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

LHC Data Science Workshop @CERN

Christian L. Müller, Simons Center for Data Analysis
SCDA at the Simons Foundation
High-dimensional regression
SCDA at the Simons Foundation
Systems biology: from data to understanding

Data

Sequencing data
Phenotypic measurements, disease type

Different living systems

Different data types
Cross section, time & spatial series

Computational modeling
using tools from numerics, machine learning, text mining, optimization...

Descriptive
Predictive
Systems biology: from data to understanding

Data
- Sequencing data
- Phenotypic measurements, disease type

Computational modeling
- using tools from numerics, machine learning, text mining, optimization…

Descriptive

Predictive

Different living systems

Different data types
- Cross section, time & spatial series

High-dimensional regression
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

![Production rate vs. Gene expressions](image)

**High-Dimensional Statistics with a View Toward Applications in Biology**

Peter Bühlmann, Markus Kalisch, and Lukas Meier

Seminar for Statistics, ETH Zurich, CH-8092 Zurich, Switzerland;
email: mbuehlma@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 linear regression

\[ n \begin{bmatrix} Y \\ \beta^* \end{bmatrix} = X \begin{bmatrix} \beta^* \\ \sigma \epsilon \end{bmatrix} + p \]
High-dimensional sparse linear regression

\[ n \{ Y \} = X \times \beta^* + \sigma \epsilon \]
High-dimensional linear regression

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

\[ Y = X \beta^* + \sigma \varepsilon, \quad \text{(Model)} \]

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.
High-dimensional linear regression with LASSO

\[
\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{\| Y - X \beta \|^2_2}{n} + \lambda \| \beta \|_1 \right\}
\]

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]
High-dimensional linear regression with LASSO

\[ \hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{\| Y - X \beta \|^2}{n} + \lambda \| \beta \|_1 \right\} \]
High-dimensional linear regression with LASSO

\[
\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{||Y - X\beta||_2^2}{n} + \lambda ||\beta||_1 \right\}
\]

Likelihood term

Sparsity

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]
High-dimensional linear regression with LASSO

\[ \widehat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{\| Y - X \beta \|^2}{n} \frac{1}{n} + \lambda \| \beta \|_1 \right\} \]

Likelihood term

Sparsity

tuning parameter
High-dimensional linear regression with LASSO

\[ \hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{\|Y - X\beta\|^2_2}{n} + \lambda \|\beta\|_1 \right\} \]

Likelihood term

Sparsity
High-dimensional linear regression with LASSO

\[ \beta_1 < c \]

L1 ball

\[ \| \beta \|_1 < c \]

L2 ball (Tikhonov)

\[ \| \beta \|_2^2 < c \]
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), warm-start
Convex vs. non-convex objective functions

LASSO-type problems

Neural networks
Evaluating estimator performance

- Prediction error: \[ \|X\beta^* - X\hat{\beta}\|_2^2 / n \]
- Estimation error: \[ \|\beta^* - \hat{\beta}\|_2 / n \]
- Variable selection/superset recovery:
  \[ \text{Hamm}(S, \hat{S}) \]

\[ S = \text{support}(\beta^*) \quad \text{i.e. the set of non-zero entries} \]
\[ \hat{S} = \text{support}(\hat{\beta}) \]
Theoretical guarantees for the LASSO

• 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
Theoretical guarantees for the LASSO

- 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 \( \lambda = O(\sigma \sqrt{n \log p}) \)
Theoretical guarantees for the LASSO

- 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 \( \lambda = O(\sigma \sqrt{n \log p}) \)

\[
\frac{1}{n} \left\| X\beta^* - X\hat{\beta} \right\|^2_2 = O\left( \sigma \sqrt{\frac{\log p}{n} \left\| \beta^* \right\|_1} \right)
\]
Theoretical guarantees for the LASSO

- Wainwright, 2008 showed a key result for exact support recovery. Assume:
  - Mutual incoherence: for some $\gamma > 0$, we have
    $$\|(X_S^T X_S)^{-1} X_S^T X_j\|_1 \leq 1 - \gamma, \quad \text{for } i \notin S,$$
  - Minimum eigenvalue: for some $C > 0$, we have
    $$\Lambda_{\min}(\frac{1}{n} X_S^T X_S) \geq C,$$
    where $\Lambda_{\min}(A)$ denotes the minimum eigenvalue of a matrix $A$
  - Minimum signal:
    $$|\beta_i^*| \geq \lambda \left( \|(X_S^T X_S)^{-1}\|_{\infty} + \frac{4\sigma}{C} \right), \quad \text{for } i \in S,$$
Theoretical guarantees for the LASSO

• Wainwright, 2008 showed a key result for exact support recovery. Assume:

  – Mutual incoherence: for some $\gamma > 0$, we have
    \[ \| (X_S^T X_S)^{-1} X_S^T X_j \|_1 \leq 1 - \gamma, \quad \text{for } i \notin S, \]

  – Minimum eigenvalue: for some $C > 0$, we have
    \[ \Lambda_{\text{min}} \left( \frac{1}{n} X_S^T X_S \right) \geq C, \]
    where $\Lambda_{\text{min}}(A)$ denotes the minimum eigenvalue of a matrix $A$

  – Minimum signal:
    \[ |\beta_i^*| \geq \lambda \left( \| (X_S^T X_S)^{-1} \|_\infty + \frac{4\sigma}{C} \right), \quad \text{for } i \in S, \]
Theoretical guarantees for the LASSO

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?

\[ \hat{\beta}_i \]
Three popular model selection choices

- **k-fold cross-validation**
- Information criteria (BIC, AIC, …)
- Stability selection (based on subsampling, bootstrapping)
Three popular model selection choices

- k-fold cross-validation
- Information criteria (BIC, AIC, …)
- Stability selection (based on subsampling, bootstrapping)
How can we get rid of tuning? The TREX

LASSO

\[ \hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{\|Y - X\beta\|^2}{n} + \lambda \|\beta\|_1 \right\} \]

From theory we know:

\[ \lambda \sim \frac{\sigma \|X^\top \epsilon\|_\infty}{n} \]
How can we get rid of tuning? The TREX

\[
\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\text{arg min}} \left\{ \frac{\|Y - X\beta\|_2^2}{n} + \lambda \|\beta\|_1 \right\}
\]

From theory we know:

\[
\lambda \sim \frac{\sigma \|X^\top \epsilon\|_\infty}{n}
\]

TREX

\[
\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\text{arg min}} \left\{ \frac{\|Y - X\beta\|_2^2}{2} + \frac{1}{2}\|X^\top (Y - X\beta)\|_\infty + \|\beta\|_1 \right\}
\]
Standard approach: The LASSO (Tibshirani, 1996)

$$\hat{\beta}_{\text{Lasso}}(\lambda) \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \| Y - X\beta \|_2^2 + \lambda \| \beta \|_1 \right\}. \text{ (Lasso)}$$

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

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

$$\hat{\beta}_{\text{TREX}} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \| X^T(Y - X\beta) \|_\infty + \| \beta \|_1 \right\}. \text{ (TREX)}$$

+ good statistical properties
+ Tuning-free method
- non-convex optimization problem
How can we solve the TREX?

- The data-fitting term \( L(\beta) = \frac{1}{2} ||Y - X\beta||_2^2 \) of the non-smooth TREX objective function \( f_{\text{TREX}} = L(\beta) + ||\beta||_1 \) is approximated by the smooth term \( \overline{L}(\beta) = \frac{1}{2} \frac{||Y - X\beta||_2^2}{||X^T(Y - X\beta)||_q} \).

- In practice, for any \( q > 10 \), the function \( \overline{L}(\beta) + ||\beta||_1 \) is a sufficient approximation to \( f_{\text{TREX}} \) and can be efficiently minimized with projected scaled sub-gradient algorithms

\[
\|x\|_p := \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p} \quad \frac{\partial}{\partial x_k} \|x\|_p = \frac{x_k |x_k|^{p-2}}{\|x\|_p^{p-1}}.
\]
How can we solve the TREX?

• The data-fitting term $L(\beta) = \frac{1}{2}||Y - X\beta||_2^2$ of the non-smooth TREX objective function $f_{TREX} = L(\beta) + ||\beta||_1$ is approximated by the smooth term $ar{L}(\beta) = \frac{1}{2}||Y - X\beta||_2^2$.

• In practice, for any $q > 10$, the function $\bar{L}(\beta) + ||\beta||_1$ is a sufficient approximation to $f_{TREX}$ and can be efficiently minimized with projected scaled sub-gradient algorithms.

\[
\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \quad \text{and} \quad \frac{\partial}{\partial x_k} \|x\|_p = \frac{x_k |x_k|^{p-2}}{\|x\|_p^{p-1}}.
\]
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).
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

Production rate

Gene expressions

High-dimensional regression
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!
High-dimensional classification

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

$$\text{Prob}(y_i | x_i) = \frac{1}{1 + \exp(-(\beta^T x_i))}$$
High-dimensional classification

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

$$\text{Prob}(y_i | x_i) = \frac{1}{1 + \exp(-(\beta^T x_i))}$$
High-dimensional classification

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

$$
\text{Prob}(y_i \mid x_i) = \frac{1}{1 + \exp(-(\beta^T x_i))}
$$

- LASSO analog is **sparse logistic regression**

$$
f(\beta) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i (\beta^T x_i))) + \lambda \|\beta\|_1,
$$
High-dimensional classification

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

$$\text{Prob}(y_i | x_i) = \frac{1}{1 + \exp(-(\beta^T x_i))}$$

- LASSO analog is **sparse logistic regression**

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

Likelihood term
High-dimensional classification

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

$$
\text{Prob}(y_i | x_i) = \frac{1}{1 + \exp(-(\beta^T x_i))}
$$

- LASSO analog is **sparse logistic regression**

$$
f(\beta) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i (\beta^T x_i))) + \lambda \| \beta \|_1.
$$
High-dimensional classification

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

$$\text{Prob}(y_i | x_i) = \frac{1}{1 + \exp(-(\beta^T x_i))}$$

- LASSO analog is **sparse logistic regression**

$$f(\beta) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i(\beta^T x_i))) + \lambda \|\beta\|_1.$$
Sparse logistic regression is the core of our recent **VIPUR framework** for protein variant prediction.
What if $p$ and $n$ are really large?
High-dimensional regression

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

No problem!
What if \( p \) and \( n \) are really large? CoCoA

“Communication-Efficient Distributed Block-Coordinate Ascent”

CoCoA+ paper (ICML 2015)

CoCoA paper (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
Machine Learning Applications

Classification
- Support Vector Machine (SVM)
  \((\text{reg.: } L_1, L_2, \text{elastic-net})\)
- Logistic Regression
  \((\text{reg.: } L_1, L_2, \text{elastic-net})\)

Structured Prediction
\((\text{reg.: } L_1, L_2, \text{elastic-net})\)

Regression
- Least Squares
  \((\text{reg.: } L_1, L_2, \text{elastic-net})\)

CoCoA+
\(D = \text{dual}\)

\(\text{prox } \text{CoCoA+}\)
\(D = \text{dual}\)

\(\text{primal prox } \text{CoCoA+}\)
\(D = \text{primal}\)

\(L_1: \text{get bounded support!}\)

slides adapted from M. Jaggi

High-dimensional regression

November 12, 2015::CERN
Experiments

Note that n>p possible

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training n</th>
<th>Features p</th>
<th>Sparsity</th>
<th>λ</th>
<th>Workers K</th>
</tr>
</thead>
<tbody>
<tr>
<td>cov</td>
<td>522,911</td>
<td>54</td>
<td>22.22%</td>
<td>1e-6</td>
<td>4</td>
</tr>
<tr>
<td>rcv1</td>
<td>677,399</td>
<td>47,236</td>
<td>0.16%</td>
<td>1e-6</td>
<td>8</td>
</tr>
<tr>
<td>imagenet</td>
<td>32,751</td>
<td>160,000</td>
<td>100%</td>
<td>1e-5</td>
<td>32</td>
</tr>
</tbody>
</table>

Time<800s

slides adapted from M. Jaggi

High-dimensional regression

5 Experiments

In this section, we compare COCOA (H=1e6) to other algorithms by using various local updates of each iteration. The improvements are robust for both regimes to avoid communication while still making significant global progress by efficiently combining the convergence speed. We see clearly that COCOA (H=1e6) is able to converge to a more accurate solution in all cases.

In the experiments, we apply these algorithms to standard hinge loss while noting that it is a benefit to avoid having to tune this data-dependent parameter.

We additionally study the effect of scaling of the average by a parameter whether the primal vector is updated locally on each inner iteration or not, and whether the resulting combination/communication of the updates is by an average over the total size of the data or a pure average of the updates.

In the theoretical analysis, we do not assume that K,n,p are related and only require K,n,p to be finite. 

For the non-locally updating mini-batch methods, (mini-batch SDCA [3] and mini-batch SGD [13]), we use the same regularization parameters as specified in [13, 14].

Our results indicate that COCOA (H=1e6) converges much faster than the other methods. On average, COCOA (H=1e6) achieves a Log Primal Suboptimality of O(1/K) at time T=KH, and decreases communication but also affects the objective value over time.

For the locally-updating methods (COCOA (H=1e6), mini-batch-CD (H=100), local-SGD (H=1e5), mini-batch-SGD (H=100)), we compare the other algorithms, when all use the same regularization parameters as specified in [13, 14]. COCOA (H=1e6) is able to converge to a more accurate solution in all cases.
Dissolve\textsuperscript{struct}

A Library for Distributed Structured Prediction

built on Spark

Structured SVM solver
Block Coordinate
Frank-Wolfe

Distributed Optimization
CoCoA

Open Source
Approximate Inference allowed!
drop-in replacement for SVM\textsuperscript{struct}

Applications:

Text
- Parsing
- POS tagging
- sentence alignment
- entity linking

Biology
Protein structure & function prediction

Vision
2d+3d Segmentation, OCR

more?
- Scene understanding
- object localization & recog.

Your Application?

dalab.github.io/dissolve-struct/

slides adapted from M. Jaggi

High-dimensional regression

November 12, 2015::CERN
Further extension: Structured prediction to include your physics knowledge

- Special regularization norms when more is known about the variables
- Examples include Group LASSO, Sparse Group LASSO, Sparse Overlapping Group LASSO, Tree-guided LASSO …
Further extension: Structured prediction to include your physics knowledge

• Special regularization norms when more is known about the variables

• Examples include Group LASSO, Sparse Group LASSO, Sparse Overlapping Group LASSO, Tree-guided LASSO ...

\[ n \begin{cases} Y = & X \times \beta^* \end{cases} + \sigma \epsilon \]
Further extension: Structured prediction to include **your physics knowledge**

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

\[
Y = X \beta^* + \sigma \epsilon
\]
Further extension: Structured prediction to include your physics knowledge

• Special regularization norms when more is known about the variables

• Examples include Group LASSO, Sparse Group LASSO, Sparse Overlapping Group LASSO, Tree-guided LASSO …

\[ Y = X \beta^* + \sigma \epsilon \]
How can we construct dependency graphs among variables in the p>>n regime?
The underdetermined regime: \( p \gg n \)

- A *synthetic* example: Sampling from a multivariate normal distribution (MVN)
- We draw \( n=100,\ldots,2000 \) samples from a \( p=600 \) dimensional normal distribution with *zero* correlation among the variables
The underdetermined regime: $p \gg n$

- 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
The p>>n problem

Spurious correlations vanish with increasing sample size

True correlation 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:

$X_1$ $X_2$ $X_3$ $X_4$ $X_5$

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:

$$K^{-1} = \frac{k}{T} \begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}$$

or

$$K = \frac{T}{k} \begin{bmatrix} 0.83 & 0.67 & 0.50 & 0.33 & 0.17 \\ 0.67 & 1.33 & 1.00 & 0.67 & 0.33 \\ 0.50 & 1.00 & 1.50 & 1.00 & 0.50 \\ 0.33 & 0.67 & 1.00 & 1.33 & 0.67 \\ 0.17 & 0.33 & 0.50 & 0.67 & 0.83 \end{bmatrix}$$
From sparse linear regression to dependency graphs: node-wise regression

\[ n \begin{pmatrix} Y \\ X \end{pmatrix} = \begin{pmatrix} p \\ \beta^* \end{pmatrix} \times \begin{pmatrix} X \\ \epsilon \end{pmatrix} + \sigma \]
From sparse linear regression to graphs: node-wise regression

\[ n \left\{ Y \right\} = \left\{ X \right\} + \sigma \epsilon \]

\[ \beta^* \]

\[ p \]

\[ \sigma \]

\[ \epsilon \]
Node-wise regression: Use each column as response

\[ n \begin{bmatrix} X_i^1 \\ \vdots \\ X_i^n \end{bmatrix} = X_n^{-i} \times \begin{bmatrix} \beta^*_1 \\ \vdots \\ \beta^*_{p-1} \end{bmatrix} + \sigma \epsilon \]

Neighborhood selection
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!
Node-wise regression/ neighborhood selection

**The Annals of Statistics**
2006, Vol. 34, No. 3, 1436–1462
DOI: 10.1214/009053606000000281
© 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
Sparse graphical model inference (GLASSO)

• 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
Sparse graphical model inference (GLASSO)

- 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

$$
C^{-1} = \arg \min_{C^{-1} \in PD} -\log \det(C^{-1}) + \text{tr}(C^{-1}S) + \lambda \|C^{-1}\|_1
$$
Sparse graphical model inference (GLASSO)

• 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

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

Likelihood term
Sparse graphical model inference (GLASSO)

• 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

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

- **Likelihood term**
- **Sparsity term**
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

**Electronic Journal of Statistics**
Vol. 5 (2011) 935–980
ISSN: 1935-7524
DOI: 10.1214/11-EJS631

**High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence**

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti and Bin Yu

Berkeley, CA 94720-1776 USA
e-mail: pradeepr@stat.berkeley.edu
wainwrig@stat.berkeley.edu
garveshr@stat.berkeley.edu
binyu@stat.berkeley.edu

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
Neighborhood selection with the TREX (GTREX)

Figure 1: Hamming distances of the true graphs to GLASSO, MB(or), and MB(and) with optimal tuning parameter $\lambda^*$ and to GTREX as a function of the sample size $n$. In the top row, examples of the corresponding graphs are displayed.
Neighborhood selection with the TREX (GTREX)

Single-Hub  |  Double-Hub  |  Four-Hub  |  Four-niches  |  Erdős-Rényi  |  Scale-free

Hamming distance

$p=100$

$p=200$
Network thinking in science
Network thinking in science

The Internet
Network thinking in science

The Internet

Power grid network
Network thinking in science

The Internet

Power grid network

Metabolic network
Network thinking in science

The Internet

Power grid network

Metabolic network

Transcriptional regulatory networks
Network thinking in science

The Internet

Power grid network

Metabolic network

Protein-protein interaction

Transcriptional regulatory networks
Network thinking in science

The Internet

Power grid network

Metabolic network

Food webs

Protein-protein interaction

Transcriptional regulatory networks
Network thinking in science

- Power grid network
- Metabolic network
- Microbial interaction networks
- Transcriptional regulatory networks
- The Internet
- Protein-protein interaction networks
Microbial ecology: from data to understanding

**Data**
- Microbial populations
- Host/environmental co-variates: transcript-omics, etc
- Different habitats
- Different data types
  - Cross section, time & spatial series

**Computational Statistics**
- Estimation of species abundance
  - Novel regression models for the specific data types
  - Estimation of species interactions

**High-dimensional regression**

Descriptive

Predictive
**Microbial ecology: from data to understanding**

**Data**
- Microbial populations
- Host/environmental co-variates: transcript-omics, etc
- Different habitats
- Different data types
  - Cross section, time & spatial series

**Computational Statistics**
- Estimation of **species abundance**
- Novel regression models for the specific data types
- Estimation of species interactions

**High-dimensional regression**

**Descriptive**

**Predictive**
The scales of our micro-universe

- $10^{30}$ Viruses
- $10^{28}$ Bacteria
- $10^{22}$ Stars
- $10^{14}$ Human Microbiota
- $10^9$ Human Population

Small World – Big Data
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.

HUMANS MICROBIOME PROJECT

- Nasal
- Oral
- Skin
- Gastro-intestinal
- Urogenital
Large-scale 16S rRNA sequencing

Community Sampling

Sample 1

Sample 2

PCR amplify 16 rRNA gene
Sequencing
Group into OTUs
Assign Taxonomy

OTUs: Operational Taxonomic Units: groups of similar taxa

OTUs: Operational Taxonomic Units: groups of similar taxa
Large-scale 16S rRNA sequencing

Community

Community Sampling

Samples

Sample site

OTUs: Operational Taxonomic Units

groups of similar taxa

PCR amplify 16 rRNA gene
Sequencing
Group into OTUs
Assign Taxonomy

Analysis Tasks
Classification
Regression
Ecology Networks
Predict Mechanism

High-dimensional regression
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

Methods:
- Pearson's coefficient
- SparCC
- CCRePE

Correlation threshold
- $\geq \pm 0.35, \pm 0.5$
- $\geq \pm 0.35, \leq 0.5$
SPIEC-EASI learns parsimonious direct microbial interaction networks from data
SPIEC-EASI learns parsimonious direct microbial interaction networks from data
SPIEC-EASI learns parsimonious direct microbial interaction networks from data

OTU data

Data processing, centered log-ratio transformation (clr)

Sparse graphical model learning

Stability-based model selection

Sparse Inverse Covariance Estimation for Ecological Association Inference
SPIEC-EASI learns parsimonious direct microbial interaction networks from data

**OTU data**

**Dependency graph, covariance matrix**

**Sparse Inverse Covariance Estimation for Ecological Association Inference**

Data processing, centered log-ratio transformation (clr)

Sparse graphical model learning

Stability-based model selection
SPIEC-EASI learns parsimonious direct microbial interaction networks from data

RESEARCH ARTICLE

Sparse and Compositionally Robust Inference of Microbial Ecological Networks

Zachary D. Kurtz¹☯, Christian L. Müller²,³☯, Emily R. Miraldi¹,²,³☯, Dan R. Littman¹, Martin J. Blaser¹, Richard A. Bonneau²,³,⁴*

1 Departments of Microbiology and Medicine, New York University School of Medicine, New York, New York, United States of America, 2 Department of Biology, Center for Genomics and Systems Biology, New York University, New York, New York, United States of America, 3 Courant Institute of Mathematical Sciences, New York University, New York, New York, United States of America, 4 Simons Center for Data Analysis, Simons Foundation, New York, New York, United States of America
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)
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

November 12, 2015::CERN
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
**Data Driven:** The New Big Science

Who's Driving?  |  Digits in the Sky  |  Revolutionary Algorithms  |  The Facts of Life  |  On Quantum Memory

---

**CHAPTER 3: REVOLUTIONARY ALGORITHMS**

**The Mathematical Shape of Things to Come**

Scientific data sets are becoming more dynamic, requiring new mathematical techniques on par with the invention of calculus.

*By: Jennifer Ouellette*

October 4, 2013

---

**NEXT IN THE SERIES**

Our Bodies, Our Data

---

**Email:** cmueller@simonsfoundation.org