Structural Sparsity in Medical Imaging

Xinwei Sun

School of Data Science, Fudan University

June 7, 2022

Outline

1. Structural Sparsity in Medical Imaging
2. Recovering Structural Sparsity via Differential Inclusion
3. False-Discovery-Rate Control
   - Split Knockoff
   - FDR Smoothing on Heterogeneous Features

Outline

1. Structural Sparsity in Medical Imaging
2. Recovering Structural Sparsity via Differential Inclusion
3. False-Discovery-Rate Control
   - Split Knockoff
   - FDR Smoothing on Heterogeneous Features

Selection of Disease-related Features

- Disease Prediction vs Feature Selection.
- Among imaging features $X_1, \ldots, X_p$ (the sample size n may be less than p), *which features are disease related*?
- **Example:** degenerated gray matter (GM) in Alzheimer's Disease.

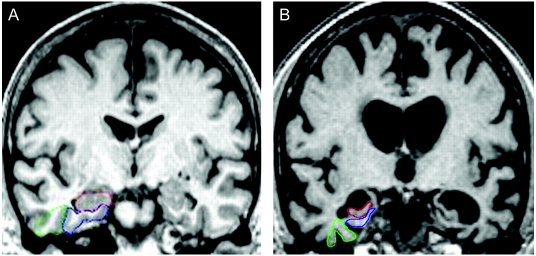


Figure: Selecting Atrophied Gray Matter Voxels in Alzheimer's Disease

Structural Sparsity

Structural Sparsity for diseased gray matter voxels:

1. sparsity 2. spatial coherence 3. positivity (with Y).

Structural Sparsity

Structural Sparsity for diseased gray matter voxels:

1. sparsity 2. spatial coherence 3. positivity (with Y).
   - Lasso [Liu'2012]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1$.

Structural Sparsity

Structural Sparsity for diseased gray matter voxels:

1 sparsity 2 spatial coherence 3 positivity (with Y).

- Lasso [Liu'2012]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1$.
- Elastic Net [Xiao'2021]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$.

Structural Sparsity

Structural Sparsity for diseased gray matter voxels:

1. sparsity 2. spatial coherence 3. positivity (with Y).

- Lasso [Liu'2012]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1$.
- Elastic Net [Xiao'2021]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$.
- Graph Net [Grosenic'2013]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|D\beta\|_2^2$.

Structural Sparsity

Structural Sparsity for diseased gray matter voxels:

1. sparsity 2. spatial coherence 3. positivity (with Y).

- Lasso [Liu'2012]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1$.
- Elastic Net [Xiao'2021]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$.
- Graph Net [Grosenic'2013]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|D\beta\|_2^2$.
- Generalized Lasso [Xin'2014]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|D\beta\|_1$.

Structural Sparsity

Structural Sparsity for diseased gray matter voxels:

1 sparsity 2 spatial coherence 3 positivity (with Y).

- Lasso [Liu'2012]: $\min_{\beta} \ell(\beta) + \lambda_1 \|\beta\|_1$.
- Elastic Net [Xiao'2021]: $\min_{\beta} \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$.
- Graph Net [Grosenic'2013]: $\min_{\beta} \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|D\beta\|_2^2$.
- Generalized Lasso [Xin'2014]: $\min_{\beta} \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|D\beta\|_1$.
- Non-negative Generalized Fused Lasso (n^2-GFL) [Xin'2015]: $\min_{\beta > 0} \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|D\beta\|_1$.

Structural Sparsity

Structural Sparsity for diseased gray matter voxels:

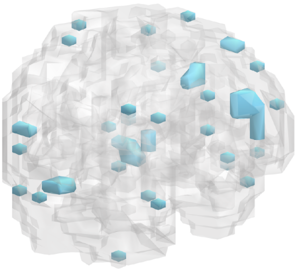
1. sparsity 2. spatial coherence 3. positivity (with Y).

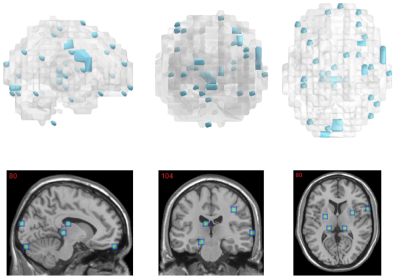
- Lasso [Liu'2012]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1$.
- Elastic Net [Xiao'2021]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$.
- Graph Net [Grosenic'2013]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|D\beta\|_2^2$.
- Generalized Lasso [Xin'2014]: $\min_\beta \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|D\beta\|_1$.
- Non-negative Generalized Fused Lasso (n^2-GFL) [Xin'2015]: $\min_{\beta>0} \ell(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|D\beta\|_1$.

Limitations. i) Incoherence condition is not easy to satisfy; ii) Overlook the gap between prediction and feature selection.

Procedural Bias

- Enlarged GM voxels in lateral ventricle features.
- Introduced in the procedure of preprocessing.
- Prediction – lesion features: helpful for disease prediction.





Selection of Degenerated Features

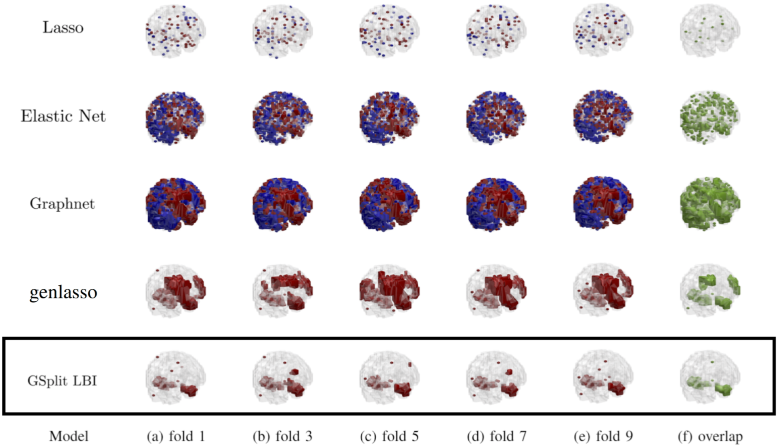


Figure: Lesion Features Selection. [1]

[1] GSplit LBI: Taming the Procedural Bias. MICCAI, 2017.

Procedural Bias

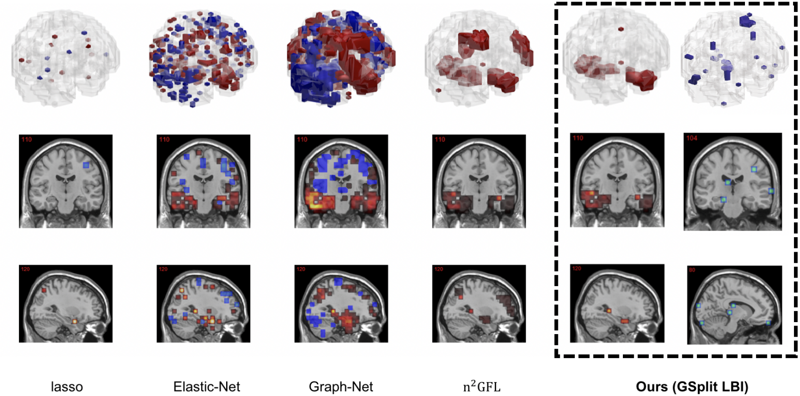


Figure: Procedural Bias Selection. [2]

[2] **GSplit LBI: Taming the Procedural Bias. MICCAI, 2017.**

Prediction Power

	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%**

Figure: Results on ADNI. [3]

[3] **GSplit LBI: Taming the Procedural Bias. MICCAI, 2017.**

How do we overcome these limitations and achieve this result?

Outline

1. Structural Sparsity in Medical Imaging
2. Recovering Structural Sparsity via Differential Inclusion
3. False-Discovery-Rate Control
   - Split Knockoff
   - FDR Smoothing on Heterogeneous Features

Sparse Regression

Considering recovering $\beta^\star$ from following model:

$$y = X\beta^\star + \varepsilon,$$

where X is the design matrix.

Structural Sparsity. $\gamma^\star := D\beta^\star, S := \mathrm{supp}(\gamma^\star)$; $|S| \ll |\mathrm{row}(D)|$.

Sparse Regression

Considering recovering $\beta^\star$ from following model:

$$y = X\beta^\star + \varepsilon,$$

where X is the design matrix.

Structural Sparsity. $\gamma^\star := D\beta^\star, S := \mathrm{supp}(\gamma^\star)$; $|S| \ll |\mathrm{row}(D)|$.

- $D := I$. Pure sparsity.
- D is wavelet basis. Wavelet smoothing.
- D is graph laplacian. Image denosing.

Sparse Regression

Considering recovering $\beta^\star$ from following model:

$$y = X\beta^\star + \varepsilon,$$

where X is the design matrix.

Structural Sparsity. $\gamma^\star := D\beta^\star, S := \mathrm{supp}(\gamma^\star)$; $|S| \ll |\mathrm{row}(D)|$.

- $D := I$. Pure sparsity.
- D is wavelet basis. Wavelet smoothing.
- D is graph laplacian. Image denosing.

How to recover $\gamma^\star$ sparse pattern (sparsisitency) and estimate true values of $\beta^\star$ ($\gamma^\star$) (consistency)?

Generalized Lasso (Review)

Generalized lasso (genlasso):

$$\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|D\beta\|_1$$

- Total Variation (Rudin, et al.'1992); Fused Lasso (Tibshirani et al.'2005); Lasso (Tibshirani'1996).

Generalized Lasso (Review)

Generalized lasso (genlasso):

$$\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|D\beta\|_1$$

- Total Variation (Rudin, et al.'1992); Fused Lasso (Tibshirani et al.'2005); Lasso (Tibshirani'1996).
- Problems of genlasso.
  1. Optimizations for several λ.
  2. Incoherence condition: **hard to satisfy** (Vaiter et al.'2013,Zhao et al.'2006).

Structural Sparsity via Differential Inclusion

Structural Sparse Regression:

$$y = X\beta^{\star} + \varepsilon, \gamma^{\star} = D\beta^{\star}, |S| \ll |\mathrm{row}(D)|.$$

Variable Splitting between $D\beta$ and γ:

$$\mathcal{L}_{\nu}(\beta, \gamma) := \frac{1}{2n}\|y - X\beta\|_2^2 + \frac{1}{2\nu}\|D\beta - \gamma\|_2^2.$$

Inverse Scale Space via Differential Inclusion

$$\begin{aligned} 0 &= -\nabla_{\beta}\mathcal{L}_{\nu}(\beta_t, \gamma_t) \\ \dot{\rho}_t &= -\nabla_{\gamma}\mathcal{L}_{\nu}(\beta_t, \gamma_t) \\ \rho_t &\in \partial\|\gamma_t\|_1 \end{aligned}$$

Differential Inclusion in Inverse Scale Space

Split Bregman Inverse Scale Space (Split Bregman ISS) [4]:

$$
\begin{aligned}
0 &= -\nabla_\beta \mathcal{L}_\nu(\beta_t, \gamma_t) && \text{-update of } \beta_t \\
\dot{\rho}_t &= -\nabla_\gamma \mathcal{L}_\nu(\beta_t, \gamma_t) && \text{-update of } \gamma_t \\
\rho_t &\in \partial \|\gamma_t\|_1 && \text{-update of } \gamma_t
\end{aligned}
$$

- At each t, β_t is solved directly.

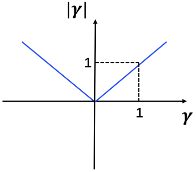
[4] **Split lbi: an iterative regularization path. NeurIPS, 2016.**

Differential Inclusion in Inverse Scale Space

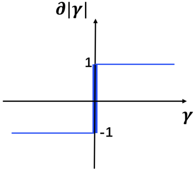
Split Bregman Inverse Scale Space (Split Bregman ISS) [4]:

$$
\begin{aligned}
0 &= -\nabla_\beta \mathcal{L}_\nu(\beta_t, \gamma_t) && \text{-update of } \beta_t \\
\dot{\rho}_t &= -\nabla_\gamma \mathcal{L}_\nu(\beta_t, \gamma_t) && \text{-update of } \gamma_t \\
\rho_t &\in \partial \|\gamma_t\|_1 && \text{-update of } \gamma_t
\end{aligned}
$$

- At each t, β_t is solved directly.
- $\rho(i) = \mathrm{sign}(\gamma(i))$ for $\gamma(i) \neq 0$; if $\rho(i) \in (-1, 1)$ then $\gamma(i) = 0$.



$f := |\cdot|$



$\partial f := \partial|\cdot|$

[4] Split lbi: an iterative regularization path. NeurIPS, 2016.

Differential Inclusion in Inverse Scale Space

Split Bregman Inverse Scale Space (Split Bregman ISS) [5]:

$$
\begin{aligned}
0 &= -\nabla_\beta \mathcal{L}_\nu(\beta_t, \gamma_t) && \text{-update of } \beta_t \\
\dot{\rho}_t &= -\nabla_\gamma \mathcal{L}_\nu(\beta_t, \gamma_t) && \text{-update of } \gamma_t \\
\rho_t &\in \partial \|\gamma_t\|_1 && \text{-update of } \gamma_t
\end{aligned}
$$

- At each t, β_t is solved directly.
- $\rho(i) = \mathrm{sign}(\gamma(i))$ for $\gamma(i) \neq 0$; if $\rho(i) \in (-1, 1)$ then $\gamma(i) = 0$.
- ρ_t : gradient descent flow, starting from 0.

[5] **Split lbi: an iterative regularization path. NeurIPS, 2016.**

Differential Inclusion in Inverse Scale Space

Split Bregman Inverse Scale Space (Split Bregman ISS) [5]:

$$0 = -\nabla_\beta \mathcal{L}_\nu(\beta_t, \gamma_t) \quad \text{-update of } \beta_t$$
$$\dot{\rho}_t = -\nabla_\gamma \mathcal{L}_\nu(\beta_t, \gamma_t) \quad \text{-update of } \gamma_t$$
$$\rho_t \in \partial \|\gamma_t\|_1 \quad \text{-update of } \gamma_t$$

- At each t, β_t is solved directly.
- $\rho(i) = \mathrm{sign}(\gamma(i))$ for $\gamma(i) \neq 0$; if $\rho(i) \in (-1, 1)$ then $\gamma(i) = 0$.
- ρ_t : gradient descent flow, starting from 0.
- $\rho_t(i)$ reaches $\pm 1 \implies \gamma_t(i) \neq 0$ is selected.

[5] **Split lbi: an iterative regularization path. NeurIPS, 2016.**

Differential Inclusion in Inverse Scale Space

Split Bregman Inverse Scale Space (Split Bregman ISS) [5]:

$$0 = -\nabla_\beta \mathcal{L}_\nu(\beta_t, \gamma_t) \quad \text{-update of } \beta_t$$
$$\dot{\rho}_t = -\nabla_\gamma \mathcal{L}_\nu(\beta_t, \gamma_t) \quad \text{-update of } \gamma_t$$
$$\rho_t \in \partial\|\gamma_t\|_1 \quad \text{-update of } \gamma_t$$

- At each t, β_t is solved directly.
- $\rho(i) = \mathrm{sign}(\gamma(i))$ for $\gamma(i) \neq 0$; if $\rho(i) \in (-1, 1)$ then $\gamma(i) = 0$.
- ρ_t : gradient descent flow, starting from 0.
- $\rho_t(i)$ reaches $\pm 1 \implies \gamma_t(i) \neq 0$ is selected.
- Regularization Solution Path. t is regularization parameter.

[5] **Split lbi: an iterative regularization path. NeurIPS, 2016.**

Damped (Linearized) Bregman ISS

- Append $\frac{1}{2\kappa}\|\gamma_t\|_2^2$ for strongly convexity (with $\kappa > 0$):

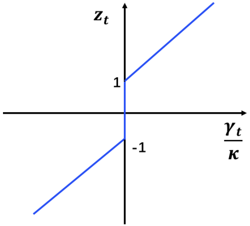
$$z_t \triangleq \rho_t + \frac{\gamma_t}{\kappa} \in \partial\left(\|\gamma_t\|_1 + \frac{1}{2\kappa}\|\gamma_t\|_2^2\right)$$

Damped (Linearized) Bregman ISS

- Append $\frac{1}{2\kappa}\|\gamma_t\|_2^2$ for strongly convexity (with $\kappa > 0$):

$$z_t \triangleq \rho_t + \frac{\gamma_t}{\kappa} \in \partial\left(\|\gamma_t\|_1 + \frac{1}{2\kappa}\|\gamma_t\|_2^2\right)$$

- γ_t can be obtained from z_t.



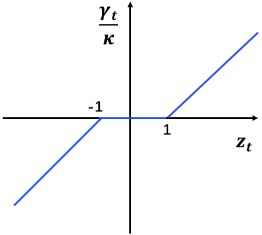
Damped (Linearized) Bregman ISS

- Append $\frac{1}{2\kappa}\|\gamma_t\|_2^2$ for strongly convexity (with $\kappa > 0$):

$$z_t \triangleq \rho_t + \frac{\gamma_t}{\kappa} \in \partial\left(\|\gamma_t\|_1 + \frac{1}{2\kappa}\|\gamma_t\|_2^2\right)$$

- γ_t can be obtained from z_t:

$$\gamma_t = \kappa * \mathrm{sign}(z_t) \odot \max(|z_t| - 1, 0).$$



Damped (Linearized) Bregman ISS

- Append $\frac{1}{2\kappa}\|\gamma_t\|_2^2$ for strongly convexity (with $\kappa > 0$):

$$z_t \triangleq \rho_t + \frac{\gamma_t}{\kappa} \in \partial\left(\|\gamma_t\|_1 + \frac{1}{2\kappa}\|\gamma_t\|_2^2\right)$$

- γ_t can be obtained from z_t:

$$\gamma_t = \kappa * \mathrm{sign}(z_t) \odot \max(|z_t| - 1, 0).$$

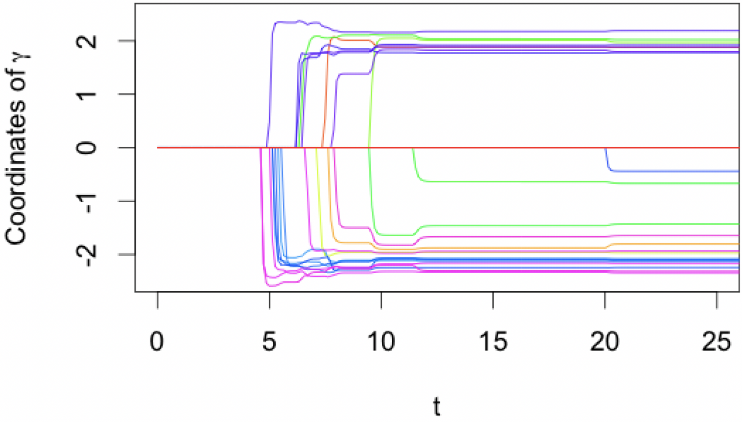
- Split Linearized Bregman Invserse Scale Space (Split LBISS):

$$\begin{aligned} \dot{\beta}_t/\kappa &= -\nabla_\beta \mathcal{L}_\nu(\beta_t, \gamma_t) \\ \dot{z}_t &= -\nabla_\gamma \mathcal{L}_\nu(\beta_t, \gamma_t) \\ \gamma_t &= \kappa * \mathrm{sign}(z_t) \odot \max(|z_t| - 1, 0) \end{aligned}$$

- Discretization. Split Linearized Bregman Iteration (Split LBI):

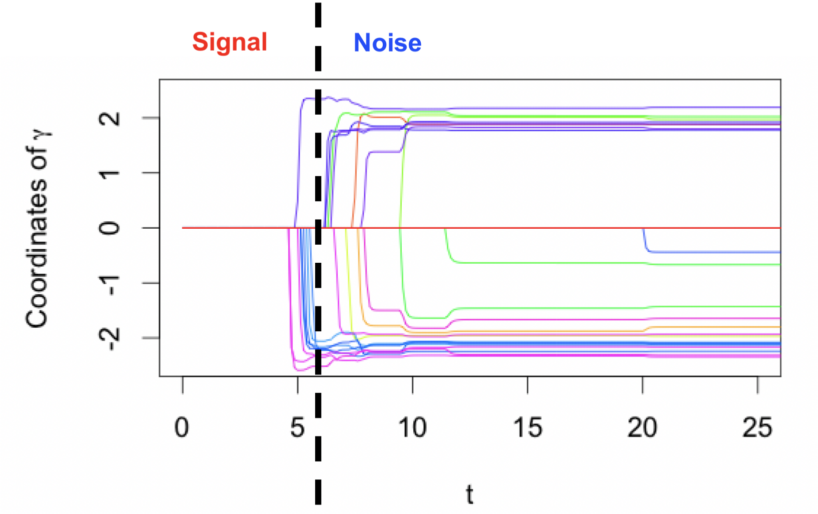
$$\begin{aligned} \beta_{k+1} &= \beta_k - \kappa\alpha\nabla_\beta \mathcal{L}(\beta_k, \gamma_k) \\ z_{k+1} &= z_k - \alpha\nabla_\gamma \mathcal{L}(\beta_k, \gamma_k) \\ \gamma_{k+1} &= \kappa * \mathrm{sign}(z_{k+1}) \odot \max(|z_{k+1}| - 1, 0) \end{aligned}$$

Solution Path



Solution Path

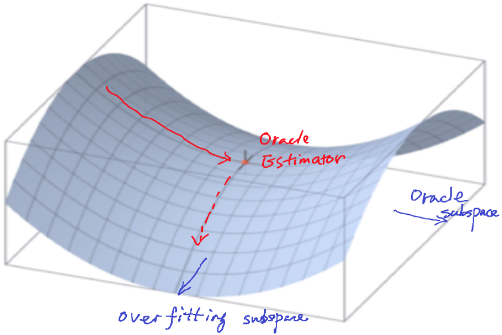
Linear model $y = X\beta^{\star} + \text{noise}$.



No-False-Positive

Do the early selected features belong to the true signal set?

Intuition



- Irrespresentable Condition $\Longrightarrow$ Loss decay on the oracle space.
- Restricted Strong Convexity $\Longrightarrow$ Unique solution.
- Early Stopping after picking up the signals.

Irrepresentable Condition

ν in Variable Splitting: relax $\gamma^\star = D\beta^\star$.

$$\mathcal{L}_\nu(\beta, \gamma) := \frac{1}{2n}\|y - X\beta\|_2^2 + \frac{1}{2\nu}\|D\beta - \gamma\|_2^2.$$

Assumption 1 (Irrepresentable Condition).

$$\mathrm{IRR}(\nu) := \|\Sigma_{S^c,S}(\nu)\Sigma_{S,S}^{-1}(\nu)\|_\infty < 1.$$

The $\Sigma(\nu) := (I - D(\nu X^* X + D^\top D)^\dagger D^\top)/\nu$.

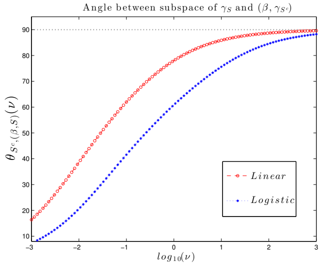
Decollinearity

The angle between (S^c) and (β, S), via $\mathrm{H}(\nu) = \nabla^2 \mathcal{L}_\nu(\beta, \gamma)$.

$$\theta^{\nu}_{S^c,(\beta,S)} := \arccos\left(\sqrt{\frac{\mathrm{trace}(H_{S^c,(\beta,S)}(\nu) H^{\dagger}_{(\beta,S),(\beta,S)}(\nu) H_{(\beta,S),S^c}(\nu))}{\mathrm{trace}(H_{S^c,S^c}(\nu))}}\right)$$

Theorem (Sun-Han-Hu-Yao-Wang'2020)

The $\lim_{\nu\to\infty} \theta^{\nu}_{S^c,(\beta,S)} \to \pi/2 \iff \ker(X) \subset \ker(D_S)$.



Irrepresentable Condition

ν in Variable Splitting: relax $\gamma^\star = D\beta^\star$.

$$\mathcal{L}_\nu(\beta,\gamma) := \frac{1}{2n}\|y - X\beta\|_2^2 + \frac{1}{2\nu}\|D\beta - \gamma\|_2^2.$$

Assumption 1 (Irrepresentable Condition).

$$\mathrm{IRR}(\nu) := \|\Sigma_{S^c,S}(\nu)\Sigma_{S,S}^{-1}(\nu)\|_\infty < 1.$$

The $\Sigma(\nu) := (I - D(\nu X^* X + D^\top D)^\dagger D^\top)/\nu$.

Theorem (Huang-Sun-Xiong-Yuan' 2016)

- $\lim_{\nu\to 0} \mathrm{IRR}(\nu) = \mathrm{IC}$ (genlasso Vaiter'13).
- $\lim_{\nu\to\infty} \mathrm{IRR}(\nu)$ exists and $= 0 < \mathrm{IC} \iff \ker(X) \subset \ker(D_S)$.

Assumptions

ν in Variable Splitting: relax $\gamma^\star = D\beta^\star$.

$$\mathcal{L}_\nu(\beta,\gamma) := \frac{1}{2n}\|y - X\beta\|_2^2 + \frac{1}{2\nu}\|D\beta - \gamma\|_2^2.$$

Assumption 1 (Irrepresentable Condition).

$$\mathrm{IRR}(\nu) := \|\Sigma_{S^c,S}(\nu)\Sigma_{S,S}^{-1}(\nu)\|_\infty < 1.$$

Assumption 2 (Restricted Strongly Convexity).

$$\Sigma_{S,S} \geq \lambda I, \text{ for some } \lambda > 0.$$

The $\Sigma(\nu) := (I - D(\nu X^* X + D^\top D)^\dagger D^\top)/\nu$.

Path Consistency

Sparse Estimator $\tilde{\beta}_t$.

$$\tilde{\beta}_t = P_{S_t}(\beta_t) := \arg \min_{D_{S^c} x = 0} \|\beta_t - x\|_2.$$

Theorem (Huang-Sun-Xiong-Yuan'2020)

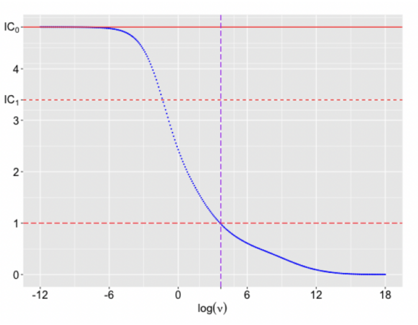
Under Irrepresentable Condition and Restricted Strongly Convexity, then there exists $\bar{\tau}$ s.t.

- No-false-positive. $\mathrm{supp}(\gamma_t) \subseteq S$, for $0 \leqslant t \leqslant \bar{\tau}$.
- Sign-Consistency. $\mathrm{sign}(\gamma_{\bar{\tau}}) = \mathrm{sign}(\gamma^\star)$ if $\gamma^\star$ is strong.
- ℓ_2 consistency of γ_t. $\|\gamma_{\bar{\tau}} - D\beta^\star\|_2 \leqslant \mathcal{O}(\sqrt{\frac{s \log m}{n}})$.
- ℓ_2 consistency of $\tilde{\beta}_t$. $\|\tilde{\beta}_t - \beta^\star\|_2 \leq \mathcal{O}(\sqrt{\frac{s \log m}{n}})$.

Simulation

Left. $\mathrm{IRR}(\nu)$ vs ν and IC.

Right. Split LBI vs genlasso in terms of AUC.



genlasso	Split LBI		
	1	5	10
.9426 (.0390)	.9845 (.0185)	.9969 (.0065)	**.9982 (.0043)**

$D = I$

genlasso	Split LBI		
	1	5	10
.9705 (.0212)	.9955 (.0056)	.9996 (.0014)	**.9998 (.0009)**

D is 1-d fused lasso matrix

Explore Beyond Sparsity

Variable Splitting. Dense and Sparse parameter.

$$\mathcal{L}_\nu(\beta,\gamma) := \frac{1}{2n}\|y - X\beta\|_2^2 + \frac{1}{2\nu}\|D\beta - \gamma\|_2^2.$$

- Sparse parameter: $\tilde{\beta} := \arg\min_{D_{S^c}x=0} \|\beta - x\|_2$.
- Dense parameter: Explore beyond sparsity.

Summary

Split Bregman Inverse Scale Space (Split Bregman ISS)

$$
\begin{aligned}
0 &= -\nabla_{\beta}\mathcal{L}_{\nu}(\beta_t, \gamma_t) \\
\dot{\rho}_t &= -\nabla_{\gamma}\mathcal{L}_{\nu}(\beta_t, \gamma_t) \\
\rho_t &\in \partial\|\gamma_t\|_1
\end{aligned}
$$

- A regularization solution path via differential inclusion.
- More and more variables are selected as iterates.
- Earlier selected features belong to the true signal set.
- Variable Splitting enables to explore beyond sparsity.

False-Discovery-Rate (FDR) Control

Piratically, the incoherence is unknown. Then,

False-Discovery-Rate (FDR) Control

Piratically, the incoherence is unknown. Then,

How to control the FDR?

Outline

1. Structural Sparsity in Medical Imaging

2. Recovering Structural Sparsity via Differential Inclusion

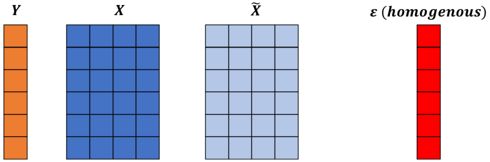
3. False-Discovery-Rate Control
    - Split Knockoff
    - FDR Smoothing on Heterogeneous Features

Knockoff (Review) for Pure Sparsity

- FDR control for pure sparsity. $y = X\beta^* + \varepsilon$.
- Control $\mathbb{E}\left[\frac{\{i \in \hat{S} : \beta^*(i) = 0\}}{|\hat{S}|}\right] \leqslant q$.

Knockoff (Review) for Pure Sparsity

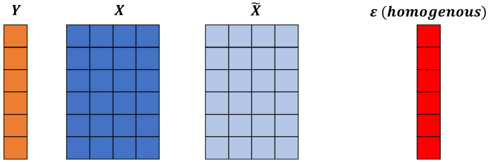
- FDR control for pure sparsity. $y = X\beta^* + \varepsilon$.
- Control $\mathbb{E}\left[\frac{\{i \in \hat{S} : \beta^*(i) = 0\}}{|\hat{S}|}\right] \leqslant q$.



- For each X_j, create a knockoff copy $\tilde{X}_j$ as a control.

Knockoff (Review) for Pure Sparsity

- FDR control for pure sparsity. $y = X\beta^* + \varepsilon$.
- Control $\mathbb{E}\left[\frac{\{i \in \hat{S} : \beta^*(i) = 0\}}{|\hat{S}|}\right] \leqslant q$.



- For each X_j, create a knockoff copy $\tilde{X}_j$ as a control.
- Implement black-box algorithm to obtain Z_j and $\tilde{Z}_j$.
  - Z_j ($\tilde{Z}_j$): the effect of X_j on Y ($\tilde{X}_j$ on Y).
- For each j, $W_j := \max(Z_j, \tilde{Z}_j) * \mathrm{sign}(Z_j - \tilde{Z}_j)$.

Knockoff (Review) for Pure Sparsity

- For each j, $W_j := \max(Z_j, \tilde{Z}_j) * \mathrm{sign}(Z_j - \tilde{Z}_j)$.
  - For non-null: $W_j > 0$ and is large.
  - For null: $P(W_j > 0) = 1/2$, *i.e.*, comparable with its copy.

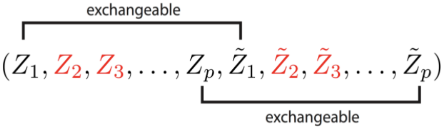
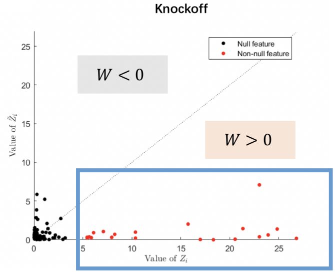


Figure: Null Statistics are **pairwise exchangeable**.

Knockoff (Review) for Pure Sparsity

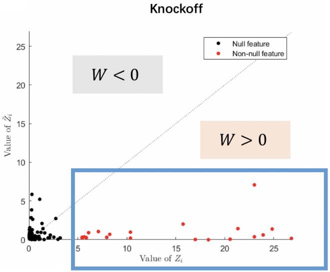
- Knockoff selects features that are clearly better than their copies.



- Knockoff selects $\hat{S} := \{i : W_i > T_q\}$, $T_q := \min_T \left\{ \frac{|\{i:W_i<-T\}|}{1 \cup |\{i:W_i \geq T\}|} \right\}$

Knockoff (Review) for Pure Sparsity

- Knockoff selects features that are clearly better than their copies.



- Knockoff selects $\hat{S} := \{i : W_i > T_q\}$, $T_q := \min_T \left\{ \frac{|\{i:W_i < -T\}|}{1 \cup |\{i:W_i \geqslant T\}|} \right\}$

Theorem (R.Barbers and Candés'15)

Given any desired level $q > 0$, knockoff has $\mathrm{FDR} \leqslant q$.

Heterogeneous Noise for Structural Sparsity

Structural sparsity: $y = X\beta^* + \varepsilon$, $\gamma^* = D\beta^*$ ($D \in \mathbb{R}^{m\times p}$) is sparse.

- If $\mathrm{rank}(D) = m$, we have $y = XD^{\dagger}\gamma^*(= \beta^*) + \varepsilon$.
- For $m > p$ or $\mathrm{rank}(D) < m$, $\tilde{y} = X_\beta\beta^* + X_\gamma\gamma^* + \varepsilon$:

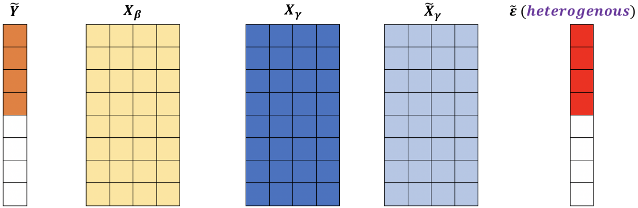
$$\tilde{y} = \begin{pmatrix} \frac{y}{\sqrt{n}} \\ 0_m \end{pmatrix}, X_\beta = \begin{pmatrix} \frac{X}{\sqrt{n}} \\ \frac{D}{\sqrt{\nu}} \end{pmatrix}, X_\gamma = \begin{pmatrix} 0_{n\times m} \\ -\frac{I_m}{\sqrt{\nu}} \end{pmatrix}, \tilde{\varepsilon} = \begin{pmatrix} \frac{\varepsilon}{\sqrt{n}} \\ 0_m \end{pmatrix}.$$

Heterogeneous Noise for Structural Sparsity

Structural sparsity: $y = X\beta^* + \varepsilon$, $\gamma^* = D\beta^*$ ($D \in \mathbb{R}^{m\times p}$) is sparse.

- If $\mathrm{rank}(D) = m$, we have $y = XD^{\dagger}\gamma^*(= \beta^*) + \varepsilon$.
- For $m > p$ or $\mathrm{rank}(D) < m$, $\tilde{y} = X_\beta\beta^* + X_\gamma\gamma^* + \varepsilon$:

$$\tilde{y} = \begin{pmatrix} \frac{y}{\sqrt{n}} \\ 0_m \end{pmatrix}, X_\beta = \begin{pmatrix} \frac{X}{\sqrt{n}} \\ \frac{D}{\sqrt{\nu}} \end{pmatrix}, X_\gamma = \begin{pmatrix} 0_{n\times m} \\ -\frac{I_m}{\sqrt{\nu}} \end{pmatrix}, \tilde{\varepsilon} = \begin{pmatrix} \frac{\varepsilon}{\sqrt{n}} \\ 0_m \end{pmatrix}.$$

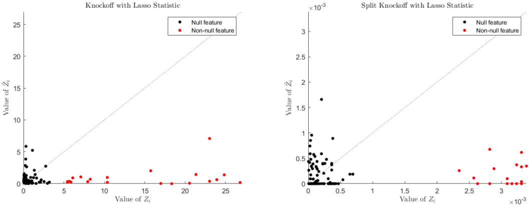


Exchangeability Fails

- For each $X_\gamma(j)$, construct a copy $\tilde{X}_\gamma(j)$.
- $Z_j := \sup\{\lambda : \gamma_j(\lambda) \neq 0\}$, $\tilde{Z}_j := \sup\{\lambda : \tilde{\gamma}_j(\lambda) \neq 0\}$.

$$\gamma(\lambda) := \arg\min_{\gamma} \frac{1}{2}\|\tilde{y} - X_\beta \beta(\lambda) - X_\gamma \gamma\|^2 + \lambda\|\gamma\|_1,$$

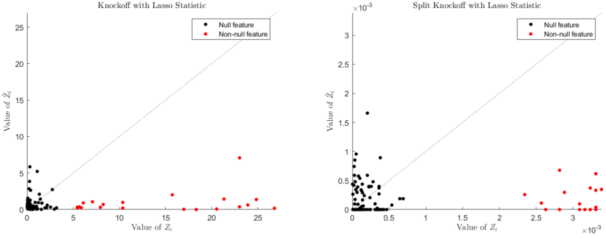
$$\tilde{\gamma}(\lambda) := \arg\min_{\tilde{\gamma}} \frac{1}{2}\|\tilde{y} - X_\beta \beta(\lambda) - \tilde{X}_\gamma \tilde{\gamma}\|^2 + \lambda\|\tilde{\gamma}\|_1.$$



(a) Knockoff (b) Split Knockoff

Figure: For null j, $P(W_j < 0) > P(W_j > 0)$.

Split Knockoff (Truncations)



(a) Knockoff (b) Split Knockoff

Solution (Truncation). Given $\{Z_j\}_j, \{\tilde{Z}_j\}_j$, we define

$$W_j = Z_j * \mathrm{sign}(Z_j - \tilde{Z}_j * 1(r_j = \tilde{r}_j)),$$

where $r_j = \lim_{t \to Z_j -} \mathrm{sign}(\gamma_j(t))$ and $\tilde{r}_j = \lim_{t \to \tilde{Z}_j -} \mathrm{sign}(\tilde{\gamma}_j(t))$.

Null Statistics W_j are not Independent

We look at the KKT condition:

$$\lambda\rho(\lambda) + \frac{\gamma(\lambda)}{\nu} = \frac{D\beta(\lambda)}{\nu} \text{ (determined by } \xi := \frac{X^\top \varepsilon}{n}\text{)},$$
$$\lambda\tilde{\rho}(\lambda) + \frac{\tilde{\gamma}(\lambda)}{\nu} = \frac{D\beta(\lambda)}{\nu} + \underbrace{\{-s\gamma^* + \tilde{X}_{\gamma,1}\varepsilon/\sqrt{n}\}}_{\zeta}.$$

Condition on $\xi := \frac{X^\top \varepsilon}{n}$, $P(W_j > 0) = P(\zeta_j < 0)$. Besides,

$$\mathbb{E}[\zeta] = 0,\ \mathrm{Var}[\zeta] = \frac{\sigma^2(2s - \nu s^2)}{n} I_m;$$
$$\mathbb{E}[\zeta|\xi] \neq 0,\ \mathrm{Var}[\zeta|\xi] = \frac{\sigma^2(2s - \nu s^2)}{n} I_m - R;$$

Solution: Data Splitting

We can split the dataset into $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$:

- $\beta(\lambda)$ is obtained from $\mathcal{D}_1$;
- $Z_j, \tilde{Z}_j$ are obtained from $\mathcal{D}_2$.

$$\lambda\rho(\lambda) + \frac{\gamma(\lambda)}{\nu} = \frac{D\beta(\lambda)}{\nu} \text{ (determined by } \xi := \frac{X^\top \varepsilon_1}{n_1}\text{)},$$
$$\lambda\tilde{\rho}(\lambda) + \frac{\tilde{\gamma}(\lambda)}{\nu} = \frac{D\beta(\lambda)}{\nu} + \zeta.$$

Then, $P(W_j > 0) = P(\zeta_j < 0)$ with

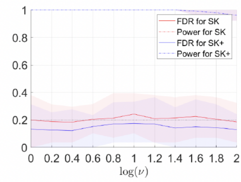
$$\mathbb{E}[\zeta] = 0, \; \mathrm{Var}[\zeta] = \frac{\sigma^2(2s - \nu s^2)}{n_2} I_m.$$

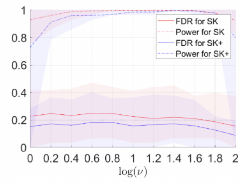
Theoretical Analysis

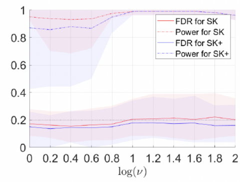
Theorem (FDR Control of Split Knockoff)

For any $q > 0$, we have $\mathrm{FDR} \leqslant q$.

Simulation Experiments [6]:



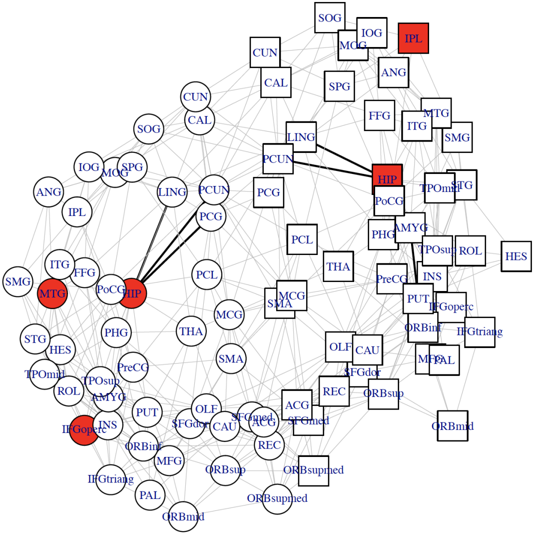




$D_1 = I_p$ D_2 is 1-d fused lasso $D_3 = [D_1; D_2]$

[6] **Controlling the False Discovery Rate in Transformational Sparsity: Split Knockoffs.**

Alzheimer's Disease



Knockoff requires $n > m + p$.

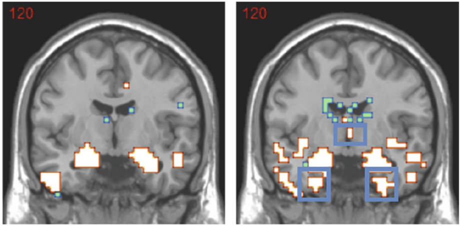
In high-dimensional analysis, the power is limited due to **multi-collinearity** problem.

From $p = 2,527$ voxels to $p = 20,091$.

Accuracy of Split LBI: V_8 (90.91%) vs V_4 (89.77%).

- V_8: $2,527$ voxels with $8 \times 8 \times 8mm^3$.
- V_4: $20,091$ voxels with $4 \times 4 \times 4mm^3$.

The performance drops with even more features!



GSplit LBI FDR-HS

How do we alleviate this problem and improve emprical utility?

Hypothesis Testing

- Two-sample T-test: $t_i = \frac{\bar{x}_{1,i} - \bar{x}_{2,i}}{\sqrt{\frac{s_{1,i}^2}{n_1} + \frac{s_{2,i}^2}{n_2}}}$, $s_{1,i}^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (x_{1,i}(j) - \bar{x}_{1,i})^2$.
  - $\mathcal{H}_0 : t_i \sim \mathcal{T}_{n_1+n_2-2}$. We obtain p_i.
  - Cannot control FDR.
- Benjamini-Hochberg Procedure [Bajamini'1995]: first rank $p_1, \ldots, p_n$ in an ascending order $p_{(1)}, \ldots, p_{(n)}$ and selects $\{i : i \leqslant k\}$ in which $k := \max\{i : p_{(i)} \leqslant \frac{i\alpha}{n}\}$.
  - Too conservative in feature selection.

Local FDR

The [Efron'2001] proposed LocalFDR, an Empirical-Bayes Method:

$$f(z) = p_0 f_0(z) + (1 - p_0) f_1(z), \text{where}$$

- p_0 denotes the prior of being null;
- $f_0(z), f_1(z)$ denote the p.d.f of null and non-null groups.

We have: $\mathrm{fdr}(z) := p(i \text{ is null } | z) = \frac{p_0 f_0(z)}{f(z)}$, where

- Central Matching to estimate $f_0(z), p_0$;
- Kernel density to estimate $f(z)$.

Limitation: does not consider spatial coherence!

FDR Smoothing

The [Tansey'2014] considers:

$$f(z) = p_0 f_0(z) + (1 - p_0) f_1(z) \implies f(z) = (1 - c_i) f_0(z) + c_i f_1(z),$$

where $c_i := \mathrm{sigmoid}(\beta_i)$. We optimize:

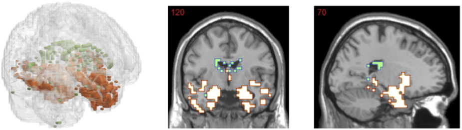
$$\ell(\beta) + \lambda \|D\beta\|_1,$$

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

Heterogeneous Smoothing

Note: The procedural bias and lesions are heterogeneous:

- Procedural bias is enlarged; lesions are atrophied.
- Different degree of spatial coherence.



Heterogeneous Smoothing

We split $G := G_{\rm pro} \cup G_{\rm les} \cup G_{\rm int}$, with

- $G_{\rm pro} = (V_{\rm pro}, E_{\rm pro})$, $V_{\rm pro} = \{i : z_i \leqslant 0\}$, $E_{\rm pro} = \{(i,j) \in E : z_i, z_j \leqslant 0\}$;
- $G_{\rm les} = (V_{\rm les}, E_{\rm les})$, $V_{\rm les} = \{i : z_i > 0\}$, $E_{\rm les} = \{(i,j) \in E : z_i, z_j > 0\}$;
- $G_{\rm int} = (V_{\rm int}, E_{\rm int})$, $V_{\rm les} = V_{\rm pro} \cup V_{\rm les}$, $E_{\rm int} = \{(i,j) \in E : z_i \leqslant 0, z_j > 0\}$.

We turn to optimize:

$$\ell(\beta) + \lambda_{\rm pro}\|D_{G_{\rm pro}}\beta\|_1 + \lambda_{\rm les}\|D_{G_{\rm les}}\beta\|_1 + \lambda_{\rm int}\|D_{G_{\rm int}}\beta\|_1,$$

with $D_G\beta(i,j) = \beta_i - \beta_j$ for $(i,j) \in E$[7].

[7] **FDR-HS: An Empirical Bayesian Identification of Heterogenous Features in Neuroimage Analysis. MICCAI, 2018**

Optimization

The loss is not convex, hence we introduce

$$s_i = \begin{cases} 1 & \text{if } z_i \sim f_1(z) \\ 0 & \text{if } z_i \sim f_0(z) \end{cases}.$$

The loss is the modified as:

$$\ell(\beta, s) := \sum_{i=1}^{N} \{\log\left(1 + \exp(\beta_i)\right) - s_i \beta_i\},$$
$$g(\beta, s) := \ell(\beta, s) + \lambda_{\text{pro}} \|D_{G_{\text{pro}}} \beta\|_1 + \lambda_{\text{les}} \|D_{G_{\text{les}}} \beta\|_1 + \lambda_{\text{int}} \|D_{G_{\text{int}}} \beta\|_1.$$

We implement Expectation-Maximization (EM) to optimize β and z[8].

[8] **FDR-HS: An Empirical Bayesian Identification of Heterogeneous Features in Neuroimage Analysis. MICCAI, 2018**

Feature Selection

After estimating $\tilde{\beta}_i$, $f_0(z)$, $f_1(z)$, we selects feature with

$$p(s_i = 0|z_i, \tilde{\beta}_i) = \frac{(1-\tilde{c}_i)f_0(z_i)}{\tilde{c}_i f_1(z_i) + (1-\tilde{c}_i)f_0(z_i)} < \gamma \left(\tilde{c}_i = \mathrm{sigmoid}(\beta_i)\right).$$

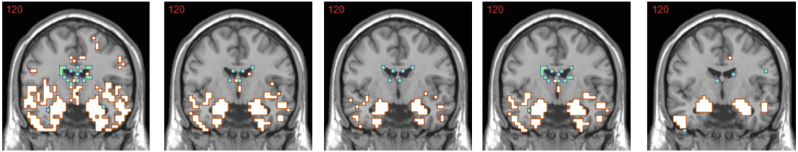
The $\gamma \in (0,1)$ is a pre-setting threshold hyper-parameter.

Results on ADNI

Table 1. Comparison between FDR-HS and others on 10-fold classification result

	Univariate + ElasticNet				Multivariate	
	T-test	BH_q [4]	LocalFDR [7]	FDR-HS	GSplit LBI [12]	Elastic Net [16]
15ADNC	89.61%	89.61%	87.01%	**90.26%**	85.06%	87.01%
15MCINC	70.50%	71.00%	73.50%	**75.00%**	72.50%	72.00%
30ADNC	88.64%	89.77%	89.77%	**91.48%**	89.77%	88.07%

Feature Selection on ADNI



T-test | BH_q | localFDR | FDR-HS | GSplit LBI

Table 2. Comparison between FDR-HS and others on stability (measured by mDC)

	T-test	BH_q	LocalFDR	FDR-HS	GSplit LBI
$\mathrm{mDC}^{(+)}$ (Lesion features)	0.6705	0.6248	0.6698	**0.6842**	0.4598
$\mathrm{mDC}^{(-)}$ (Procedural Bias)	0.6267	0.5541	0.5127	**0.6540**	0.3033

$$\mathrm{mDC} := \frac{K\left|\cap_{k=1}^{K} S^{\pm}(k)\right|}{\sum_{k=1}^{K}\left|S^{\pm}(k)\right|}, \quad S^{\pm}(k): \text{ selected lesions and procedural bias.}$$

Thank You!

sunxinwei@fudan.edu.cn