Differential Inclusion Method in High Dimensional Statistics

Yuan Yao

HKUST

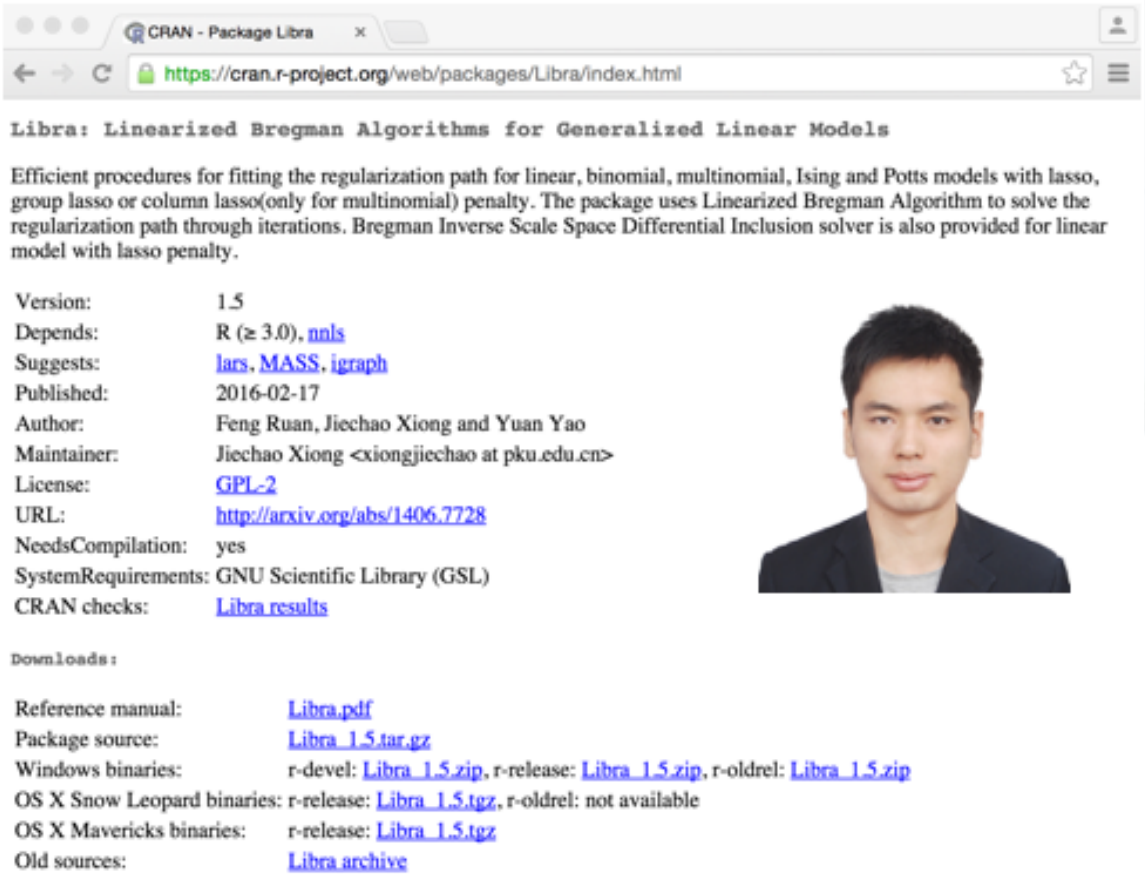
June 30, 2018

Acknowledgements

- Theory
  - *Stanley Osher*, *Wotao Yin* (UCLA)
  - *Feng Ruan* (Stanford & PKU)
  - *Jiechao Xiong*, *Chendi Huang* (PKU)
- Applications:
  - *Qianqian Xu*, *Jiechao Xiong*, *Chendi Huang*, *Xinwei Sun* (PKU)
  - *Lingjing Hu* (BCMU)
  - *Yifei Huang*, *Weizhi Zhu* (HKUST)
  - *Ming Yan*, *Zhimin Peng* (UCLA)
- Grants:
  - National Basic Research Program of China (973 Program), NSFC

Cran R package: Libra

http://cran.r-project.org/web/packages/Libra/


CRAN - Package Libra

https://cran.r-project.org/web/packages/Libra/index.html

Libra: Linearized Bregman Algorithms for Generalized Linear Models

Efficient procedures for fitting the regularization path for linear, binomial, multinomial, Ising and Potts models with lasso, group lasso or column lasso(only for multinomial) penalty. The package uses Linearized Bregman Algorithm to solve the regularization path through iterations. Bregman Inverse Scale Space Differential Inclusion solver is also provided for linear model with lasso penalty.

Version:	1.5
Depends:	R (≥ 3.0), nnls
Suggests:	lars, MASS, igraph
Published:	2016-02-17
Author:	Feng Ruan, Jiechao Xiong and Yuan Yao
Maintainer:	Jiechao Xiong <xiongjiechao at pku.edu.cn>
License:	GPL-2
URL:	http://arxiv.org/abs/1406.7728
NeedsCompilation:	yes
SystemRequirements:	GNU Scientific Library (GSL)
CRAN checks:	Libra results

Downloads:

Reference manual:	Libra.pdf
Package source:	Libra_1.5.tar.gz
Windows binaries:	r-devel: Libra_1.5.zip, r-release: Libra_1.5.zip, r-oldrel: Libra_1.5.zip
OS X Snow Leopard binaries:	r-release: Libra_1.5.tgz, r-oldrel: not available
OS X Mavericks binaries:	r-release: Libra_1.5.tgz
Old sources:	Libra archive

Libra (1.6) currently includes

Sparse statistical models:

- linear regression: ISS (differential inclusion), LB
- logistic regression (binomial, multinomial): LB
- graphical models (Gaussian, Ising, Potts): LB

Two types of regularization:

- LASSO: l_1-norm penalty
- Group LASSO: $l_2 - l_1$ penalty

`Libra` computes regularization paths via Linearized Bregman Iteration (LB)

for $\theta_0 = z_0 = \mathbf{0}$ and $k \in \mathbb{N}$,

$$z_{k+1} = z_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla_\theta \ell(x_i, \theta_k) \tag{1a}$$

$$\theta_{k+1} = \kappa \cdot \text{prox}_{\|\cdot\|_*}(z_{k+1}) \tag{1b}$$

where

- $\ell(x, \theta)$ is the *loss* function to minimize
- $\text{prox}_{\|\cdot\|_*}(z) := \arg\min_u \left(\frac{1}{2}\|u - z\|^2 + \|z\|_* \right)$
- $\alpha_k > 0$ is step-size
- $\kappa > 0$ while $\alpha_k \kappa \|\nabla_\theta^2 \hat{\mathbb{E}} \ell(x, \theta)\| < 2$
- as simple as ISTA, easy to parallel implementation

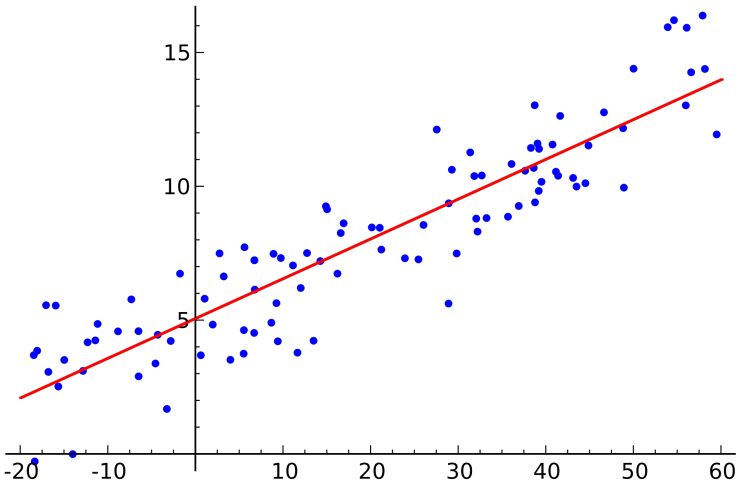
Linear Regression

Linear Regression:

$$y = X\beta + \epsilon$$

β is sparse or group sparse, with two types of penalty:

- "ungrouped": $\sum_i |\beta_i|$
- "grouped": $\sum_g \sqrt{\sum_{g_i=g} \beta_i^2}$

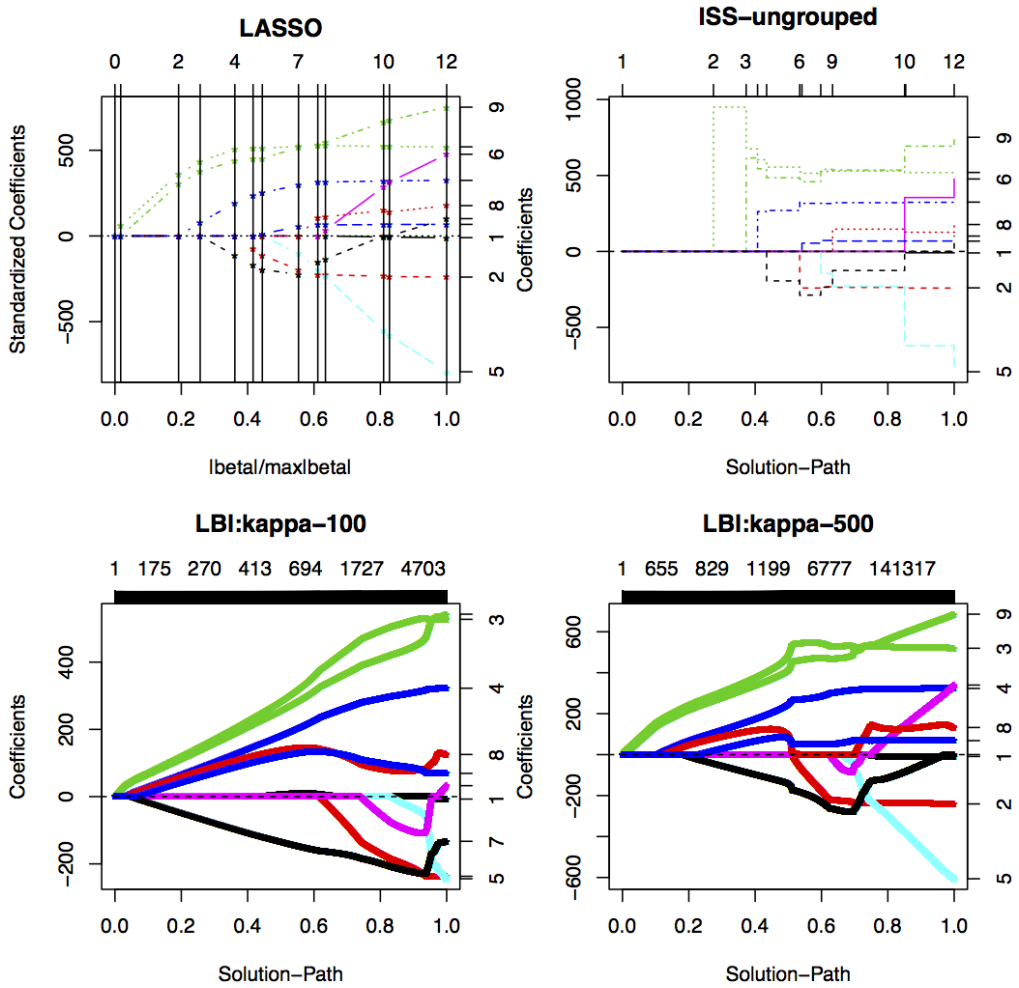


Linear Regression Example: Diabetes Data

```
data('diabetes')
attributes(x)
#$dim
# [1] 442  10
#$dimnames[[2]]
# [1] "age" "sex" "bmi" "map" "tc"  "ldl" "hdl" "tch" "ltg" "glu"

lassopath = lars(x,y)
isspath = iss(x,y)
lb(x,y,kappa=100,alpha=0.005,family="gaussian",group="ungrouped",
    intercept=FALSE,normalize=FALSE)
lb(x,y,kappa=500,alpha=0.001,family="gaussian",group="ungrouped",
    intercept=FALSE,normalize=FALSE)
```

LB generates iterative regularization paths



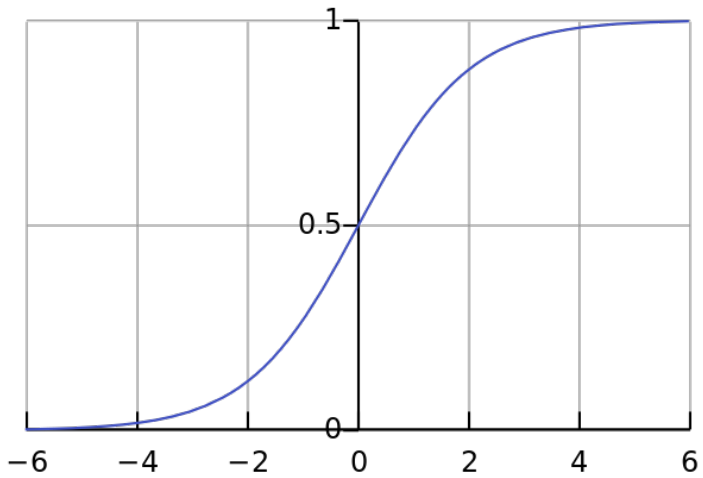
Logistic Regression

Logistic Regression:

$$\log \frac{P(y=1|X)}{P(y=-1|X)} = X\beta \Leftrightarrow P(y=1|X) = \frac{e^{X\beta}}{1+e^{X\beta}} =: \sigma(X\beta)$$

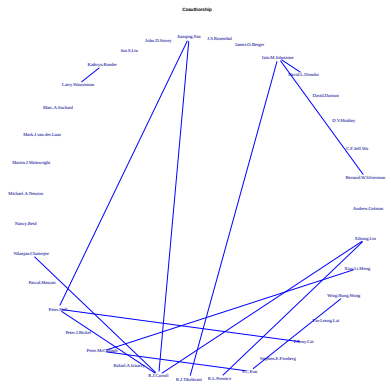
β is sparse or group sparse, with two types of penalty:

- "ungrouped": $\sum_i |\beta_i|$
- "grouped": $\sum_g \sqrt{\sum_{g_i=g} \beta_i^2}$

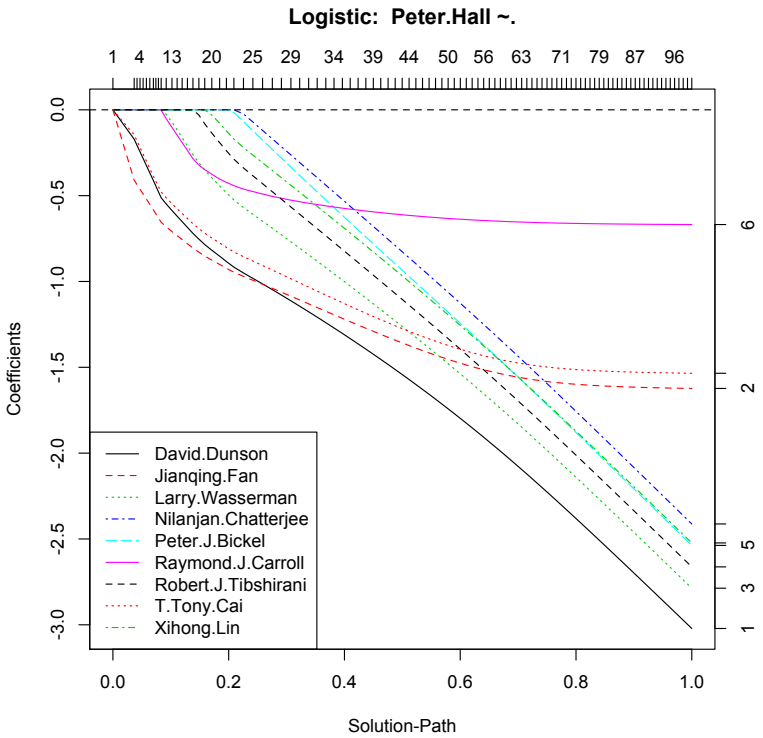


Example: Publications of COPSS Award Winners

- dataset is provided by Prof. Jiashun Jin @CMU
- 3248 papers by 3607 authors between 2003 and the first quarter of 2012 from:
  - the Annals of Statistics, Journal of the American Statistical Association, Biometrika and Journal of the Royal Statistical Society Series B
- a subset of 382 papers by 35 COPSS award winners
- Question: can we model the coauthorship structure to predict the out-of-sample behavior?



A logistic regression path with early stopping regularization



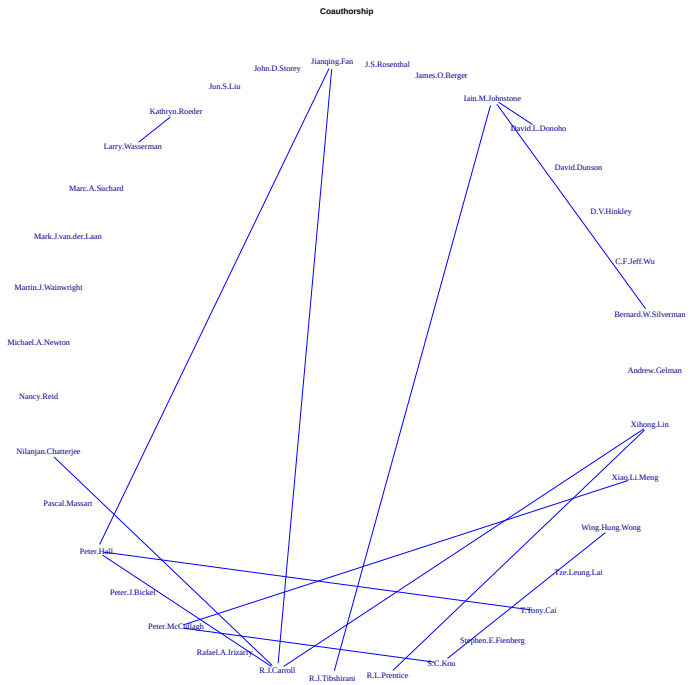


Figure: Peter Hall vs. other COPSS award winners in sparse logistic regression [papers from AoS/JASA/Biometrika/JRSSB, 2003-2012]: true coauthors are merely Tony Cai, R.J. Carroll, and J. Fan

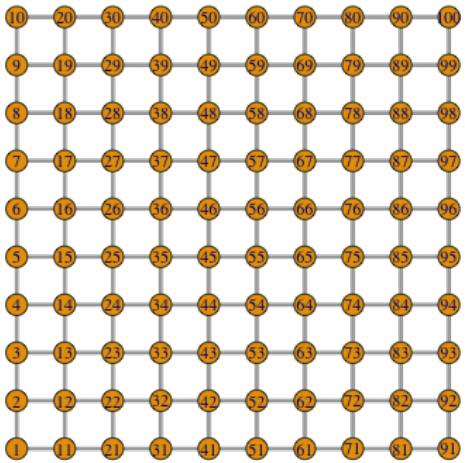
Sparse Ising Model

All models are wrong, but some are useful (George Box):

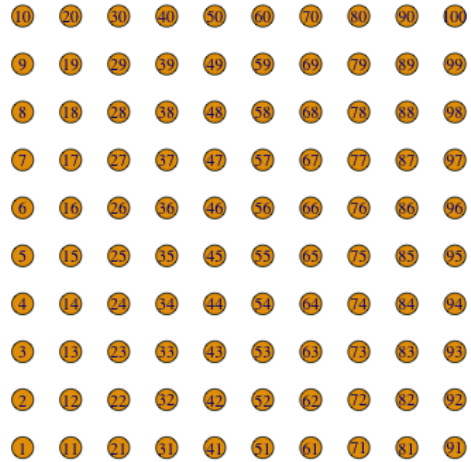
$$P(x_1, \ldots, x_p) \sim \exp\left(\sum_i H_i x_i + \sum_{i,j} J_{ij} x_i x_j\right)$$

- Ising model: $x_i = 1$ if author i appears in a paper, otherwise 0
- H_i describes the mean publication rate of author i
- J_{ij} describes the interactions between author i and j
  - $J_{ij} > 0$: author i and j collaborate more often than others
  - $J_{ij} < 0$: author i and j collaborate less frequently than others
  - sparsity: $J_{ij} = 0$ mostly, a model of collaboration network
  - learned by maximum composite conditional likelihood with LB

Early stopping against overfitting in sparse Ising model learning



a true Ising model of 2-D grid



a movie of LB path

Application: Sparse Ising Model of COPSS Award Winners

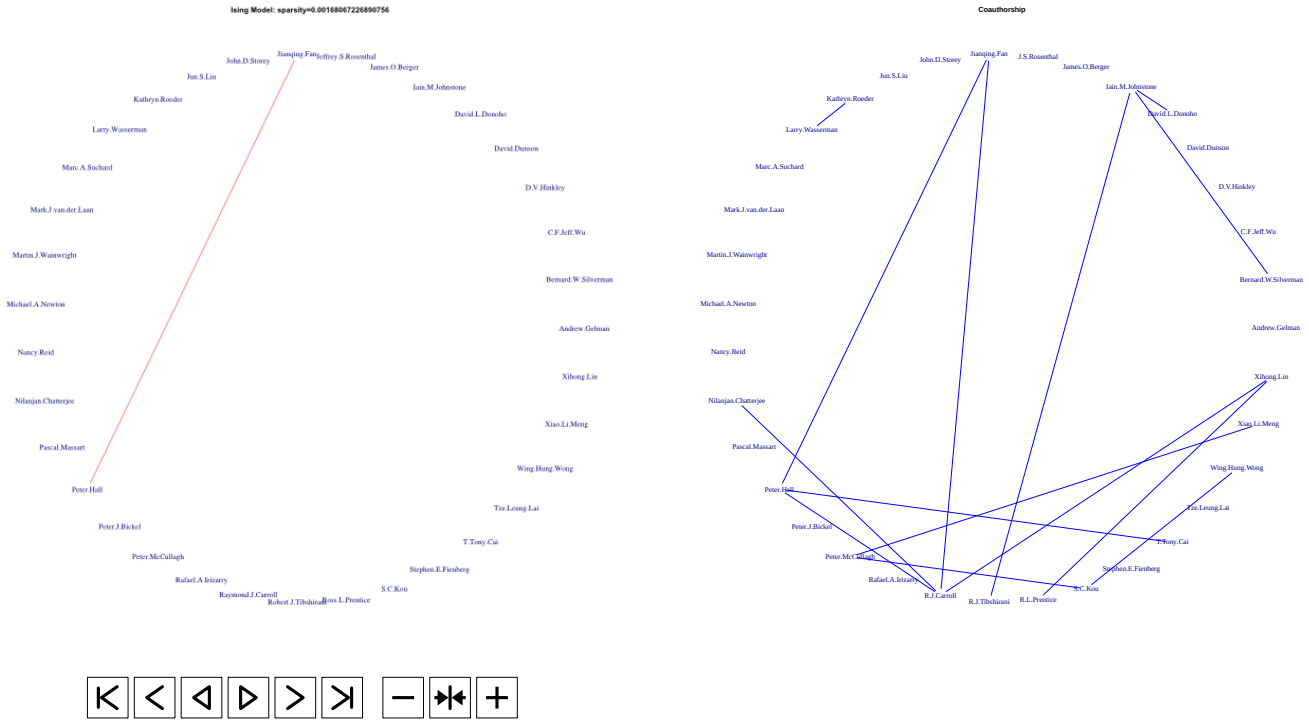
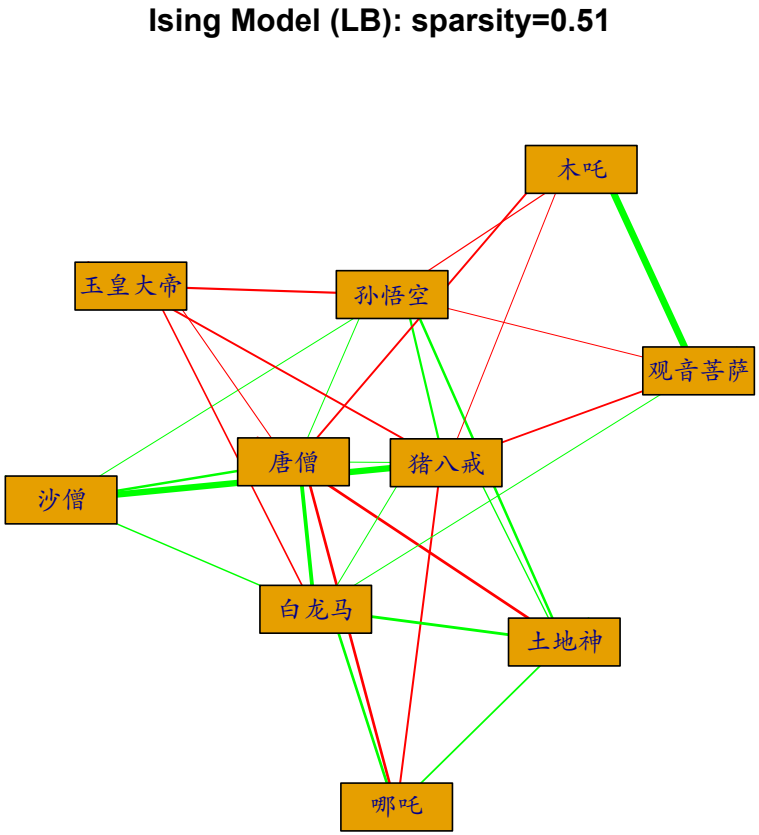


Figure: Left: LB path of Ising Model learning; Right: coauthorship network of existing data. Typically COPSS winners do not like working together; Peter Hall (1951-2016) is the hub of statisticians, like Erdös for mathematicians

Example: Ising Model of Journey to the West



Example: Dream of the Red Mansion (Xueqin Cao vs. E. Gao)

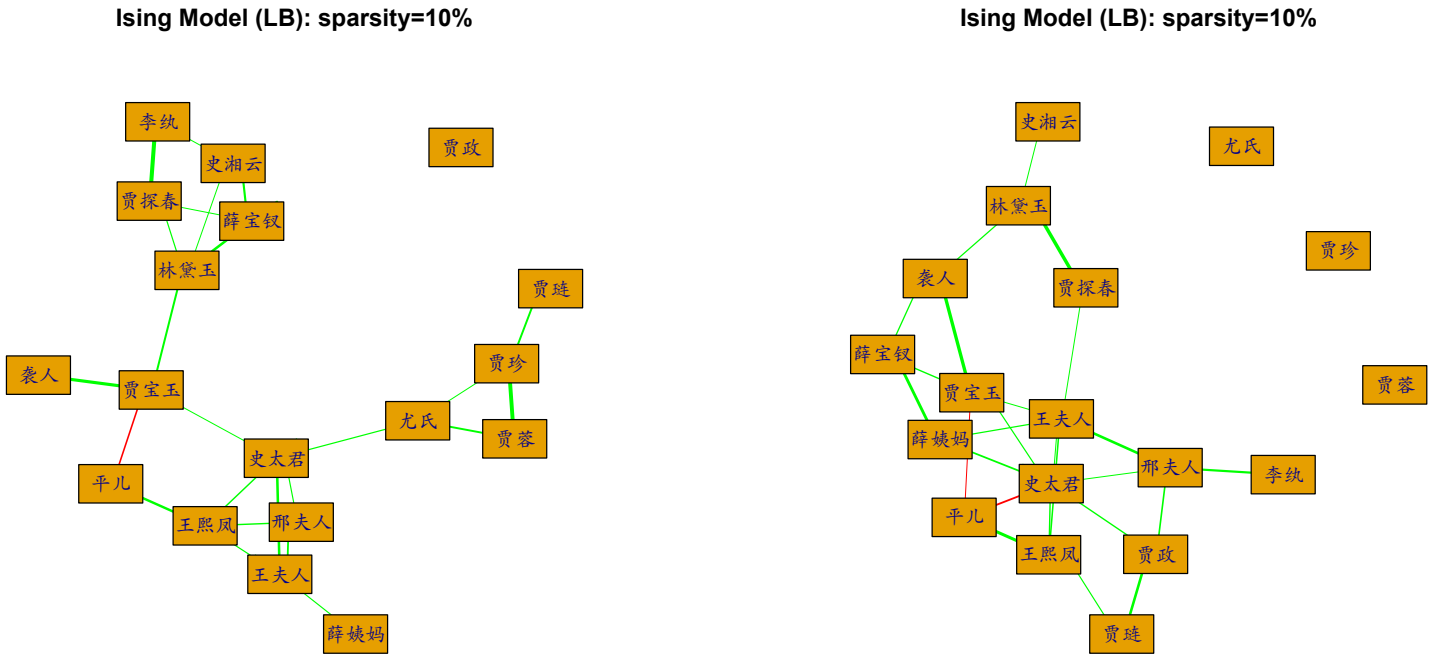


Figure: Left: main characters net in the first 80 chapters at sparsity 10%; Right: the remaining 40 chapters.

How does it work?

A story behind the R-package is in the following:

- The simple iterative algorithm shadows a particular kind of dynamics: differential inclusions, which are restricted gradient descent flows
- Simple discretized algorithm, amenable for parallel implementation
- Under nearly the same condition as LASSO, it reaches variable selection consistency
- but may incur less bias than LASSO
- Equipped with variable splitting, it weakens the conditions of generalized LASSO in variable selection

Sparse Linear Regression

Assume that $\beta^* \in \mathbb{R}^p$ is sparse and unknown. Consider recovering β^* from n linear measurements

$$y = X\beta^* + \epsilon, \qquad y \in \mathbb{R}^n$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is **noise**.

- Basic Sparsity: $S := \mathrm{supp}(\beta^*)$ ($s = |S|$) and T be its complement.
  - X_S (X_T) be the columns of X with indices restricted on S (T)
  - X is n-by-p, with $p \gg n \geq s$.
- Or Structural Sparsity: $\gamma^* = D\beta^*$ is sparse, where D is a linear transform (wavelet, gradient, etc.), $S = \mathrm{supp}(\gamma^*)$
- *How to recover β^* (or γ^*) sparsity pattern (sparsistency) and estimate values with variations (consistency)?*

Best Possible in Basic Setting: The Oracle Estimator

Had God revealed S to us, the *oracle estimator* was the subset least square solution (MLE) with $\tilde{\beta}^*_T = 0$ and

$$\tilde{\beta}^*_S = \beta^*_S + \frac{1}{n}\Sigma_n^{-1} X_S^T \epsilon, \quad \text{where } \Sigma_n = \tfrac{1}{n} X_S^T X_S \tag{2}$$

"Oracle properties"

- **Model selection consistency**: $\mathrm{supp}(\tilde{\beta}^*) = S$;
- **Normality**: $\tilde{\beta}^*_S \sim \mathcal{N}(\beta^*, \frac{\sigma^2}{n}\Sigma_n^{-1})$.

So $\tilde{\beta}^*$ is unbiased, i.e. $\mathbb{E}[\tilde{\beta}^*] = \beta^*$.

Recall LASSO

LASSO:

$$\min_{\beta} \|\beta\|_1 + \frac{t}{2n}\|y - X\beta\|_2^2.$$

optimality condition:

$$\frac{\rho_t}{t} = \frac{1}{n}X^T(y - X\beta_t), \tag{3a}$$

$$\rho_t \in \partial\|\beta_t\|_1, \tag{3b}$$

where $\lambda = 1/t$ is often used in literature.

- Chen-Donoho-Saunders'1996 (BPDN)
- Tibshirani'1996 (LASSO)

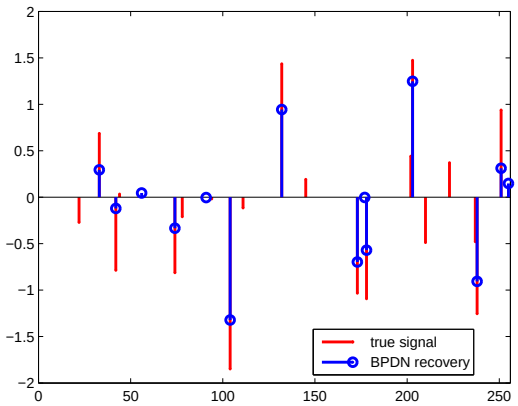
The Bias of LASSO

LASSO is biased, i.e. $\mathbb{E}(\hat{\beta}) \neq \beta^*$

- e.g. $X = Id$, $n = p = 1$, LASSO is soft-thresholding

$$\hat{\beta}_\tau = \begin{cases} 0, & \text{if } \tau < 1/\tilde{\beta}^*; \\ \tilde{\beta}^* - \frac{1}{\tau}, & \text{otherwise,} \end{cases}$$

- e.g. $n = 100$, $p = 256$, $X_{ij} \sim \mathcal{N}(0,1)$, $\epsilon_i \sim \mathcal{N}(0, 0.1)$



True vs LASSO (t hand-tuned)

LASSO Estimator is Biased at Path Consistency

Even when the following **path consistency** (conditions given by Zhao-Yu'06, Zou'06, Yuan-Lin'07, Wainwright'09, etc.) is reached at τ_n:

$$\exists \tau_n \in (0, \infty) \text{ s.t. } \operatorname{supp}(\hat{\beta}_{\tau_n}) = S,$$

LASSO estimate is biased away from the oracle estimator

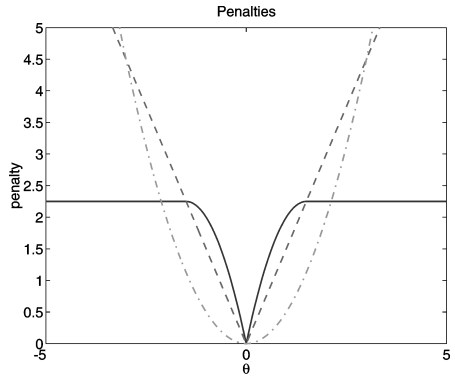
$$(\hat{\beta}_{\tau_n})_S = \tilde{\beta}^*_S - \frac{1}{\tau_n} \Sigma^{-1}_{n,S} \operatorname{sign}(\beta^*_S), \quad \tau_n > 0.$$

How to remove the bias and return the Oracle Estimator?

Nonconvex Regularization?

- To reduce bias, **non-convex** regularization was proposed (Fan-Li's SCAD, Zhang's MPLUS, Zou's Adaptive LASSO, l_q $(q < 1)$, etc.)

$$\min_{\beta} \sum_i p(|\beta_i|) + \frac{t}{2n}\|y - X\beta\|_2^2.$$



- Yet it is generally hard to locate the global optimizer
- *Any other simple scheme?*

New Idea

- LASSO:

$$\min_{\beta} \|\beta\|_1 + \frac{t}{2n}\|y - X\beta\|_2^2.$$

New Idea

- LASSO:

$$\min_{\beta} \|\beta\|_1 + \frac{t}{2n}\|y - X\beta\|_2^2.$$

- KKT optimality condition:

$$\Rightarrow \rho_t = \frac{1}{n}X^T(y - X\beta_t)t$$

New Idea

- LASSO:

$$\min_{\beta} \|\beta\|_1 + \frac{t}{2n}\|y - X\beta\|_2^2.$$

- KKT optimality condition:

$$\Rightarrow \rho_t = \frac{1}{n}X^T(y - X\beta_t)t$$

- Taking derivative (assuming differentiability) w.r.t. t

$$\Rightarrow \dot{\rho}_t = \frac{1}{n}X^T(y - X(\dot{\beta}_t t + \beta_t)), \quad \rho_t \in \partial\|\beta_t\|_1$$

New Idea

- LASSO:

$$\min_{\beta} \|\beta\|_1 + \frac{t}{2n}\|y - X\beta\|_2^2.$$

- KKT optimality condition:

$$\Rightarrow \rho_t = \frac{1}{n}X^T(y - X\beta_t)t$$

- Taking derivative (assuming differentiability) w.r.t. t

$$\Rightarrow \dot{\rho}_t = \frac{1}{n}X^T(y - X(\dot{\beta}_t t + \beta_t)), \quad \rho_t \in \partial\|\beta_t\|_1$$

- Assuming sign-consistency in a neighborhood of τ_n,

$$\text{for } i \in S,\ \rho_{\tau_n}(i) = \text{sign}(\beta^*(i)) \in \pm 1 \Rightarrow \dot{\rho}_{\tau_n}(i) = 0,$$

$$\Rightarrow \dot{\beta}_{\tau_n}\tau_n + \beta_{\tau_n} = \tilde{\beta}^*$$

New Idea

- LASSO:

$$\min_{\beta} \|\beta\|_1 + \frac{t}{2n}\|y - X\beta\|_2^2.$$

- KKT optimality condition:

$$\Rightarrow \rho_t = \frac{1}{n} X^T (y - X\beta_t) t$$

- Taking derivative (assuming differentiability) w.r.t. t

$$\Rightarrow \dot{\rho}_t = \frac{1}{n} X^T (y - X(\dot{\beta}_t t + \beta_t)), \quad \rho_t \in \partial \|\beta_t\|_1$$

- Assuming sign-consistency in a neighborhood of τ_n,

$$\text{for } i \in S,\ \rho_{\tau_n}(i) = \text{sign}(\beta^*(i)) \in \pm 1 \Rightarrow \dot{\rho}_{\tau_n}(i) = 0,$$

$$\Rightarrow \dot{\beta}_{\tau_n} \tau_n + \beta_{\tau_n} = \tilde{\beta}^*$$

- Equivalently, the blue part removes bias of LASSO automatically

$$\beta^{lasso}_{\tau_n} = \tilde{\beta}^* - \frac{1}{\tau_n} \Sigma_n^{-1} \text{sign}(\beta^*) \Rightarrow \dot{\beta}^{lasso}_{\tau_n} \tau_n + \beta^{lasso}_{\tau_n} = \tilde{\beta}^* (oracle)!$$

Differential Inclusion: Inverse Scaled Spaces (ISS)

Differential inclusion replacing $\dot{\beta}^{lasso}_{\tau_n}\tau_n + \beta^{lasso}_{\tau_n}$ by β_t

$$\dot{\rho}_t = \frac{1}{n}X^T(y - X\beta_t), \tag{4a}$$

$$\rho_t \in \partial\|\beta_t\|_1. \tag{4b}$$

starting at $t = 0$ and $\rho(0) = \beta(0) = \mathbf{0}$.

- Replace ρ/t in LASSO KKT by $\mathrm{d}\rho/\mathrm{d}t$

$$\frac{\rho_t}{t} = \frac{1}{n}X^T(y - X\beta_t)$$

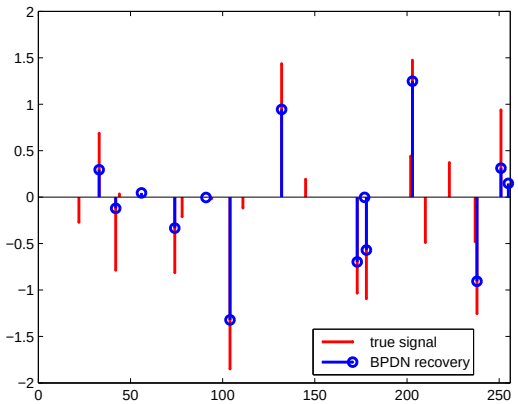
- Burger-Gilboa-Osher-Xu'06 (in image recovery it recovers the objects in an inverse-scale order as t increases (larger objects appear in β_t first))

Examples

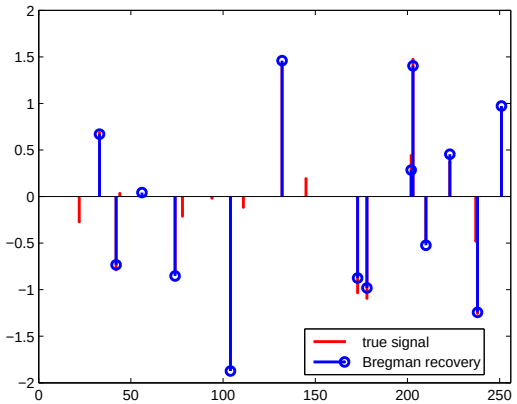
- e.g. $X = Id$, $n = p = 1$, hard-thresholding

$$\beta_\tau = \begin{cases} 0, & \text{if } \tau < 1/(\tilde{\beta}^*); \\ \tilde{\beta}^*, & \text{otherwise,} \end{cases}$$

- the same example shown before



True vs LASSO



True vs ISS

Solution Path: Sequential Restricted Maximum Likelihood Estimate

- ρ_t is piece-wise linear in t,

$$\rho_t = \rho_{t_k} + \frac{t - t_k}{n} X^T (y - X\beta_{t_k}), \quad t \in [t_k, t_{k+1})$$

where $t_{k+1} = \sup\{t > t_k : \rho_{t_k} + \frac{t-t_k}{n} X^T (y - X\beta_{t_k}) \in \partial \|\beta_{t_k}\|_1\}$

- β_t is piece-wise constant in t: $\beta_t = \beta_{t_k}$ for $t \in [t_k, t_{k+1})$ and $\beta_{t_{k+1}}$ is the **sequential restricted Maximum Likelihood Estimate** by solving nonnegative least square (Burger et al.'13; Osher et al.'16)

$$\begin{aligned} \beta_{t_{k+1}} = \arg\min_\beta \quad & \|y - X\beta\|_2^2 & & \\ \text{subject to} \quad & (\rho_{t_{k+1}})_i \beta_i \geq 0 & \forall\, i \in S_{k+1}, & \\ & \beta_j = 0 & \forall\, j \in T_{k+1}. & \end{aligned} \tag{5}$$

- Note: Sign consistency $\rho_t = \mathrm{sign}(\beta^*) \Rightarrow \beta_t = \tilde{\beta}^*$ the oracle estimator

Example: Regularization Paths of LASSO vs. ISS

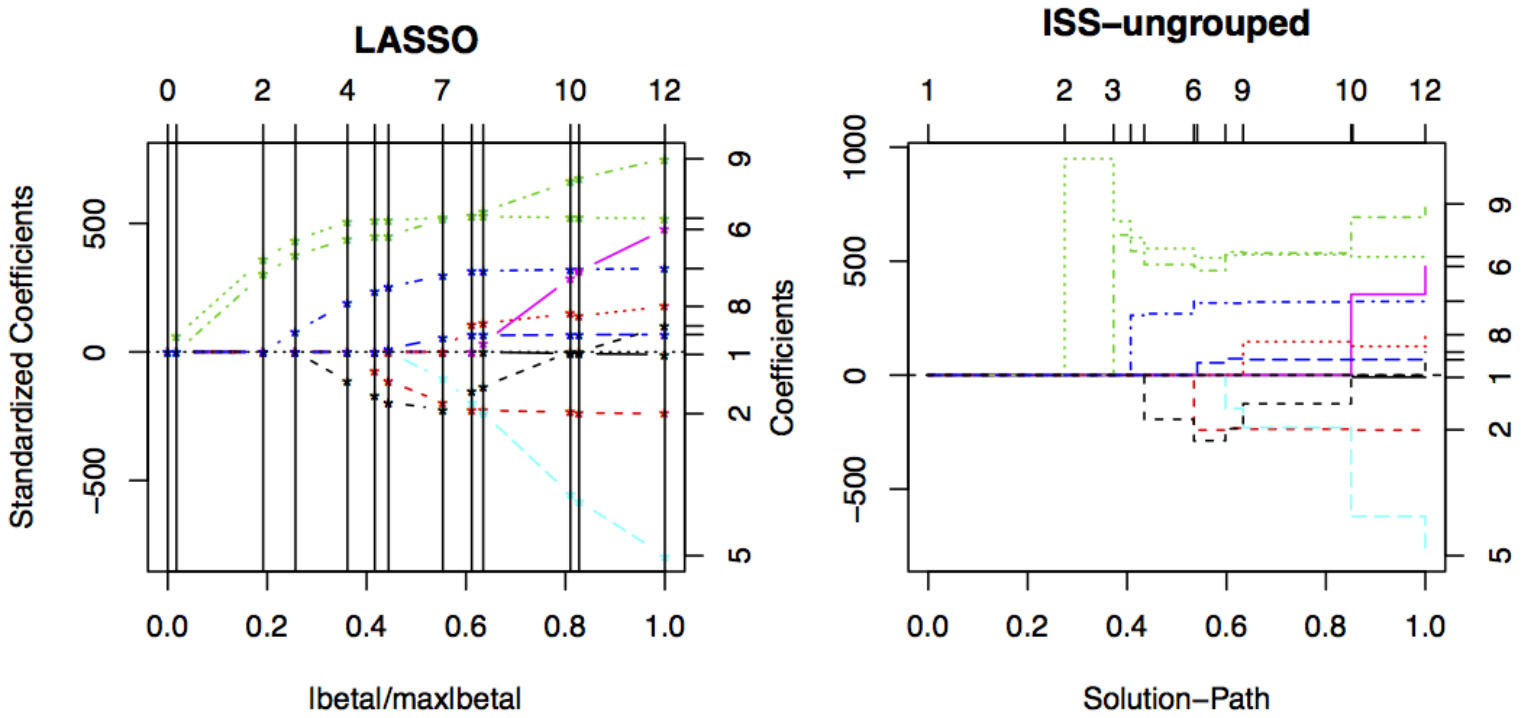


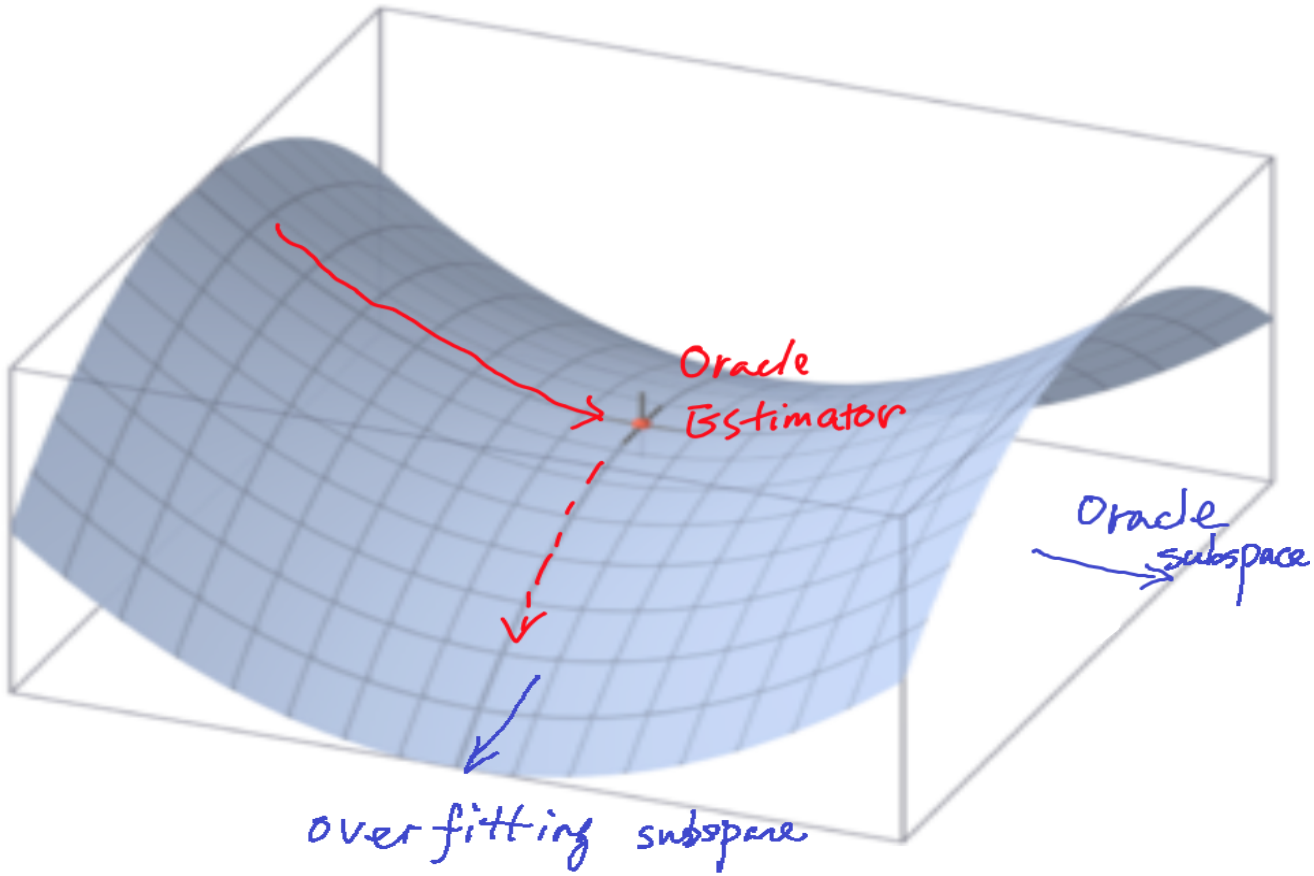
Figure: Diabetes data (Efron et al.'04) and regularization paths are different, yet bearing similarities on the order of parameters being nonzero

How does it work? A Path Consistency Theory

Our aim is to show that under nearly the same conditions for sign-consistency of LASSO, there exists points on their paths $(\beta(t), \rho(t))_{t \geq 0}$, which are

- sparse
- sign-consistent (the same sparsity pattern of nonzeros as true signal)
- the oracle estimator which is unbiased, better than the LASSO estimate.
- Early stopping regularization is necessary to prevent overfitting noise!

Intuition



History: two traditions of regularizations

- Penalty functions
  - ℓ_2: Ridge regression/Tikhonov regularization: $\frac{1}{n}\sum_{i=1}^n \ell(y_i, x_i^T\beta) + \lambda\|\beta\|_2^2$
  - ℓ_1 (sparse): Basis Pursuit/LASSO (ISTA): $\frac{1}{n}\sum_{i=1}^n \ell(y_i, x_i^T\beta) + \lambda\|\beta\|_1^2$
- Early stopping of dynamic regularization paths
  - ℓ_2-equivalent: Landweber iterations/gradient descent/ℓ_2-Boost

$$\frac{d\beta_t}{dt} = -\frac{1}{n}\sum_{i=1}^n \nabla_\beta \ell(y_i, x_i^T\beta), \quad \beta_t = \nabla\left\{\frac{1}{2}\|\beta_t\|^2\right\}$$

  - ℓ_1 (sparse)-equiv.: Orthogonal Matching Pursuit, Linearized Bregman Iteration (sparse Mirror Descent) (not ISTA! – later)

$$\frac{d\rho_t}{dt} = -\frac{1}{n}\sum_{i=1}^n \nabla_\beta \ell(y_i, x_i^T\beta), \quad \rho_t \in \partial\|\beta_t\|_1$$

Assumptions

(A1) **Restricted Strongly Convex**: $\exists \gamma \in (0, 1]$,

$$\frac{1}{n} X_S^T X_S \geq \gamma I$$

(A2) **Incoherence/Irrepresentable** Condition: $\exists \eta \in (0, 1)$,

$$\left\| \frac{1}{n} X_T^T X_S^\dagger \right\|_\infty = \left\| \frac{1}{n} X_T^T X_S \left(\frac{1}{n} X_S^T X_S \right)^{-1} \right\|_\infty \leq 1 - \eta$$

- "Irrepresentable" means that one can not represent (regress) column vectors in X_T by covariates in X_S.
- The incoherence/irrepresentable condition is used independently in Tropp'04, Yuan-Lin'05, Zhao-Yu'06, and Zou'06, Wainwright'09, etc.

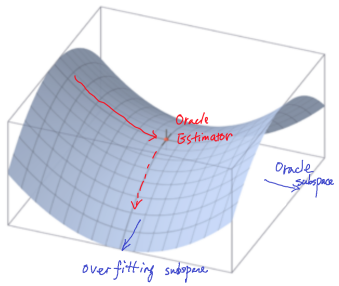
Understanding the Dynamics

ISS as restricted gradient descent:

$$\dot{\rho}_t = -\nabla L(\beta_t) = \frac{1}{n} X^T (y - X\beta_t), \quad \rho_t \in \partial \|\beta_t\|_1$$

such that

- incoherence condition and strong signals ensure it firstly evolves on index set S to reduce the loss
- strongly convex in subspace restricted on index set $S \Rightarrow$ fast decay in loss
- early stopping after all strong signals are detected, before picking up the noise



Path Consistency

Theorem (Osher-Ruan-Xiong-Y.-Yin'2016)

Assume (A1) and (A2). Define an early stopping time

$$\bar{\tau} := \frac{\eta}{2\sigma}\sqrt{\frac{n}{\log p}}\left(\max_{j\in T}\|X_j\|\right)^{-1},$$

and the smallest magnitude $\beta^*_{\min} = \min(|\beta^*_i| : i \in S)$. *Then*

- *No-false-positive: for all* $t \le \bar{\tau}$, *the path has no-false-positive with high probability,* $\mathrm{supp}(\beta(t)) \subseteq S$;
- *Consistency: moreover if the signal is strong enough such that*

$$\beta^*_{min} \ge \left(\frac{4\sigma}{\gamma^{1/2}} \vee \frac{8\sigma(2+\log s)\left(\max_{j\in T}\|X_j\|\right)}{\gamma\eta}\right)\sqrt{\frac{\log p}{n}},$$

there is $\tau \le \bar{\tau}$ *such that solution path* $\beta(t)) = \tilde{\beta}^*$ *for every* $t \in [\tau, \bar{\tau}]$.

Note: equivalent to LASSO with $\lambda^* = 1/\bar{\tau}$ (Wainwright'09) up to $\log s$.

Large scale algorithm: Linearized Bregman Iteration

Damped Dynamics: continuous solution path

$$\dot{\rho}_t + \frac{1}{\kappa}\dot{\beta}_t = \frac{1}{n}X^T(y - X\beta_t), \quad \rho_t \in \partial\|\beta_t\|_1. \tag{6}$$

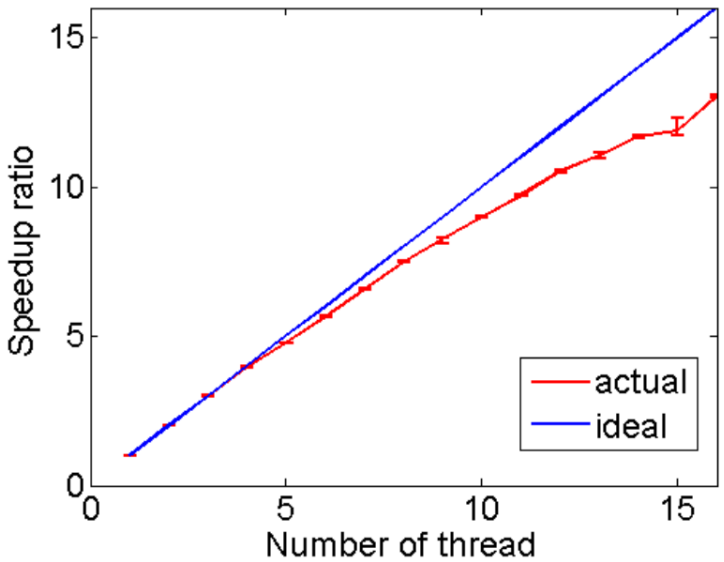
Linearized Bregman Iteration as forward Euler discretization proposed even earlier than ISS dynamics (Osher-Burger-Goldfarb-Xu-Yin'05, Yin-Osher-Goldfarb-Darbon'08): for $\rho_k \in \partial\|\beta_k\|_1$,

$$\rho_{k+1} + \frac{1}{\kappa}\beta_{k+1} = \rho_k + \frac{1}{\kappa}\beta_k + \frac{\alpha_k}{n}X^T(y - X\beta_k), \tag{7}$$

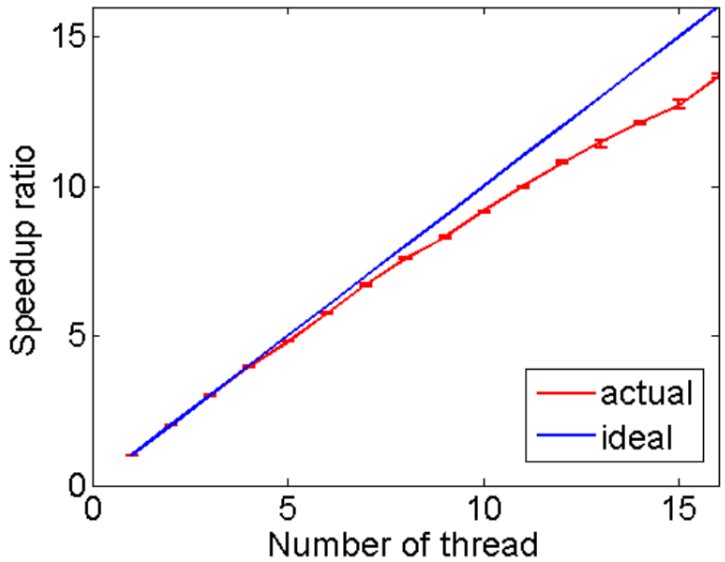
where

- Damping factor: $\kappa > 0$
- Step size: $\alpha_k > 0$ s.t. $\alpha_k\kappa\|\Sigma_n\| \leq 2$
- Moreau Decomposition: $z_k := \rho_k + \frac{1}{\kappa}\beta_k \Leftrightarrow \beta_k = \kappa \cdot \mathit{Shrink}(z_k, 1)$

Easy for Parallel Implementation



(a) n=1000



(b) n=2000

Figure: Linear speed-ups on a 16-core machine with synchronized parallel computation of matrix-vector products.

Comparison with ISTA

Linearized Bregman (LB) iteration:

$$z_{t+1} = z_t - \alpha_t X^T(\kappa X \mathit{Shrink}(z_t, 1) - y)$$

which is not **ISTA**:

$$z_{t+1} = \mathit{Shrink}(z_t - \alpha_t X^T(Xz_t - y), \lambda).$$

Comparison:

- **ISTA**:
  - as $t \to \infty$ solves **LASSO**: $\frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$
  - parallel run ISTA with $\{\lambda_k\}$ for LASSO regularization paths
- **LB**: a single run generates the whole regularization path at same cost of ISTA-LASSO estimator for a fixed regularization

LB generates regularization paths

$n = 200$, $p = 100$, $S = \{1, \ldots, 30\}$, $x_i \sim N(0, \Sigma_p)$ ($\sigma_{ij} = 1/(3p)$ for $i \neq j$ and 1 otherwise)

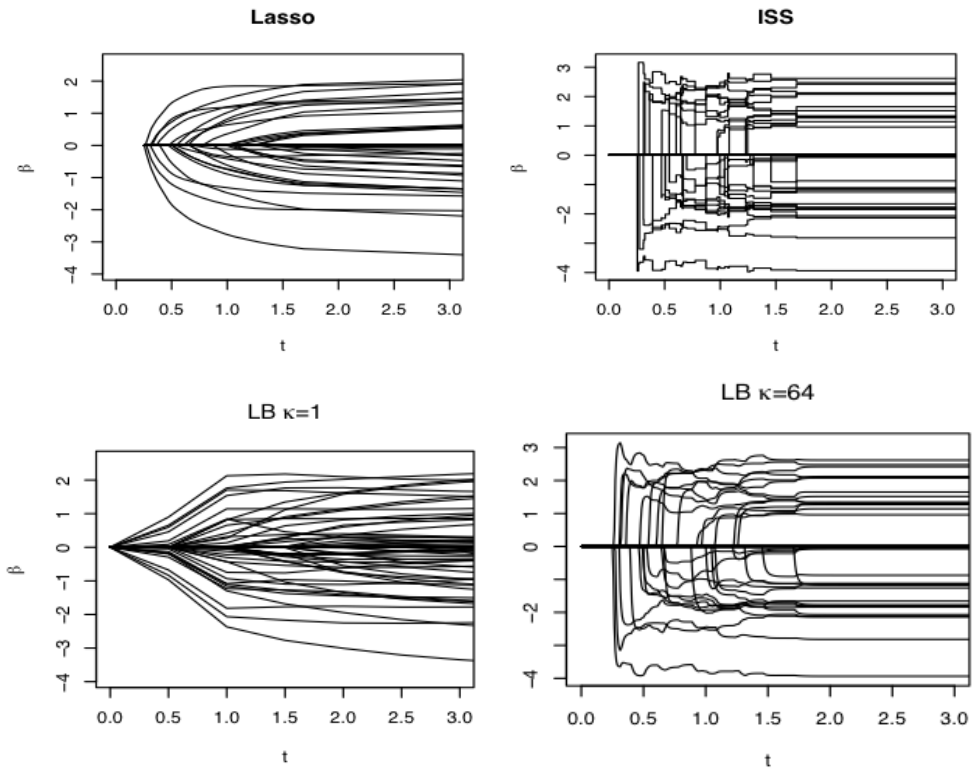
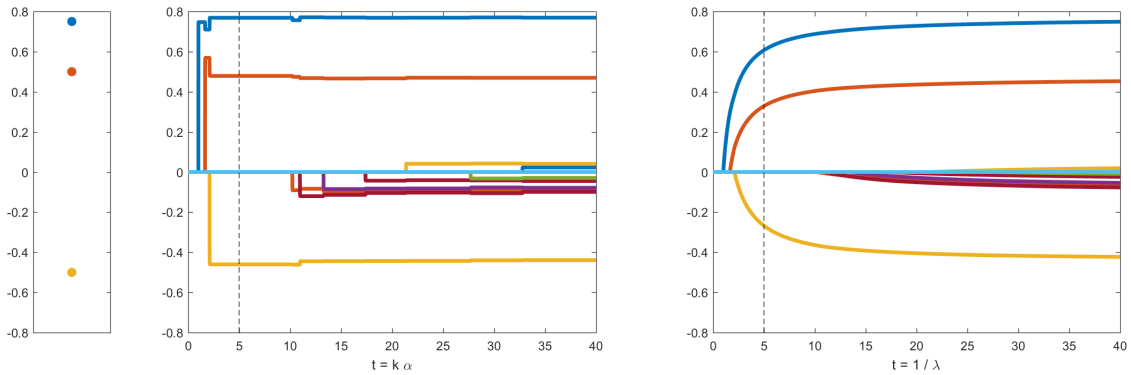


Figure: As $\kappa \to \infty$, LB paths have a limit as piecewise-constant ISS path

Accuracy: LB may be less biased than LASSO



- Left shows (the magnitudes of) nonzero entries of $\beta^\star$.
- Middle shows the regularization path of LB.
- Right shows the regularization path of LASSO vs. $t = 1/\lambda$.

Path Consistency in Discrete Setting

Theorem (Osher-Ruan-Xiong-Y.-Yin'2016)

Assume that κ is large enough and α is small enough, with $\kappa\alpha\|X_S^* X_S\| < 2$,

$$\overline{\tau} := \frac{(1 - B/\kappa\eta)\eta}{2\sigma}\sqrt{\frac{n}{\log p}}\left(\max_{j\in T}\|X_j\|\right)^{-1}$$

$$\beta^*_{max} + 2\sigma\sqrt{\frac{\log p}{\gamma n}} + \frac{\|X\beta^*\|_2 + 2s\sqrt{\log n}}{n\sqrt{\gamma}} \triangleq B \leq \kappa\eta,$$

then all the results for ISS can be extended to the discrete algorithm.

Note: it recovers the previous theorem as $\kappa \to \infty$ and $\alpha \to 0$, so LB can be less biased than LASSO.

General Loss and Regularizer

$$\dot{\eta}_t = -\frac{\kappa_0}{n}\sum_{i=1}^{n}\nabla_\eta \ell(x_i, \theta_t, \eta_t) \tag{8a}$$

$$\dot{\rho}_t + \frac{\dot{\theta}_t}{\kappa_1} = -\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta \ell(x_i, \theta_t, \eta_t) \tag{8b}$$

$$\rho_t \in \partial\|\theta_t\|_* \tag{8c}$$

where

- $\ell(x_i, \theta)$ is a loss function: negative logarithmic likelihood, non-convex loss (neural networks), etc.
- $\|\theta_t\|_*$ is the Minkowski-functional (gauge) of dictionary convex hulls:

$$\|\theta\|_* := \inf\{\lambda \geq 0 : \theta \in \lambda K\}, \quad K \text{ is a symmetric convex hull of } \{a_i\}$$

  - it can be generalized to non-convex regularizers

Linearized Bregman Iteration Algorithms

Differential inclusion (8) admits the following Euler Forward discretization

$$\eta_{t+1} = \eta_t - \frac{\alpha_k \kappa_0}{n} \sum_{i=1}^{n} \nabla_\eta \ell(x_i, \theta_t, \eta_t) \tag{9a}$$

$$z_{t+1} = z_t - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla_\theta \ell(x_i, \theta_t, \eta_t) \tag{9b}$$

$$\theta_{t+1} = \kappa_1 \cdot \text{prox}_{\|\cdot\|_*}(z_{t+1}) \tag{9c}$$

where (9c) is given by Moreau Decomposition with

$$\text{prox}_{\|\cdot\|_*}(z_t) = \arg\min_x \frac{1}{2}\|x - z_t\|^2 + \|x\|_*,$$

and

- $\alpha_k > 0$ is step-size while $\alpha_k \kappa_i \|\nabla^2_\theta \hat{\mathbb{E}} \ell(x, \theta)\| < 2$
- as simple as ISTA, easy to parallel implementation

More reference

- **Logistic Regression**: loss – conditional likelihood, regularizer – l_1 (Shi-Yin-Osher-Saijda'10,Huang-Yao'18)
- **Graphical Models** (Gaussian/Ising/Potts Model): loss – likelihood, composite conditional likelihood, regularizer – l_1 and group l_1 (Huang-Yao'18)
- **Fused LASSO/TV**: split Bregman with composite l_2 loss and l_1 gauge (Osher-Burger-Goldfarb-Xu-Yin'06, Burger-Gilboa-Osher-Xu'06, Yin-Osher-Goldfarb-Darbon'08, Huang-Sun-Xiong-Yao'16)
- **Matrix Completion/Regression**: gauge – the matrix nuclear norm (Cai-Candès-Shen'10)

Split LB vs. Generalized LASSO

Structural Sparse Regression:

$$y = X\beta^\star + \epsilon,\ \gamma^\star = D\beta^\star\ \ (S = \mathrm{supp}\,(\gamma^\star)\,,\ s = |S| \ll p)\,, \tag{10}$$

Loss that splits prediction vs. sparsity control

$$\ell\,(\beta,\gamma) := \frac{1}{2n}\,\|y - X\beta\|_2^2 + \frac{1}{2\nu}\,\|\gamma - D\beta\|_2^2 \quad (\nu > 0). \tag{11}$$

Split LBI:

$$\beta_{k+1} = \beta_k - \kappa\alpha\nabla_\beta \ell(\beta_k,\gamma_k), \tag{12a}$$

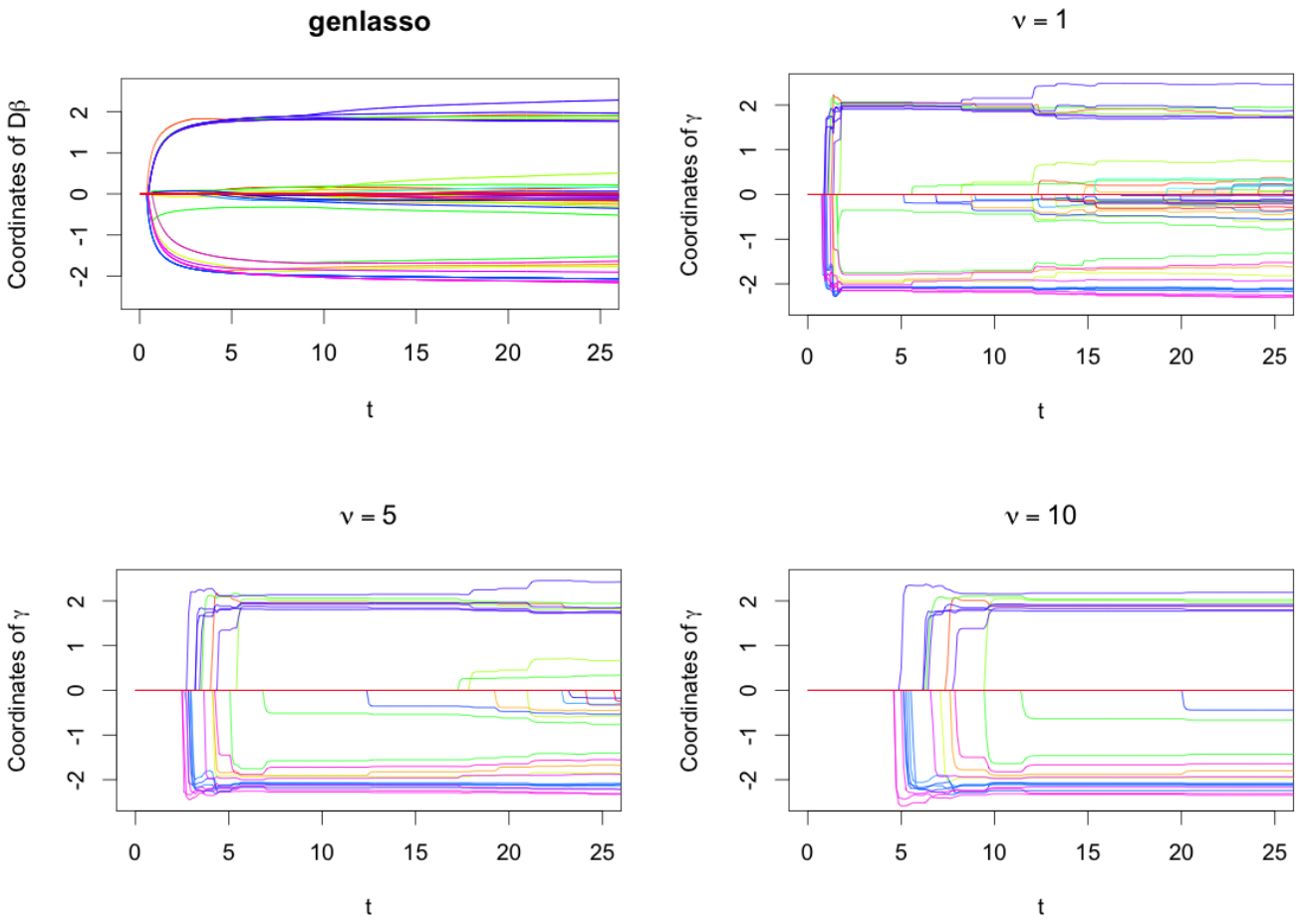
$$z_{k+1} = z_k - \alpha\nabla_\gamma \ell(\beta_k,\gamma_k), \tag{12b}$$

$$\gamma_{k+1} = \kappa\cdot \mathrm{prox}_{\|\cdot\|_1}(z_{k+1}), \tag{12c}$$

Generalized LASSO (`genlasso`):

$$\arg\min_\beta \left(\frac{1}{2n}\,\|y - X\beta\|_2^2 + \lambda\,\|D\beta\|_1\right). \tag{13}$$

Split LBI vs. Generalized LASSO paths



Split LB may beat Generalized LASSO in Model Selection

genlasso	Split LBI			genlasso	Split LBI		
	$\nu = 1$	$\nu = 5$	$\nu = 10$		$\nu = 1$	$\nu = 5$	$\nu = 10$
.9426	.9845	.9969	**.9982**	.9705	.9955	.9996	**.9998**
(.0390)	(.0185)	(.0065)	(**.0043**)	(.0212)	(.0056)	(.0014)	(**.0009**)

- Example: $n = p = 50$, $X \in \mathbb{R}^{n \times p}$ with $X_j \sim N(0, I_p)$, $\epsilon \sim N(0, I_n)$
- (Left) $D = I$ (LASSO vs. Split LB)
- (Right) 1-D fused (generalized) LASSO vs. Split LB (next page).
- In terms of Area Under the ROC Curve (AUC), LB has less false discoveries than `genlasso`
- *Why*? Split LB may need weaker irrepresentable conditions than generalized LASSO...

Structural Sparsity Assumptions

- Define $\Sigma(\nu) := (I - D(\nu X^* X + D^T D)^\dagger D^T)/\nu$.
- **Assumption 1**: Restricted Strong Convexity (RSC).

$$\Sigma_{S,S}(\nu) \succeq \lambda \cdot I. \tag{14}$$

- **Assumption 2**: Irrepresentable Condition (IRR).

$$\mathrm{IRR}(\nu) := \|\Sigma_{S^c,S}(\nu) \cdot \Sigma_{S,S}^{-1}(\nu)\|_\infty \leq 1 - \eta. \tag{15}$$

- $\nu \to 0$: RSC and IRR above reduce to the RSC and IRR neccessary and sufficient for consistency of `genlasso` (Vaiter'13,LeeSunTay'13).
- $\nu \neq 0$: by allowing variable splitting in proximity, IRR above can be weaker than literature, bringing better variable selection consistency than `genlasso` (observed before)!

Identifiable Condition (IC) and Irrepresentable Condition (IRR)

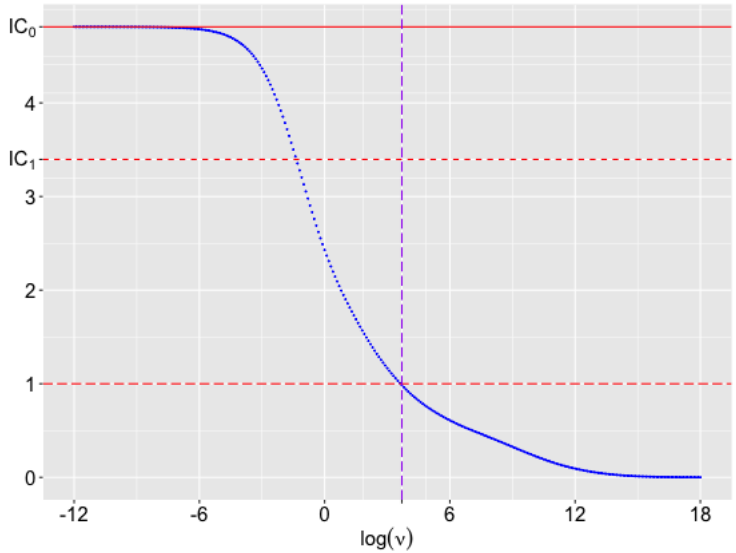
- Let the columns of W form an orthogonal basis of $\ker(D_{S^c})$.

$$\Omega^S := \left(D_{S^c}^{\dagger}\right)^T \left(X^*XW\left(W^TX^*XW\right)^{\dagger}W^T - I\right)D_S^T, \tag{16}$$

$$\mathrm{IC}_0 := \left\|\Omega^S\right\|_\infty, \ \mathrm{IC}_1 := \min_{u\in\ker(D_{S^c})}\left\|\Omega^S\mathrm{sign}\left(D_S\beta^\star\right) - u\right\|_\infty. \tag{17}$$

- The sign consistency of `genlasso` has been proved, under $\mathrm{IC}_1 < 1$ (Vaiter et al. 2013).
- We will show the sign consistency of Split LBI, under $\mathrm{IRR}(\nu) < 1$.
- If $\mathrm{IRR}(\nu) < \mathrm{IC}_1$, then our IRR is easier to be met?

Split LB improves Irrepresentable Condition (Huang-Sun-Xiong-Y.'16)



Theorem (Huang-Sun-Xiong-Y.'2016)

- $\mathrm{IC}_0 \geq \mathrm{IC}_1$.
- $\mathrm{IRR}(\nu) \to \mathrm{IC}_0$ $(\nu \to 0)$.
- $\mathrm{IRR}(\nu) \to C$ $(\nu \to \infty)$. $C = 0 \Longleftrightarrow \ker(X) \subseteq \ker(D_S)$.

Consistency

Theorem (Huang-Sun-Xiong-Y.'2016)

Under RSC and IRR, with large κ and small δ, there exists K such that with high probability, the following properties hold.

- *No-false-positive property: γ_k $(k \le K)$ has no false-positive, i.e.* $\mathrm{supp}(\gamma_k) \subseteq S = \mathrm{supp}(\gamma^\star)$.
- *Sign consistency of γ_k: If $\gamma^\star_{\min} := \min(|\gamma^\star_j| : j \in S)$ (the minimal signal) is not weak, then* $\mathrm{supp}(\gamma_K) = \mathrm{supp}(\gamma^\star)$.
- *ℓ_2 consistency of γ_k:* $\|\gamma_K - \gamma^\star\|_2 \le C_1\sqrt{s \log m/n}$.
- *ℓ_2 "consistency" of β_k:* $\|\beta_K - \beta^\star\|_2 \le C_2\sqrt{s \log m/n} + C_3\nu$.

- Issues due to variable splitting (despite benefit on IRR):
  - $D\beta_K$ does not follow the sparsity pattern of $\gamma^\star = D\beta^\star$.
  - β_K incurs an additional loss $C_3\nu$ ($\nu \sim \sqrt{s \log m/n}$ minimax optimal).

Consistency

Theorem (Huang-Sun-Xiong-Y.'2016)

Define

$$\tilde{\beta}_k := \mathrm{Proj}_{\ker(D_{S_k^c})}(\beta_k) \ \ (S_k = \mathrm{supp}(\gamma_k)) \tag{18}$$

Under RSC and IRR, with large κ and small δ, there exists K such that with high probability, the following properties hold, if $\gamma^\star_{\min}$ is not weak.

- *Sign consistency of $D\tilde{\beta}_K$:* $\mathrm{supp}(D\tilde{\beta}_K) = \mathrm{supp}(D\beta^\star)$.
- *ℓ_2 consistency of $\tilde{\beta}_K$:* $\left\|\tilde{\beta}_K - \beta^\star\right\|_2 \le C_4\sqrt{s\log m/n}$.

Application: Alzheimer's Disease Detection

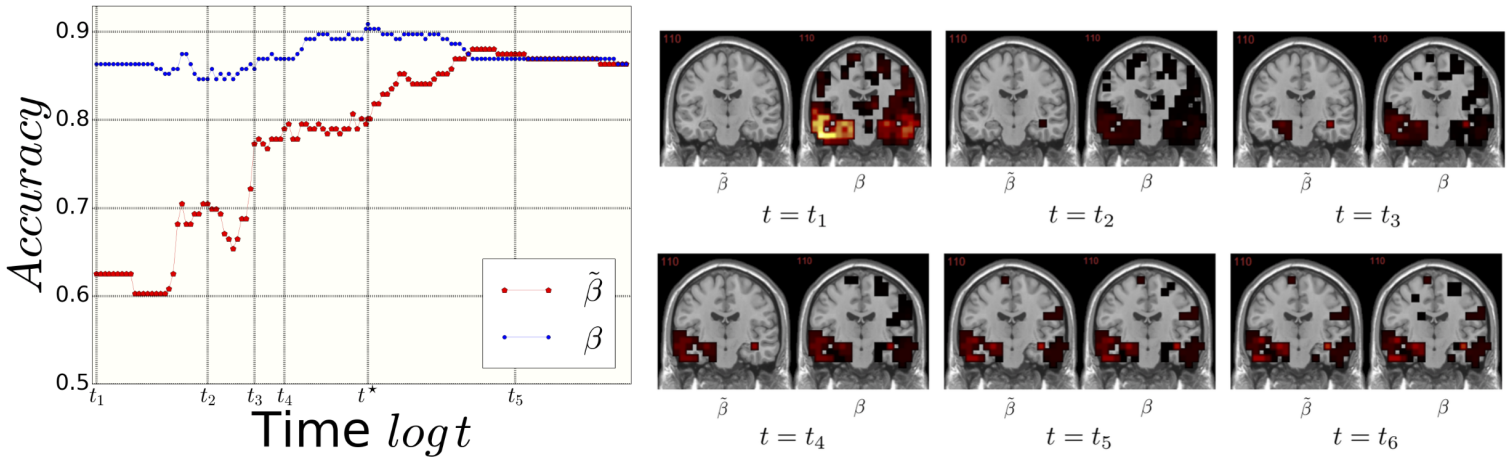


Figure: [Sun-Hu-Y.-Wang'17] A split of prediction (β) vs. interpretability ($\tilde{\beta}$): $\tilde{\beta}$ corresponds to the degenerate voxels interpretable for AD, while β additionally leverages the procedure bias to improve the prediction

Application: Partial Order of Basketball Teams

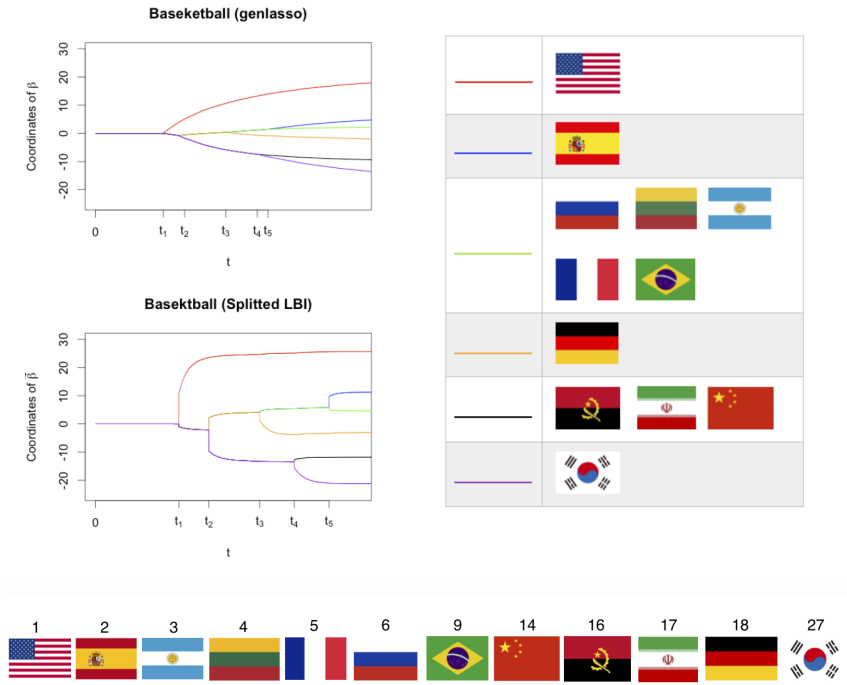


Figure: Partial order ranking for basketball teams. Top left shows $\{\beta_\lambda\}$ $(t = 1/\lambda)$ by `genlasso` and $\tilde{\beta}_k$ $(t = k\alpha)$ by Split LBI. Top right shows the same grouping result just passing t_5. Bottom is the FIBA ranking of all teams.

Summary

We have seen:

- The limit of Linearized Bregman iterations follows a restricted gradient flow: differential inclusions dynamics
- It passes the unbiased *Oracle Estimator* under sign-consistency
- Sign consistency under nearly the same condition as LASSO
  - Restricted Strongly Convex + Irrepresentable Condition
- Split extension: sign consistency under a weaker condition than generalized LASSO
  - under a provably weaker Irrepresentable Condition
- Early stopping regularization is exploited against overfitting under noise

A Renaissance of Boosting as restricted gradient descent ...

Some Reference

- Osher, Ruan, Xiong, Yao, and Yin, "Sparse Recovery via Differential Equations", *Applied and Computational Harmonic Analysis*, 2016
- Xiong, Ruan, and Yao, "A Tutorial on Libra: R package for Linearized Bregman Algorithms in High Dimensional Statistics", *Handbook of Big Data Analytics*, Eds. by Wolfgang Karl Härdle, Henry Horng-Shing Lu, and Xiaotong Shen, Springer, 2017. `https://arxiv.org/abs/1604.05910`
- Xu, Xiong, Cao, and Yao, "False Discovery Rate Control and Statistical Quality Assessment of Annotators in Crowdsourced Ranking", *ICML 2016*, arXiv:1604.05910
- Huang, Sun, Xiong, and Yao, "Split LBI: an iterative regularization path with structural sparsity", *NIPS 2016*, `https://github.com/yuany-pku/split-lbi`
- Sun, Hu, Wang, and Yao, "GSplit LBI: taming the procedure bias in neuroimaging for disease prediction", *MICCAI 2017*
- Huang and Yao, "A Unified Dynamic Approach to Sparse Model Selection", *AISTATS 2018*
- Huang, Sun, Xiong, and Yao, "Boosting with Structural Sparsity: A Differential Inclusion Approach", *Applied and Computational Harmonic Analysis*, 2018, arXiv: 1704.04833
- R package:
  - `http://cran.r-project.org/web/packages/Libra/index.html`