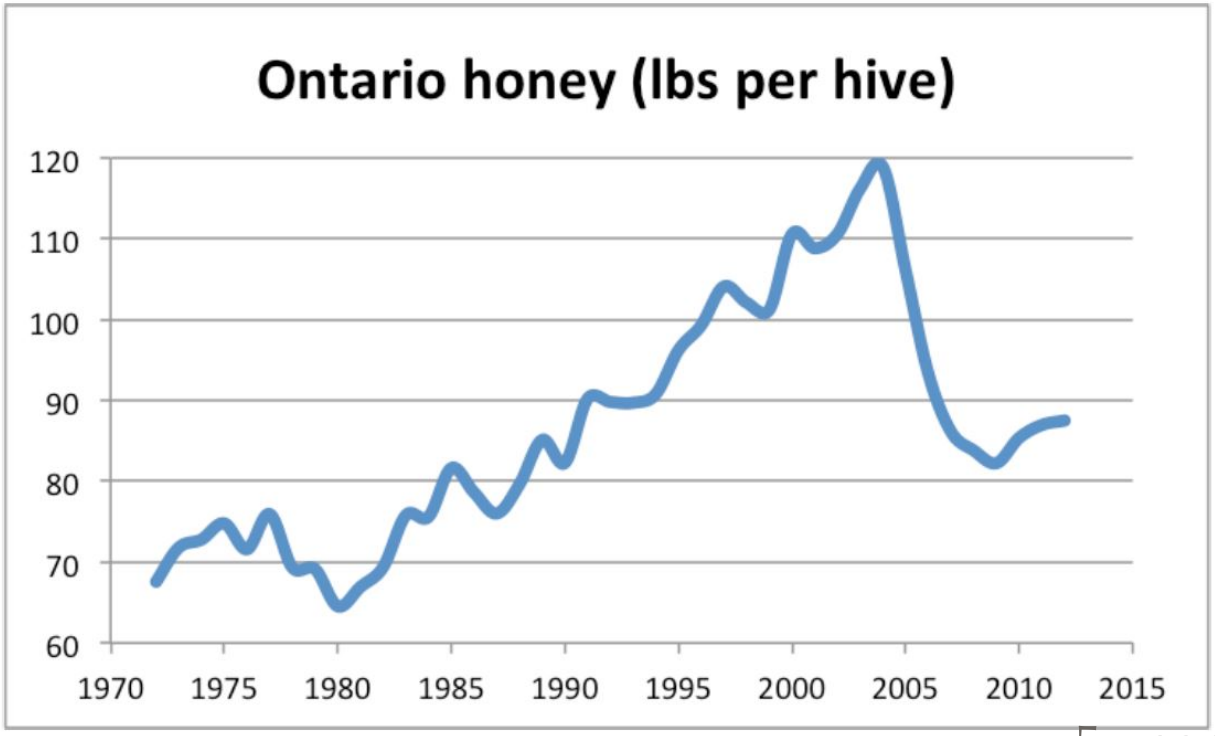
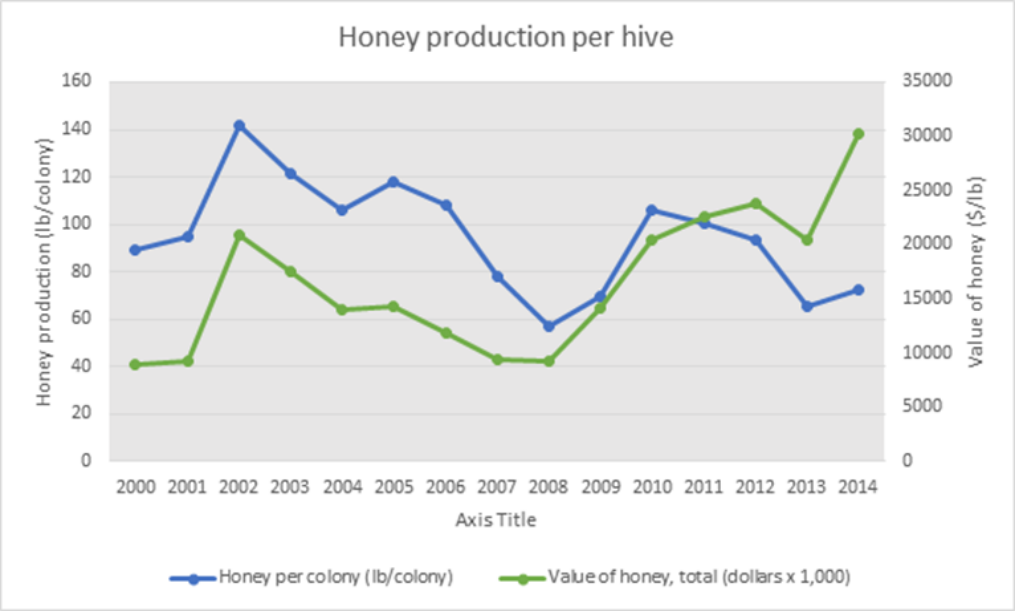
Topics covered in Part 2

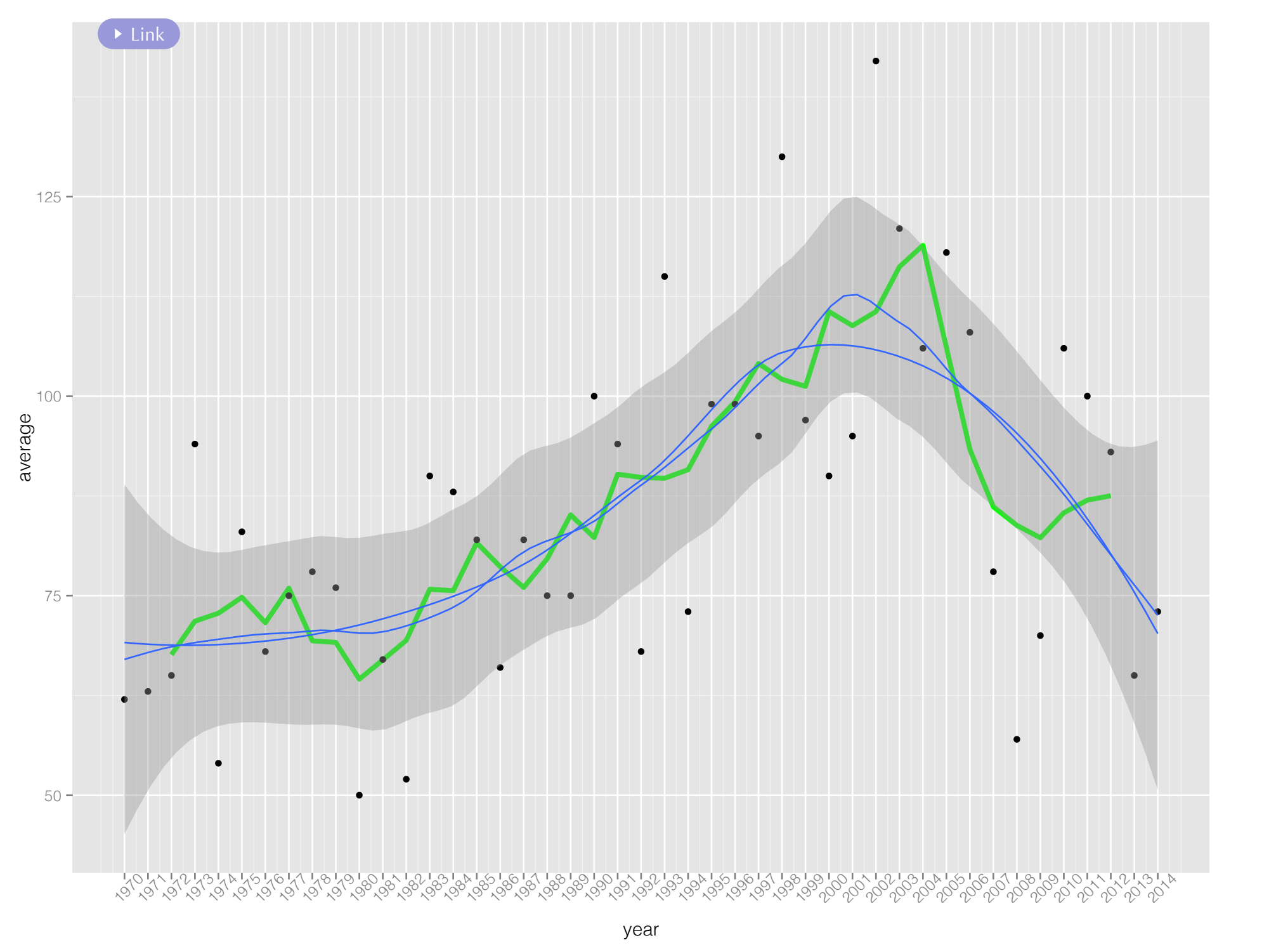
- visualization
- inference
- planning of studies
- privacy



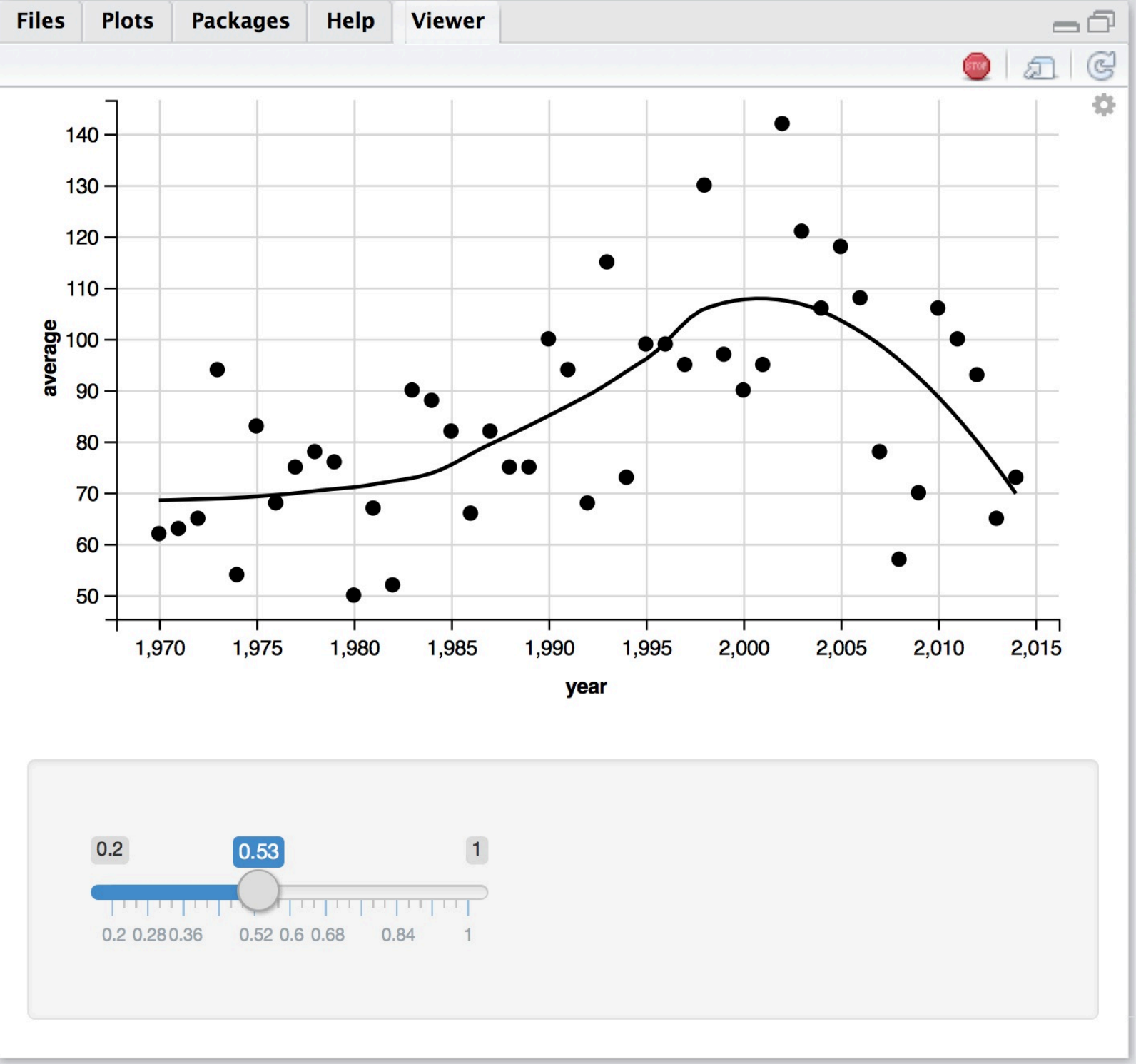
5-year running mean



annual averages


Link
125
100
75
50
average
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
year

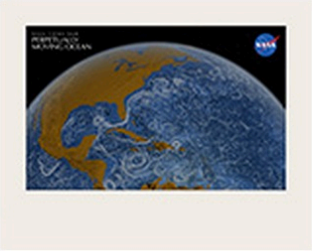
ggvis


Files
Plots
Packages
Help
Viewer
average
140
130
120
110
100
90
80
70
60
50
1,970
1,975
1,980
1,985
1,990
1,995
2,000
2,005
2,010
2,015
year
0.2
0.53
1
0.2 0.28 0.36
0.52 0.6 0.68
0.84
1

Link to Katy's home page

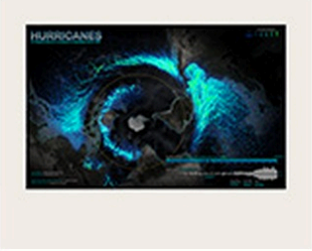
Iteration IX (2013): Science Maps Showing Trends and Dynamics

Four Existing Maps



IX.1

IX.3

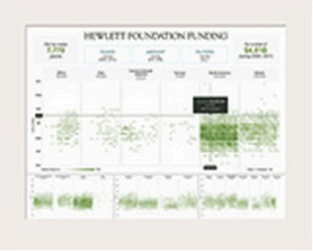


IX.2

IX.4

Six Science Maps

IX.5



IX.7

IX.9

IX.6

IX.8

IX.10

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.

Link to Places and Spaces

Inference for Lasso $n < p$

Bühlmann et al.

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

- 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': $\{j : \beta_j^0 \neq 0\}$
- 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 $n < p$

Bühlmann 2014 Ann. Rev.; van de Geer et al. 2014 AOS

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

- what about estimated standard errors for $\hat{\beta}_{Lasso}$?
- corrected ridge or lasso estimators:
  1. for $j = 1, \ldots, p$ construct a linear estimator $\hat{b}_j = Z_j^{\mathrm{T}} Y / Z_j^{\mathrm{T}} X_j$
  2. $\mathrm{E}(\hat{b}_j) = \beta_j + \sum_{k \neq j} P_{jk} \beta_k = \sum_{k \neq j} Z_j^{\mathrm{T}} X_k / Z_j^{\mathrm{T}} X_j$
  3. estimate this:
  $$\hat{\beta}_{j,corr} = \hat{b}_j - \sum_{k \neq j} P_{jk} \hat{\beta}_{k,Lasso}$$
- result
$$\sqrt{n}(\hat{\beta}_{j,corr} - \beta_j)/\hat{\sigma} \xrightarrow{\mathcal{L}} N(0,1)$$
- sparsity assumption $|S| \log p / \sqrt{n} \to 0$

Inference for Lasso $n < p$

Bühlmann et al., Lockhart et al.

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

First, we define some needed quantities. Let A be the active set just before λ_k, and suppose that predictor j enters at λ_k. Denote by $\hat{\beta}(\lambda_{k+1})$ the solution at the next knot in the path λ_{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,

$$\tilde{\beta}_A(\lambda_{k+1}) = \underset{\beta_A \in \mathbb{R}^{|A|}}{\text{argmin}} \frac{1}{2}\|y - X_A\beta_A\|_2^2 + \lambda_{k+1}\|\beta_A\|_1. \tag{4}$$

We propose the *covariance test statistic* defined by

$$T_k = (\langle y, X\hat{\beta}(\lambda_{k+1})\rangle - \langle y, X_A\tilde{\beta}_A(\lambda_{k+1})\rangle)/\sigma^2. \tag{5}$$

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

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 $S \subset \{1, \ldots, p\}; \quad \hat{\beta}_S(\lambda)$
- Covariance test

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

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

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

- asymptotically exponentially distributed
- "fixing k, the test is a conditional test for H_S given that $\widehat{A}_{k-1} = S$ ... a non-observable event"

Planning of Studies

Propensity Scores, Peter Austin; MSM, Erica Moodie

- 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
- $Y_i = (Y_i^{obs}, Y_i^{mis})$
- Missing Completely at Random $\Pr(R_i = r_i \mid Y_i, X_i) = \Pr(R_i = r_i)$
- Missing At Random $\Pr(R_i = r_i \mid Y_i, X_i) = \Pr(R_i = r_i \mid Y_i^{obs}, X_i)$
- 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, Snehelata 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 Statistica
  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 Meinshausen
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 Institutue

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 Institutue

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/
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 andother tree spaces that arise naturally in the study of latent treegraphical 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 somebasic 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.

4. Parameter estimation.
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 Graphicalmodels. In this lecture I discuss these examples.

Allied Activity:

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