# Inverse Scale Space Method for Sparse Learning and Statistical Properties

Yuan Yao

HKUST

June 26, 2022

# Acknowledgements


- Theory
  - *Stanley Osher*, *Wotao Yin* (Alibaba & UCLA)
  - *Feng Ruan* (Stanford & PKU)
  - *Jiechao Xiong*, *Chendi Huang*, *Xinwei Sun* (PKU)
  - *Yang Cao*, *Jinshan Zeng* (HKUST)
- Applications:
  - *Xinwei Sun* (Fudan/PKU)
  - *Qianqian Xu*, *Jiechao Xiong*, *Chendi Huang* (PKU)
  - *Yanwei Fu* (Fudan University)
  - *Chen Liu*, *Donghao Li* (Fudan and HKUST)
  - *Wenqi Zeng* (HKUST)
  - *Lingjing Hu* (BCMU)

1. Libra (R) and DessiLBI (Python)
   - Libra: Linear/Logistic Regression, Ising graphical models
   - DessiLBI: Deep structurally splitting Linearized Bregman Iteration
2. From LASSO to Inverse Scale Space
   - LASSO and Bias
   - Differential Inclusion of Inverse Scale Space
   - Statistical Path Consistency with Early Stopping
   - Large Scale Algorithm: Linearized Bregman Iteration (LBI)
3. Variable Splitting: Split LBI
   - A Weaker Irrepresentable/Incoherence Condition
   - Applications: Alzheimer's Disease, Deep Learning, and Ranking
   - Data Adaptive Early Stopping Rule: Split Knockoffs
4. Summary

# 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), LBI
- logistic regression (binomial, multinomial): LBI
- graphical models (Gaussian, Ising, Potts): LBI

Two types of regularization:

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

# Libra computes regularization paths via Linearized Bregman Iteration (LBI)

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

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

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

where

- $\ell(x, \theta)$ is the *loss* function to minimize
- $\mathrm{prox}_{\|\cdot\|_*}(z) := \arg\min_u \left(\frac{1}{2}\|u - z\|^2 + \|u\|_*\right)$
- $\alpha_k > 0$ 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), yet different limit dynamics

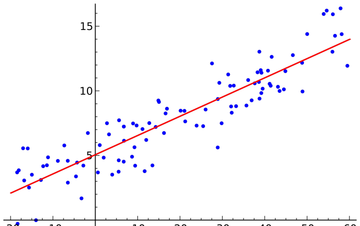
# Linear Regression

Linear Regression:

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

$\beta$ is sparse or group sparse, with two types of penalty:

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

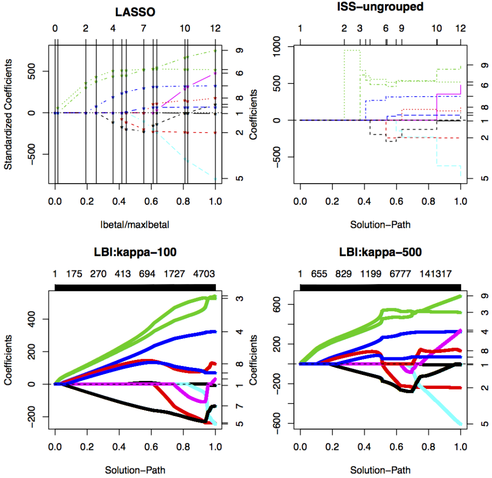


# Linear Regression Example: Diabetes Data

```
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)
```

# LBI generates iterative regularization paths



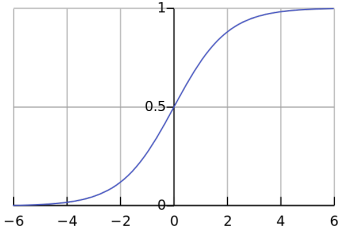
# Logistic Regression

Logistic Regression:

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

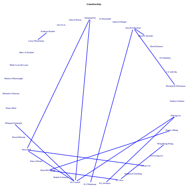
$\beta$ is sparse or group sparse, with two types of penalty:

- "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

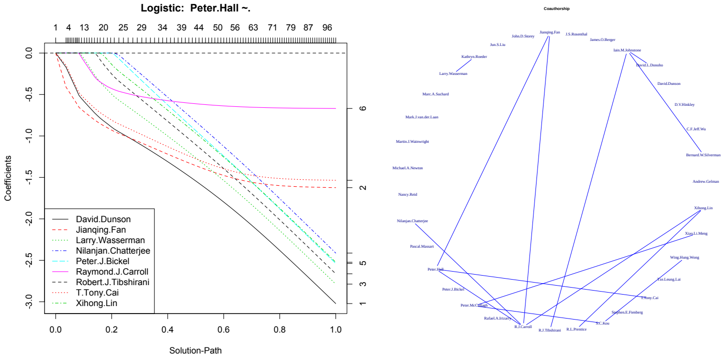



Figure: Peter Hall vs. other COPSS award winners in sparse logistic regression [papers from AoS/JASA/Biometrika/JRSSB, 2003-2012]: true coauthors are merely Tony Cai, R.J. Carroll, and J. Fan

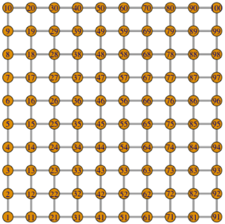
# Sparse Ising Model

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

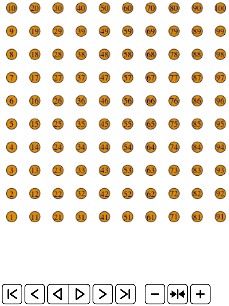
$$P(x_1, \ldots, x_p) \sim \exp\left(\sum_i H_i x_i + \sum_{i,j} J_{ij} x_i x_j\right)$$

- 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

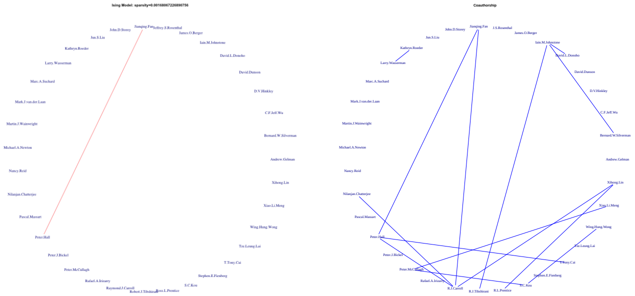



Figure: Left: LB path of Ising Model learning; Right: coauthorship network of existing data. Typically COPSS winners do not like working together; Peter Hall (1951-2016) is the hub of statisticians, like Erdös for mathematicians

# DessiLBI: Sparse Filters Learned on MNIST

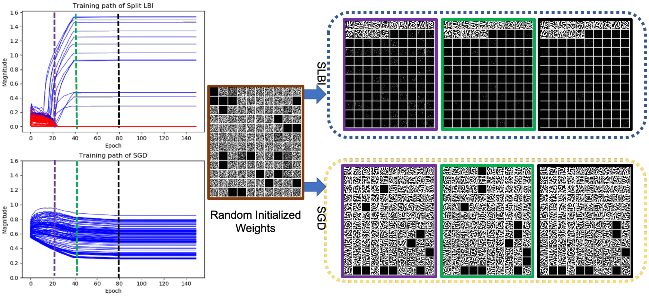



Figure: [see Yanwei Fu's talk] Visualization of solution path and filter patterns in the third convolutional layer (i.e., conv.c5) of LetNet-5, trained on MNIST, showing a sparse selection of filters without sacrificing accuracy. From Fu et al. DessiLBI, ICML 2020, https://github.com/DessiLBI2020/DessiLBI.

# DessiLBI: Non-semantic Features Learned on ImageNet

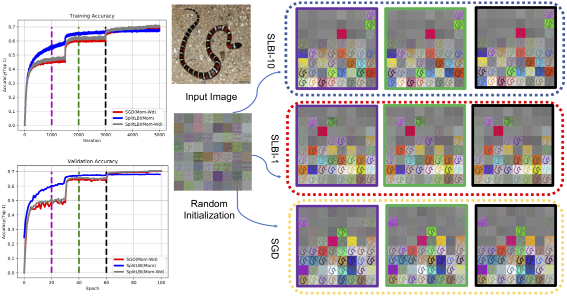



Figure: [see Yanwei Fu's talk] Visualization of the first convolutional layer filters of ResNet-18 trained on ImageNet-2012, where texture features are more important than colour/shapes. Given the input image and initial weights visualized in the middle, filter response gradients at 20 (purple), 40 (green), and 60 (black) epochs are visualized. SGD with Momentum (Mom) and Weight Decay (WD), is compared with SLBI.

# How does it work?

In the sequel, we shall see a story on statistical model selection consistency with early stopping:

- The simple iterative algorithm shadows a particular kind of dynamics: differential inclusions of inverse scale spaces, as special cases of Mirror Descent, where important features are learned fast
- Simple discretized algorithm, amenable for parallel implementation
- Under nearly the same condition as LASSO, it reaches model selection consistency with early stopping
- but may incur less bias than LASSO
- Equipped with variable splitting, it weakens the conditions of generalized LASSO in feature selection

# Sparse Linear Regression

Assume that $\beta^* \in \mathbb{R}^p$ is sparse and unknown. Consider recovering $\beta^*$ 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 $p \gg n \geq s$.
- Generalized Structural/Transformational Sparsity: $\gamma^* = D\beta^*$ is sparse, where $D$ is a linear transform (wavelet, gradient, etc.), $S = \mathrm{supp}(\gamma^*)$
- *How to recover $\beta^*$ (or $\gamma^*$) 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)

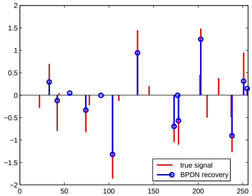
# The Bias of LASSO

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, courtesy of Wotao Yin)

# 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 $\tau_n$:

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

LASSO estimate is biased away from the oracle estimator

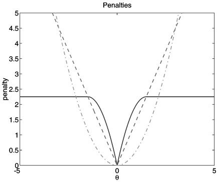
$$(\hat{\beta}_{\tau_n})_S = \tilde{\beta}^*_S - \frac{1}{\tau_n} \Sigma_{n,S}^{-1} \mathrm{sign}(\beta^*_S), \quad \tau_n > 0.$$

*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 $\tau_n$,

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

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

# 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 $\tau_n$,

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

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

- Equivalently, the blue part removes bias of LASSO automatically

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

# Differential Inclusion: Inverse Scaled Spaces (ISS)

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

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

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

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

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

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

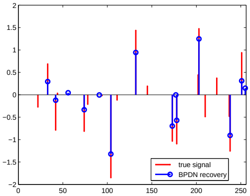
- Burger-Gilboa-Osher-Xu'06 (in image recovery it recovers the objects in an inverse-scale order as $t$ increases (larger objects appear in $\beta_t$ first))

# Examples

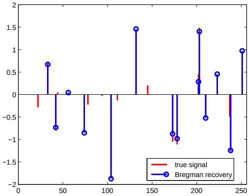
- e.g. $X = Id$, $n = p = 1$, hard-thresholding

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

- the same example shown before (figures by courtesy of Wotao Yin)




True vs LASSO




True vs ISS

## Solution Path: Sequential Restricted Maximum Likelihood Estimate

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

- $\beta_t$ 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

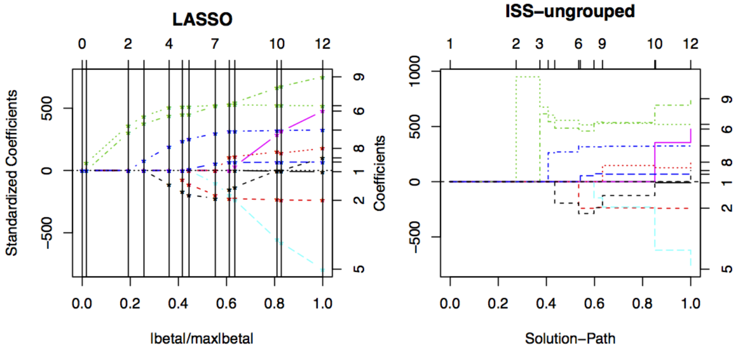



Figure: Diabetes data (Efron et al.'04) and regularization paths are different, yet bearing similarities on the order of parameters being nonzero

# Why? 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!

# 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, Zou'06, Wainwright'09, etc.

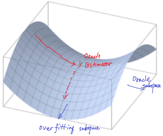
# Understanding the Dynamics

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 $S$ (Oracle Subspace) 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 overfitting 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: $z_k := \rho_k + \frac{1}{\kappa}\beta_k \Leftrightarrow \beta_k = \kappa \cdot Shrink(z_k, 1)$

# Comparison with ISTA

**Linearized Bregman (LB)** iteration:

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

which is not **ISTA**:

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

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

# LBI 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)

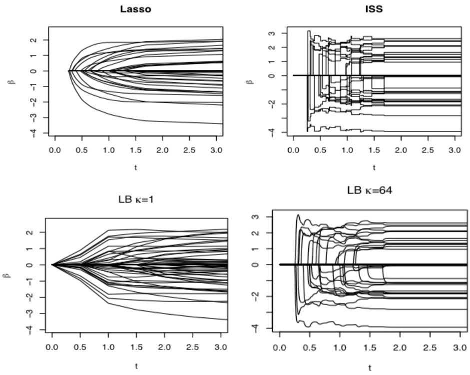
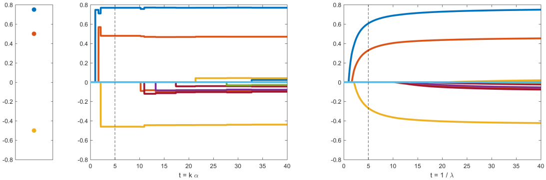



Figure: As $\kappa \to \infty$, LB paths have a limit as piecewise-constant ISS path

# 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 $\kappa$ is large enough and $\alpha$ 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

# More reference on generalizations

- **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)

# Structural or Transformational Sparsity

Structural/Transformational Sparse Regression:

$$y = X\beta^\star + \epsilon, \tag{9a}$$

$$\gamma^\star = D\beta^\star, \tag{9b}$$

where

$$S = \mathrm{supp}\left(\gamma^\star\right), \quad s := |S| \ll p.$$

# Split LBI vs. Generalized LASSO

- Generalized LASSO (genlasso):

$$\arg\min_{\beta} \left( \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|D\beta\|_1 \right). \tag{10}$$

- Split LBI: Loss that splits prediction vs. sparsity control

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

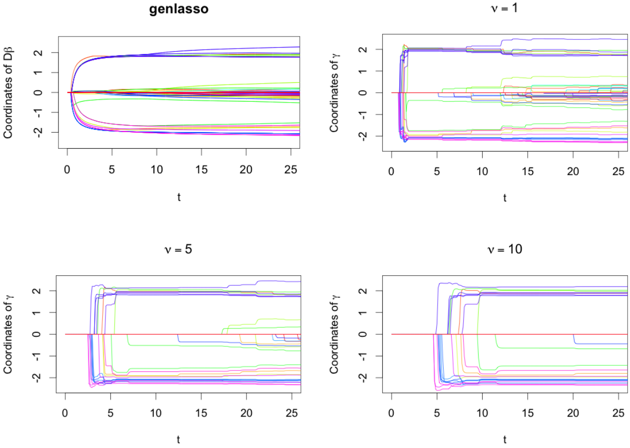
Algorithm [Huang-Sun-Xiong-Y. 2016] :

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

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

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

# Split LBI vs. Generalized LASSO paths



# Split LBI 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<br>(.0390) | .9845<br>(.0185) | .9969<br>(.0065) | **.9982**<br>(**.0043**) | .9705<br>(.0212) | .9955<br>(.0056) | .9996<br>(.0014) | **.9998**<br>(**.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 LBI)
- (Right) 1-D fused (generalized) LASSO vs. Split LBI
- In terms of Area Under the ROC Curve (AUC), Split LBI has less false discoveries than genlasso
- *Why*? Split LBI 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).

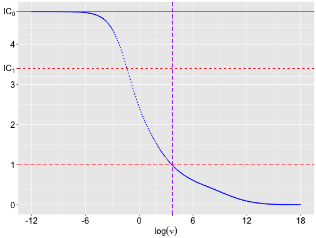
$$\Sigma_{S,S}(\nu) \succeq \lambda \cdot I. \tag{13}$$

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

$$\mathrm{IRR}(\nu) := \|\Sigma_{S^c,S}(\nu) \cdot \Sigma_{S,S}^{-1}(\nu)\|_{\infty} \leq 1 - \eta. \tag{14}$$

- $\nu \to 0$: RSC and IRR above reduce to the 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)!

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




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

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

# Remark: Identifiable Conditions (IC)

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

$$\Omega^S := \left(D_{S^c}^{\dagger}\right)^T \left(X^* X W \left(W^T X^* X W\right)^{\dagger} W^T - I\right) D_S^T, \tag{15}$$

$$\mathrm{IC}_0 := \left\|\Omega^S\right\|_{\infty}, \ \mathrm{IC}_1 := \min_{u \in \ker(D_{S^c})} \left\|\Omega^S \mathrm{sign}\left(D_S \beta^{\star}\right) - u\right\|_{\infty}. \tag{16}$$

- The sign consistency of `genlasso` has been proved, under $\mathrm{IC}_1 < 1$ [Vaiter et al. 2013].
- The sign consistency of Split LBI is proved under $\mathrm{IRR}(\nu) < 1$ [Huang-Sun-Xiong-Y.'2016].
- As $\mathrm{IRR}(\nu) < \mathrm{IC}_1$ when $\nu$ grows, our IRR is easier to meet.

# Consistency

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

*Under RSC and IRR, with large $\kappa$ and small $\delta$, there exists $K$ such that with high probability, the following properties hold.*

- *No-false-positive property: $\gamma_k$ $(k \leq K)$ has no false-positive, i.e. $\mathrm{supp}(\gamma_k) \subseteq S = \mathrm{supp}(\gamma^\star)$.*
- *Sign consistency of $\gamma_k$: If $\gamma^\star_{\min} := \min(|\gamma^\star_j| : j \in S)$ (the minimal signal) is not weak, then $\mathrm{supp}(\gamma_K) = \mathrm{supp}(\gamma^\star)$.*
- *$\ell_2$ consistency of $\gamma_k$: $\|\gamma_K - \gamma^\star\|_2 \leq C_1\sqrt{s\log m/n}$.*
- *$\ell_2$ "consistency" of $\beta_k$: $\|\beta_K - \beta^\star\|_2 \leq C_2\sqrt{s\log m/n} + C_3\nu$.*

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

# Consistency

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

*Define*

$$\tilde{\beta}_k := \mathrm{Proj}_{\ker(D_{S_k^c})}(\beta_k) \ \ (S_k = \mathrm{supp}(\gamma_k)) \tag{17}$$

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

- *Sign consistency of $D\tilde{\beta}_K$:* $\mathrm{supp}(D\tilde{\beta}_K) = \mathrm{supp}(D\beta^\star)$.
- *$\ell_2$ consistency of $\tilde{\beta}_K$:* $\left\|\tilde{\beta}_K - \beta^\star\right\|_2 \leq C_4\sqrt{s\log m/n}$.

# Application: Partial Order of Basketball Teams

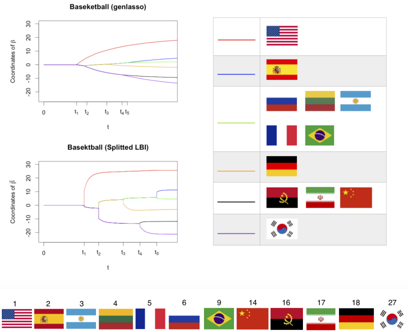



Figure: Partial order ranking for basketball teams. Top left shows $\{\beta_\lambda\}$ ($t = 1/\lambda$) by genlasso and $\tilde{\beta}_k$ ($t = k\alpha$) by Split LBI. Top right shows the same grouping result just passing $t_5$. Bottom is the FIBA ranking of all teams.

# Application: Sparse Neural Nets in Early Stopping (Lottery Tickets)

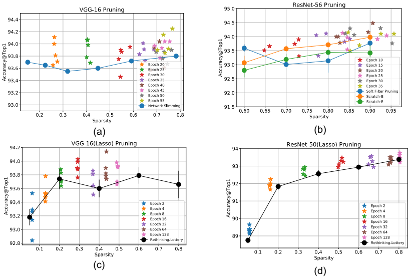



Figure: [see Yanwei Fu's talk] SplitLBI with early stopping finds sparse subnets whose test accuracies (stars) after retrain are comparable or even better than the baselines (Network Slimming, Soft-Filter Pruning, Scratch-B, Scratch-E, and "Rethinking-Lottery" as reported in Rethink the Value of Pruning. Sparse filters of VGG-16 and ResNet-56 are show in (a) and (b), while sparse weights of VGG-16 and ResNet-50 are shown in (c) and (d).

# Application: Alzheimer's Disease Detection

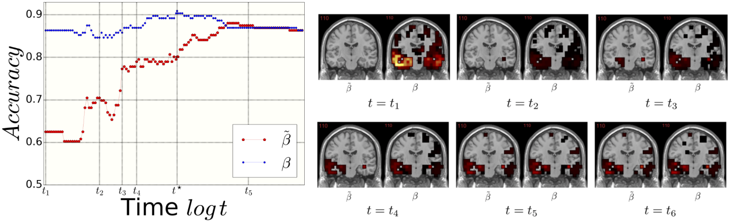



Figure: [Sun-Hu-Y.-Wang'17] A split of prediction ($\beta$) vs. interpretability ($\tilde{\beta}$): $\tilde{\beta}$ corresponds to the degenerate voxels interpretable for AD, while $\beta$ additionally leverages the procedure bias to improve the prediction (c.f. Xinwei Sun talk).

# Controlling the False Discovery Rates via Knockoffs

- The early stopping rule $\bar{\tau}$ in theory (for power) is unknown in applications;
- Knockoff (Barber-Candès (2015)) gives a data adaptive early stopping rule with FDR control:

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\#\text{false discoveris}}{1 \vee \#\text{discoveries}}\right].$$

- The method makes a fake Knockoff feature $\tilde{X}$ as the control group:

$$\tilde{X}^T \tilde{X} = X^T X, \; X^T \tilde{X} = X^T X - \mathrm{diag}(s),$$

where $s$ is some proper non-negative vector. The Knockoff features mimic the original feature $X$, but decoupled with the original feature.

# Knockoff Methods

Knockoff is extended to:

1. Group sparse and multi-task regression model (Dai-Barber, ICML 2016).
2. Huber's robust regression with LBI (Xu et al. ICML 2016).
3. High dimensional setting (Barber-Candes 2019).
4. Model-X Knockoff for random design (Candes et al. 2016).
5. Deep Knockoff for nonparametric random designs (Romano et al. 2019).
6. Split Knockoffs for structural/transformational sparsity (Cao-Sun-Y. 2022).

# Summary

- The limit of Linearized Bregman iteration follows differential inclusion of inverse scale space, where significant features emerge earlier on solution paths
- It renders 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 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*, 48(1):1-45, 2020, arXiv: 1704.04833
- Qianqian Xu, Xinwei Sun, Zhiyong Yang, Qingming Huang, and Yuan Yao. "iSplit LBI: Individualized Partial Ranking with Ties via Split LBI". *NeurIPS 2019*, arXiv:1910.05905
- Yanwei Fu, Chen Liu, Donghao Li, Xinwei Sun, Jinshan Zeng, and Yuan Yao. DessiLBI: Exploring Structural Sparsity of Deep Neural Networks via Differential Inclusion Paths. *ICML 2020*, arXiv:2007.02010.
- Yanwei Fu, Chen Liu, Donghao Li, Zuyuan Zhong, Xinwei Sun, Jinshan Zeng, and Yuan Yao. Exploring Structural Sparsity of Deep Networks via Inverse Scale Spaces. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2022.
- Yang Cao, Xinwei Sun, Yuan Yao. "Controlling the False Discovery Rate in Structural Sparsity: Split Knockoffs". arXiv:2103.16159.
- R package: `http://cran.r-project.org/web/packages/Libra/index.html`
- Pytorch package for deep learning: `https://github.com/DessiLBI2020/DessiLBI`