Localization of Heterogeneous Disease Features in Neuroimage: A Differential Inclusion Approach

Yuan Yao

HKUST

March 23, 2019

Acknowledgements

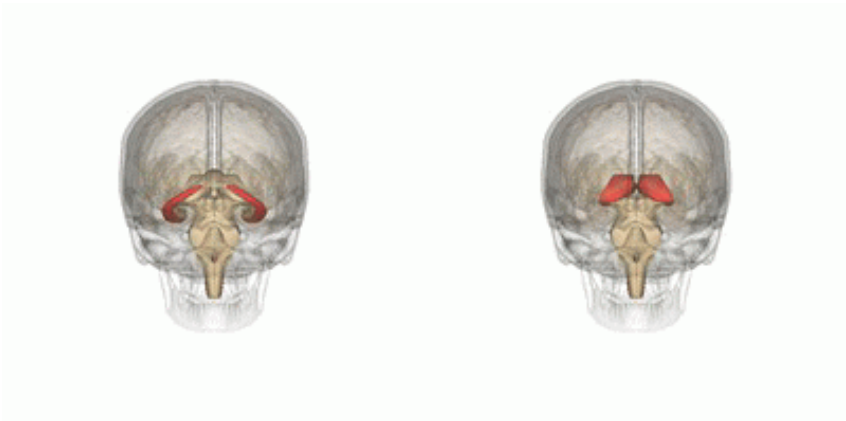
- Applications:
  - *Xinwei Sun*, *Yizhou Wang*, *Chendi Huang* (PKU)
  - *Lingjing Hu* (Capital University of Medical Sciences)
  - *Qianqian Xu* (CAS)
- Theory
  - *Chendi Huang*, *Xinwei Sun*, *Jiechao Xiong* (PKU & Tecent AI Lab)
- Discussions:
  - *Stan Osher*, *Wotao Yin* (UCLA), *Feng Ruan* (Stanford & PKU)
- Grants:
  - National Basic Research Program of China (973 Program), NSFC

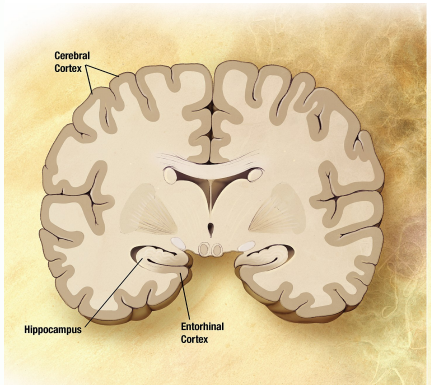
What is Alzheimer's Disease (AD)

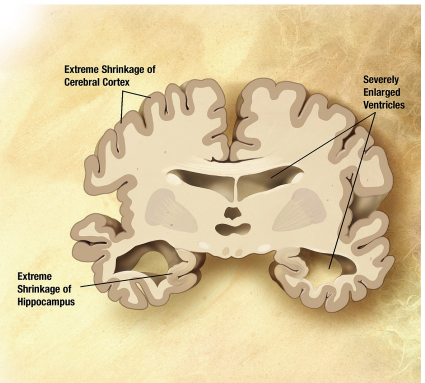
- A complex chronically progressive neurodegenerative disease and the most common form of dementia in elderly people worldwide.
- In 2015, about 29.8 million AD people worldwide and dementia resulted in 1.9 million deaths.
- In Hong Kong alone, dementia was estimated to affect approximately 103,433 people over the age of 60 years in 2009 and is expected to increase 222% by 2039.
- The patient with AD is transitioned from Normal Control (NC), to Mild Cognitive Impairment (MCI) then to AD.
- There are no clear rules to define status of the diseases.
- However, the brain structure changes a lot during the process before functional changes detected.

Biomarkers from sMRI

- Symptoms can be alleviated or inhibited by drugs, it's hard to say whether it has structurally improvement.
- Biomarkers processed by structural Magnetic Resonance Imaging (sMRI) (PET etc.) as predictors used to classify NC, MCI, and AD, e.g. atrophy of Hippocampus and Thalamus.







Heterogeneity of Voxel-based Features

The preprocessed features on structural Magnetic Resonance Imaging (sMRI) images contain the following voxel-wise features:

- Lesion features that are contributed to the disease
- Procedural bias introduced during the preprocessing steps and shown to be helpful in classification
- Irrelevant or null features that are not due to disease status

Heterogeneity of Voxel-based Features

The preprocessed features on structural Magnetic Resonance Imaging (sMRI) images contain the following voxel-wise features:

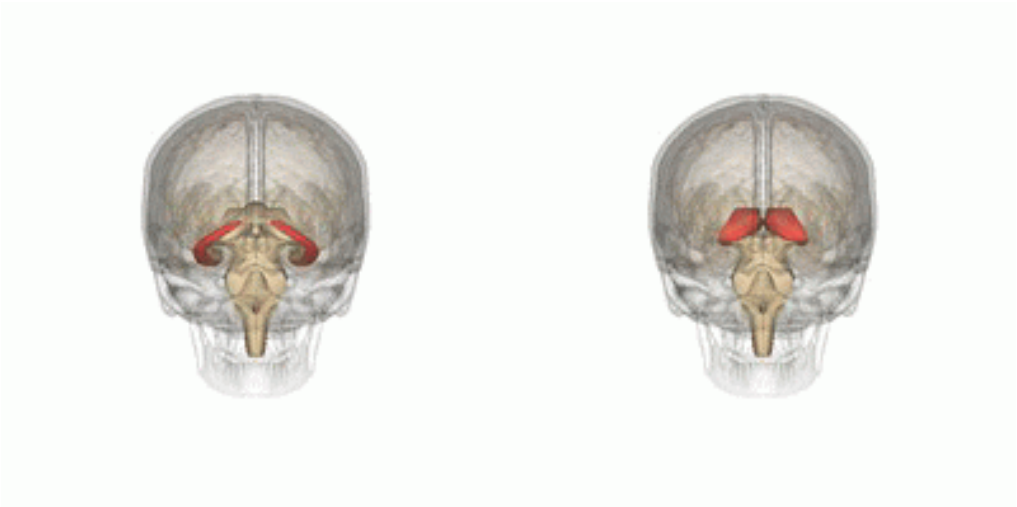
- Lesion features that are contributed to the disease
- Procedural bias introduced during the preprocessing steps and shown to be helpful in classification
- Irrelevant or null features that are not due to disease status

Our two goals in voxel-based neuroimage analysis for disease prediction:

- Accurate Classification: NC, MCI, AD
- Stable feature selection of lesion features and procedural bias, with high recall and low false discovery rate (FDR).

Lesion Features

- Have been the main focus in disease prediction.
- Only a few number of gray matter voxels are correlated with the disease.
- Geometrically clustered in atrophied regions in dementia disease such as Alzheimer's Disease (AD).



Procedural Bias

- Introduced during the preprocessing steps, are found to be helpful for disease prediction.
- Refer to the mistakenly enlarged Gray Matter (GM) voxels surrounding locations with cerebral spinal fluid (CSF) spaces enlarged, e.g. lateral ventricle in AD. [Sun et al. MICCAI 2017]

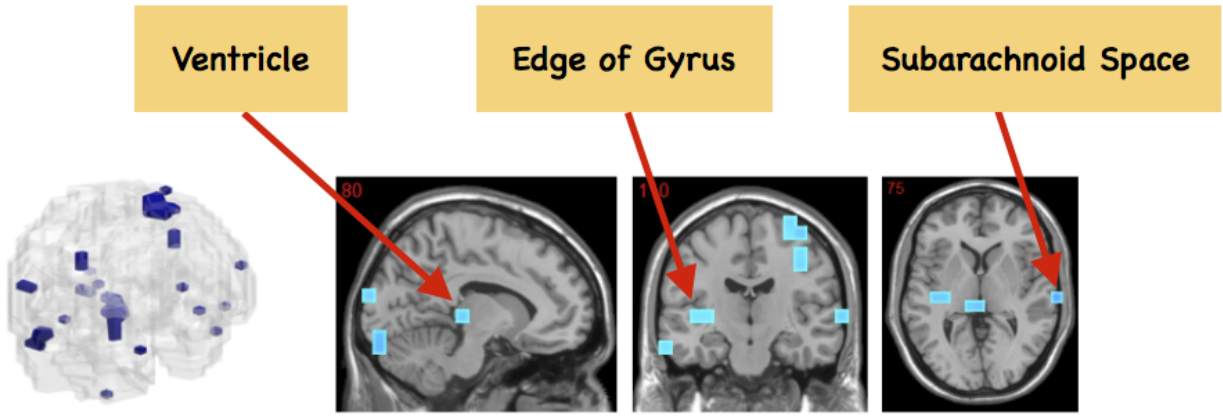
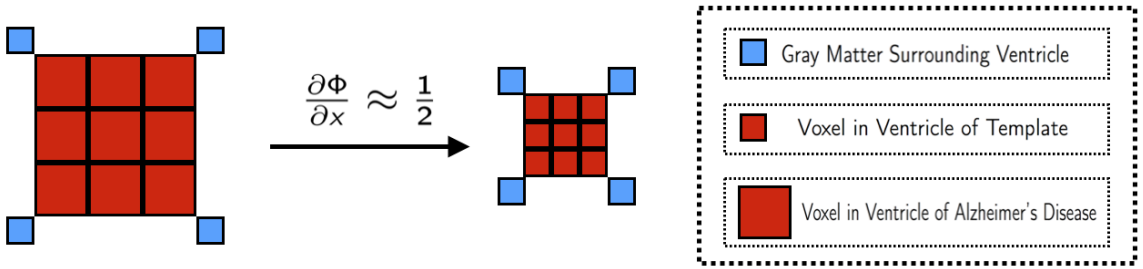


Fig. 1. The overlapped voxels among top 150 negative value voxels in each fold of β_{pre} at the time corresponding to the best average prediction result in the path of GSplit LBI using 10-fold cross-validation. For subjects with AD, they represent enlarged GM voxels surrounding lateral ventricle, subarachnoid space, edge of gyrus, etc.

Illustration of why "procedural bias" introduced

1. Spatial Normalization



2. Modulation
   - Corrects for changes in volume resulting from the normalization step.
   - Multiplying intensity of each voxel ($\mathrm{I}(\cdot)$) with determinant of inverse Jacobean matrix $\left(\frac{\partial \Phi}{\partial x}\right)$

$$\mathrm{I}_{\mathrm{after}}(\blacksquare) = \left|\left(\frac{\partial \Phi}{\partial x}\right)^{-1}\right| \mathrm{I}_{\mathrm{before}}(\blacksquare)$$

$$\mathrm{I}_{\mathrm{after}}(\blacksquare) = \left|\left(\frac{\partial \Phi}{\partial x}\right)^{-1}\right| \mathrm{I}_{\mathrm{before}}(\blacksquare)$$

3. Ref: Ashburner-Friston (2001) Neuroimage 14(6): 1238–1243.

Limitation of Existing Models

- The lesion features have been the only focus of existing models for their medical interpretability.
- In VBM analysis, procedure biases are introduced on some features during the commonly used preprocessing procedure of T_1 weighted image (e.g. DARTEL [Ashburner 2007, Neuroimage])
- Procedure bias can be helpful for classification, however they are ignored in the literature.

ADNI Dataset

- Consider AD/NC classification (namely ADNC) and MCI (Mild Cognitive Impairment)/NC (namely MCINC) classification
- The data are obtained from ADNI[1] database, which is split into 1.5T and 3.0T (namely 15 and 30) MRI scan magnetic field strength datasets.
- DARTEL VBM pipeline [Ashburner (2007) Neuroimage] is implemented to preprocess the data.
- The input features consist of 2,527 $8\times8\times8$ mm^3 size voxels with average values in GM population template greater than 0.1.
- Experiments are designed on 15ADNC, 30ADNC and 15MCINC tasks.

[1] http://adni.loni.ucla.edu

The efficacy of Procedural bias in prediction

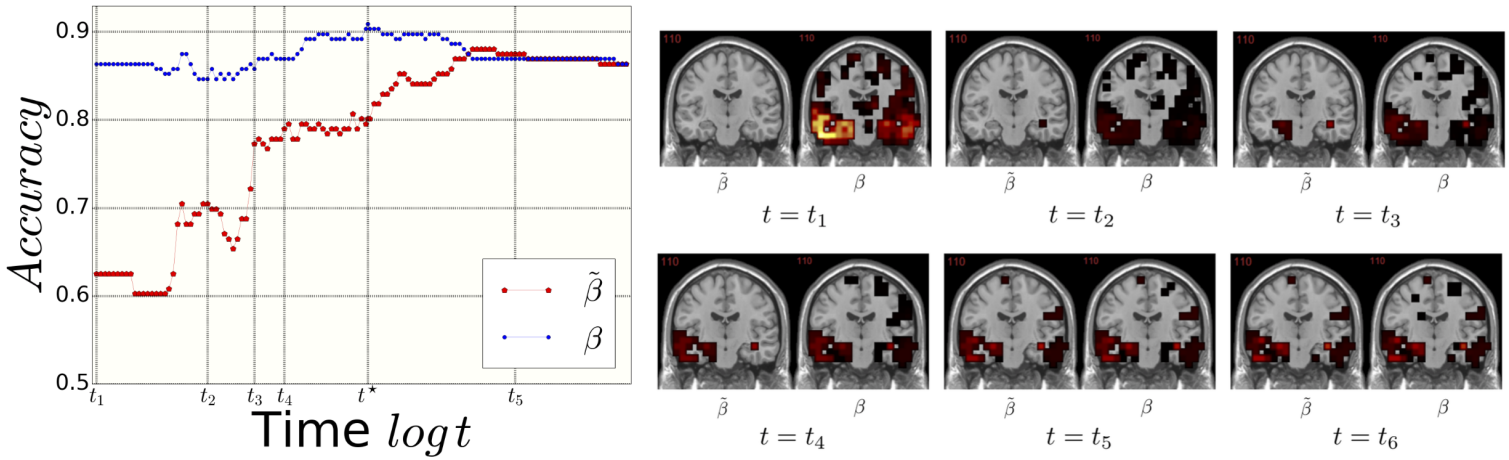
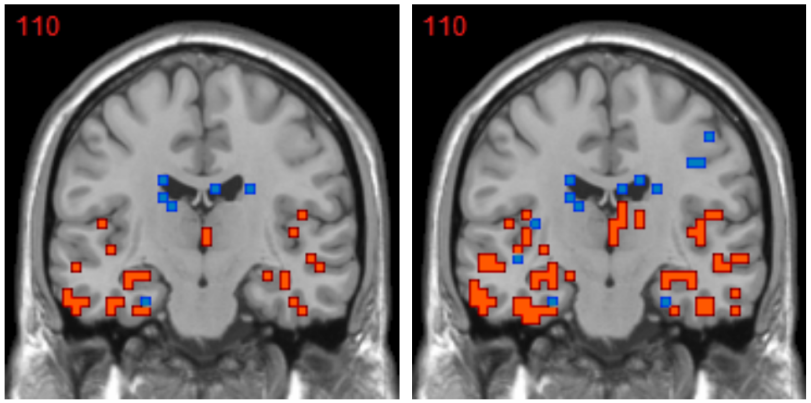


Figure: Exploitation of both lesion features and procedure bias (blue) improves prediction accuracy by dominating the curve merely by lesion features (red). $\tilde{\beta}$ corresponds to the lesion features interpretable for AD, while β additionally leverages the procedure bias to improve the prediction.

Lesion Features vs. Procedure Bias



Local FDR T-test($p < 0.001$)

Figure: Leision features (red) as degenerate GM voxels, while Procedure Bias (blue) contains 'enlarging' GM voxels in preprocessing due to the enlarging cerebral spinal fluid (CSF), surrounding lateral ventricle and subarachnoid space etc.

Results of Prediction

Exploiting both the procedure bias and lesion features, our method (GSplitLBI [Sun-Hu-Y.-Wang 2017]) achieves the state-of-the-art performance in the 10-fold cross-validation error:

Table 1. Comparison of GSplit LBI with other models

	MLDA	SVM	Lasso	Graphnet	Elastic Net	TV + l_1	n^2GFL	GSplit LBI (β_{pre})
15ADNC	85.06%	83.12%	87.01%	86.36%	88.31%	83.77%	86.36%	**88.96%**
30ADNC	86.93%	87.50%	87.50%	88.64%	89.20%	87.50%	87.50%	**90.91%**
15MCINC	61.41%	70.13%	69.80%	72.15%	70.13%	73.83%	69.80%	**75.17%**

Stability of Feature Selection

Our method (GSplitLBI [Sun-Hu-Y.-Wang 2017]) also champions in stability of heterogeneous feature selection:

Table 2. mDC comparison between GSplit LBI and other models

	Lasso	Elastic Net	Graphnet	TV + l_1	n^2 GFL	GSplit LBI (β_{les})
Accuracy	87.50%	89.20%	88.64%	87.50%	87.50%	88.64%
mDC	0.1992	0.5631	0.6005	0.5824	0.5362	**0.7805**
$\sum_{k=1}^{10} \lvert S(k)\rvert/10$	50.2	777.8	832.6	712.6	443.9	129.4

Where

$$mDC := \frac{10|\cap_{k=1}^{10} S(k)|}{\sum_{k=1}^{10} |S(k)|}$$

with $S(k)$ denoting the support set of β_{les} in k-th fold.

How do we reach this?

Below we are going to introduce a new methodology:

- Boosting with Structural Sparsity
- mathematically, a differential inclusion method as restricted gradient flows, whose discretization meets the sparse mirror descent algorithm
- variable splitting enables us to exploit both lesion features and procedure bias effectively
- it can be generalized to FDR heterogeneous smoothing for new algorithms in Empirical Bayes

Structural Sparse Regression

Consider recovering $\beta^* \in \mathbb{R}^p$ from n linear measurements

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

where ϵ_i is i.i.d. sub-Gaussian noise.

- Structural Sparsity:

  $$\gamma^* = D\beta^*$$

  is unknown yet sparse, $S = \mathrm{supp}(\gamma^*)$ with capacity $s = |S| \ll \min(n, p)$.
  - sparse linear regression: $D = I$; TV: D is graph gradient; Wavelet: $D = W^T$, etc.
- *How to recover β^* and γ^*'s sparsity pattern (sparsistency) and estimate values with variations (consistency)?*

Generalized LASSO

Generalized LASSO (Tibshirani-Taylor'11):

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

- a.k.a. TV minimization (ROF'92), fused LASSO (Tibshirani et al. 05)
- $D = I$ reduces to LASSO (Tibshirani'96)
- for **fixed** λ, use ADMM or Split Bregman to find an optimizer (Goldstein-Osher'09, Ye-Xie'11, Zhu'15)
- full regularization path: $\hat{\beta}(\lambda) : \mathbb{R}_+ \to \mathbb{R}^p$?
  - Efron-Hastie-Johnstone-Tibshirani'04 (`lars`) for LASSO
  - Tibshirani-Taylor'11, Arnold-Tibshirani'16 (`genlasso`): expensive

Best Possible: The Oracle Estimator

Consider the simple case of Lasso with $D = I$. 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)

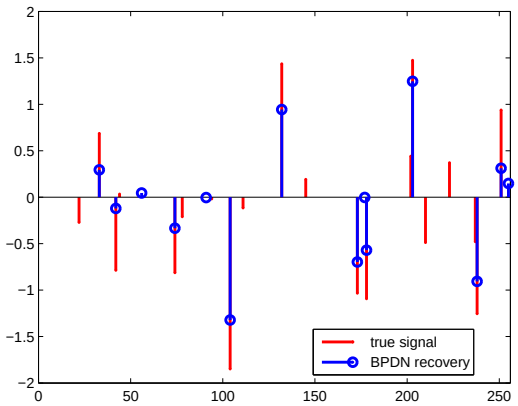
The Bias of LASSO

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

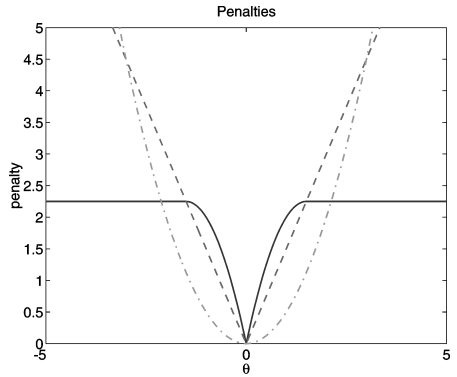
$$(\hat{\beta}_{\tau_n})_S = \tilde{\beta}^*_S - \frac{1}{\tau_n} \Sigma^{-1}_{n,S} \operatorname{sign}(\beta^*_S), \quad \tau_n > 0.$$

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 $\mathrm{d}\rho/\mathrm{d}t$

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

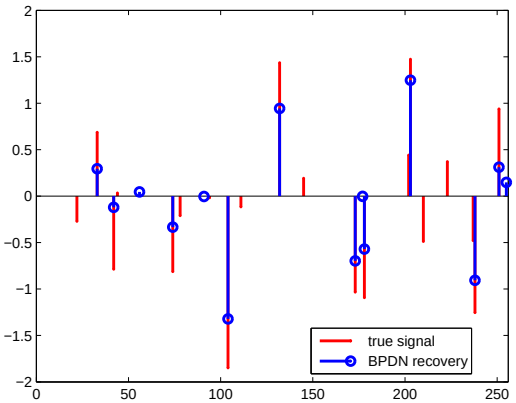
- Burger-Gilboa-Osher-Xu'06 (in image recovery it recovers the objects in an inverse-scale order as t increases (larger objects appear in β_t first))

Examples

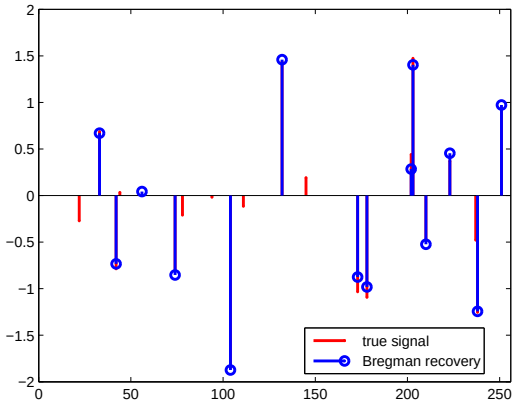
- e.g. $X = Id$, $n = p = 1$, hard-thresholding

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

- 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

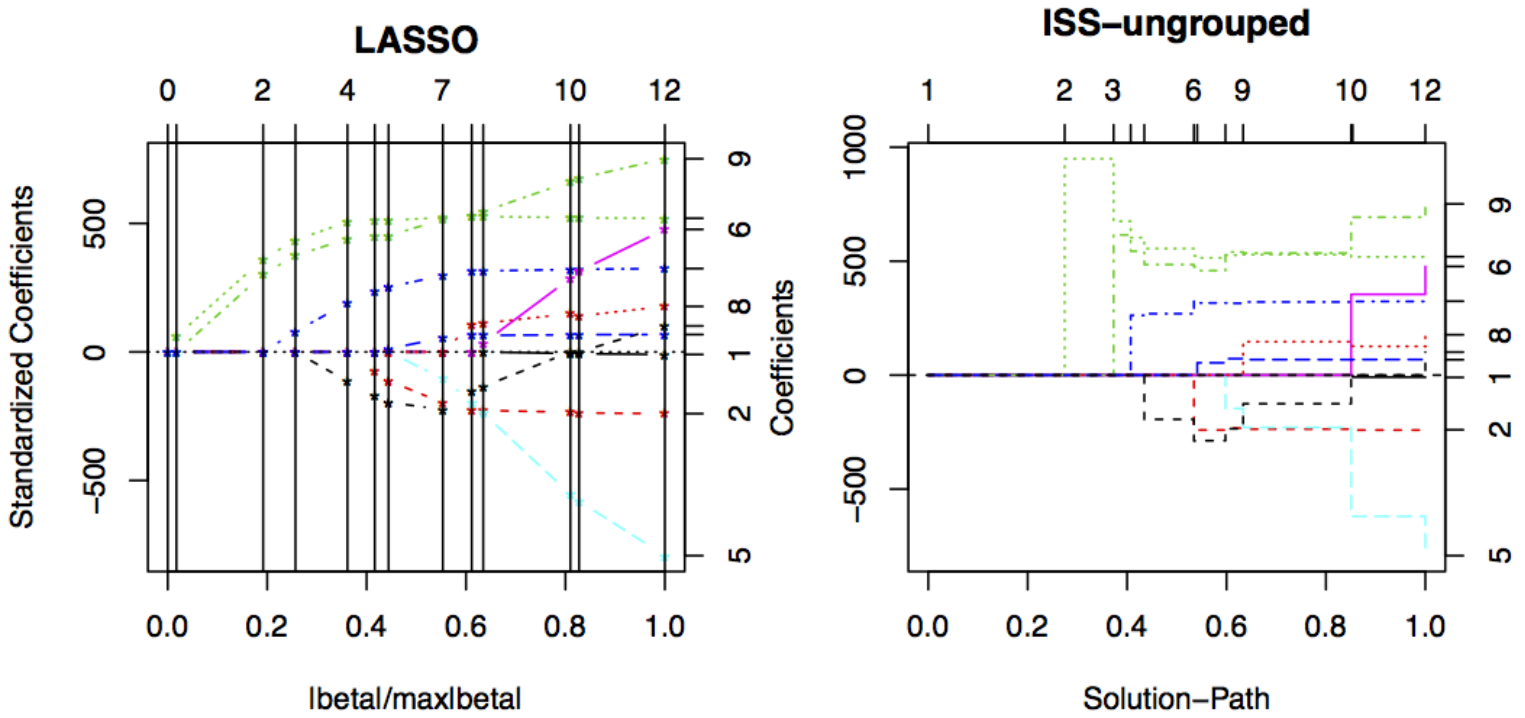


Figure: Diabetes data (Efron et al.'04) and regularization paths are different, yet bearing similarities on the order of parameters being nonzero

How does it work?

A Path Consistency Theory is given in [Osher-Ruan-Xiong-Y.-Yin 2016], 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!

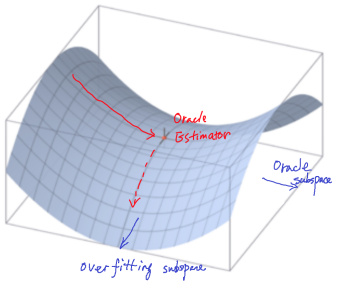
Understanding the Dynamics

ISS as restricted gradient descent:

$$\dot{\rho}_t = -\nabla \ell(\beta_t), \quad \rho_t \in \partial \|\beta_t\|_1$$

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



Large scale algorithm: Linearized Bregman Iteration

Damped Dynamics: continuous solution path

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

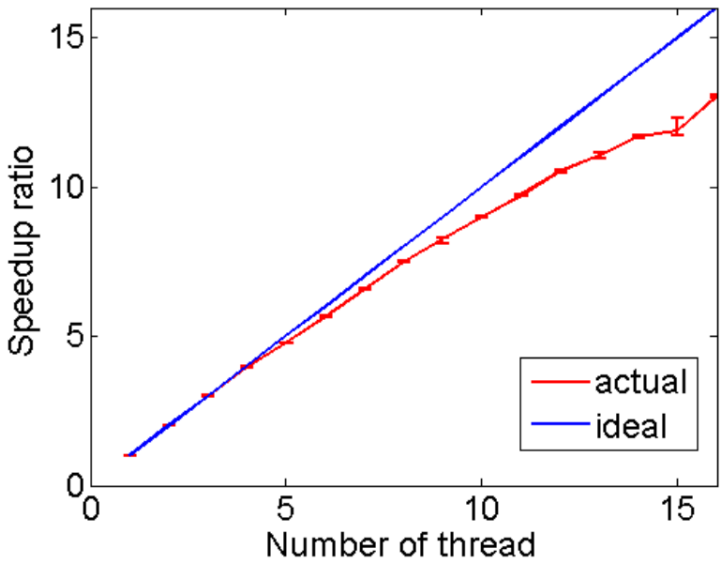
Linearized Bregman Iteration as forward Euler discretization proposed even earlier than ISS dynamics (Osher-Burger-Goldfarb-Xu-Yin'05, Yin-Osher-Goldfarb-Darbon'08): for $\rho_k \in \partial\|\beta_k\|_1$,

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

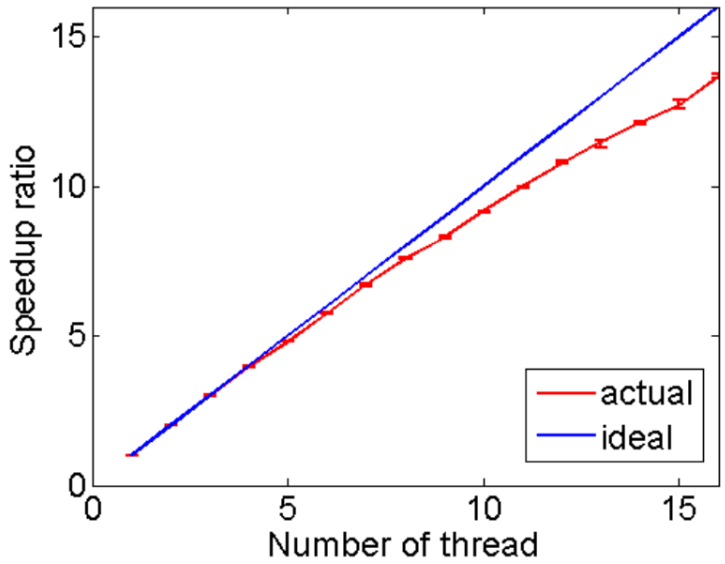
where

- 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.

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 $\alpha_k \kappa_i \|\nabla_\theta^2 \hat{\mathbb{E}} \ell(x, \theta)\| < 2$
- as simple as ISTA, easy to parallel implementation

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

Generalized LASSO

Generalized LASSO (Tibshirani-Taylor'11):

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

- a.k.a. TV minimization (ROF'92), fused LASSO (Tibshirani et al. 05)
- $D = I$ reduces to LASSO (Tibshirani'96)
- full regularization path: $\hat{\beta}(\lambda) : \mathbb{R}_+ \to \mathbb{R}^p$?
  - Efron-Hastie-Johnstone-Tibshirani'04 (`lars`) for LASSO
  - Tibshirani-Taylor'11, Arnold-Tibshirani'16 (`genlasso`): expensive

Variable Splitting LBI

Loss that splits prediction vs. sparsity control

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

Split LBI in 3 lines:

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

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

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

- Eq. (12a) is ℓ_2-Boost (Buhlman-Yu'02) or Landweber Iteration (Yao-Rosasco-Caponnetto'07)
- others w.r.t. γ_k are Linearized Bregman Iteration (Yin et al.'08), or Mirror Descent with Elastic Net (Nemirovski-Yudin'83)

How does it work?

- **Sign-Consistency**: there is a necessary and sufficient condition for the existence of a point on `genlasso` path s.t.

$$\mathrm{sign}(D\hat{\beta}_\lambda) = \mathrm{sign}(D\beta^\star), \quad \exists \lambda \in \mathbb{R}_+$$

  - [Vaiter et al.'13; Lee et al.'13]
- A new family of conditions are presented for **Split LBI**
  - weaker than that for `genlasso`
  - [Huang-Sun-Xiong-Y.'16]
- Technique:
  - differential inclusions: limits of Split LBI (restricted gradient flow)
  - Early stopping regularization against overfitting noise

A Renaissance of Boosting as restricted gradient descent ...

A Model of AD

In the following we take Alzheimer's Disease (AD) as an example to illustrate our method.

- Our dataset consists of N samples $\{x_i, y_i\}_1^N$.
- $x_i \in \mathbb{R}^p$ collects the i^{th} neuroimaging data with p voxels and $y_i = \{\pm 1\}$ indicates the disease status (-1 denotes AD).
- $X \in \mathbb{R}^{N \times p}$ and $y \in \mathbb{R}^p$ are concatenations of $\{x_i\}_i$ and $\{y_i\}_i$.

The most common classification method is logistic regression (LR), i.e.

$$\min_{\beta, \beta_0} \ell(\beta, \beta_0) := \sum_{i=1}^{N} \log \left(1 + e^{-y_i (x_i^T \beta + \beta_0)}\right) \tag{13}$$

Prior Knowledge of Lesion features in Alzheimer's Disease

- The number of voxels involved in the disease prediction is small, suggesting sparse structure of β.
- Geometrically clustered or 3D-smooth, suggesting a TV-type sparsity on $D_G\beta$ where D_G is a graph difference operator. [2]
- Degenerate, suggesting nonnegative constraint in β.

[2] Here $D_G : \mathbb{R}^V \to \mathbb{R}^E$ denotes a graph difference operator on $G = (V, E)$, where V is the node set of voxels, E is the edge set of voxel pairs in neighbour (e.g. 3-by-3-by-3), such that $D_G(\beta)(i, j) := \beta(i) - \beta(j)$.

Choice of Penalty or Regularization Function

To sum up, it suggests a penalty or regularization function

$$\psi(\beta) = \|D\beta\|_1 + \frac{1}{2\kappa}\|D\beta\|_2^2 + \mathcal{X}_{\text{non-neg}}(\beta) \tag{14}$$

where

- $D = \begin{bmatrix} I_{p\times p} & D_G^T \end{bmatrix}^T$
- $\mathcal{X}_{\text{non-neg}}(\beta) = \begin{cases} 0 & \text{if } \beta \geq 0 \\ +\infty & \text{others} \end{cases}$
- $\frac{1}{2\kappa}\|D\beta\|_2^2$ is to make sure the strong convexity of ψ where κ can be large

Solving $\min_{\beta,\beta_0} \ell(\beta, \beta_0) + \lambda\psi(\beta)$?

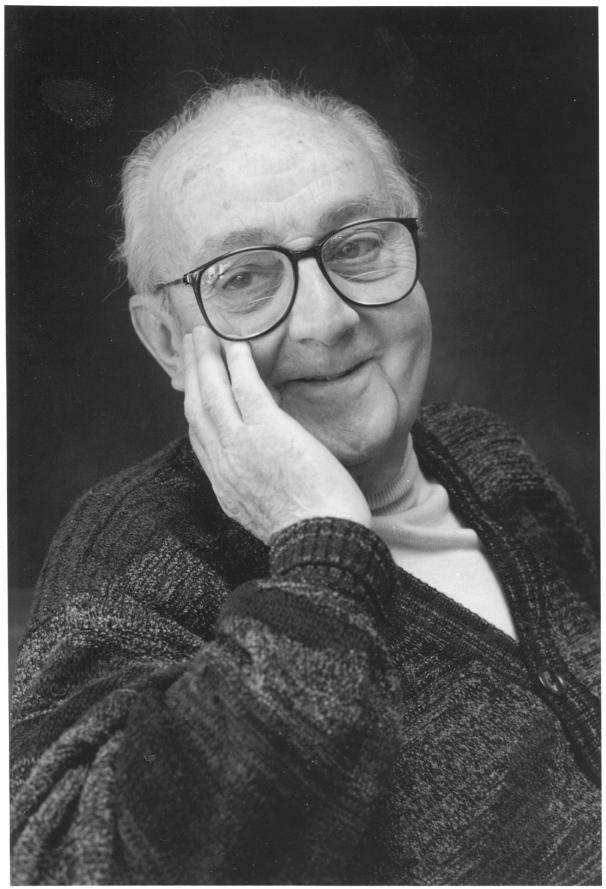


Figure: George Box: "Essentially, all models are wrong, but some are useful."

Issues of such a model

It can pursue the sparsity structure of lesion features, however, note that

- It ignores the procedural bias, which are mistakenly enlarged gray matter voxels here.
- It's hard to solve the path for the existence of D in $\partial\psi(\beta)$.

Variable Splitting

We adopt variable splitting scheme, i.e.

$$S_{\rho,\nu}(\beta,\gamma) = \frac{1}{2\nu}\|D_\rho\beta - \gamma\|_2^2$$

where

- $D_\rho = \begin{bmatrix} I_{p\times p} & \rho D_G \end{bmatrix}^T$ with $\rho > 0$
- $\gamma = \begin{bmatrix} \gamma_V^T & \gamma_G^T \end{bmatrix}^T$. $\gamma_V \in \mathbb{R}^p$ corresponds to β; γ_G corresponds to $\rho D_G\beta$
- We want γ to be sparse and control the distance of it from $D\beta$. β is a dense estimator and hence has the capability to capture other features that are correlated with the label.
- $\nu > 0$ controls the distance between $D\beta$ and γ

Generalized Split ISS

The new loss function is

$$\ell(\beta, \gamma) = \ell(\beta) + S_{\rho,\nu}(\beta, \gamma)$$

and new penalty

$$\psi(\gamma) = \|\gamma\|_1 + \frac{1}{2\kappa}\|\gamma\|_2^2 + \mathcal{X}_{\text{non-pos}}(\gamma_1).$$

We aim to solve the differential inclusion:

$$\frac{d(\beta^t, \rho^t)}{dt} = -\nabla_{(\beta,\gamma)}\ell(\beta^t, \gamma^t)$$
$$\text{s.t.} \quad \rho^t \in \partial\psi(\gamma^t)$$

Generalized Split LBI

Algorithm

Initialize: $k = 0$, $t^k = 0$, $\beta_0^k = 0$, $\beta^k = 0$, $\beta^k = 0$, $\gamma_V^k = 0_p$, $\gamma_G^k = 0_m$, $z_V^k = 0_p$, $z_G^k = 0_m$ and $S_k := \mathrm{supp}(\gamma^k) = \emptyset$

For $k = 1, 2, ...$

- $\beta_0^{k+1} = \beta_0^k - \kappa\alpha\nabla_{\beta_0}\ell(\beta_0^k, \beta^k, \gamma^k)$
- $\beta^{k+1} = \beta^k - \kappa\alpha\nabla_{\beta}\ell(\beta_0^k, \beta^k, \gamma^k)$
- $z^{k+1} = z^k - \alpha\nabla_{\gamma}\ell(\beta_0^k, \beta^k, \gamma^k)$
- $\gamma_V^{k+1} = \kappa \cdot \mathcal{S}^+(z_V^{k+1}, 1)$, where $\mathcal{S}^+(x, 1) = \max(x - 1, 0)$
- $\gamma_G^{k+1} = \kappa \cdot \mathcal{S}(z_G^{k+1}, 1)$, where $\mathcal{S}(x, 1) = \mathrm{sign}(x) \cdot \max(|x| - 1, 0)$
- $\tilde{\beta}^{k+1} = P_{S_{k+1}}\beta^{k+1}$, where $P_S = P_{ker(D_{S^c})} = I - D_{S^c}^{\dagger} D_{S^c}$

End

where $\gamma^{k+1} = \begin{bmatrix} \gamma_V^{k+1} \\ \gamma_G^{k+1} \end{bmatrix}$ and $z^{k+1} = \begin{bmatrix} z_V^{k+1} \\ z_G^{k+1} \end{bmatrix}$

Limitation of Such Models

1. For Multivariate models, they suffer from multicollinearity problem under high dimension data.
2. For Univariate models such as two-sample T-test, BHq, LocalFDR, they assume independence of features and hence can not capture spatial correlation among features.
3. The lesion features and procedural bias are heterogenous in terms of the level of spatial coherence and volumetric change:
   - Lesion features are with higher level of spatial coherence than procedural bias
   - Lesion features are degenerate voxels; while procedural bias refer to mistakenly enlarged voxels.

From Multivariate Method to Empirical Bayes

Assuming for each voxel $i \in \{1, ..., p\}$, the statistic z_i is sampled from the following mixture:

$$z_i \sim \sum_{k=0}^{1} \mathbb{P}(s_i = k)\mathbb{P}(z_i|s_i = k) = c_i f_1(z_i) + (1 - c_i)f_0(z_i), \tag{16}$$

where

- $z_i = \Phi^{-1}\left(F_{N-2}(t_i)\right)$ with t_i computed by two-sample t-test
- $z_i \sim f_0(z)$ when $s_i = 0$; $z_i \sim f_1(z)$ when $s_i = 1$.
- $f_0(\cdot)$ is density function of nulls and $f_1(\cdot)$ is that of non-nulls
- $c_i = \mathbb{P}(s_i = 1) = \mathrm{sigmoid}(\beta_i) = e^{\beta_i} / \left(1 + e^{\beta_i}\right)$

Posterior Likelihood of Empirical Bayes

One can thus define the loss function as the negative log-likelihood of z_i:

$$\ell(\beta) = -\sum_{i=1}^{N} \log\left(\frac{e^{\beta_i}}{1+e^{\beta_i}} f_1(z_i) + \frac{1}{1+e^{\beta_i}} f_0(z_i)\right) \tag{17}$$

Different from logistic regression function,

- $f_{t=0,1}(z_i)$ is not binary and hence (17) is not convex.
- The sigmoid in normal logistic function is $\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$ with design matrix X. Here $X = I$, which means it proceeds voxel-by-voxel. Hence, it does not have multicollinearity problem.
- Spatial smoothness can be imposed on β: $\gamma = D\beta$.

EM algorithm via Split LB (FDR LBI)

Loss function

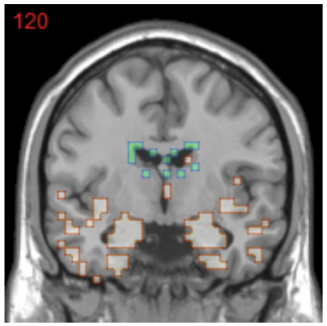
$$\ell(\beta, \gamma, s) = \sum_{i=1}^{p} \left(\log \left(1 + e^{\beta_i} \right) - s_i \beta_i \right) + \frac{1}{2\nu} \|D\beta - \gamma\|_2^2$$

Iteration:

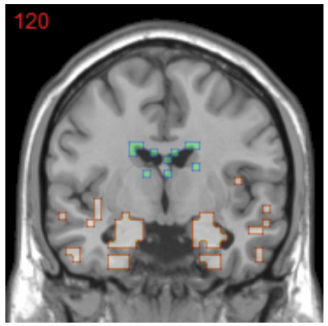
- **E-step:** $s_i^{k+1} = E(s_i | \beta_i^k, z) = \frac{\exp(\beta_i^k) f_1(z_i)}{\exp(\beta_i^k) f_1(z_i) + f_0(z_i)}$
- **M-step:**
  1. $\beta_i^{k+1} = \beta_i^{k+1} - \kappa\alpha \nabla_\beta \ell(\beta^k, \gamma^k | s^{k+1})$
  2. $z_i^{k+1} = z_i^{k+1} - \alpha \nabla_\gamma \ell(\beta^k, \gamma^k | s^{k+1})$
  3. $\gamma_i^{k+1} = \kappa \cdot \mathrm{prox}_{\|\cdot\|_1}(z_i^{k+1})$

30ADNC

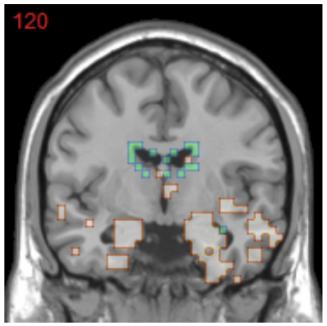
FDR LBI. path: 20
accuracy: 90.91
#selected features: 1328



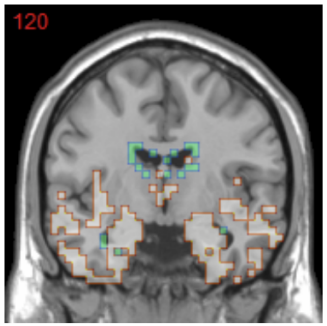
FDR LBI. path: 200
accuracy: 92.05
#selected features: 757



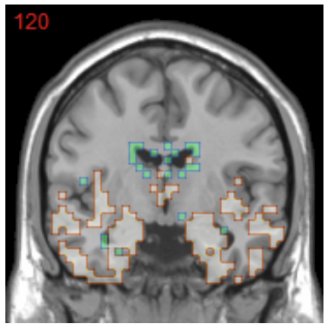
FDR LBI. path: 240
accuracy: 88.07
#selected features: 1372



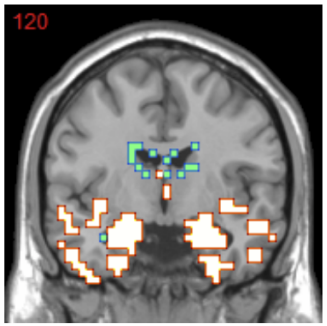
FDR LBI. path: 360
accuracy: 89.20
#selected features: 2827



FDR LBI. path: 500
accuracy: 86.93
#selected features: 3211



FDR TV. accuracy:91.48
#selected features: 1196



Some Reference

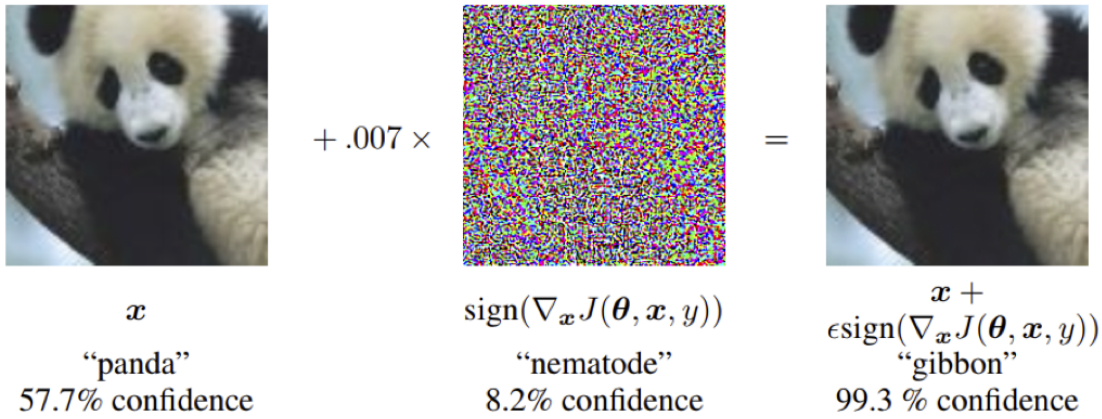
- Osher, Ruan, Xiong, Yao, and Yin, "Sparse Recovery via Differential Equations", *Applied and Computational Harmonic Analysis*, 2016
- Xiong, Ruan, and Yao, "A Tutorial on Libra: R package for Linearized Bregman Algorithms in High Dimensional Statistics", *Handbook of Big Data Analytics*, Eds. by Wolfgang Karl Härdle, Henry Horng-Shing Lu, and Xiaotong Shen, Springer, 2017
- 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
- Sun, Hu, Yao, and Wang, "GSplit LBI: taming the procedure bias in neuroimaging for disease prediction", MICCAI 2017
- Huang, Yao, "A Unified Dynamic Approach to Sparse Model Selection", AISTATS 2018
- Sun, Hu, Zhang, Yao, and Wang, "FDR-HS: An Empirical Bayesian Identification of Heterogenous Features in Neuroimage Analysis", MICCAI 2018
- Huang, Sun, Xiong, and Yao, "Boosting with Structural Sparsity: A Differential Inclusion Approach", *Applied and Computational Harmonic Analysis*, 2018
- R package:
  - `http://cran.r-project.org/web/packages/Libra/index.html`
- Matlab package:
  - `https://github.com/yuany-pku/split-lbi`
- Python package to appear

Summary: A Renaissance of Boosting?

- Boosting as gradient descent is arguably the **best off-the-shelf machine learning algorithm** (Leo Breiman): **AdaBoost** (Freund-Schapire), **L2Boost** (Buhlman-Yu), **LogitBoost** (Friedman)
- The (sparsity) restricted gradient descent dynamics: **differential inclusions**
  - has a simple discretized algorithm to follow the regularization path (LBI), amiable for parallel realizations in **big data analytics**
  - improves model selection or prediction, **better than generalized LASSO**
  - is widely adapted to **various sparsity** constraints
- This is a gift of Applied Mathematics to High Dimensional Statistics

Towards a Deeper Understanding of Deep Learning I

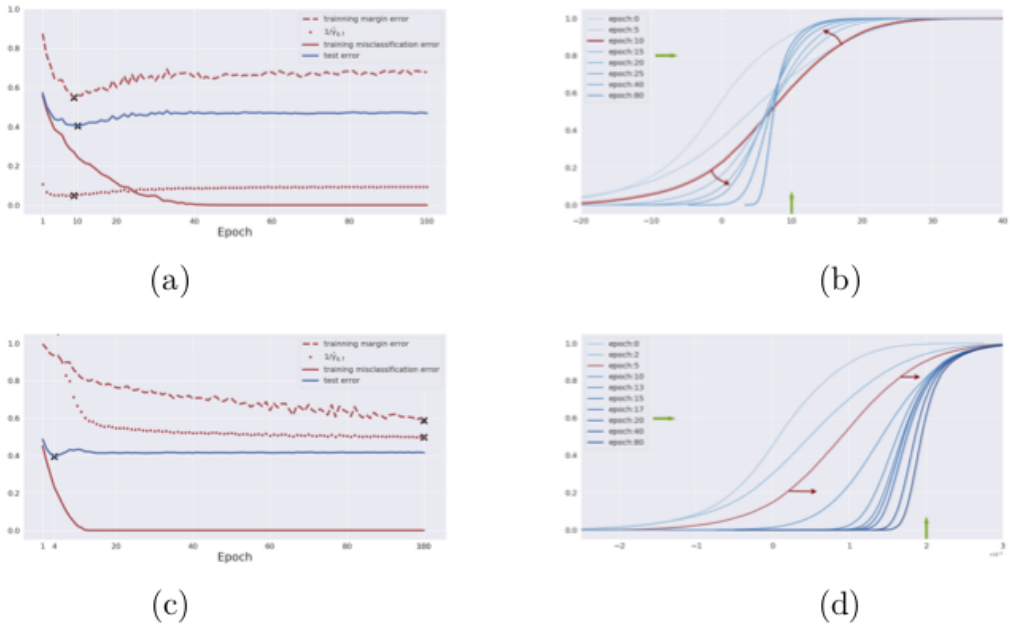
- How can one achieve robustness against adversarial?
  - Gao-Liu-Y.-Zhu'18 (arXiv:1810.02030): under Huber's ϵ-contamination model in robust statistics, Generative Adversarial Networks (GANs) provably achieve robust estimates with statistical optimality



[Goodfellow et al., 2014]

Towards a Deeper Understanding of Deep Learning II

- How do over-parameterized models generalize well?
  - Margin theory (Bartlett et al.1997, 1998, 2017)?
  - Breiman's Dilemma (1999): margin improvement leads to overfitting
  - Zhu-Huang-Y.'18 (arXiv:1810.03389): ubiquitous in neural networks, *phase transitions of margin dynamics*,



The END


thank you
danke
謝謝
gracias
merci
obrigado
dziękuję
bedankt
go raibh maith agat
teşekkür ederim
спасибо
ngiyabonga
dank je
mochchakkeram
sukriya
kop khun krap
terima kasih
grazie
arigatō
dakujem
tapadh leat
mauruuru
hvala
sagolun
takk
merci
рахмат
Баярлалаа
mersi
barka
welalin
tack
spas
mahalo
vinaka
faafetai lava
kiitos
dankie
nandri
nanni
dhanyavad
köszönöm
bayarlalaa
gràcie
enkosi
akun
tankon
aciū
misaotra
matondo
paldies
grazzi
djiere dieuf
tau
chnorakaloutioun
gratias ago
gràcies
sulpáy
дякую
mamnun
chokrane
murakoze
tenki
obrigada
хвала
asante
manana
sobodi
dēkuji
mesi
didi madloba
najis tuke
kam sah hamnida
rahmat
তোমাকে ধন্যবাদ
ありがとう
tanemirt
rahmet
diolch
dhanyavadagalu
shukriya
mercè
мерси
trugarez
xiexie
감사합니다
ευχαριστώ