High-dimensional sparse linear models: Estimation, Variable Selection, and Graphs



SCDA at the Simons Foundation







High-dimensional regression

SCE SIMONS FOUNDATION Advancing Research in Basic Science and Mathematics ation





SIMONS SOCIETY OF FELLOW

MATHEMATICS & PHYSICAL SCIENCES

LIFE SCIENCES

Home » Mathematics and Physical Sciences

Mathematics and Physical Sciences

Grants to Individuals

Simons Investigators

Simons Fellows (Deadline: September 30, 2015)

Targeted Grants in MMLS (LOI Deadline: September 30, 2015)

Collaboration Grants for Mathematicians (Deadline: January 28, 2016)

Targeted Grants in MPS (LOI Deadline: Rolling)

Simons Award for Graduate Students **AMS - Simons Travel Grants**

Simons Collaborations

Simons Collaborations in MPS RFA (LOI Deadline: October 1, 2015)

Algorithms and Geometry

Many Electron Problem



November 12, 2015::CERN

High-dimensional regression

SCDA at the Simons Foundation



High-dimensional regression

Systems biology: from data to understanding



November 12, 2015::CERN

Systems biology: from data to understanding



High-dimensional regression

November 12, 2015::CERN

Systems biology: three key statistical problems

- Find relations between outcome (e.g., specific patient phenotype) and measurements (e.g., genes)
- Classify severity of disease state based on gene measurements
- Find relationships among variables (genes, microbes,...)

Systems biology: three key statistical problems

Regression

Classification

Graph learning

Properties of many biological data

- Cross-sectional data
- Noisy with uncertain error distributions
- Number of samples n << p (number of predictors (e.g. genes))
- Number of samples n is O(1e2)
- Number of samples **p** is O(1e3)





Real-world example: Riboflavin production in B. subtilis



р

High-Dimensional Statistics with a View Toward Applications in Biology

Peter Bühlmann, Markus Kalisch, and Lukas Meier

Seminar for Statistics, ETH Zürich, CH-8092 Zürich, Switzerland; email: buhlmann@stat.math.ethz.ch, kalisch@stat.math.ethz.ch, meier@stat.math.ethz.ch

Real-world example: Riboflavin production in B. subtilis



Can we identify a subset of genes that is related to riboflavin production? High-Dimensional Statistics with a View Toward Applications in Biology

Peter Bühlmann, Markus Kalisch, and Lukas Meier

Seminar for Statistics, ETH Zürich, CH-8092 Zürich, Switzerland; email: buhlmann@stat.math.ethz.ch, kalisch@stat.math.ethz.ch, meier@stat.math.ethz.ch

High-dimensional linear regression



High-dimensional sparse linear regression



High-dimensional linear regression

We aim at variable selection in linear regression. We therefore consider models of the form

$$Y = X\beta^* + \sigma\epsilon, \qquad (\text{Model})$$

where $Y \in \mathbb{R}^n$ is a response vector, $X \in \mathbb{R}^{n \times p}$ a design matrix, $\sigma > 0$ a constant, and $\varepsilon \in \mathbb{R}^n$ a noise vector.

J. R. Statist. Soc. B (1996) 58, No. 1, pp. 267–288

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]

$$\widehat{eta} \in rgmin_{eta \in \mathbb{R}^p} \left\{ rac{\|Y - Xeta\|_2^2}{n} + \lambda \|eta\|_1
ight\}$$

J. R. Statist. Soc. B (1996) 58, No. 1, pp. 267–288

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]



High-dimensional regression

J. R. Statist. Soc. B (1996) 58, No. 1, pp. 267–288

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]



J. R. Statist. Soc. B (1996) 58, No. 1, pp. 267–288





High-dimensional regression



 $\|\beta\|_1 < c$

L2 ball (Tikhonov) $\|\beta\|_2^2 < c$

 β_1

High-dimensional regression

Algorithmic approaches to solve the LASSO

- The LASSO is a non-smooth **convex** optimization problems
- Many algorithms available (efficiency dependent on **p** and **n**)
- Coordinate descent, Least-angle regression (LARS), projected sub-gradient, path-following algorithms (over lambda), warmstart

Convex vs. non-convex objective functions



LASSO-type problems

Neural networks

Evaluating estimator performance

- Prediction error:
- Estimation error:
- Variable selection/ support recovery:

 $||X\beta^* - X\hat{\beta}||_2^2/n$ $||\beta^* - \hat{\beta}||_2/n$ $\operatorname{Hamm}(S, \hat{S})$

 $S = \text{support}(\beta^*)$ $\hat{S} = \text{support}(\hat{\beta})$

i.e. the set of non-zero entries

- Extensive theoretical results known regarding estimation and prediction error with respect to sample complexity, variance, and design matrices (see Bühlmann and van de Geer, 2011)
- Most basic result: Set

- Extensive theoretical results known regarding estimation and prediction error with respect to sample complexity, variance, and design matrices (see Bühlmann and van de Geer, 2011)
- Most basic result: Set $\lambda = O(\sigma \sqrt{n \log p})$

- Extensive theoretical results known regarding estimation and prediction error with respect to sample complexity, variance, and design matrices (see Bühlmann and van de Geer, 2011)
- Most basic result: Set $\ \lambda = O(\sigma \sqrt{n \log p})$

$$\frac{1}{n} \|X\beta^* - X\hat{\beta}\|_2^2 = O(\sigma \sqrt{\frac{\log p}{n}} \|\beta^*\|_1)$$

High-dimensional regression

- Wainwright, 2008 showed a key result for exact support recovery. Assume:
 - Mutual incoherence: for some $\gamma > 0$, we have

 $\|(X_S^T X_S)^{-1} X_S^T X_j\|_1 \le 1 - \gamma, \quad \text{for } i \notin S,$

- Minimum eigenvalue: for some C > 0, we have

$$\Lambda_{\min}\left(\frac{1}{n}X_S^T X_S\right) \ge C,$$

where $\Lambda_{\min}(A)$ denotes the minimum eigenvalue of a matrix A

– Minimum signal:

$$|\beta_i^*| \ge \lambda \Big(\| (X_S^T X_S)^{-1} \|_{\infty} + \frac{4\sigma}{C} \Big), \quad \text{for } i \in S,$$

High-dimensional regression

- Wainwright, 2008 showed a key result for exact support recovery. Assume:
 - Mutual incoherence: for some $\gamma > 0$, we have

$$\|(X_S^T X_S)^{-1} X_S^T X_j\|_1 \le 1 - \gamma, \quad \text{for } i \notin S,$$

– Minimum eigenvalue: for some C > 0, we have

$$\Lambda_{\min}\left(\frac{1}{n}X_S^T X_S\right) \ge C,$$

where $\Lambda_{\min}(A)$ denotes the minimum eigenvalue of a matrix A – Minimum signal:

$$|\boldsymbol{\beta}_i^*| \ge \lambda \Big(\| (X_S^T X_S)^{-1} \|_{\infty} + \frac{4\sigma}{C} \Big), \quad \text{for } i \in S,$$

High-dimensional regression

Under these conditions on the design X and predictors and

$$\lambda \geq 2\sigma \sqrt{2n\log p}/\gamma$$

then the LASSO will recover the correct support (and sign) with **high probability**.

How do we find the correct regularization?



Three popular model selection choices

k-fold cross-validation

- Information criteria (BIC,AIC,...)
- Stability selection (based on subsampling, bootstrapping)

Three popular model selection choices



High-dimensional regression



How an we get rid of tuning? The TREX





How an we get rid of tuning? The TREX



Standard approach: The LASSO (Tibshirani, 1996)

$$\widehat{\beta}_{\text{Lasso}}(\lambda) \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{\|Y - X\beta\|_2^2}{n} + \frac{\lambda}{\|\beta\|_1} \right\}. \quad \text{(Lasso)}$$

- + convex optimization problem
- + good statistical properties
- Tuning of regularization parameter required

Novel proposition: The TREX (Lederer and M., AAAI 2015)

$$\widehat{\beta}_{\text{TREX}} \in$$

$$\widehat{\beta}_{\text{TREX}} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{\|Y - X\beta\|_2^2}{\frac{1}{2} \|X^\top (Y - X\beta)\|_{\infty}} + \|\beta\|_1 \right\}.$$
(TREX)

- + good statistical properties
- + Tuning-free method
- non-convex optimization problem

How can we solve the TREX?

- The data-fitting term $L(\beta) = \frac{||\mathbf{Y} \mathbf{X}\beta||_2^2}{\frac{1}{2}||\mathbf{X}^T(\mathbf{Y} \mathbf{X}\beta)||_{\infty}}$ of the non-smooth TREX objective function $f_{\text{TREX}} = L(\beta) + ||\beta||_1$ is approximated by the smooth term $\overline{L}(\beta) = \frac{||\mathbf{Y} \mathbf{X}\beta||_2^2}{\frac{1}{2}||\mathbf{X}^T(\mathbf{Y} \mathbf{X}\beta)||_q}$.
- In practice, for any q > 10, the function $\overline{L}(\beta) + ||\beta||_{1}$ is a sufficient approximation to f_{TREX} and can be efficiently minimized with projected scaled sub-gradient algorithms

$$\|\mathbf{x}\|_{p} := \left(\sum_{i=1}^{n} |x_{i}|^{p}\right)^{1/p} \cdot \frac{\partial}{\partial x_{k}} \|\mathbf{x}\|_{p} = \frac{x_{k} |x_{k}|^{p-2}}{\|\mathbf{x}\|_{p}^{p-1}} \cdot \frac{\partial}{\partial x_{k}} \|\mathbf{x}\|_{p} = \frac{\|\mathbf{x}_{k}\|_{p}^{p-2}}{\|\mathbf{x}\|_{p}^{p-1}} \cdot \frac{\partial}{\partial x_{k}} \|\mathbf{x}\|_{p} = \frac{\|\mathbf{x}_{k}\|_{p}^{p-2}}{\|\mathbf{x}\|_{p}^{p-1}} \cdot \frac{\partial}{\partial x_{k}} \|\mathbf{x}\|_{p} = \frac{\partial}{\partial x_{k}} \|\mathbf{x}\|_{p} = \frac{\partial}{\partial x_{k}} \|\mathbf{x}\|_{p}^{p-1} \cdot \frac{\partial}{\partial x_{k}} \|\mathbf{x}\|_{p} = \frac{\partial}{\partial x_{k}} \|\mathbf{x}\|_$$
How can we solve the TREX?

- The data-fitting term $L(\beta) = \frac{||Y-X\beta||_2^2}{\frac{1}{2}||X^T(Y-X\beta)||_{\infty}}$ of the non-smooth TREX objective function $f_{\text{TREX}} = L(\beta) + ||\beta||_1$ is approximated by the smooth term $\overline{L}(\beta) = \frac{||Y-X\beta||_2^2}{\frac{1}{2}||X^T(Y-X\beta)||_q}$.
- In practice, for any q > 10, the function $\overline{L}(\beta) + ||\beta||_{1}$ is a sufficient approximation to f_{TREX} and can be efficiently minimized with projected scaled sub-gradient algorithms

$$\|\mathbf{x}\|_{p} := \left(\sum_{i=1}^{n} |x_{i}|^{p}\right)^{1/p} \cdot \frac{\partial}{\partial x_{k}} \|\mathbf{x}\|_{p} = \frac{x_{k} |x_{k}|^{p-2}}{\|\mathbf{x}\|_{p}^{p-1}} \cdot$$

Numerical illustration



Figure 1. Hamming distances (to true support) on synthetic normal data generated using (Model) with parameters n=100, $p=500, \beta^* = [1,1,1,1,1,0,...,0]$) and off-diagonal correlation matrix entries $\kappa = 0$ (first column), $\kappa = 0.5$ (second column), and $\kappa = 0.9$ (third column).

High-dimensional regression

November 12, 2015::CERN

Real-world example: Riboflavin production in B. subtilis



Can we identify a subset of genes that is related to riboflavin production?

Yes, we can!



Figure 3. Left: Best fit of different models Lasso-CV with 38 genes, TREX with 20 genes, and B-TREX with three genes. Right: Top 10 list of genes, found by B-TREX (and the three genes found by stability selection [6]).

Thought experiment: Riboflavin production in B. subtilis



Thought experiment: Riboflavin production in B. subtilis



Image that only measured high (+1)/low(-1) production rate!

Thought experiment: Riboflavin production in B. subtilis



Image that only measured high (+1)/low(-1) production rate!

This is a classification task!

- The data Y are discrete labels y_i (-1, 1)
- Simplest solution is logistic regression instead of linear regression

$$\operatorname{Prob}(y_i | \mathbf{x}_i) = \frac{1}{1 + \exp(-(\beta^T \mathbf{x}_i))}$$

The data Y are discrete labels y_i (-1,1)



Simplest solution is logistic regression instead of linear regression

$$\operatorname{Prob}(y_i | \mathbf{x}_i) = \frac{1}{1 + \exp(-(\beta^T \mathbf{x}_i))}$$

The data Y are discrete labels y_i (-1,1)



Simplest solution is logistic regression instead of linear regression

$$\operatorname{Prob}(y_i | \mathbf{x}_i) = \frac{1}{1 + \exp(-(\beta^T \mathbf{x}_i))}$$

LASSO analog is sparse logistic regression

$$f(\beta) = 1/n \sum_{i=1}^{n} \log(1 + \exp(-y_i(\beta^T \mathbf{x}_i))) + \lambda \|\beta\|_1,$$

High-dimensional regression

High-c

High-dimensional classification

The data Y are discrete labels y_i (-1,1)



Simplest solution is logistic regression instead of linear regression

$$\operatorname{Prob}(y_i | \mathbf{x}_i) = \frac{1}{1 + \exp(-(\beta^T \mathbf{x}_i))}$$

LASSO analog is sparse logistic regression

$$f(\beta) = \underbrace{1/n \sum_{i=1}^{n} \log(1 + \exp(-y_i(\beta^T \mathbf{x}_i))) + \lambda \|\beta\|_1,$$

Likelihood term

The data Y are discrete labels y_i (-1,1)



Simplest solution is logistic regression instead of linear regression

$$\operatorname{Prob}(y_i | \mathbf{x}_i) = \frac{1}{1 + \exp(-(\beta^T \mathbf{x}_i))}$$

LASSO analog is sparse logistic regression



The data Y are discrete labels y_i (-1,1)



Simplest solution is logistic regression instead of linear regression

$$\operatorname{Prob}(y_i | \mathbf{x}_i) = \frac{1}{1 + \exp(-(\beta^T \mathbf{x}_i))}$$

LASSO analog is sparse logistic regression



Sparse logistic regression is the core of our recent **VIPUR framework** for protein variant prediction

bioRxiv preprint first posted online October (14, 2015; doi 1 http://dx.doi org/10.#101/029041; The copyright holder for this preprint is the author/funder. It is made available under a CC-BY-ND 4.0 International license.

Published online when published 2015

Nucleic Acids Research, 2015, Vol. ???, No. ? 1–28 doi:10.1093/nar/gkn000

Robust Classification of Protein Variation Using Structural Modeling and Large-Scale Data Integration

Evan H. Baugh^{1,3,*}, Riley Simmons-Edler^{1,3}, Christian L. Müller^{2,3,5}, Rebecca F. Alford^{6,7}, Natalia Volfovsky⁴, Alex E. Lash⁴, Richard Bonneau ^{1,2,3,4,5 *}

¹ Department of Biology, New York University, NY, NY 10003 ² Computer Science Department, New York University, NY, NY 10003 ³ New York University Center for Genomics and Systems Biology, NY, NY 10003 ⁴ Simons Foundation, NY, NY 10010 ⁵ Simons Center for Data Analysis, Simons Foundation, NY, NY 10010 ⁶ Carnegie Mellon University Department of Chemistry, 5000 Forbes Ave, Pittsburgh, PA, 15289 ⁷ Commack High School, Commack NY, 11725

What if **p** and **n** are really large?

What if **p** and **n** are really large?

No problem!

Volkan Cevher, Stephen Becker, and Mark Schmidt

Convex Optimization for Big Data



Scalable, randomized, and parallel algorithms for big data analytics

November 12, 2015::CERN

What if **p** and **n** are really large? CoCoA

"Communication-Efficient Distributed Block-Coordinate Ascent"

<u>*CoCoA+ paper*</u> (ICML 2015)

<u>CoCoA paper</u> (NIPS 2014)

prox CoCoA / primal CoCoA: on arXiv soon

code is available on github

Martin Jaggi, Simone Forte, Virginia Smith, Martin Takáč, Chenxin Ma, Tribhuvanesh Orekondy, Aurelien Lucchi, Peter Richtarik, Thomas Hofmann, Michael I. Jordan

slides adapted from M. Jaggi

High-dimensional regression

November 12, 2015::CERN

Machine Learning Applications

Classification

Support Vector Machine (SVM) (reg.: L1, L2, elastic-net) Logistic Regression (reg.: L1, L2, elastic-net) Structured Prediction (reg.: L1, L2, elastic-net) Regression Least Squares (reg.: L1, L2, elastic-net)



slides adapted from M. Jaggi

Experiments Note that n>p possible





Time<800s



slides adapted from M. Jaggi

High-dimensional regression



slides adapted from M. Jaggi <u>dalab.github.io/dissolve-struct/</u> High-dimensional regression

November 12, 2015::CERN

- Special regularization norms when more is known about the variables
- Examples include Group LASSO, Sparse Group LASSO, Sparse Overlapping Group LASSO, Tree-guided LASSO



- Special regularization norms when more is known about the variables
- Examples include Group LASSO, Sparse Group LASSO, Sparse Overlapping Group LASSO, Tree-guided LASSO



- Special regularization norms when more is known about the variables
- Examples include Group LASSO, Sparse Group LASSO, Sparse Overlapping Group LASSO, Tree-guided LASSO



- Special regularization norms when more is known about the variables
- Examples include Group LASSO, Sparse Group LASSO, Sparse Overlapping Group LASSO, Tree-guided LASSO



How can we construct dependency graphs among variables in the p>>n regime?



The underdetermined regime: p>>n

- A synthetic example: Sampling from a multivariate normal distribution (MVN)
- We draw n=100,...,2000 samples from a p=600 dimensional normal distribution with zero correlation among the variables







High-dimensional regression

November 12, 2015::CERN

November 12, 2015::CERN

The underdetermined regime: p>>n

- A synthetic example: Sampling from a multivariate normal distribution (MVN)
- We draw n=100,...,2000 samples from a p=600 dimensional normal distribution with zero correlation among the variables



Sample correlation matrix: n=100



High-dimensional regression

The p>>n problem

Spurious correlations vanish with increasing sample size



n=500

n=2000





NOVERIDER 12, 2015: CERIN

Conditional independence and sparsity

David MacKay's Gaussian Quiz. Assume a simple system of springs where you observe the position x of the five masses:



inverse-covariance matrix

or

covariance matrix?

Conditional independence and sparsity

David MacKay's Gaussian Quiz. Assume a simple system of springs where you observe the position x of the five masses:



High-dimensional regression

From sparse linear regression to dependency graphs: node-wise regression



From sparse linear regression to graphs: node-wise regression



Node-wise regression: Use each column as response



High-dimensional regression

Patch the neighborhoods together (and/or rule)



Patch the neighborhoods together (and/or rule)



This algorithm approximately recovers the non-zero entries of the **inverse** covariance matrix!

High-dimensional regression

Node-wise regression/ neighborhood selection

The Annals of Statistics 2006, Vol. 34, No. 3, 1436–1462 DOI: 10.1214/00905360600000281 © Institute of Mathematical Statistics, 2006

HIGH-DIMENSIONAL GRAPHS AND VARIABLE SELECTION WITH THE LASSO

BY NICOLAI MEINSHAUSEN AND PETER BÜHLMANN

ETH Zürich

Theoretical results about exact conditions on when recovery is possible!

Knowledge of optimal lambda is necessary
- Sparsity of the underlying network means that the inverse C⁻¹ of the correlation (covariance) matrix C is sparse: sparse Gaussian graphical model.
- Given: the sample correlation (covariance) matrix S
- Goal: Finding a sparse C⁻¹ by convex optimization

- Sparsity of the underlying network means that the inverse C⁻¹ of the correlation (covariance) matrix C is sparse: sparse Gaussian graphical model.
- Given: the sample correlation (covariance) matrix S
- Goal: Finding a sparse C⁻¹ by convex optimization

$$C^{-1} = \arg\min_{C^{-1} \in PD} - \operatorname{logdet}(C^{-1}) + \operatorname{tr}(C^{-1}S) + \lambda \|C^{-1}\|_{1}$$

- Sparsity of the underlying network means that the inverse C⁻¹ of the correlation (covariance) matrix C is sparse: sparse Gaussian graphical model.
- Given: the sample correlation (covariance) matrix S
- Goal: Finding a sparse C⁻¹ by convex optimization

$$C^{-1} = \arg\min_{C^{-1} \in PD} -\log\det(C^{-1}) + \operatorname{tr}(C^{-1}S) + \lambda \|C^{-1}\|_1$$

Likelihood term

- Sparsity of the underlying network means that the inverse C⁻¹ of the correlation (covariance) matrix C is sparse: sparse Gaussian graphical model.
- Given: the sample correlation (covariance) matrix S
- Goal: Finding a sparse C⁻¹ by convex optimization



Efficient algorithms exist for this optimization problem even in very high dimensions

- Neighborhood selection (Meinshausen and Buehlmann, 2006)
- Graphical LASSO (Yuan et al., 2007, Friedman et al. 2008,2011)
- Alternating Linearization (Scheinberg et al., 2010), QUadratic Inverse Covariance (Hsieh et al, 2010)
- Beautiful theoretical results!
- Extensions to **non-normal** data through non-parametric approaches
- Scalable to very high dimensions (BIG and QUiC...)

Further reading with theoretical results



Graph recovery performance depends on graph topology

Neighborhood selection with the TREX (GTREX)

- Replace the LASSO with the standard TREX estimator
- Bootstrapping the TREX estimator to get edge probabilities
- Compare estimators using varying graph topologies

Topology Adaptive Graph Estimation in High Dimensions

Johannes Lederer Cornell University Christian L. Müller Simons Center for Data Analysis, Simons Foundation

High-dimensional regression

November 12, 2015::CERN

Neighborhood selection with the TREX (GTREX)



Neighborhood selection with the TREX (GTREX)



High-dimensional regression

November 12, 2015::CERN

The Internet



Power grid network

The Internet





Power grid network

The Internet





Metabolic network



The Internet



Power grid network



Metabolic network



Transcriptional regulatory networks



November 12, 2015::CERN

The Internet



Power grid network



Metabolic network



Transcriptional regulatory networks



November 12, 2015::CERN

Protein-protein interaction



The Internet



Power grid network



Food webs

Metabolic network



Transcriptional regulatory networks



November 12, 2015::CERN

Protein-protein interaction





The Internet

Network thinking in science

Power grid network

Metabolic network

Microbial interaction networks ----ranscriptional **Protein-protein inter** latory networks E. coli

November 12, 2015::CERN

November 12, 2015::CERN

Microbial ecology: from data to understanding



Microbial ecology: from data to understanding



The scales of our micro-universe



Large-scale efforts in 16S rRNA sequencing in microbiology



The Earth Microbiome Project is a systematic attempt to characterize the global microbial taxonomic and functional diversity for the benefit of the planet and mankind



Large-scale 16S rRNA sequencing

Community



OTUs: Operational Taxonomic Units: groups of similar taxa

High-dimensional regression

November 12, 2015::CERN

Large-scale 16S rRNA sequencing



High-dimensional regression

November 12, 2015::CERN

Key hypothesis for network inference

Key hypothesis for network inference

The network of interactions between the different microbes (and the host) is sparse.





Key hypothesis for network inference

Build the most **parsimonious** network model that explains the data accurately and robustly.





Direct microbial interaction networks from estimating conditional dependence



Direct microbial interaction networks from estimating conditional dependence



Direct microbial interaction networks from estimating conditional dependence



November 12, 2015::CERN

Sparse InversE Covariance Estimation for Ecological ASsociation Inference



November 12, 2015::CERN







Large-scale learning of microbial interaction networks across multiple habitats

We collected **>300** 16S rRNA data sets with sufficient sample size across multiple habitats (gut, oral, skin, fresh water, soil, sea water,)



Large-scale learning of microbial interaction networks across multiple habitats

We collected **>300** 16S rRNA data sets with sufficient sample size across multiple habitats (gut, oral, skin, fresh water, soil, sea water,)



The data set comprises >10^5 different OTUs from >10^4 samples
Large-scale learning of microbial interaction networks across multiple habitats

We collected **>300** 16S rRNA data sets with sufficient sample size across multiple habitats (gut, oral, skin, fresh water, soil, sea water,)



The data set comprises >10^5 different OTUs from >10^4 samples

We focus on **301** networks of sufficient size (Number of edges **m > 50)**

NEW YORK UNIVERSITY

Visualization of the inferred interaction networks

HMP Microbe Networks from 4 body sites: Tongue Dorsum



HMP Microbe Networks from 4 body sites: Stool



HMP Microbe Networks from 4 body sites: Mid-vagina



HMP Microbe Networks from 4 body sites: Left Retroauricular crease



High-dimensional regression

Key questions for ecological network analysis

Can we find network **properties** that are **common to all** (or most) networks across habitats?

How do **microbial** ecological network **relate** to other ecological networks (**food webs**, etc.)?

Key questions for ecological network analysis

Can we find network **properties** that are **common to all** (or most) networks across habitats?

How do **microbial** ecological network **relate** to other ecological networks (**food webs**, etc.)?

Is there a simple **generative network model** that fits the inferred networks?

From an evolutionary perspective: are there **co**evolutionary models that can explain the networks?

Summary

Reviewed LASSO and introduced the (tuning-free) TREX for regression and variable selection

Pointed to implementation and useful extensions

Showed how to use these methods for dependency graph recovery

Proposed SPIEC-EASI for inference of microbial interactions from 16S rRNA abundance data

What can sparse learning do for HEP?

Let's discuss this today and tomorrow!

The TREX team



Johannes Lederer, UW Jacob Bien, Cornell Irina Gaynanova, Texas A&M

The SPIEC-EASI team



Zachary Kurtz, Emily Miraldi

Rich Bonneau Martin Blaser Dan Littman

SIMONS FOUNDATION





High-dimensional regression

SIMONS FOUNDATION



High-dimensional regression

NEW YORK UNIVERSITY