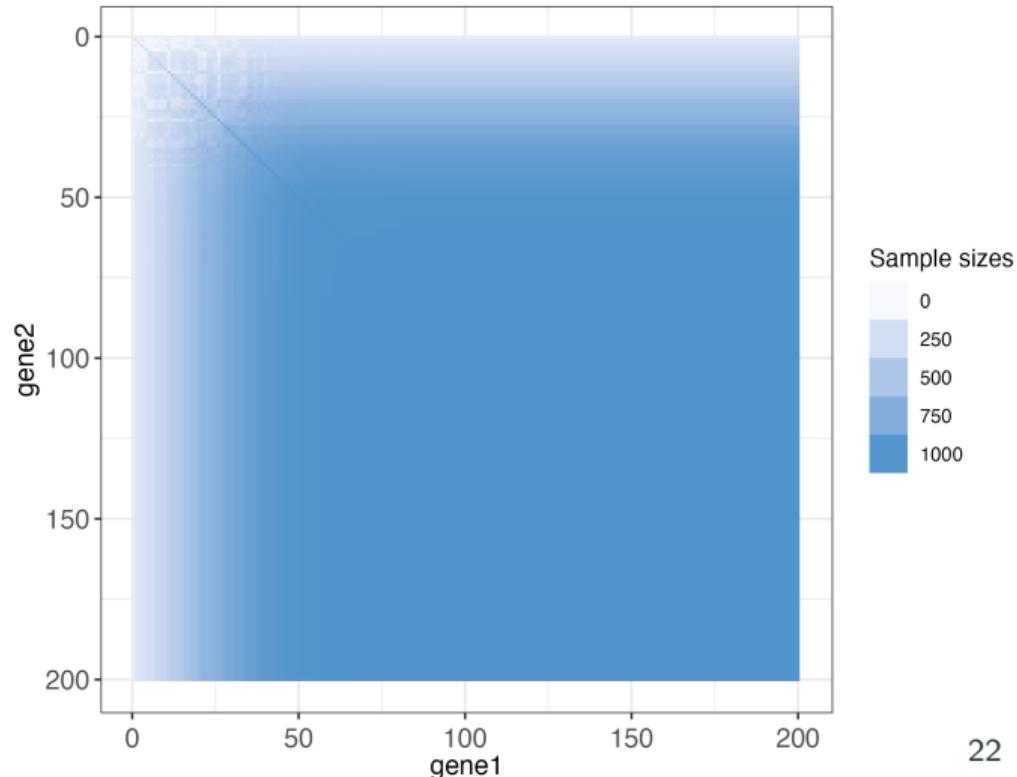
1. $n_1^{(a,b)}$ is sufficiently large for all (a, b) (**holds even if $\max n_{j,k} \gg \min n_{j,k}$**);
2. Most edge pairs $(a, b), (a', b')$ are only **weakly correlated** (satisfied by most sparse graphs).

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

Empirical Studies

Simulation: Graph Selection Comparison

- Simulate data from a **scale-free graph** with 200 nodes
- **Real measurement pattern** in a **real single-cell RNA sequencing** data set (Chu et al., 2016)

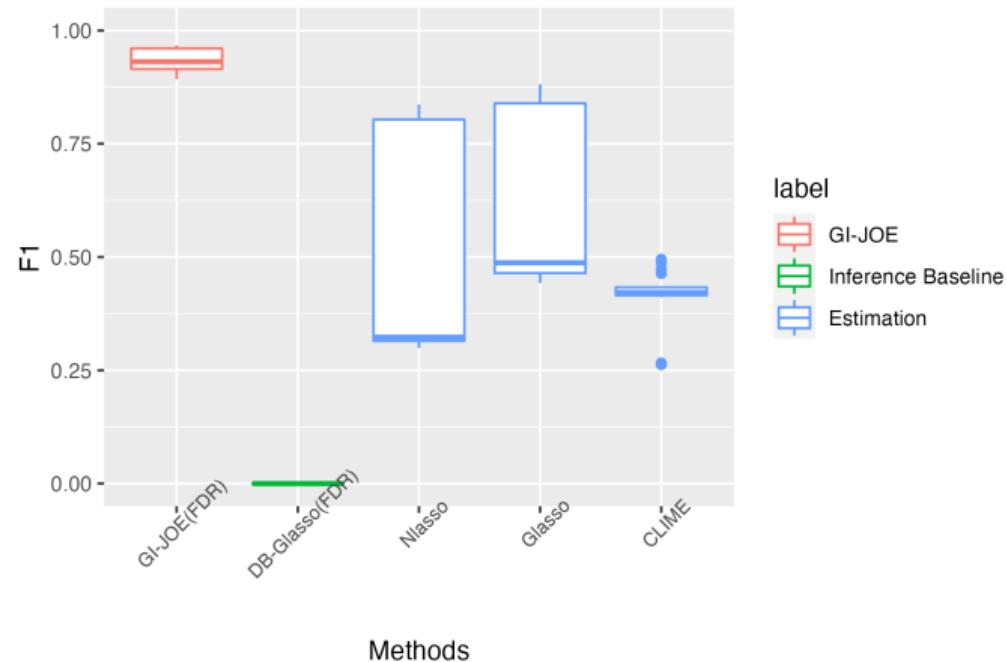


Simulation: Graph Selection Comparison

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

the higher the better

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

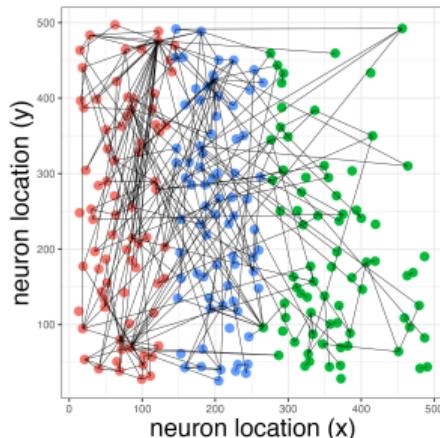


Application to Neuronal Functional Data

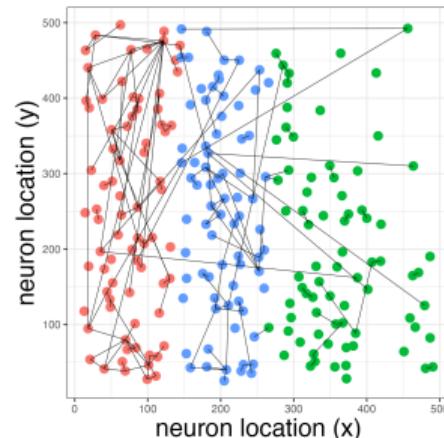
- Neuronal functional recordings of a mouse's visual cortex from Allen Brain Atlas
- Firing activities of $p = 227$ neurons, $n = 8931$ time points
- Goal: learn **functional connectivity amongst these neurons**
- Data is fully observed; we test how our method performs on **manually masked data**

Application to Neuronal Functional Data

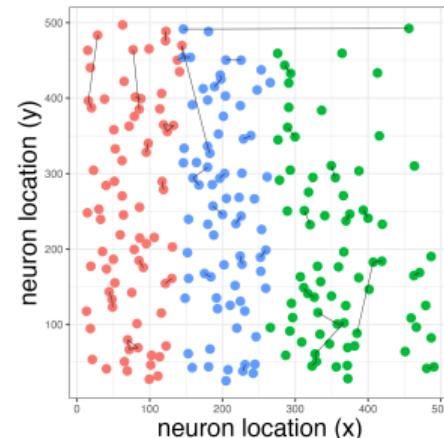
Manually mask functional data; three sets of neurons observed with high, median, and low probabilities.



FDR-selected graph with
full data (oracle)



GI-JOE (FDR), applied to
erase data



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

Summary

- Erose measurements: challenge for reliable graph learning

Summary

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

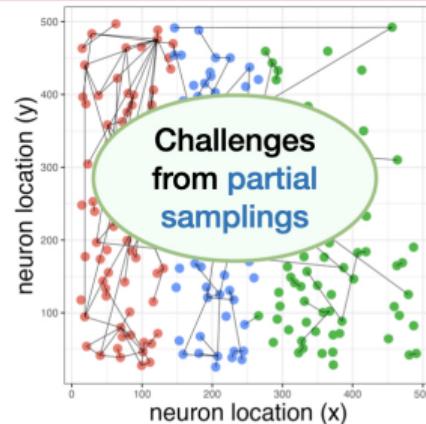
Summary

- 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!
- Future directions: more reliable **feature selection / causal structural learning** from erosely measured data under dependency?
- 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

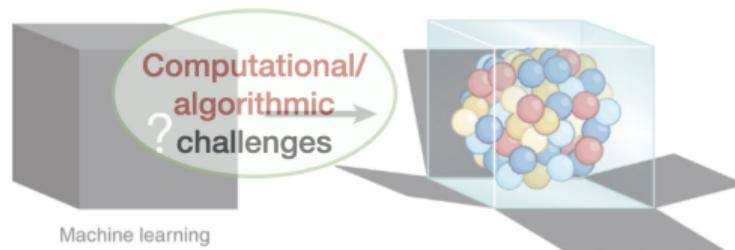
First part:

UQ for reliable
scientific
discoveries



Second part:

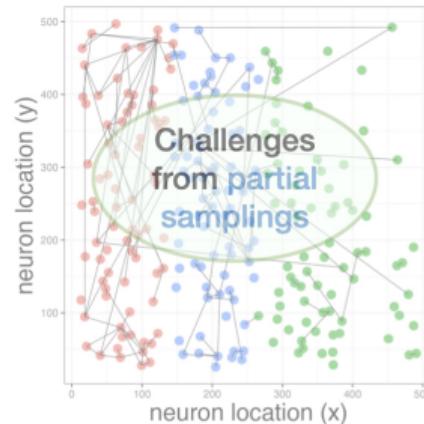
ML in the society
UQ for model-
agnostic ML
interpretations



From Complex Data Collection to Complex Machine Learning Systems

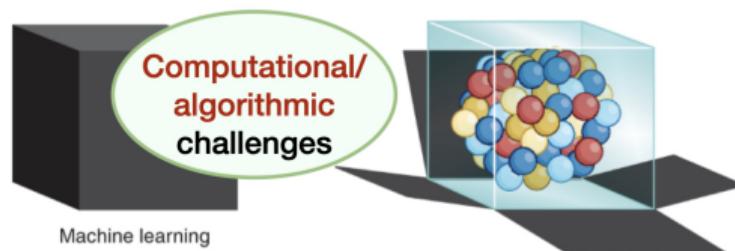
First part:

UQ for reliable
scientific
discoveries



Second part:

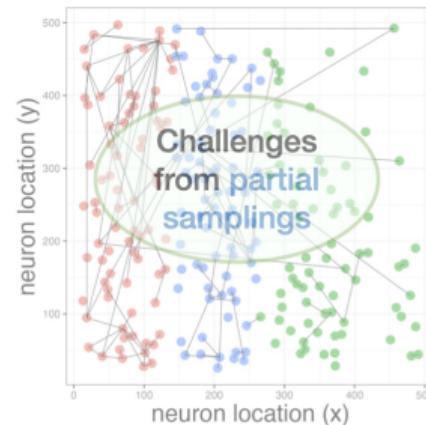
ML in the society
UQ for model-
agnostic ML
interpretations



From Complex Data Collection to Complex Machine Learning Systems

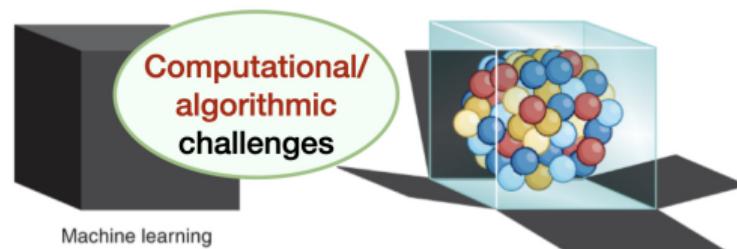
First part:

UQ for reliable
scientific
discoveries



Second part:

ML in the society
UQ for model-
agnostic ML
interpretations



Uncertainty Quantification for Model-agnostic Machine Learning Interpretations

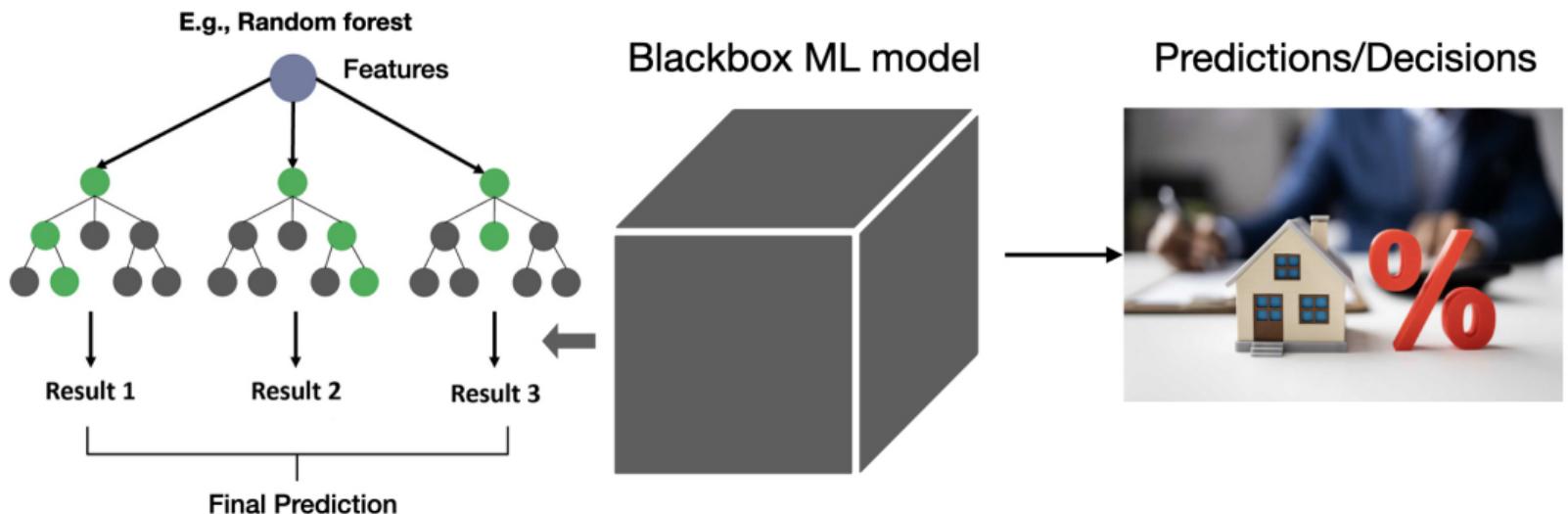
Interpreting Black-box Machine Learning Models

Machine learning is widely applied in **high-stakes applications**:



Can we trust machine learning? Make it interpretable!

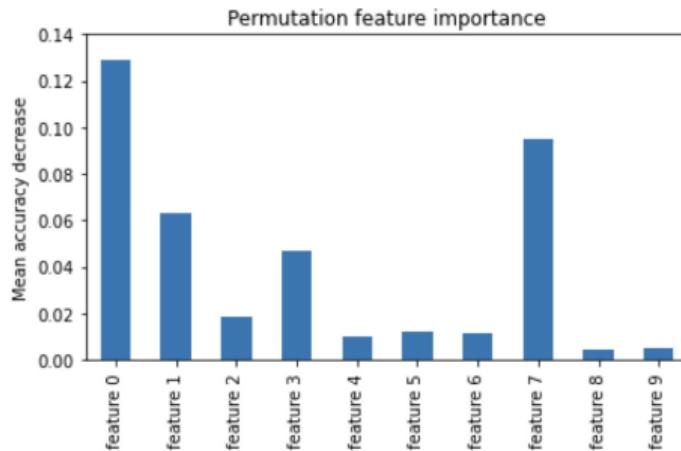
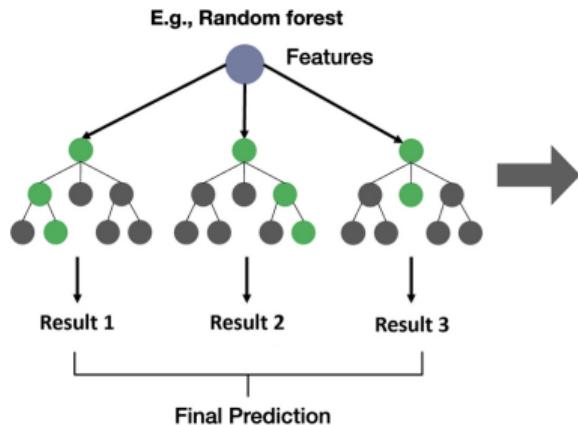
Interpreting Black-box Machine Learning Models



Why is this ML system rejecting my mortgage application?

Feature Importance for Interpretable Machine Learning

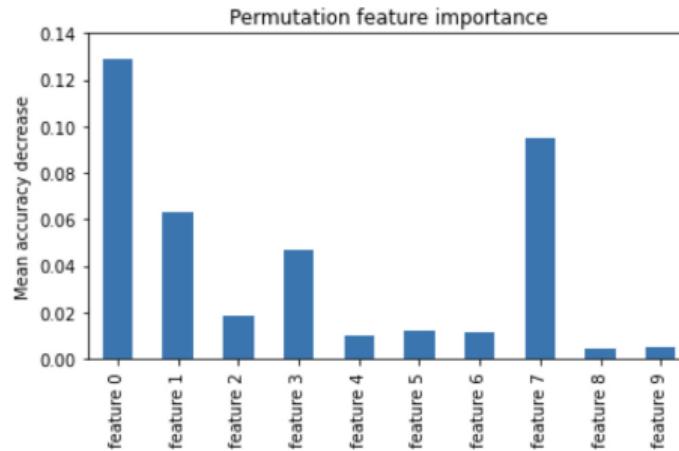
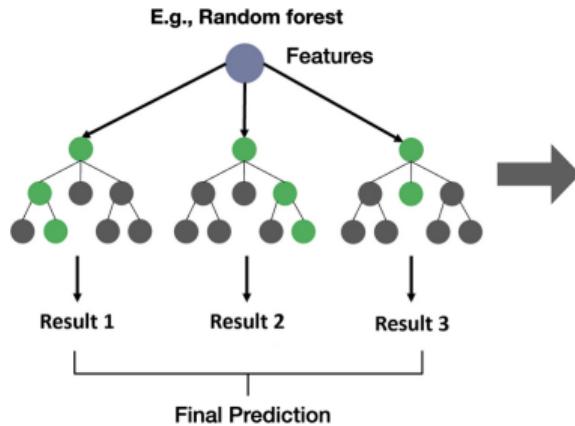
Feature importance: How does my model's prediction rely on each feature?



- Model-specific: defined for random forest, linear models, deep learning, etc.
- **Model-agnostic:** feature occlusion (Covert et al., 2021), permutation (König et al., 2021), Shapley values (Sundararajan and Najmi, 2020), etc.

Feature Importance for Interpretable Machine Learning

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



Can we trust feature importance? Uncertainty quantification?

Two Types of Feature Importance

Population feature importance

- Assume a [data-generating model](#); infer about the population
- E.g., [Conditional independence test](#), knockoff
- ML models are only tools
- [Impossible without strong assumptions about the data or model](#) (Shah and Peters, 2020)!

Two Types of Feature Importance

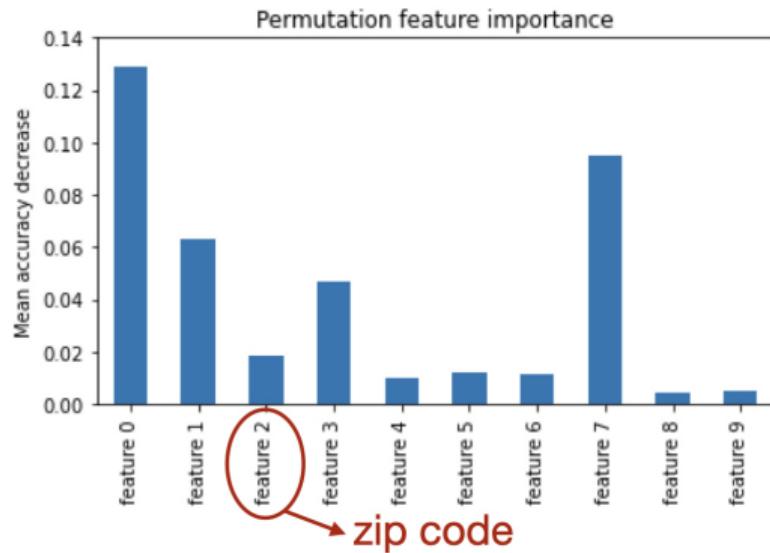
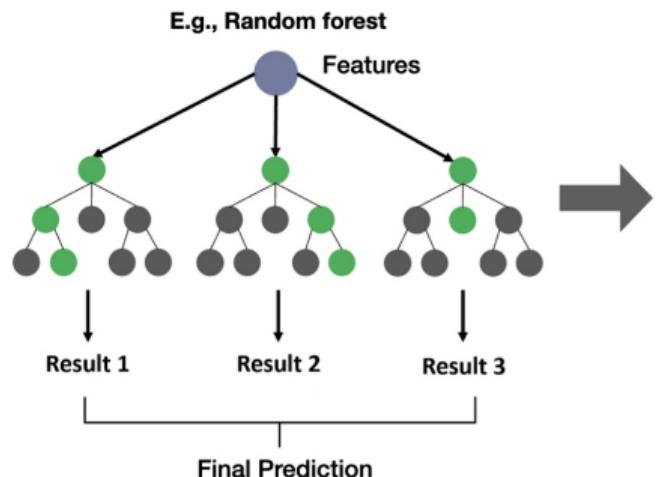
Population feature importance

- Assume a **data-generating model**; infer about the population
- E.g., **Conditional independence test**, knockoff
- ML models are only tools
- **Impossible without strong assumptions about the data or model** (Shah and Peters, 2020)!

ML feature importance

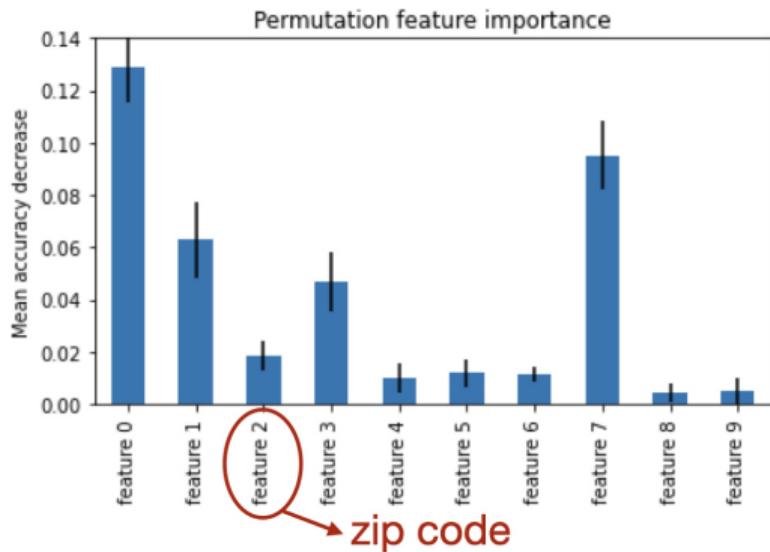
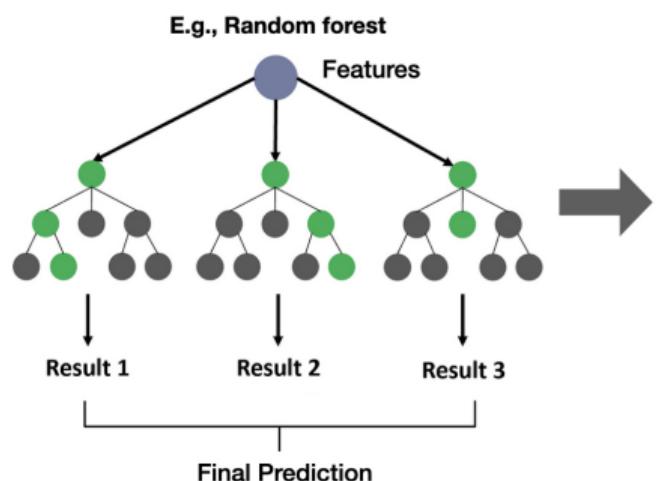
- Property of the **model**
- Which feature does my ML model rely on for decisions?
- Desired for **model diagnostics, auditing, and deployment**

ML Feature Importance



- Example: mortgage decision unfair w.r.t. races
- Why? Model is race-blind
- Check feature importance: zip code is a proxy of race?

ML Feature Importance



UQ for ML feature importance:

- **important societal consequences but understudied!**

Population feature importance

- Inference for Lasso (Lee et al., 2016; Van de Geer et al., 2014)
- Conditional independence test for random forest (Chi et al., 2022)
- Model-agnostic methods: Floodgate (Zhang and Janson, 2020), GCM (Shah and Peters, 2020), VIMP (Williamson et al., 2021)

ML feature importance

- Only a few works (Fisher et al., 2019; Lei et al., 2018; Rinaldo et al., 2019; Watson and Wright, 2021)
- Many are heuristic
- Most face computational challenges
- Efficient and rigorous UQ for ML feature importance?

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

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



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

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

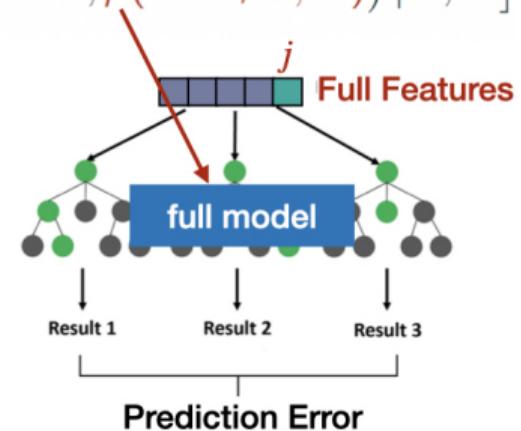


Inference target: Predictive power without feature j vs. with feature j .

Prior Work: LOCO Inference

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

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



Prior Work: LOCO Inference

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

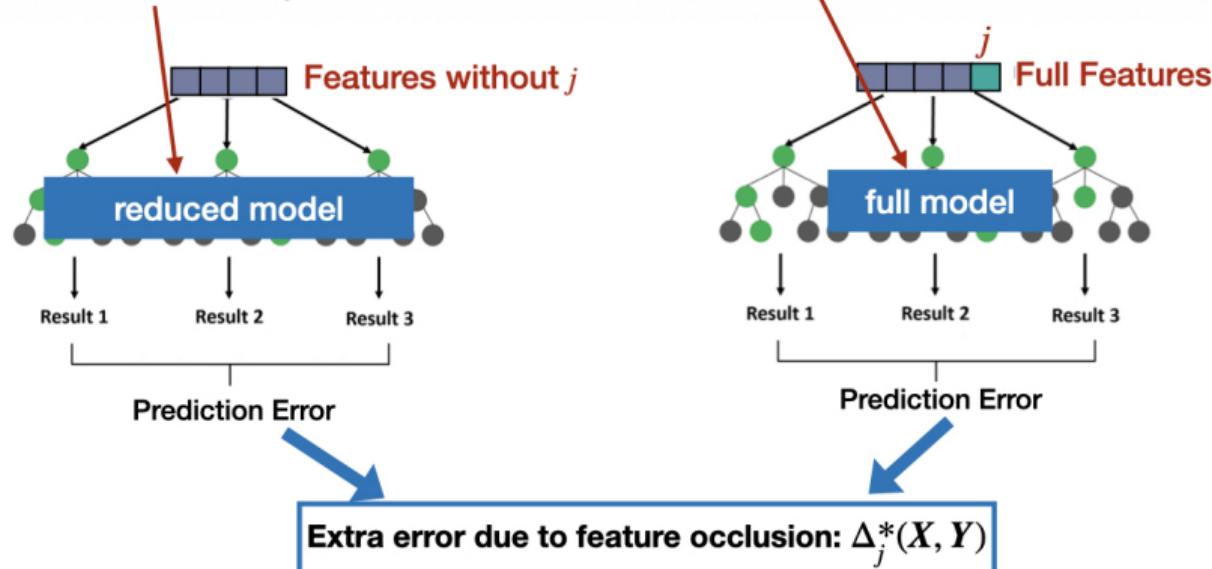
$$\Delta_j^*(\mathbf{X}, \mathbf{Y}) = \mathbb{E} [\text{Error}(Y^{\text{test}}, \mu_{-j}(X_{-j}^{\text{test}}; \mathbf{X}_{-j}, \mathbf{Y})) - \text{Error}(Y^{\text{test}}, \mu(X^{\text{test}}; \mathbf{X}, \mathbf{Y})) | \mathbf{X}, \mathbf{Y}]$$



Prior Work: LOCO Inference

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

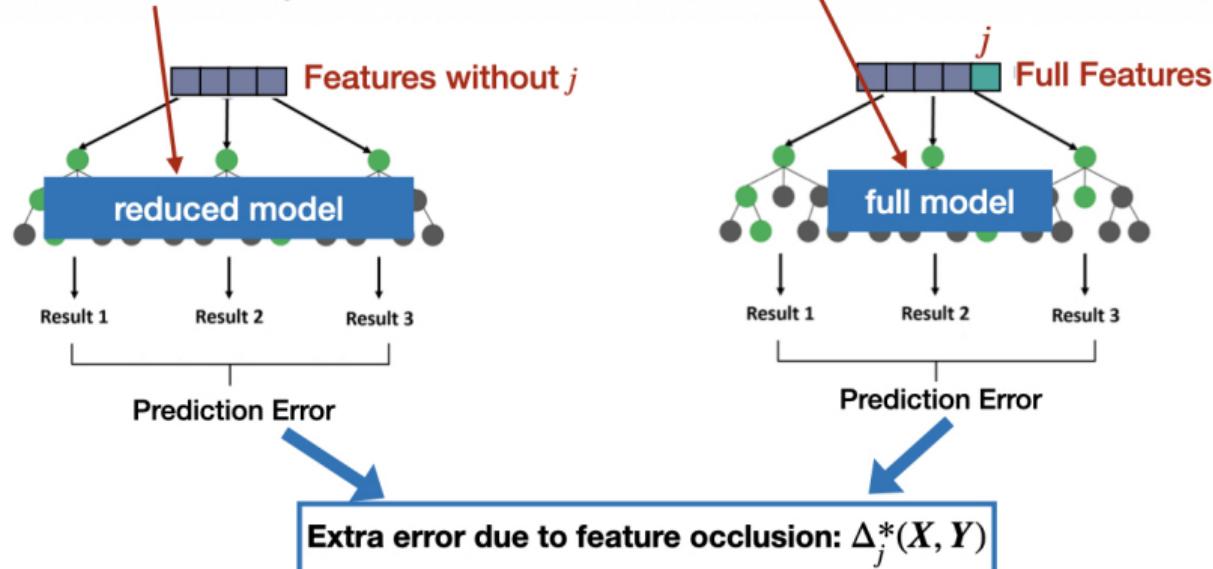
$$\Delta_j^*(\mathbf{X}, \mathbf{Y}) = \mathbb{E} [\text{Error} (Y^{\text{test}}, \mu_{-j}(X_{-j}^{\text{test}}; \mathbf{X}_{-j}, \mathbf{Y})) - \text{Error} (Y^{\text{test}}, \mu(X^{\text{test}}; \mathbf{X}, \mathbf{Y})) | \mathbf{X}, \mathbf{Y}]$$



Prior Work: LOCO Inference

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

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



Property of the current models; model-agnostic

LOCO inference approach:

- Split data; fit full and reduced models to training data
- Feature occlusion scores on test data \Rightarrow confidence intervals

Prior Work: LOCO Inference

LOCO inference approach:

- Split data; fit full and reduced models to training data
- Feature occlusion scores on test data \Rightarrow confidence intervals

Advantages

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

LOCO inference approach:

- Split data; fit full and reduced models to training data
- Feature occlusion scores on test data ⇒ confidence intervals

Advantages

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

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

LOCO inference approach:

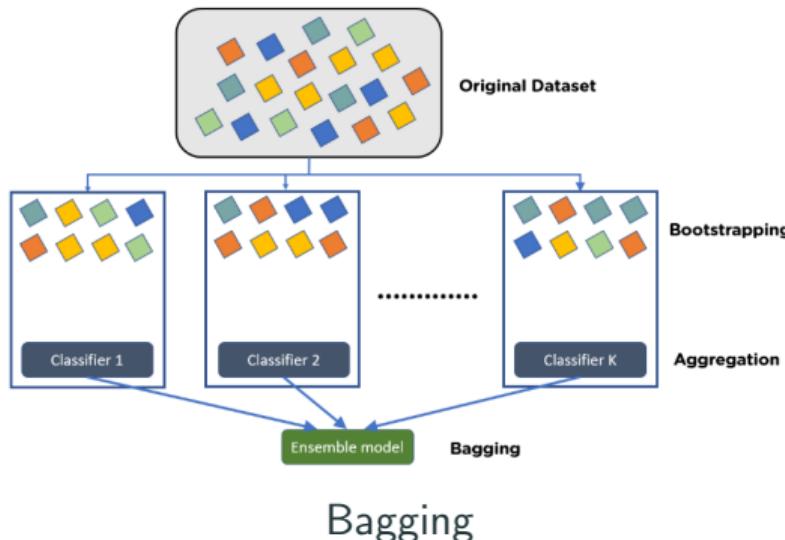
- Split data; fit full and reduced models to training data
- Feature occlusion scores on test data ⇒ confidence intervals

Our Goal

Can we utilize the general LOCO framework to perform ML feature importance inference, while [avoiding data splitting and model refitting](#)?

Our Approach: LOCO Inference for an Ensemble Framework

LOCO Inference for Ensemble Learning



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

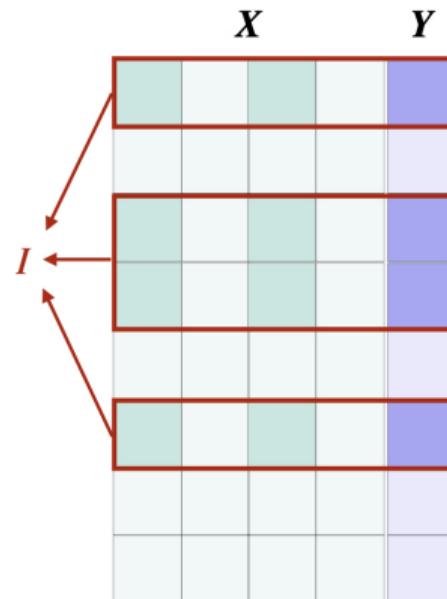
Inspiration: Jackknife+ After Bootstrap
(Kim et al., 2020).

- Many ensemble methods are good predictors
- Conformal inference (Jackknife+) for bagging is **computationally free with no data-splitting!**

Idea: Minipatch Ensembles.

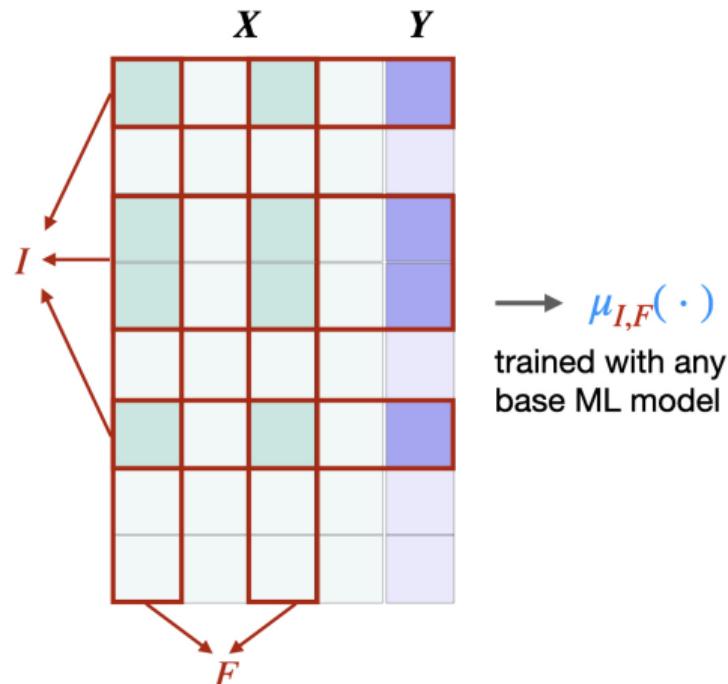
Minipatch Ensemble Learning

Minipatch ensembles: like bagging, but double-subsampling for both observations and features (Yao and Allen, 2020).



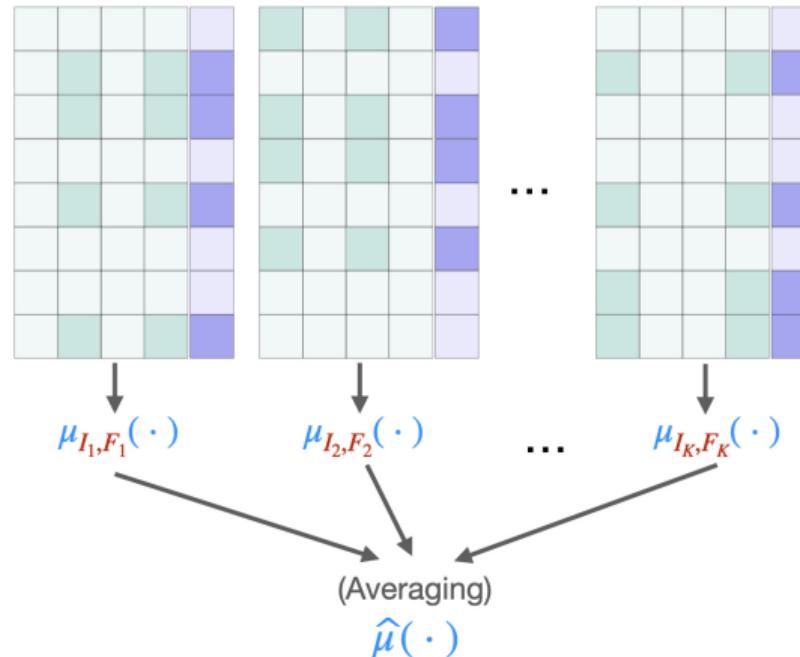
Minipatch Ensemble Learning

Minipatch ensembles: like bagging, but double-subsampling for both observations and features (Yao and Allen, 2020).



Minipatch Ensemble Learning

Minipatch ensembles: like bagging, but double-subsampling for both observations and features (Yao and Allen, 2020).

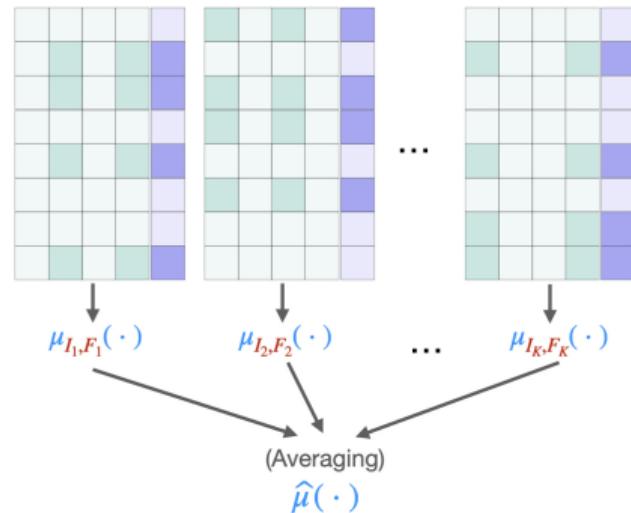


Minipatch Ensemble Learning

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

Advantages:

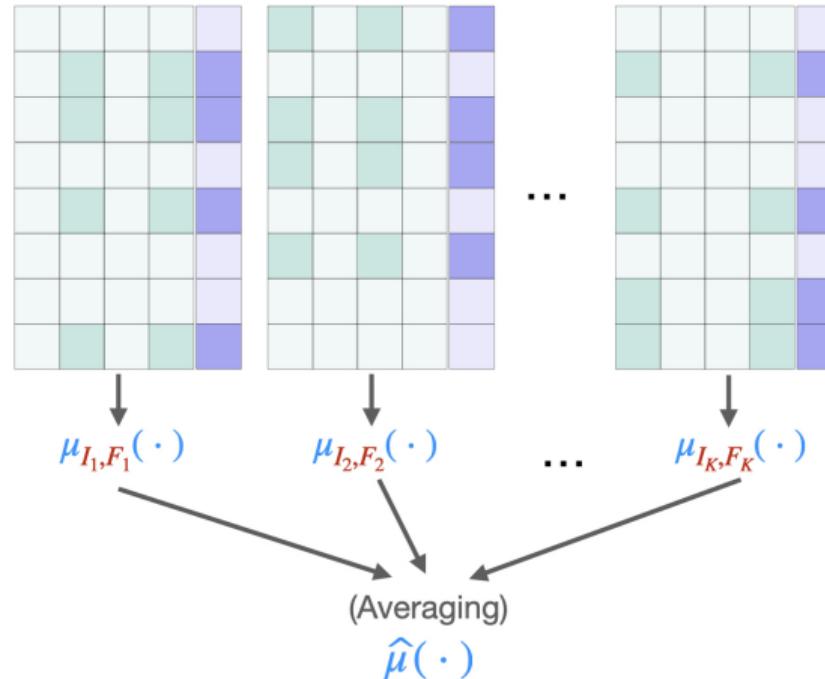
- Fast and easily parallelizable
- Ensemble diversity; **implicit regularization** (LeJeune et al., 2020; Yao et al., 2021)



LOCO Inference for Minipatch Ensembles?

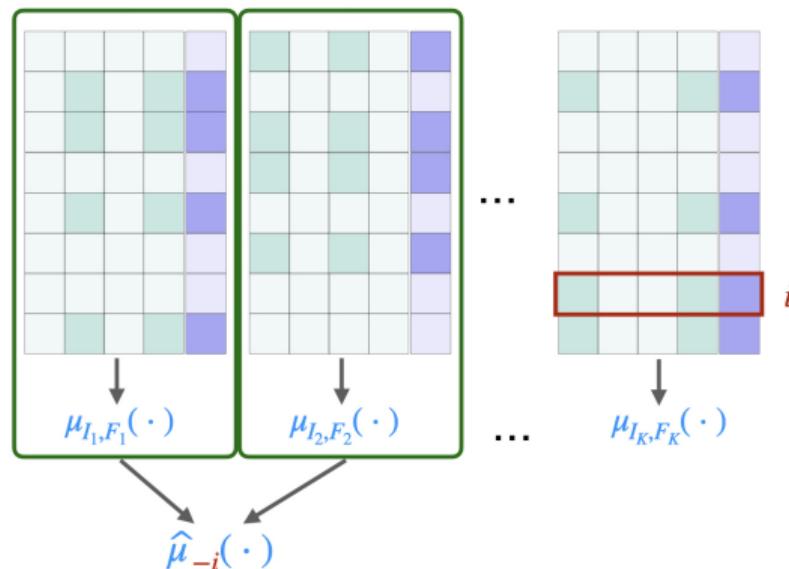
Algorithm: LOCO for Minipatch

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



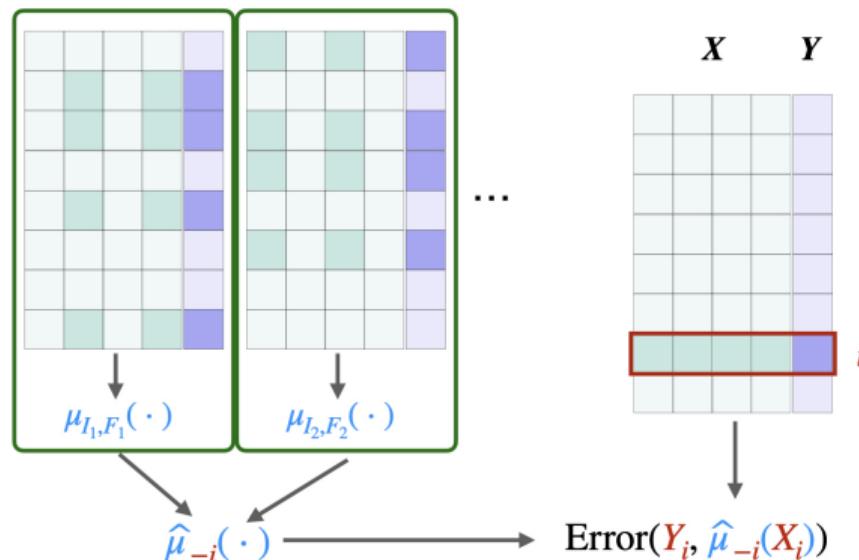
Algorithm: LOCO for Minipatch

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



Algorithm: LOCO for Minipatch

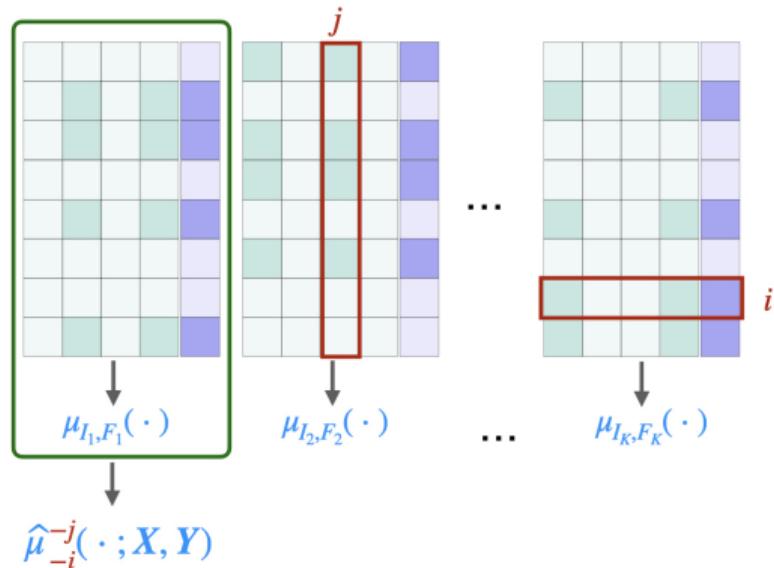
- Step 2. LOO (leave-one-observation-out) predictor: $\hat{\mu}_{-i}(X_i)$.
 - Ensemble minipatches **without observation i** .
 - Compute test error on sample i .



No data-splitting!
Simple model averaging;
Free computationally!

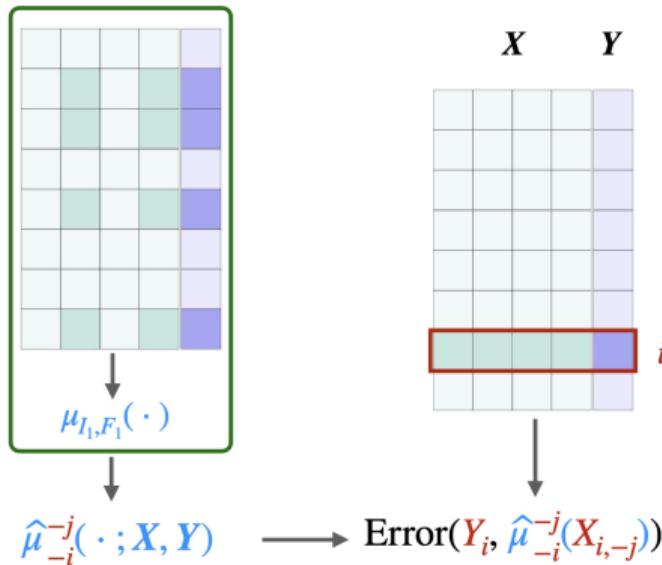
Algorithm: LOCO for Minipatch

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



Algorithm: LOCO for Minipatch

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



Simple model averaging;
Free computationally!

Algorithm: LOCO for Minipatch

- **Step 4.** Compute feature occlusion score for observation $1 \leq i \leq N$:

$$\hat{\Delta}_j(X_i, Y_i) = \text{Error}(\textcolor{red}{Y}_i, \hat{\mu}_{-i}^{-j}(\textcolor{red}{X}_i)) - \text{Error}(\textcolor{red}{Y}_i, \hat{\mu}_{-i}(\textcolor{red}{X}_i)).$$

Importance of feature j for predicting sample i .

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

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

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

Algorithm: LOCO for Minipatch

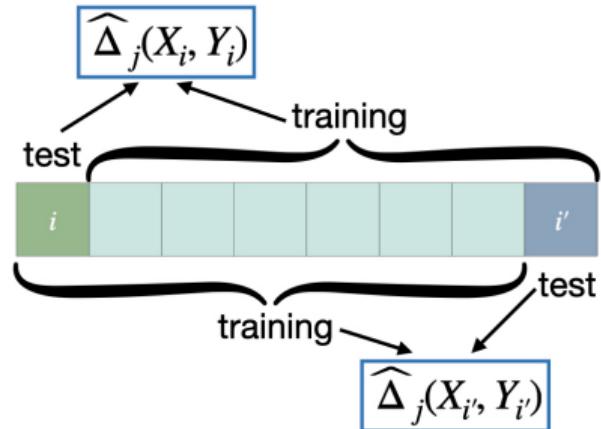
Algorithmic Advantages

- **No data-splitting** ⇒ **powerful**; feature importance inference **for the current model at hand!**
- **No model-refitting** ⇒ 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 $\{\widehat{\Delta}_j(X_i, Y_i)\}_{i=1}^N$!
- $\widehat{\Delta}_j(X_i, Y_i)$ and $\widehat{\Delta}_j(X_{i'}, Y_{i'})$ switches i and i' for training and testing; **share $N - 2$ training samples**.
- Central limit theorem not applicable!



Theoretical Guarantees

- A1. Smoothness of $\text{Error}()$.
- A2. Minipatch predictors have bounded difference ([automatically hold for classification](#)).
- A3. Small proportion of samples within each MP.
- A4. Sufficient number of random MPs.

Theoretical Guarantees

- A1. Smoothness of Error().
- A2. Minipatch predictors have bounded difference (**automatically hold for classification**).
- A3. Small proportion of samples within each MP.
- A4. Sufficient number of random MPs.

Theorem

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

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

Valid asymptotic coverage under mild assumptions; applicable to almost **any data distributions and base ML models**.

Theoretical Guarantees

Main ideas behind the theory:

- **Algorithmic stability:** changing one training sample \Rightarrow negligible perturbation of the prediction.
- Stability facilitates **statistical inference under dependency** (Bayle et al., 2020)!
- Minipatch ensembles are **stable with any base model and any data distribution!** (independent interests).
- Stability also helps with conformal inference, selective inference; but hard to test for complex algorithms (Kim and Barber, 2023).

Predictive inference is also free after training!

- Leave-one-observation-out residuals are free to compute
- Use quantiles of LOO residuals to construct distribution-free predictive intervals (**conformal inference**)
- Similar to Jackknife+ after bootstrap (Kim et al., 2020)

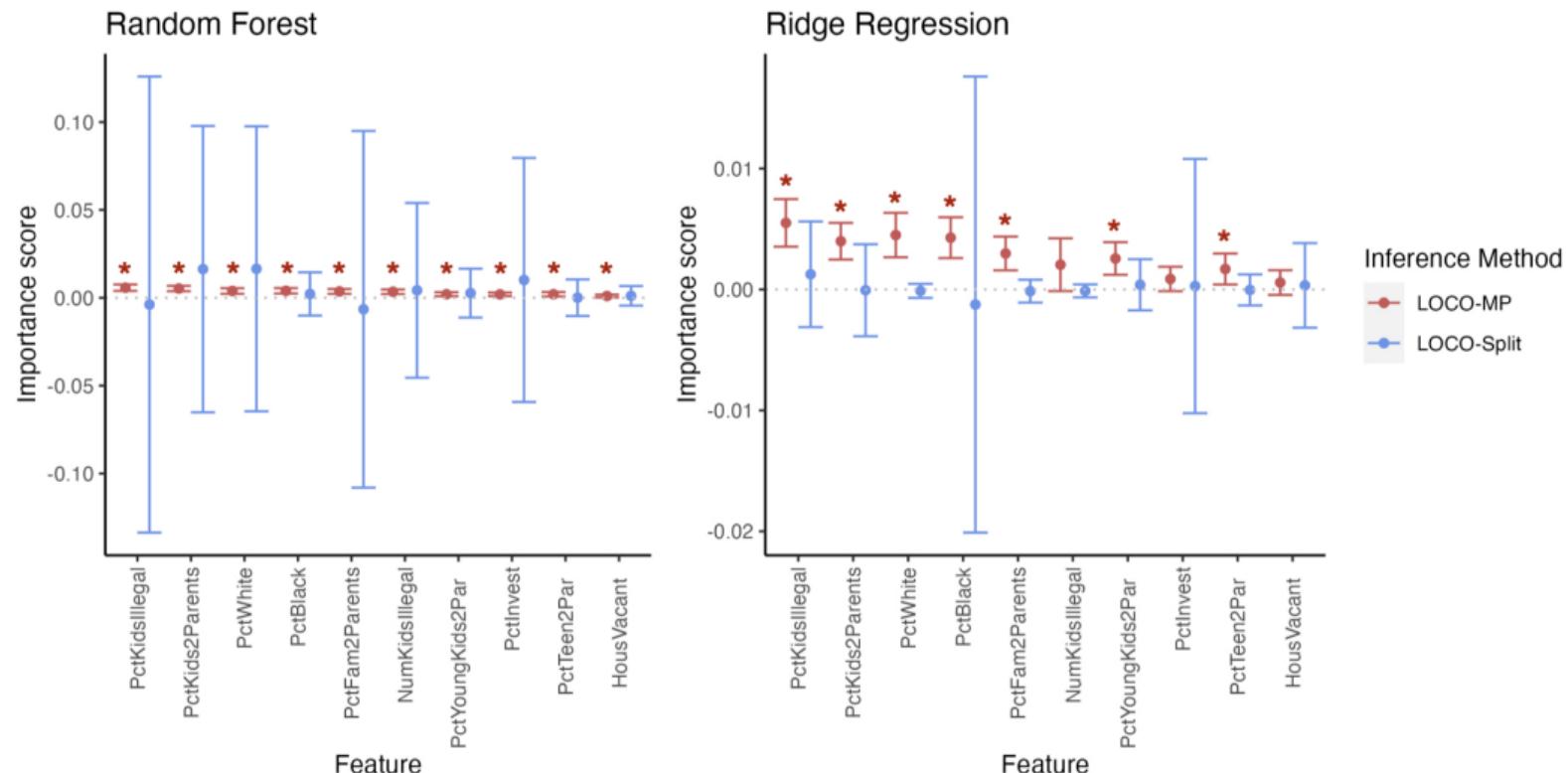
Simultaneous, immediate inference for both feature importance & prediction ⇒ convenient safety check for ML systems

Empirical Studies

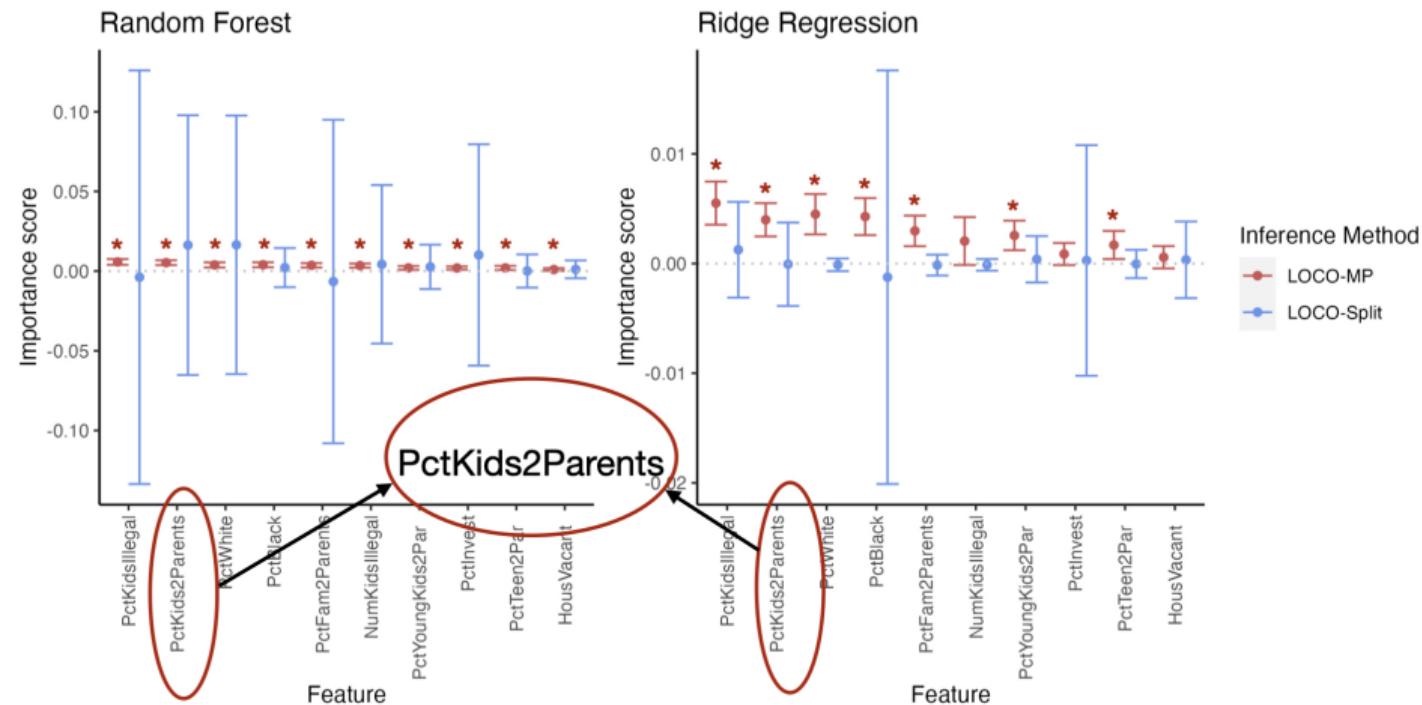
Real Data Example

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

Real Data Example



Real Data Example



Conclusion

- Uncertainty quantification for ML feature importance for minipatch ensembles

Conclusion

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

Conclusion

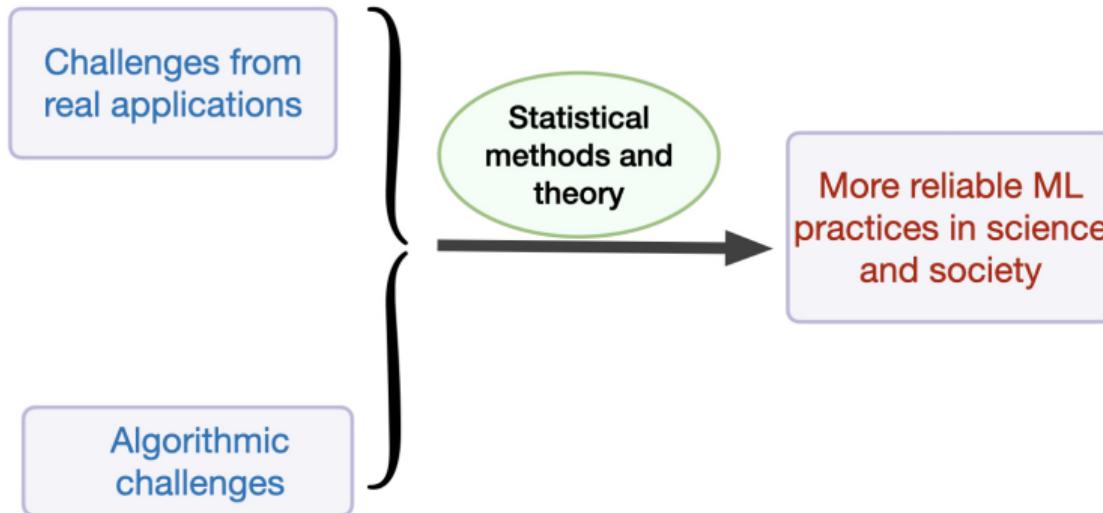
- **Uncertainty quantification for ML feature importance for minipatch ensembles**
 - Free computationally (after minipatch learning).
 - Also (free) predictive intervals.
 - Statistically powerful; assumption-light.
- **Going beyond model diagnostics: LOCO-MP for discovery?**
 - Relationship to population feature importance?
 - Correlated features?

Conclusion

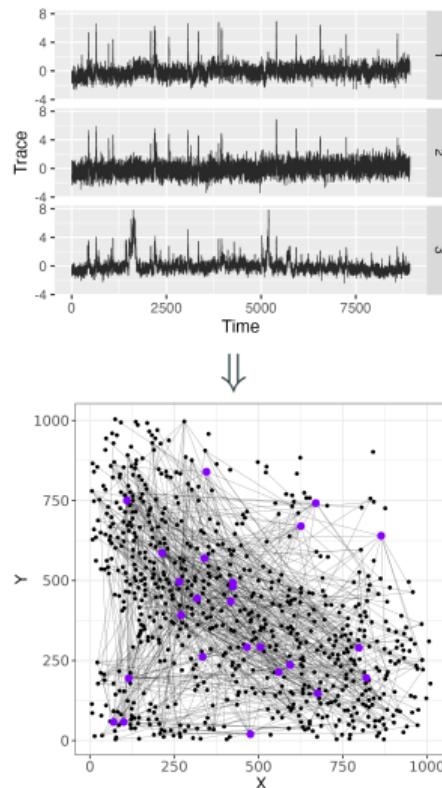
- **Uncertainty quantification for ML feature importance for minipatch ensembles**
 - Free computationally (after minipatch learning).
 - Also (free) predictive intervals.
 - Statistically powerful; assumption-light.
- **Going beyond model diagnostics: LOCO-MP for discovery?**
 - Relationship to population feature importance?
 - Correlated features?
- 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



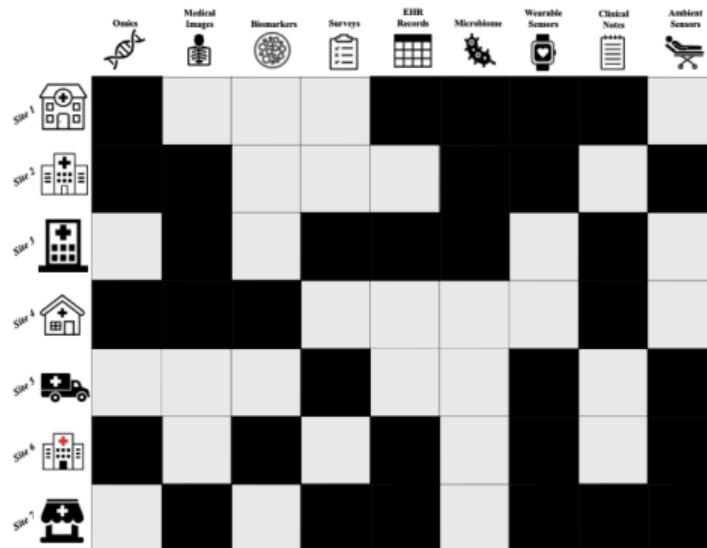
Learning Functional Connectivity in Neuroscience



- Uncertainty quantification: GI-JOE
- Low-rank covariance completion for graph quilting
under revision at JASA, Applications and Case Studies
- Nonparanormal graph quilting
STAT, 2023.

Reliable Statistical Learning in Real Applications

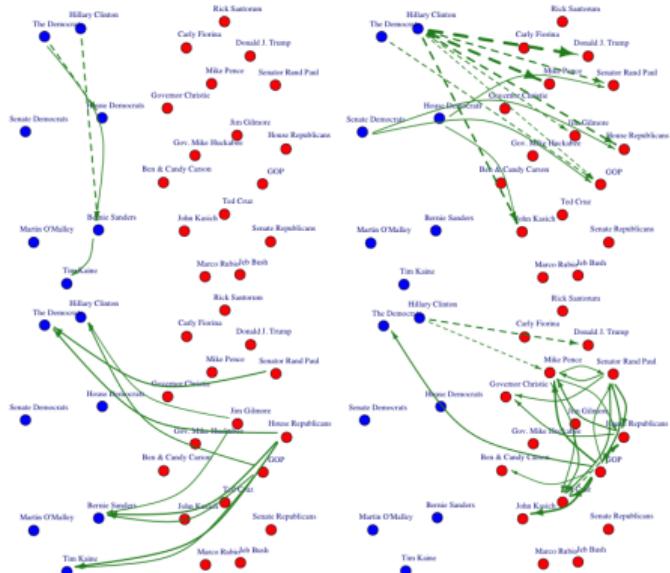
Clustering for Patchwise Multi-modal Healthcare Data



- Provable spectral clustering for patchwork learning
- PCA for patchwork learning?

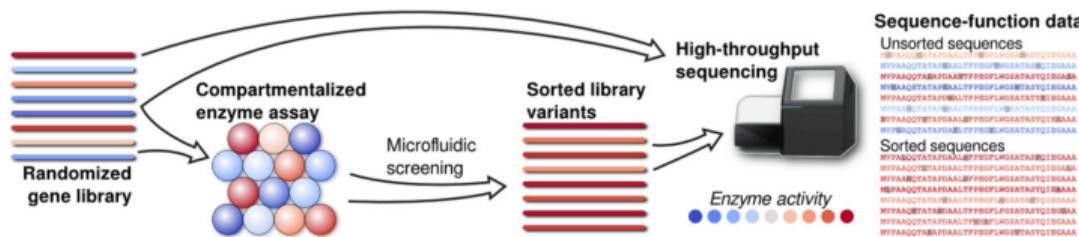
Reliable Statistical Learning in Real Applications

Granger Causal Network on Social Media and Stock Market



- 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

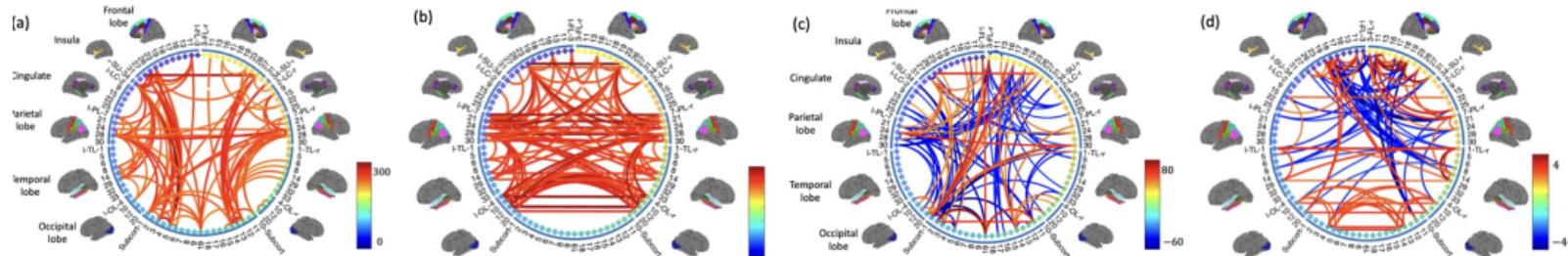
Protein Engineering from Label-contaminated Data



High-dimensional classification with **positive-unlabeled data**

under revision at Electronic Journal of Statistics

Joint Analysis of Functional & Structural Brain Connectivity in Neuroimaging



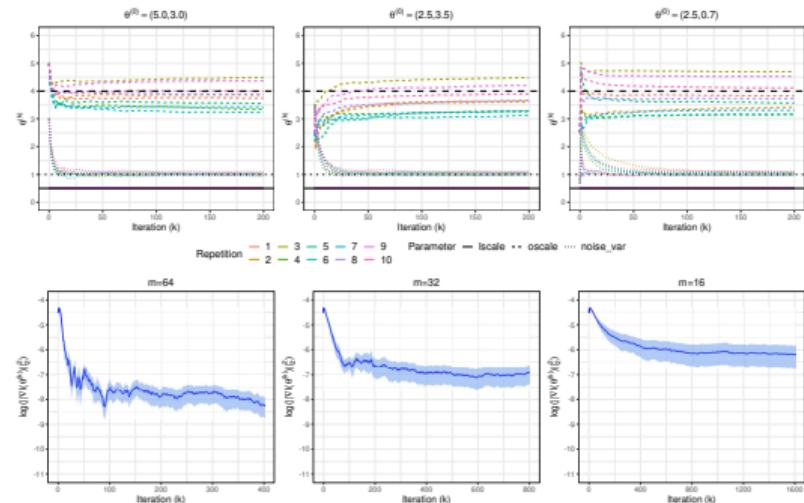
Joint Tensor PCA for Multi-modal Populations of Networks

Addressing Algorithmic Challenges for Large-scale Machine Learning

Subsampling helps both computationally and statistically

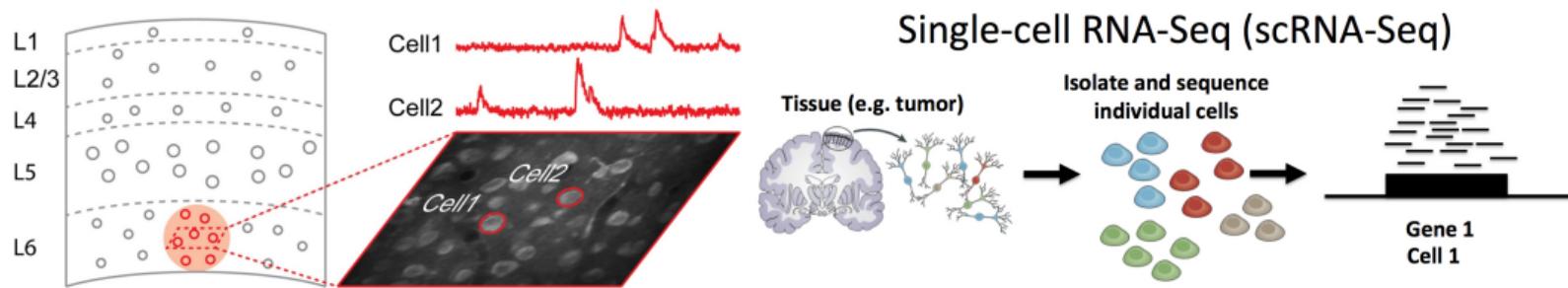
- LOCO-MP for free inference of ML interpretation
- **Provable Convergence: Stochastic Gradient Descent can Speed up Gaussian Processes!**

Journal of Machine Learning Research, 2022



Future Directions

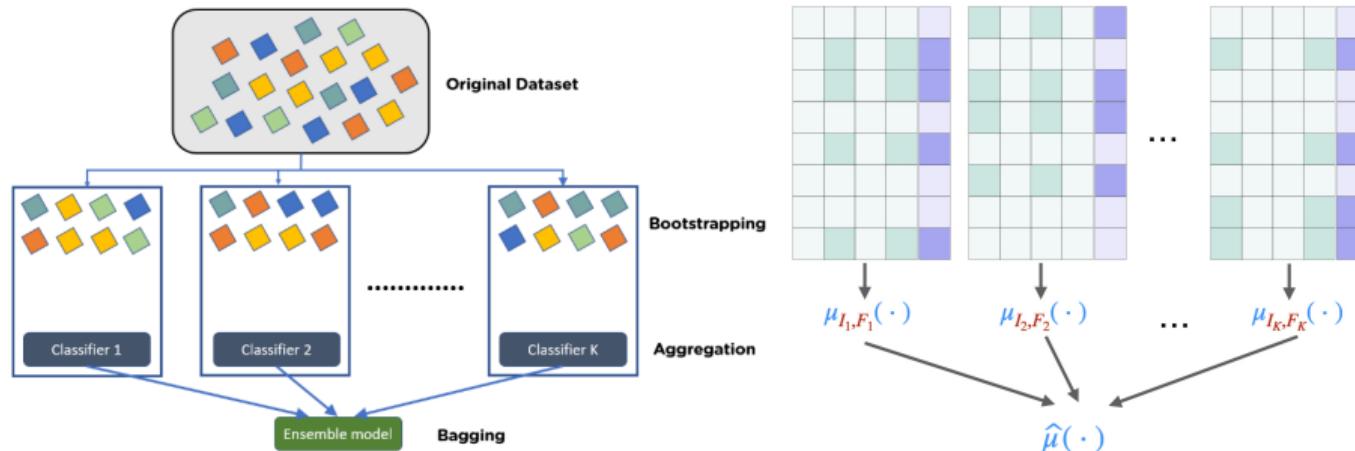
Reliable statistical learning from messy biomedical data?



Picture source: <https://speakerdeck.com/stephaniehicks/welcome-to-the-world-of-single-cell-rna-sequencing?slide=3>

latent neurons; imputation for frequent dropouts

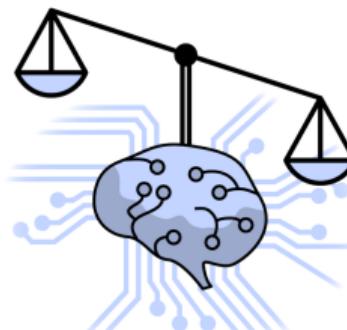
Exploit ensemble learning for statistical & computational advantages, and convenient uncertainty quantification?



Fairness in machine learning interpretations?

ML interpretations can

- inherit bias from data
- only depict the majority group



Picture source:

<https://sitn.hms.harvard.edu/uncategorized/2020/fairness-machine-learning/>

Acknowledgments

Coauthors



Luqin Gan



Genevera I. Allen

Thank you!

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 (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

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

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.

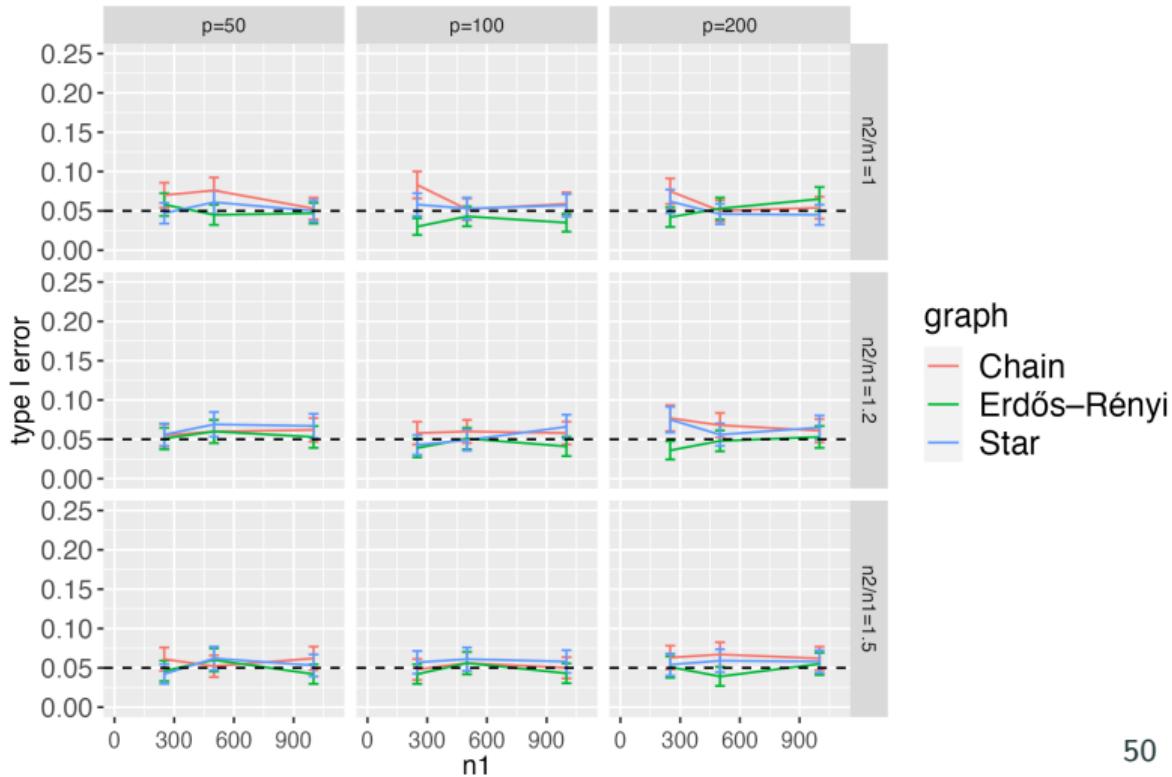
Proof Sketch for GI-JOE: Edgewise Testing

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

- 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 $j \in \mathcal{N}_a, k \in \mathcal{N}_b$.

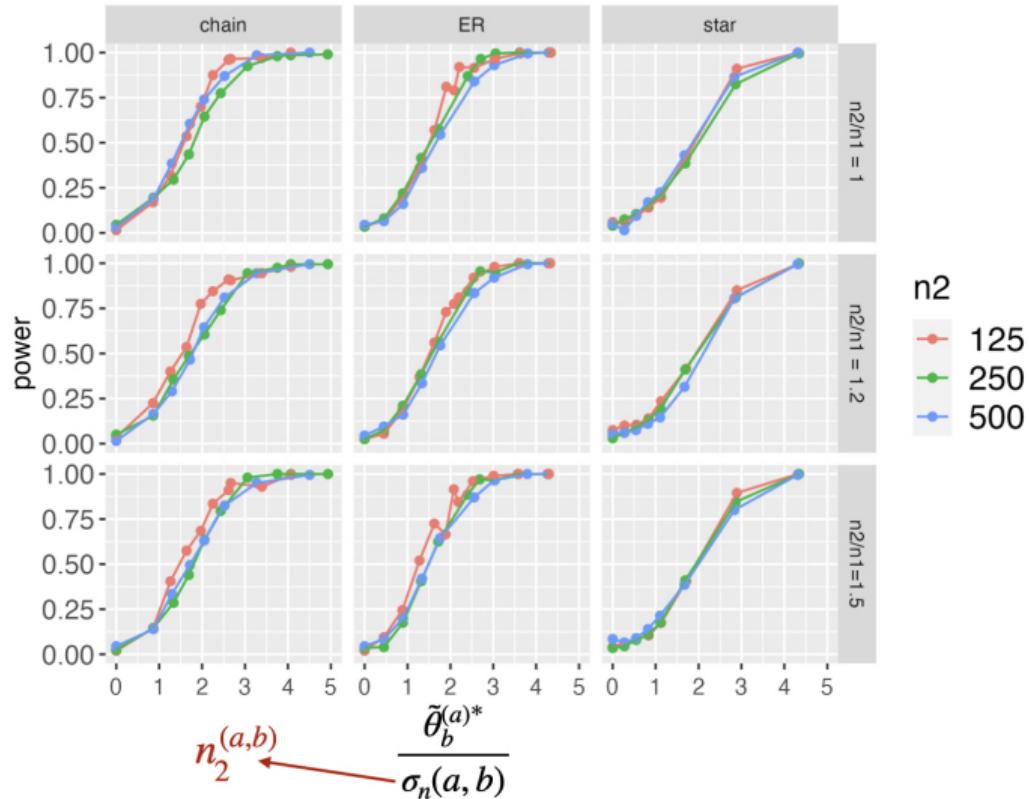
GI-JOE Simulation: Edge-wise Testing

Valid type I error with
differing pairwise sample
sizes!



GI-JOE Simulation: Edge-wise Testing

Power depends on signal strength & **localized sample size** $n_2^{(a,b)}$



Assumptions for Validity of LOCO-MP

- A1. $\text{Error}()$ is Lipschitz continuous.
- A2. Bounded difference in MP predictions $||\hat{\mu}_{I,F}(X) - \hat{\mu}_{I',F'}(X)|| \leq B$.

Assumptions for Validity of LOCO-MP

- A1. $\text{Error}()$ is Lipschitz continuous.
- 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: $n = o\left(\frac{\sigma_j}{LB}\sqrt{N}\right)$.
- A4. $K \gg \left(\frac{L^2B^2N}{\sigma_j^2} + \frac{LB\sqrt{N}}{\sigma_j} + 1\right)\log(N)$.

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 $K = 10,000$.

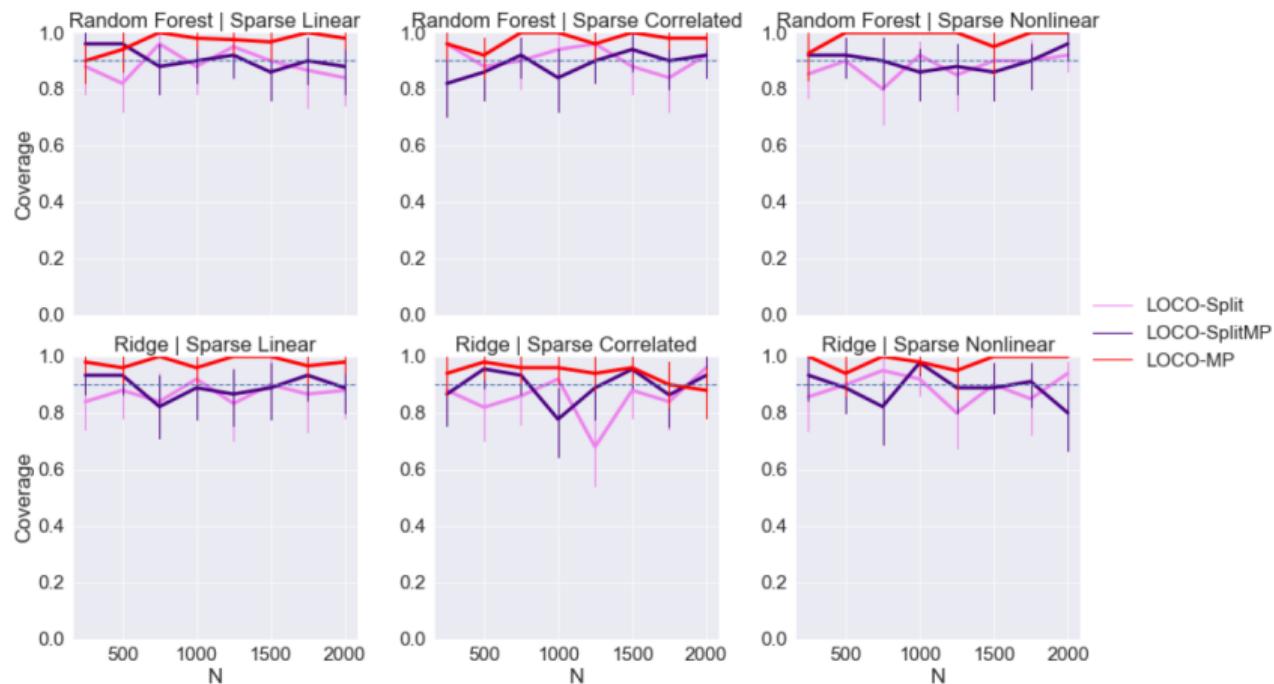
LOCO-MP Simulations: Validate Coverage

Simulation Set-up:

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

LOCO-MP Simulations: Comparative Results

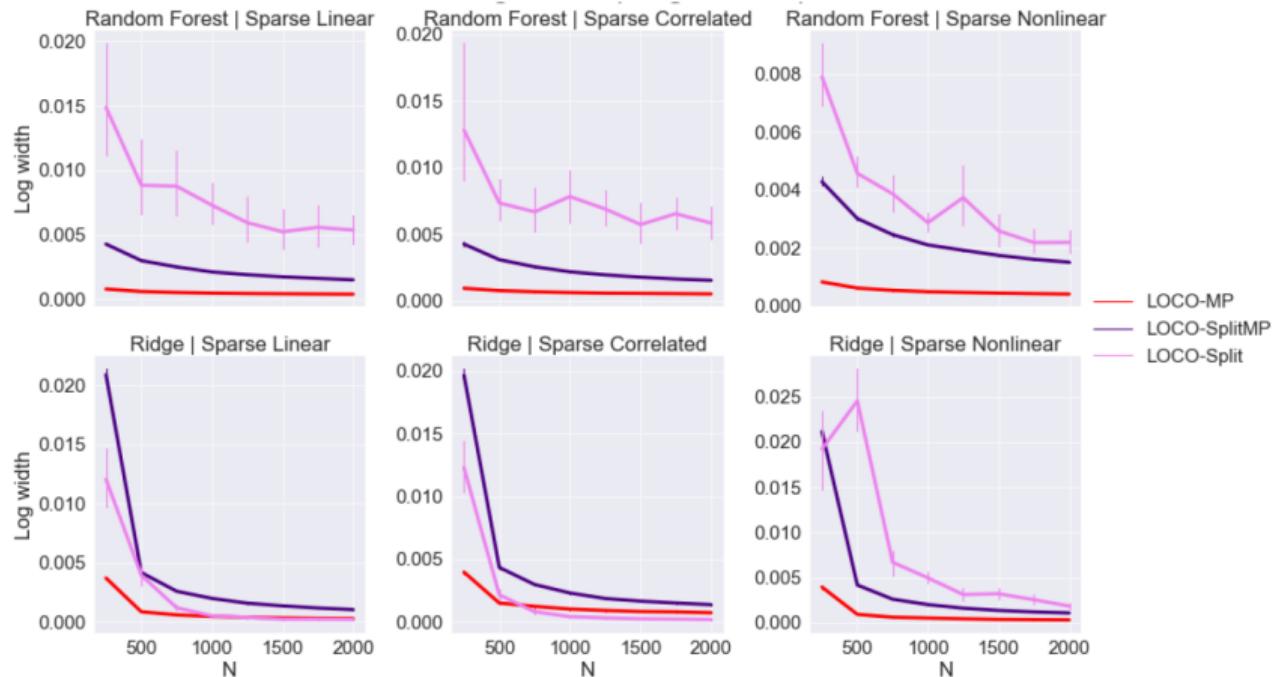
Theory Validation: Coverage.



Coverage for regression simulations for a null feature.

LOCO-MP Simulations: Comparative Results

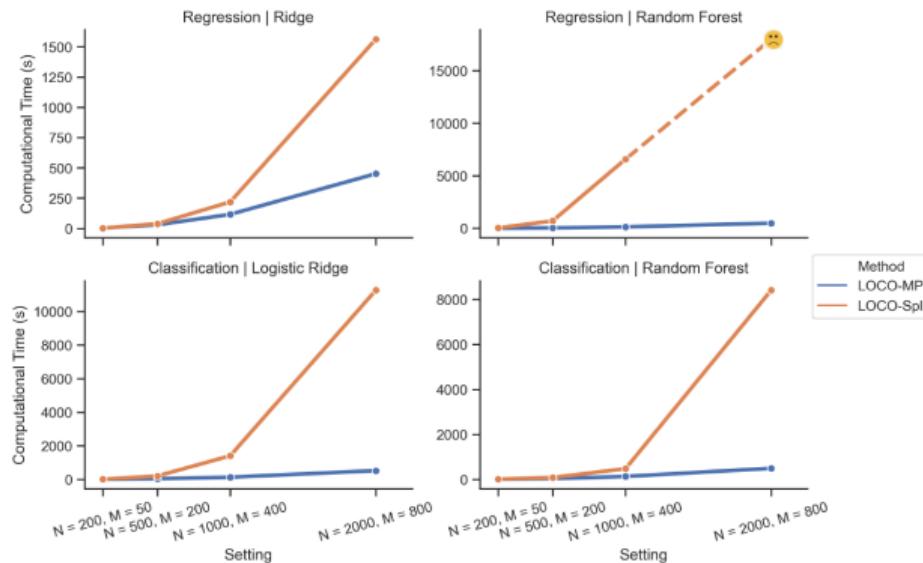
Interval Width:



Log interval width for regression simulations for a null feature.

LOCO-MP Simulations: Comparative Results

Computational Time:



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

References

- Bayle, P., Bayle, A., Janson, L., and Mackey, L. (2020). Cross-validation confidence intervals for test error. *Advances in Neural Information Processing Systems*, 33:16339–16350.
- Birkner, A., Tischbirek, C. H., and Konnerth, A. (2017). Improved deep two-photon calcium imaging in vivo. *Cell calcium*, 64:29–35.
- Chi, C.-M., Fan, Y., and Lv, J. (2022). Fact: High-dimensional random forests inference. *arXiv preprint arXiv:2207.01678*.

- 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.
- Covert, I., Lundberg, S., and Lee, S.-I. (2021). Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90.
- 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.

- Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81.
- Javanmard, A. and Javadi, H. (2019). False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1):1212–1253.
- Kim, B. and Barber, R. F. (2023). Black-box tests for algorithmic stability. *Information and Inference: A Journal of the IMA*, 12(4):2690–2719.
- Kim, B., Xu, C., and Barber, R. F. (2020). Predictive inference is free with the jackknife+-after-bootstrap. *arXiv preprint arXiv:2002.09025*.
- Kolar, M. and Xing, E. P. (2012). Estimating sparse precision matrices from data with missing values.

- König, G., Molnar, C., Bischl, B., and Grosse-Wentrup, M. (2021). Relative feature importance. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9318–9325. IEEE.
- Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.
- Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.
- LeJeune, D., Javadi, H., and Baraniuk, R. (2020). The implicit regularization of ordinary least squares ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 3525–3535. PMLR.

- Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978.
- Louppe, G. and Geurts, P. (2012). Ensembles on random patches. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 346–361. Springer.
- 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*.
- Redmond, M. (2009). Communities and Crime. UCI Machine Learning Repository.
DOI: <https://doi.org/10.24432/C53W3X>.
- Rinaldo, A., Wasserman, L., and G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438–3469.

- Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538.
- Sundararajan, M. and Najmi, A. (2020). The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR.
- 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*.

- Watson, D. S. and Wright, M. N. (2021). Testing conditional independence in supervised learning algorithms. *Machine Learning*, 110(8):2107–2129.
- Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2021). A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, (just-accepted):1–38.
- Yao, T. and Allen, G. I. (2020). Feature selection for huge data via minipatch learning. *arXiv preprint arXiv:2010.08529*.
- Yao, T., LeJeune, D., Javadi, H., Baraniuk, R. G., and Allen, G. I. (2021). Minipatch learning as implicit ridge-like regularization. In *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 65–68. IEEE.

- Zhang, L. and Janson, L. (2020). Floodgate: inference for model-free variable importance. *arXiv preprint arXiv:2007.01283*.