Topics covered in Part 2

- visualization
- inference
- planning of studies
- privacy
5-year running mean
annual averages
Iteration IX (2013): Science Maps Showing Trends and Dynamics

Four Existing Maps

Six Science Maps

This iteration features science maps that show general trends and dynamics in science and technology. The maps communicate:

- Ocean circulation estimates to quantify the ocean’s role in the global carbon cycle and to monitor its evolving heat, water, and chemical exchanges over time.
Inference for Lasso \( n < p \)

\[
\hat{\beta}_{\text{Lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}
\]

- **prediction:** \( \|X(\hat{\beta}_{\text{Lasso}} - \beta^0)\|_2^2/n \) should be ‘small’

- **estimation:** \( \|\hat{\beta}_{\text{Lasso}} - \beta^0\|_q, \quad q \in 1, 2 \) ‘small’

- **selection:** \( \mathbb{P}(\hat{S} = S_0) \) ‘large’
  
  \( S_0 \) is the ‘active set’: \( \{ j : \beta^0_j \neq 0 \} \)

- under restricted eigenvalue conditions on \( X \), can get results like

\[
\|\hat{\beta}_{\text{Lasso}} - \beta^0\|_1 = O_p(s_0 \sqrt{\log(p)/n}), \quad \lambda \approx \sqrt{\log(p)/n}
\]
Inference for Lasso $n < p$


$$\hat{\beta}_{\text{Lasso}} = \arg\min_\beta \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- what about estimated standard errors for $\hat{\beta}_{\text{Lasso}}$?
- corrected ridge or lasso estimators:
  1. for $j = 1, \ldots, p$ construct a linear estimator $\hat{b}_j = Z_j^T Y / Z_j^T X_j$
  2. $E(\hat{b}_j) = \beta_j + \sum_{k \neq j} P_{jk} \beta_k = \sum_{k \neq j} Z_j^T X_k / Z_j^T X_j$
  3. estimate this:
    $$\hat{\beta}_{j,\text{corr}} = \hat{b}_j - \sum_{k \neq j} P_{jk} \hat{\beta}_{k,\text{Lasso}}$$

- result
  $$\sqrt{n}(\hat{\beta}_{j,\text{corr}} - \beta_j) / \hat{\sigma} \xrightarrow{\mathcal{L}} N(0, 1)$$
- sparsity assumption $|S| \log p / \sqrt{n} \rightarrow 0$
Inference for Lasso $n < p$

$\hat{\beta}_{\text{Lasso}}(\lambda) = \arg\min_{\beta} (\|y - X\beta\|_2^2 + \lambda\|\beta\|_1)$

First, we define some needed quantities. Let $A$ be the active set just before $\lambda_k$, and suppose that predictor $j$ enters at $\lambda_k$. Denote by $\tilde{\beta}(\lambda_{k+1})$ the solution at the next knot in the path $\lambda_{k+1}$, using predictors $A \cup \{j\}$. Finally, let $\tilde{\beta}_A(\lambda_{k+1})$ be the solution of the lasso problem using only the active predictors $X_A$, at $\lambda = \lambda_{k+1}$. To be perfectly explicit,

$$\tilde{\beta}_A(\lambda_{k+1}) = \arg\min_{\beta_A \in \mathbb{R}^{|A|}} \frac{1}{2}\|y - X_A\beta_A\|_2^2 + \lambda_{k+1}\|\beta_A\|_1.$$

We propose the covariance test statistic defined by

$$T_k = (\langle y, X\tilde{\beta}(\lambda_{k+1}) \rangle - \langle y, X_A\tilde{\beta}_A(\lambda_{k+1}) \rangle) / \sigma^2.$$

$$T_k \overset{\mathcal{L}}{\longrightarrow} \text{Exp}(1)$$

Taylor et al. 2014 arxiv for more recent work
Inference for Lasso $n < p$

Lockhart et al. 2014 AoS; Bühlmann 2014 Discussion

- **Lasso solution path** $\{\hat{\beta}(\lambda); \lambda > 0\}$
- **Subset** $S \subset \{1, \ldots, p\}$; $\hat{\beta}_S(\lambda)$
- **Covariance test**

$$T(S, \lambda) = \{ ||y - X\hat{\beta}_S(\lambda)||_2^2 + \lambda ||\hat{\beta}_S(\lambda)||_1$$
$$- ||y - X\hat{\beta}(\lambda)||_2^2 + \lambda ||\hat{\beta}(\lambda)||_1 \} / \sigma^2$$

- if large then the correct active set seems not a subset of $S$
- **covariance test statistic**

$$T_k = T(\hat{A}_{k-1}, \hat{\lambda}_{k+1}); \; \infty = \hat{\lambda}_0 \geq \hat{\lambda}_1 \geq \hat{\lambda}_2 \ldots ;$$

- asymptotically exponentially distributed
- “fixing $k$, the test is a conditional test for $H_S$ given that $\hat{A}_{k-1} = S$ ... a non-observable event”
• Cox & Donnelly, Principles of Applied Statistics, Ch.1: “It often aids in the interpretation of an observational study to consider the question: what would have been done in an otherwise comparable experiment”
• **Propensity score**: Probability of treatment assignment, conditional on observed baseline characteristics

\[ e(x) = \Pr(T = 1 \mid x) \]

• in a stratum of subjects who are matched on propensity score, treatment assignment is not confounded with measured baseline covariates
• thus by comparing treated and control subjects with the same propensity score, one can reduce the effects of observed confounding
• assumptions: (i) no unmeasured confounding, (ii) \(0 < \Pr(T = 1 \mid x) < 1\)
• one method of analysis weights each observation by the inverse probability of the treatment received
Missing Data

- complete data $Y_i = (Y_{i1}, \ldots, Y_{in_i})$, say
- following Heagerty, missing responses in a longitudinal study
- missing data indicator $R_i = (R_{i1}, \ldots, R_{in_i})$, where $R_{ij} = 1$ if $Y_{ij}$ is observed, else 0
- $Y_i = (Y_{i\text{obs}}, Y_{i\text{mis}})$
- **Missing Completely at Random** $\Pr(R_i = r_i \mid Y_i, X_i) = \Pr(R_i = r_i)$
- **Missing At Random** $\Pr(R_i = r_i \mid Y_i, X_i) = \Pr(R_i = r_i \mid Y_{i\text{obs}}, X_i)$

- **MCAR**: available case analysis is valid
- **MAR**: available case analysis is biased, but can be fixed
- **MNAR**, also called NMAR, also called non-ignorable: available case analysis is biased

Heagerty 2006 slides

missingdata.org.uk
Privacy
Upcoming

Conferences and Workshops

- **January 12 – 23, 2015**  
  **Opening Conference and Boot Camp**  
  Organizing committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Bin Yu

- **January 26 – 30, 2015**  
  **Workshop on Big Data and Statistical Machine Learning**  
  Organizing committee: Ruslan Salakhutdinov (Chair), Dale Schuurmans, Yoshua Bengio, Hugh Chipman, Bin Yu

- **February 9 – 11, 2015**  
  **Workshop on Optimization and Matrix Methods in Big Data**  
  Organizing Committee: Stephen Vavasis Chair; Anima Anandkumar, Petros Drineas, Michael Friedlander, Nancy Reid, Martin Wainwright.

- **February 23 – 27, 2015**  
  **Workshop on Visualization for Big Data: Strategies and Principles**  
  Organizing Committee: Nancy Reid (Chair), Susan Holmes, Snehalata Huzurbazar, Hadley Wickham, Leland Wilkinson

- **March 23-27, 2015**  
  **Workshop on Big Data in Health Policy**  
  Organizing Committee: Lisa Lix (Chair), Constantine Gatsonis, Sharon-Lise Normand, Therese Stukel

- **April 13 – 16, 2015**  
  **Workshop on Big Data for Social Policy**  
  Organizing Committee: Sallie Keller (chair), Robert Groves, Mary Thompson

- **June 12–13, 2015**  
  **Closing Conference**  
  Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Hugh Chipman, Rus Salakhutdinov, Yoshua Bengio, Richard Lockhart  
  to be held at AARMS of Dalhousie University,
  Held in conjunction with the Annual Meeting of the Canadian Statistical Sciences Institute, in the two days preceding the Annual Meeting of the Statistical Society of Canada. Overview lectures by members of the organizing committee will highlight the research generated by the thematic program.
Allied Activity

January - April, 2015
**Joint Big Data Program-Statistics Department Colloquia**

July 21 – August 15, 2014
**Summer School: Statistical Learning in Big Data**
Instructors: Hugh Chipman, Acadia; Sunny Wang, St. Francis Xavier
held at AARMS

April 7-9, 2015
**Coxeter Lecture Series**
**Michael Jordan** (University of California, Berkeley)
Room 230, Fields Institute

April 9-10, 2015
**Distinguished Lecture Series in Statistical Science**
**Terry Speed** (Walter and Eliza Hall Institute for Medical Research, Melbourne)
Room 230, Fields Institute

April 20–24, 2015
**Workshop Statistical Inference for Large Scale Data**
with Richard Lockhart (Chair), Nicolai Meinhausen
held at PIMS, Simon Fraser

April 23–24, 2015
**Distinguished Lecture Series in Statistical Science**
**Bin Yu**, University of California, Berkeley
Room 230, Fields Institute

April 21-24, 2015
**CANSSI Workshop on Complex spatio-temporal data structures: Methods and applications**
held at the Fields Institute

April 29-30, 2015
**Big Data in Commercial and Retail Banking**
with Mark Reesor, (Western); Matt Davison, (Western); Adam Metzler, (Wilfrid Laurier)
held at the Fields Institute

May 4 – 8, 2015
**Workshop and Short Course on Statistical and computational challenges in networks, web mining and cybersecurity**
with Hugh Chipman (Chair), François Théberge (U Ottawa)
held at CRM, Montreal

May 11–15, 2015
**Workshop on Big Data in Environmental Science**
with Richard Lockhart (Chair), James Zidek (UBC)
held at PIMS, University of British Columbia

July 31- August 9, 2015
**Deep Learning Summer School**
[https://sites.google.com/site/deeplearningsummerschool/](https://sites.google.com/site/deeplearningsummerschool/)
Organizing Committee: Yoshua Bengio, Chair
held at CRM, Montreal
Short Course on Latent Tree graphical models

April 27, 2015 at 10:00 a.m. - 12:00 p.m.
April 28, 2015 at 10:00 a.m. - 12:00 p.m.
April 29, 2015 at 10:00 a.m. - 11:00 a.m.

Stewart Library, The Fields Institute
Instructor: Piotr Zwiernik, University of Genoa

Description:

1. Trees, tree metrics and the space of trees.
I will introduce basic graph-theoretic tree concepts, tree metrics and other tree spaces that arise naturally in the study of latent tre graphical models.

2. Latent tree graphical models.
I will define the model and discuss the basic links to Bayesian networks and undirected graphical models on trees. I will present some basic results concerning identifiability and moment structure.

3. Inference.
In many application the main interest is in learning the underlying tree. I will give an overview of some methods of learning the tree and show how the idea of tree metrics provides a natural estimator.

I will introduce the structural EM algorithm for the MLE estimation and discuss some other approximate methods.

5. Special submodels: Hidden Markov model, symmetric models and models used in phylogenetics.
Many popular models arise as special cases of latent tree Graphical models. In this lecture I discuss these examples.

Industrial Problem Solving Workshop – May
Statistics Graduate Student Research Day

STATISTICS GRADUATE STUDENT RESEARCH DAY – APRIL 17, 2015

Machine Learning for Big Data at The Fields Institute

Speakers: Dr. Robert Bell, Researcher, Google  Dr. Alekh Agarwal, Researcher, Microsoft  Dr. Kevin Patrick Murphy, Researcher, Google

For more information, please visit: www.fields.utoronto.ca/programs/scientific/14-15/gradresearch