A Differential Inclusion Approach for Learning Heterogeneous Sparsity in Neuroimaging Analysis

Wenjing Han
Communication University of China

Yueming Wu
Peking University

Xinwei Sun*
Fudan University

Lingjing Hu*
Capital Medical University

Yizhou Wang
Peking University

Abstract

In voxel-based neuroimaging disease prediction, it was recently found that in addition to lesion features, there exists another type of feature called "Procedural Bias", which is introduced during preprocessing and can further improve the prediction power. However, traditional sparse learning methods fail to simultaneously capture both types of features due to their heterogeneity in sparsity types. Specifically, the lesion features are spatially coherent and suffer from volumetric degeneration, while the procedural bias refers to enlarged voxels that are dispersedly distributed. In this paper, we propose a new method based on differential inclusion, which generates a sparse regularized solution path on multiple parameters that are enforced with heterogeneous sparsity to capture lesion features and the procedural bias separately. Specifically, we employ Total Variation with a non-negative constraint for the parameter associated with degenerated and spatially coherent lesions; on the other hand, we impose ℓ_1 sparsity with a non-positive constraint on the parameter related to enlarged and scatterly distributed procedural bias. We theoretically show that our method enjoys model selection consistency and ℓ_2 consistency in estimation. The utility of our method is demonstrated by improved prediction power and interpretability in the early prediction of Alzheimer's Disease.

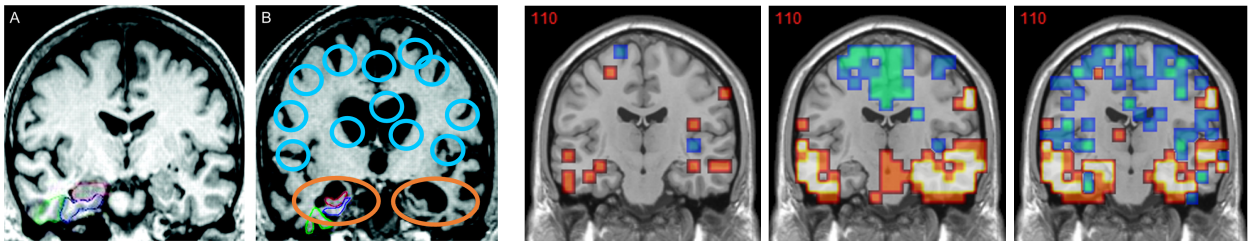
Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.
*Corresponding author

1 INTRODUCTION

Early diagnosis is important for progressive dementias (Chouliaras and O'Brien, 2023), such as Alzheimer's Disease (AD), and Vascular Dementia. To achieve this goal, many machine learning methods based on voxel-based neuroimaging data (*e.g.*, structural Magnetic Resonance Imaging (sMRI) images) have been proposed (De Martino et al., 2008; Liu et al., 2012; Xin et al., 2014). Most of these methods with sparse linear models focus on lesion features that are related to pathological changes. A typical example is Alzheimer's Disease, in which lesions refer to degenerated voxels in early damaged regions such as the two-sided hippocampus (Mu and Gage, 2011).

In the study by Sun et al. (2017), procedural bias was identified as a result of widely used but imperfect segmentation and spatial normalization, leading to results from improperly enlarged gray matter voxels after preprocessing. Although mixed with information about gray matter atrophy, this bias occurs frequently in regions with enlarged lateral ventricles. The sparsity distribution of procedural bias differs from the sparsity distribution of lesions: lesions tend to cluster geometrically, whereas procedural bias is dispersed. Take Alzheimer's Disease in Fig. 1 (a) as an example. The lesions are spatially clustered into the two-sided hippocampus (marked by orange circles) which is believed to degenerate during the progression of AD (Mu and Gage, 2011). On the other hand, the procedural bias is located at the edges/space of the gyrus (marked by blue circles), which is caused by the enlargement of cerebral spinal fluid (CSF) space (Ashburner and Friston, 2001). The lesions are spatially coherent, whereas they are more dispersedly distributed.

The presence of such heterogeneous sparsity poses a challenge for existing sparse optimization methods, as they may struggle to accurately identify either lesions or procedural bias, resulting in interpretability issues and a decline in prediction performance. Specifically, as shown in Fig. 1 (b), the ℓ_1-sparse method such as Lasso (Tibshirani, 1996; Liu et al., 2012) or Elastic



(a) Normal vs. AD (b) Lasso (c) GenLasso (d) Ours

Figure 1: Visualization of lesions and the procedural bias in AD. (a) Left: normal; right: AD, with the locations of lesions and the procedural bias respectively marked by orange and blue circles. (b)-(d): selected lesions (orange) and procedural bias (blue) of Lasso, GenLasso, and our method.

Net lacks the ability to capture lesions with spatially clustered structure (Zou and Hastie, 2005); while the Total Variation based method such as GenLasso (Xin et al., 2014, 2015; Tibshirani and Taylor, 2011) may be too spatially smoothed to accurately identify the procedural bias, as shown in Fig. 1 (c).

To well capture both features, we propose a general framework in neuroimaging analysis, dubbed *Heterogeneous Linearized Bregman Iteration* from the perspective of differential inclusion, which generates a regularized solution path on a couple of parameters enforced with heterogeneous sparsity. Motivated by the belief that lesions and procedural bias are positively and negatively correlated, respectively, with the disease status (Sun et al., 2017), we model these feature types as the positive and negative components of the ground-truth coefficient vector in the generalized linear model. To simultaneously estimate these features, we split the original single parameter into the addition of two separate parameters, applying Total Variation sparsity to capture spatial smoothness among lesions and ℓ_1 sparsity to represent sparse procedural bias. To implement this, we propose a differential inclusion-based method that is provably consistent in estimation.

To demonstrate the utility, we apply our method to the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Petersen et al., 2010). In addition to the improved prediction power, we achieve better interpretability by observing in Fig. 1 (d) that the lesions and the procedural bias selected by our methods are respectively located in early degeneration regions and enlarged locations such as the edges of the gyrus, the surrounding of the ventricular, etc.

Our contribution is summarized as follows:

- **Methodologically**, we propose a differential inclusion-based method called *Heterogeneous Linearized Bregman Iteration* (Heter-LBI) that enforces heterogeneous sparsity to capture lesion features and procedural bias.
- **Theoretically**, we demonstrate the existence and uniqueness within our differential inclusion framework. Based on this foundation, we further establish consistency results in estimation.
- **Experimentally**, we achieve better prediction power and interpretability on Alzheimer's Disease.

2 RELATED WORKS

Early Prediction in Neurodegenerative Disease. Sparse learning-based methods have been widely applied in the diagnosis and prediction of neurological diseases. For example, ASD-SAENet (Almuqhim and Saeed, 2021) uses a sparse autoencoder to classify autism spectrum disorder from fMRI data by simultaneously optimizing both reconstruction and classifier errors during training. Similarly, HB-DFL introduced by Ke et al. (2023), a deep factor learning model that uses CNNs to extract stable, low-dimensional features from tensor-structured neuroimaging data, achieving superior classification of Parkinson's Disease and ADHD compared to state-of-the-art methods.

Notably, prior to the onset of clinical symptoms, many neurodegenerative diseases such as AD experience degeneration in gray matter voxels, typically regarded as imaging biomarkers in existing methods like Lasso, Elastic Net and GenLasso. As reported by Sun et al. (2017), "Procedural Bias" refers to mistakenly enlarged gray matter voxels during the preprocessing stage, which are believed to be scatterly distributed and can further enhance prediction power. However, existing methods failed to account for the heterogeneity between degenerated voxels and procedural bias, leading to inaccurate feature selection and prediction loss. GenLasso only considered to learn lesion features, which are believed to be spatial coherent and negatively correlated with the disease label. By incorporating total variation regularization and non-

negativity constraints, it selects lesions that are more interpretable than those identified by Lasso. Without non-negativity constraint, the Lasso and GSplit LBI (Sun et al., 2017) can capture both procedure bias and lesion features. However, both of these methods employ uniform sparsity constraint for procedural bias and lesion features, thus it is not efficient in capturing heterogeneity. Specifically, the ℓ_1 sparsity of Lasso is not appropriate for capturing spatially coherent lesion features; while the total variation regularization in GSplit LBI may overly smooth out the procedural bias that is not as spatially coherent as lesions. **In this paper**, we propose a unified sparse learning framework, where heterogeneous sparse penalties are enforced to capture both types of features.

Differential Inclusion in Sparse Recovery. Our method is based on differential inclusion with sparse penalties, which has been proposed in Osher et al. (2016) for variable selection and can be viewed as the dynamic limit of *Linearized Bregman Iteration* (Yin et al., 2008; Osher et al., 2005) that was originally proposed in image denoising. Similar to Lasso, it enjoys model selection consistency in sparse recovery but is more efficient than Lasso in generating the regularized solution path. Further, Huang et al. (2016) extended this method to Total Variation. **Different from these works** that focus on a single type of sparsity, our method can simultaneously enforce heterogeneous sparse penalty types and unify them into a single model. Furthermore, our method also enjoys consistency in estimation.

3 METHODOLOGY

Problem Setup of Heterogeneous Sparsity. Suppose we have $\{x_i, y_i\}_{i=1}^n$ with y_i denoting the cognitive status (larger y indicates better cognitive status) and $x_i \in \mathbb{R}^p$ representing gray matter volume of p voxels, which are *i.i.d* generated from the following generalized linear model:

$$p(y|x,\beta^\star) \propto \exp\left(\frac{\langle x,\beta^\star\rangle \cdot y - \psi\left(\langle x,\beta^\star\rangle\right)}{d(\sigma)}\right), \quad (1)$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is link function and $d(\sigma)$ is known parameter related to the variance of distribution. Eq. (1) degenerates to the linear model when y denotes the continuous cognitive score, or to the logistic model when y denotes disease label. Denote $\beta^{*,+} := \max(\beta^*, 0)$ and $\beta^{*,-} := \max(-\beta^*, 0)$ as the positive and negative parts of β^*, which respectively correspond to lesion features and procedural bias (Sun et al., 2017). That is, a decrease in the volume of lesion voxels (or an increase in the volume of procedural bias voxels) is associated with a lower cognitive score. Besides, they are believed to be heterogeneous in spatial coherence. Formally speaking, we assume

$$\textbf{\textit{Lesion:}}\ \gamma^{*,+,I} := \beta^{*,+} \text{ and } \gamma^{*,+,D} := D_G\beta^{*,+} \text{ are sparse,}$$
$$S^{+,I} := \mathrm{supp}(I_p\beta^{*,+}),\ S^{+,D} := \mathrm{supp}(D_G\beta^{*,+}); \quad (2a)$$
$$\textbf{\textit{Procedural bias:}}\ \gamma^{*,-} := -\beta^{*,-} \text{ is sparse,}$$
$$S^{-} := \mathrm{supp}(\gamma^{*,-}), \quad (2b)$$

where $D_G \in \mathbb{R}^{m\times p}$ denotes the graph difference matrix according to $G := (V, E)$ ($m := |E|$, $(i,j) \in E$ if voxels i and j are adjacent) such that $(D_G\beta)_{i,j} = \beta_i - \beta_j$. Let $D := (I_p^\top, D_G^\top)^\top$, we denote $\gamma^{*,+} := D\beta^{*,+} = \left((\gamma^{*,+,I})^\top, (\gamma^{*,+,D})^\top\right)^\top$. We first assume that both $\beta^{*,+}$ and $\beta^{*,-}$ are sparse since only a few of them are correlated with the disease. Additionally, the lesion vector $\beta^{*,+}$ satisfies spatial smoothness: $D_G\beta^{*,+}$ is sparse. Besides, we assume the lesion parameter $\beta^{*,+}$ (*resp.* the procedural bias parameter $-\beta^{*,-}$) is non-negative because the decrease (*resp*, increase) of the volume in lesion voxels (*resp*, procedural bias) leads to a lower cognitive score. Note that such priors generally hold for many neurological diseases. **Our goal** is recover S^+ and S^- and estimate $\beta^{*,+}$ and $-\beta^{*,-}$.

Notations. We denote $\gamma^* := \left((\gamma^{*,+})^\top, (\gamma^{*,-})^\top\right)^\top$ as the concatenation of $\gamma^{*,+}$ and $\gamma^{*,-}$ in Eq. (2). Correspondingly, we denote $S := \mathrm{supp}(\gamma^*)$. Similarly, we denote $\widehat{\beta}^* := \left((\beta^{*,+})^\top, (-\beta^{*,-})^\top\right)^\top$ as the concatenation of $\beta^{*,+}$ and $-\beta^{*,-}$. For simplicity, we use β^* to represent $\widehat{\beta}^*$. Besides, we denote $X := (x_1, ..., x_n)^\top$ and $y := (y_1, y_2, ..., y_n)^\top$ as the concatenation of $\{x_i\}_i$ and $\{y_i\}_i$. Let λ_A, Λ_A as the non-zero minimum and maximum singular value of some matrix A. We use $\overline{a}$ to denote the average of the vector a. Let $[m_1 : m_2] := \{m_1, m_1+1, ..., m_2\}$ for any integers $m_2 > m_1$. We then define $\mathcal{G}^{+,I} := [1 : p]$, $\mathcal{G}^{+,D} := [p+1 : p+m]$ and $\mathcal{G}^- := [m+p+1 : m+2p]$ as the set of row indexes in $\gamma^{+,I}$, $\gamma^{+,D}$ and γ^-. We denote $S_1 \backslash S_2 := S_1 \cap S_2^c$ for two sets S_1, S_2.

3.1 Heterogeneous Linearized Bregman Iteration.

To learn (β^*, γ^*) in Eq. (1), (2), we consider the following objective function:

$$\ell_\nu(\beta,\gamma) := \ell(\tilde{X}\beta) + \frac{1}{2\nu}\|\tilde{D}\beta - \gamma\|_2^2 \ (\nu > 0),$$

where $\ell(\tilde{X}\beta) := -\frac{1}{n}\sum_{i=1}^n \log p(y_i|\tilde{x}_i^\top\beta)$ denotes the negative log-likelihood of the model in Eq. (1) with

$$\tilde{X} = (X, -X) \in \mathbb{R}^{n\times 2p},\ \tilde{D} := \begin{pmatrix} D & 0 \\ 0 & -I_p \end{pmatrix} \in \mathbb{R}^{(m+2p)\times p}$$

such that $\beta := \left((\beta^l)^\top, (\beta^p)^\top\right)^\top \in \mathbb{R}^{2p}$ and $\gamma := \left((\gamma^l)^\top, (\gamma^p)^\top\right)^\top \in \mathbb{R}^{m+2p}$, where (β^l, β^p) and (γ^l, γ^p)

are expected to estimate $(\beta^{*,+}, \beta^{*,-})$ and $(\gamma^{*,+}, \gamma^{*,-})$, respectively. The $\frac{1}{2\nu}\|\tilde{D}\beta - \gamma\|_2^2 := \frac{1}{2\nu}\|D\beta^l - \gamma^l\|_2^2 + \frac{1}{2\nu}\|\beta^p - \gamma^p\|_2^2$ denotes the variable splitting term, which has been introduced in ADMM (Boyd et al., 2011) and Split Bregman (Ye and Xie, 2011; Huang et al., 2016) to implement sparsity on γ. Particularly, Huang et al. (2016) and Huang et al. (2020) showed that this term can improve model selection consistency. Further, Sun et al. (2017) showed that this variable splitting can give β more degree of freedom to capture the procedural bias. To implement, we consider the following differential inclusion:

$$0 = -\nabla_\beta \ell(\beta(t), \gamma(t)), \tag{3a}$$
$$\dot{v}(t) = -\nabla\gamma\ell(\beta(t), \gamma(t)), \tag{3b}$$
$$v(t) \in \partial_\gamma f(\gamma(t)), \tag{3c}$$

where $\beta(0) = 0, v(0) = \gamma(0) = 0$ and $f(\gamma) := \|\gamma\|_1 + \mathbb{1}(\gamma^{l,I} \geq 0) + \mathbb{1}(\gamma^p \leq 0)$. Eq. (3) can be viewed as an extension of Split Bregman Inverse Scale Space (Split ISS) in Huang et al. (2016) to the *heterogeneous sparsity* of $\gamma^{*,+} := D\beta^{*,+}$ and $\gamma^{*,-} = -\beta^{*,-}$ that are additionally constrained with $\gamma^{l,I} \geq 0$ and $\gamma^p \leq 0$ to well capture $\beta^{*,+}$ and $\beta^{*,-}$, instead of only $f(\gamma) := \|\gamma\|_1$ in Huang et al. (2016). Therefore, we call Eq. (3) as *Heterogeneous Inverse Scale Space* (Heter-ISS). With such a newly defined $f(\gamma)$, the form of $v \in \partial f(\gamma)$ can be written as:

$$v = \rho + g: \ \rho \in \partial\|\gamma\|_1 \text{ and} \tag{4}$$
$$g \in \partial\left(\mathbb{1}(\gamma^{l,I} \geq 0) + \mathbb{1}(\gamma^p \leq 0)\right), \tag{5}$$

where $\rho \in [-1, 1]^{m+2p}$ and $g := \left((g^{l,I})^\top, 0_m^\top, (g^p)^\top\right)^\top$ such that $g^{l,I} \leq 0$ and $g^p \geq 0$ according to the definition of subgradient. Starting from $\beta(0) = 0$ and $\gamma(0) = 0$, Eq. (3) generates a regularization solution path $(\beta(t), \gamma(t))$ from sparse to dense. Specifically, as t grows until some $t_i > 0$ when $v_i^{l,I}(t_i)$ (*resp.* $|v_i^{l,D}(t_i)|$ and $v_i^p(t_i)$) reaches 1 (*resp.* $|v_i^{l,D}(t_i)| = 1$ and $v_i^p(t_i) = -1$), the corresponding element $\gamma_i^{l,I}(t_i)$ (*resp.* $\gamma_i^{l,D}(t_i)$ and $\gamma_i^p(t_i)$) will be selected as non-zero. Besides, as will be shown in Sec. 4, this solution path enjoys model selection consistency in estimation. Further, equipped with an early-stopping strategy, the solution can recover γ^* and S. Since $\beta^{*,+} = \gamma^{*,+}$ and $-\beta^{*,-} = \gamma^{*,-}$, this means we can recover the lesion features and the procedural bias.

To implement efficiently, we adopt the *Linearization* method (Osher et al., 2016; Yin et al., 2008) for discretization.

Heter-LBI via Discretization of Eq. (3). Specifically, we first append an ℓ_2 norm to $f(\gamma)$ make the penalty strongly convex:

$$f_\kappa(\gamma) := \|\gamma\|_1 + \mathbb{1}(\gamma^{l,I} \geq 0) + \mathbb{1}(\gamma^p \leq 0) + \frac{1}{2\kappa}\|\gamma\|_2^2,$$

where κ is large enough to approximate $f(\gamma)$. With such a $f_\kappa(\gamma)$, we have the following differential inclusion:

$$\frac{\dot{\beta}(t)}{\kappa} = -\nabla_\beta \ell(\beta(t), \gamma(t)), \tag{6a}$$
$$\dot{z}(t) = -\nabla\gamma\ell(\beta(t), \gamma(t)), \tag{6b}$$
$$z(t) \in \partial_\gamma f_\kappa(\gamma). \tag{6c}$$

According to $f_\kappa(\gamma)$, we have $z(t) := v(t) + \frac{\gamma(t)}{\kappa}$. Therefore, we can obtain γ from z as:

$$\gamma = \kappa \text{prox}_{f(\gamma)}(z) := \kappa \arg\min_\gamma \frac{1}{2}\|\gamma - z\|_2^2 + f(\gamma). \tag{7}$$

The solution to Eq. (7) is $\gamma^{l,I} = \kappa \max(z^{l,I} - 1, 0)$, $\gamma^{l,D} = \kappa \text{sign}(z^{l,D}) \max(|z^{l,D}| - 1, 0)$, and $\gamma^p = \kappa \max(-z^{p,I} + 1, 0)$. We call Eq. (6) as *Heterogeneous Linearized Bregman Inverse Scale Space* (Heter-LBISS), which can be discretized with step size α, leading to the iteration referred to as *Heter Linearized Bregman Iteration* (Heter-LBI) below:

$$\beta_{k+1} - \beta_k = -\alpha\kappa\nabla_\beta \ell(\beta_k, \gamma_k), \tag{8a}$$
$$z_{k+1} - z_k = -\alpha\nabla_\gamma \ell(\beta_k, \gamma_k), \tag{8b}$$
$$\gamma_{k+1} = \kappa \text{prox}_{f(\gamma)}(z_{k+1}). \tag{8c}$$

Sparse estimators via graph projection. To obtain sparse parameter $\tilde{\beta}_k := [(\tilde{\beta}_k^l)^\top, (\tilde{\beta}^p)_k^\top]^\top$ at k-th step, we project the dense parameter β_k onto the subspace of $\text{supp}(\gamma_k)$ via the following optimization:

$$\tilde{\beta}_k^l := \underset{\substack{x \geq 0, x(\mathcal{G}^{+,I} \setminus S_k^{l,I}) = 0, \\ D_G(\mathcal{G}^{+,D} \setminus S_k^{l,D}, :)x = 0}}{\text{argmin}} \|x - \beta_k^l\|_2, \tag{9}$$

$$\tilde{\beta}_k^p := \underset{x \geq 0, x(\mathcal{G}^- \setminus S_k^p) = 0}{\text{argmin}} \|x - \beta^p(t)\|_2. \tag{10}$$

It is easy to solve $\tilde{\beta}_k^p(i)$ from Eq. (10) as $\tilde{\beta}_k^p(i) = 0$ for all $i \in \mathcal{G}^- \setminus (S_k^p \cap \{i : \beta_k^p(i) \geq 0\})$; $= \beta_k^p(i)$ otherwise. Regarding $\tilde{\beta}_k^l$, the following proposition offers an effective approach by leveraging the connected components of the graph $G := \left(V, E_{\mathcal{G}^{+,D} \setminus S_k^{l,D}}\right)$.

Proposition 3.1. *Denote $G := (V, E_{\mathcal{G}^{+,D} \setminus S_k^{l,D}})$ where $V := \{1, ..., p\}$ and $E_{\mathcal{G}^{+,D} \setminus S_k^{l,D}}$ denotes the edge set induced by $\mathcal{G}^{+,D} \setminus S_k^{l,D}$: $(i, j) \in \mathcal{G}^{+,D} \setminus S_k^{l,D} \implies \tilde{\beta}_k^l(i) = \tilde{\beta}_k^l(j)$. If $G := G_1 \cup ... \cup G_C$ contains C connected components with $G_k := (V_c, E_c)$, then for each V_c and $j \in V_c$:*

$$\tilde{\beta}_k^l(j) = \begin{cases} 0 & V_c \cap \left(\mathcal{G}^{+,I} \setminus S_k^{l,I}\right) \neq \emptyset \\ \overline{\beta_k^l(V_c)} \mathbb{1}\left(\overline{\beta_k^l(V_c)} \geq 0\right) & V_c \cap \left(\mathcal{G}^{+,I} \setminus S_k^{l,I}\right) = \emptyset \end{cases}. \tag{11}$$

Remark 3.2. The proof of Prop. 3.1 is left to Appx. 6. Intuitively, it exploits the graph structure related

to D_G. Specifically, we first note that for each $(i,j) \in E_{\mathcal{G}^{+,D}\setminus S_k^{l,D}}$, $D_G(\mathcal{G}^{+,D}\setminus S_k^{l,D},:)x = 0$ means $(D_G x)_{(i,j)} = x_i - x_j = 0$. Therefore, the nodes in the same connected component (say V_c) share the same value, which is either 0 if V_c is overlapped with $\mathcal{G}^{+,I}\setminus S_k^{l,I}$ (which is the set that enforced to be 0 according to Eq. (9)) otherwise is the average of $\beta_k^l(V_c)$ if this average is greater than 0.

The complexity of calculating $\tilde{\beta}^l(t)$ is $O(p^2+m)$, which is comparable to the gradient descent step in Eq. (8a). Therefore, the projection given by Eq. (33) can be efficiently incorporated in Eq. (8). However, for the resulting solution path $(\beta_k, \tilde{\beta}_k, \gamma_k)$, several questions remain regarding estimation properties. First, **i)** given that $\tilde{X} := [X, -X]$ is not full-column rank, are $\beta^{*,+}$ and $-\beta^{*,-}$ identifiable? Additionally, **ii)** does a solution of Heter (LB)ISS (Eq. (3), (6)) exists, and is it unique? Finally, **iii)** if the answer of **ii)** is yes, then can this unique solution well estimate the lesion features $\beta^{*,+}$ and the procedural bias $-\beta^{*,-}$? We provide our answers to these questions in the subsequent section.

4 THEORETICAL ANALYSIS

In this section, we show the consistency of *Heter-LBI* in estimating β^* under the linear model when $\ell(X\beta) := \frac{1}{2n}\|y - X\beta\|_2^2$. Specifically, we first introduce some basic assumptions to ensure the identifiability of β^* in Sec. 4.1. Then we establish the solution existence and uniqueness of Heter-ISS (Eq. (3)) and Heter-LBISS (Eq. (6)) in Sec. 4.2, followed by the consistency results in Sec. 4.3.

4.1 Assumptions for Identifiability of β^*

In this section, we introduce some basic assumptions to identify $(\beta^*, \gamma^{*,+,I}, \gamma^{*,+,D}, \gamma^{*,-})$. Note that γ^* only being nonzeros on $S := \mathrm{supp}(\gamma^*)$. Therefore, the identifiability of (β^*, γ^*) is equivalent to the identifiability of the $S := \mathrm{supp}(\gamma^*)$ and parameters restricted on the sparse subspace (β, γ_S), which can be respectively guaranteed by the *Irrepresentable Condition* (IRC) (assum. 4.1) and the *Restricted Convexity Condition* (RSC) (assum. 4.3).

Denote $H(\nu) := \nabla^2 \ell_\nu(\beta, \gamma)$ as the Hessian matrix of $\ell_\nu(\beta, \gamma)$ on the vector $(\beta, \gamma^{l,I}, \gamma^{l,D}, \gamma^p)$:

$$H(\nu) = \begin{pmatrix} \tilde{X}^*\tilde{X} + \tilde{D}^\top\tilde{D}/\nu & -\tilde{D}^\top/\nu \\ -\tilde{D}/\nu & I_m/\nu \end{pmatrix},$$

where $H_{\beta,:}(\nu)$, $H_{\mathcal{G}^{+,I},:}$, $H_{\mathcal{G}^{+,D},:}$, $H_{\mathcal{G}^-,:}$ (*resp.*, $H_{:,\beta}(\nu)$, $H_{:,\mathcal{G}^{+,I}}$, $H_{:,\mathcal{G}^{+,D}}$, $H_{:,\mathcal{G}^-}$) respectively denotes the rows (*resp.*, columns) of $H(\nu)$ corresponding to β, $\gamma^{l,I}$, $\gamma^{l,D}$ and γ^p. Besides, let $H_{(\beta,S),(\beta,S)}(\nu)$ denote the submatrix of $H(\nu)$ restricted on $(\beta, S = S^{+,I} \cup S^{+,D} \cup S^-)$.

Now we are ready to introduce the IRC condition as follows:

Assumption 4.1 (Irrepresentable Condition (IRC))**.** Denote $H_0 := H_{(\beta,S),(\beta,S)}$. There is a constant $\eta \in (0, 1]$ such that

$$\begin{aligned}
\mathrm{IRC}^{+,I}(\nu) &:= \sup_{v\in\mathcal{V}} \max_{i\in\mathcal{G}^{+,I}\setminus S^{+,I}} H_{i,(\beta,S)}(\nu) H_0^\dagger(\nu) \begin{pmatrix} 0_p \\ v \end{pmatrix} \\
&\le 1-\eta, \\
\mathrm{IRC}^{+,D}(\nu) &:= \sup_{v\in\mathcal{V}} \left\| H_{\mathcal{G}^{+,D}\setminus S^{+,D},(\beta,S)}(\nu) H_0^\dagger(\nu) \begin{pmatrix} 0_p \\ v \end{pmatrix} \right\|_\infty \\
&\le 1-\eta, \\
\mathrm{IRC}^{-}(\nu) &:= \sup_{v\in\mathcal{V}} \min_{i\in\mathcal{G}^{-}\setminus S^{-}} H_{i,(\beta,S)}(\nu) H_0^\dagger(\nu) \begin{pmatrix} 0_p \\ v \end{pmatrix} \\
&\ge -1+\eta,
\end{aligned}$$

where $\mathcal{V} := (\infty, 1]^{|s^{+,I}|} \times [-1, 1]^{|s^{+,D}|} \times [-1, \infty)^{|s^-|}$ as the dual space s.t. $v_S \in \mathcal{V}$ for each $v \in f(\gamma)$.

Remark 4.2. We will show in Appx. 1 that this IRC condition can hold as long as ν is large enough.

Compared to the irrepresentable condition with single type of sparse penalty (Zhao and Yu, 2006; Osher et al., 2016; Vaiter et al., 2012), our condition is not only imposed on $S^{+,I}$ and $S^{+,D}$ that are required to be identifiable from $\mathcal{G}^{+,I}\setminus S^{+,I}$, $\mathcal{G}^{+,D}\setminus S^{+,D}$ to identify lesion features that are related to S^+; but also on S^- that should be identifiable from $\mathcal{G}^-\setminus S^-$ to identify procedural bias in S^-. Moreover, different from the previous methods that define over the supreme of subgradient of ℓ_1 norm, our assumption defines over $\mathcal{V}$ as dual space given by $\partial f(\gamma)$.

After identifying the true signal set $S := S^{+,I} \cup S^{+,D} \cup S^-$, we additionally need the following RSC condition to ensure the identifiability of (β^*, γ^*).

Assumption 4.3 (Restricted Strongly Convexity (RSC))**.** There is a constant $\lambda_H > 0$ such that

$$\left(\beta^T, \gamma_S^T\right) H_{(\beta,S),(\beta,S)} \begin{pmatrix} \beta \\ \gamma_S \end{pmatrix} \ge \lambda_H \left\| \begin{pmatrix} \beta \\ \gamma_S \end{pmatrix} \right\|_2^2. \quad (12)$$

This RSC assumption has been similarly assumed in sparse learning (Osher et al., 2016; van de Geer and Bühlmann, 2009), implies that $\ell(\beta, \gamma)$ is strongly convex restricted to the sparse subspace of γ. Indeed, we can show that this assumption is equivalent to the one in the linear model where the Hessian is $X^\top X$ (Osher et al., 2016).

Proposition 4.4. *The RSC condition in assum. 4.3 holds if and only if there exists a constant $\lambda_X > 0$ such that $X_{S^\beta}^\top X_{S^\beta} \succeq \lambda_X$, with $S^\beta := \mathrm{supp}(\beta^*)$.*

Remark 4.5. The proof is left in Appx. 2. Intuitively, Prop. 4.4 holds because D contains the identity matrix, which hence mimics the RSC behavior of ℓ_1 sparsity.

4.2 Solution Existence and Uniqueness

In this section, we show that the solution of Heter ISS (Eq. (3)) and Heter LBISS (Eq. (6)) exists and is unique.

Proposition 4.6 (Solution existence and uniqueness of Heter (LB)ISS)**.** *With $v(t), \rho(t), g(t)$ in Eq.* (4)*, we have*

1. *For Heter-ISS in Eq.* (3)*, assume that $\rho(t), g(t)$ are right continuously differentiable; and $\beta(t), \gamma(t)$ are right continuous. Then the solution exists. Besides, $v(t)$ is unique. If additionally $H_{(\beta,S(t)),(\beta,S(t))} \succ 0$ with $S(t) := \mathrm{supp}(\gamma(t))$ for some $0 \leq t \leq \tau$, $(\beta(t), \gamma(t))$ is unique when $0 \leq t \leq \tau$.*
2. *For Heter-LBISS in Eq.* (6)*, assume that $v(t), \beta(t)$ are right continuously differentiable, then there is a unique solution for $t \geq 0$.*

According to assum. 4.3, Prop. 4.6 implies that $(\beta(t), \gamma(t))$ is unique under *no-false-positive* property, *i.e.*, $S(t) \subset S$ when $t < \tau$ for some $\tau > 0$. In the next section, we will first show the *no-false-positive* of Heter (LB) ISS and moreover the recovery of S at some $\bar{\tau} < \tau$; then, once S is recovered, we further show that ℓ_2 consistency of $(\beta(\bar{\tau}), \gamma_S(\bar{\tau}))$.

4.3 Consistency of Heter-LBI

In this section, we establish the consistency of HLBI in Eq. (8), including *no-false-positive*, *sign-consistency* and ℓ_2*-consistency* in the following theorem.

Theorem 4.7 (Consistency of Decomposed LBI)**.** *Under IRC 4.1 and RSC 4.3 assumptions and $\tilde{\beta}$ defined in Eq.* (33)*, suppose κ is large enough to satisfy*

$$\kappa \geq \frac{2}{\eta}\left(1+\frac{2}{\lambda_{\tilde{D}}}\right)\left(1+\sqrt{\frac{2\left(1+\nu\Lambda_{\tilde{X}}^2+\Lambda_{\tilde{D}}^2\right)}{\lambda_H \nu}}\right) \cdot \left(\left(1+\Lambda_{\tilde{D}}\right)\|\beta^\star\|_2 + \frac{2\sigma}{\lambda_H}\left(\frac{\Lambda_{\tilde{X}}}{\lambda_{\tilde{D}}}+\frac{\Lambda_{\tilde{X}}}{\lambda_{\tilde{D}}^2}\right)\right),$$

and the step size α satisfies $\kappa\alpha\|H\|_2 < 2$. Define $\bar{\tau} := \frac{\eta}{8\sigma} \cdot \frac{\lambda_{\tilde{D}}}{\Lambda_{\tilde{X}}}\sqrt{\frac{n}{\log(m+2p)}}$. Then at $K := \lfloor \frac{\bar{\tau}}{\alpha} \rfloor$, with probability not less than $1 - \frac{8}{m+2p} - 3\exp(-4n/5)$, we have:

1. ***No-false-positive:*** *The solution has no false-positive, i.e.* $\mathrm{supp}(\gamma_k) \subseteq S$*, for $0 \leq k\alpha \leq \bar{\tau}$.*
2. ***Sign consistency of γ_K:*** *Once the signal is strong enough such that ($s := |S|$)*

$$\gamma_{\min}^\star \succeq O\left(\frac{M\log s}{\lambda_H}\sqrt{\frac{\log(m+2p)}{n}}\right) \tag{13}$$

for some constant $M > 0$, then γ_k has sign consistency at K, i.e. $\mathrm{sign}(\gamma_K) = \mathrm{sign}\left(\tilde{D}\beta^\star\right)$.

3. *ℓ_2* ***consistency of β_K:***

$$\left\|\tilde{\beta}_K - \beta^\star\right\|_2 \preceq O\left(\frac{M}{\lambda_H}\sqrt{\frac{s\log(m+2p)}{n}}\right).$$

Since $\gamma^{*,+}$ is composed of $\gamma^{*,+,I} = \beta^{*,+}$ and $\gamma^{*,-} = -\beta^{*,-}$, item 2 means we can identify the lesion and the procedural bias. By projecting β_K into the sparse subspace of γ_K, item 3 further shows the ℓ_2 consistency of $\tilde{\beta}_K$. Besides, this theorem informs us about the hyperparameter selection that will be discussed in the following.

Choosing the hyperparameters. Since the solution path generated by Eq. (6) is from sparse to dense, the early stopping mechanism is required to avoid overfitting. To choose the stopping time t, we implement cross-validation. For ν, it is a trade-off between precision and the recall of feature selection. On the one hand, the IRC condition can be easier to satisfy as ν increases (see remark 4.2); on the other, according to Eq. (13), the $\gamma_{\min}^*$ is inversely scaled λ_H that increases w.r.t. ν; therefore, the overshooting of ν can miss some features. Empirically, we find $0 < \nu < 1$ can work well. The step size α and κ should satisfy $\kappa\alpha\|H\|_2 < 1$ to satisfy the consistency.

5 NUMERICAL EXPERIMENTS

In this section, we apply our method to synthetic data. To measure the model selection consistency of Heter-LBI, we calculate the Area Under Curve (AUC).

Data Generation. We set $n = 500$, $p = h \times w = 1600$ with $h = w = 40$. We first generate the matrix form of the signal, *i.e.*, $B^* \in \mathbb{R}^{h\times w}$ such that $B^*(i,j) = 1$ if $21 \leq i \leq 27$, $21 \leq j \leq 27$; $= -1$ if i, j are odd numbers less than 15; otherwise $= 0$. For D_G with $G := (\mathcal{V}, \mathcal{E})$, we set $\mathcal{E} := \{((i_1,j_1),(i_2,j_2)) : i_1, i_2 \leq h, j_1, j_2 \leq w, |i_1 - i_2| + |j_1 - j_2| = 1\}$. We then vectorize B^* as $\beta^* \in \mathbb{R}^{p\times 1}$, with $\beta^{*,+} := \max(\beta^*, 0)$ and $\beta^{*,-} := \max(-\beta^*, 0)$. Then we set $\gamma^{*,+} := [I, D_G^\top]^\top \beta^{*,+}$ and $\gamma^{*,-} := \beta^{*,-}$. We generate $X \in \mathbb{R}^{n\times p}$ where $X_{i,j} \sim_{i.i.d} \mathcal{N}(0,1)$. We fix X and generate $y = X\beta^* + \varepsilon$, where $\varepsilon_i \sim_{i.i.d} \mathcal{N}(0,1)$. We also used Lasso and GenLasso as baseline methods.

Implementation Details. We run Heter-LBI with $\kappa = 10$, $\nu = 0.2$ and $\alpha = \frac{1}{\kappa\|H\|_2}$. To calculate AUC,

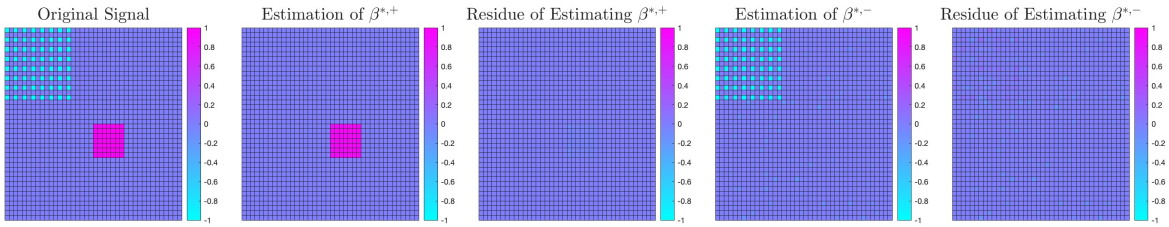


Figure 2: Illustration of estimation errors of $\beta^{*,+}$ and $\beta^{*,-}$. From left to right: original feature β^*, with $\beta^{*,+}$ marked by red and $\beta^{*,-}$ marked by red; estimated lesions $\tilde{\beta}^l$; residue of lesions $\beta^{*,+} - \tilde{\beta}^l$; estimated procedural bias $\tilde{\beta}^p$; residue of procedural bias $\beta^{*,-} - \tilde{\beta}^p$.

we record the selection time $t_\gamma := [t_1, ..., t_m]$ with $t_i := \min_{t>0}\{\gamma_i(t) \neq 0\}$ and the selection order of each element according to t_γ, *i.e.*, $\pi_\gamma := [i_1, ..., i_m]$ such that $t_{i_1} \leq t_{i_2} \leq ... \leq t_{i_m}$. Then the AUC of π_γ and the Oracle selection vector $[\mathbb{1}(i_1 \in S), ..., \mathbb{1}(i_m \in S)]$ defined by S can measure the goodness of variable selection in the solution path. That is, the higher the AUC implies that variables in S are selected earlier than variables in S^c.

Table 1: Mean and the std of AUC over 20 times for Heter-LBI.

	S^+	S^-	S	S^β
mean	0.9998	0.9849	0.9936	0.9608
std	0.0000	0.0005	0.0002	0.0016

Results Analysis. To remove the randomness effect, we repeat for 20 times. We report the mean and the standard deviation (std) of AUC for $S^+ := \text{supp}(\gamma^{*,+})$, $S^- := \text{supp}(\gamma^{*,-})$, S and $S^\beta := \text{supp}(\beta^*)$ in Tab. 1. As shown, our methods can selects true signals for both γ and β.

To further highlight the advantages of our method, we compared it against Lasso and GenLasso as baselines. Through 20 repeated experiments, we computed the mean and standard deviation of the AUC for S^β. The average AUC values for Lasso and GenLasso are 0.9460 and 0.7728, with standard deviations of 0.0023 and 0.0075, respectively. These results clearly demonstrate the superior performance of our method in variable selection compared to both.

Visualization of Reconstruction. To illustrate the estimation error, we also visualize the original "lesion features" $\beta^{*,+}$ (marked by red) and "procedural bias" $-\beta^{*,-}$ (marked by blue) in Fig. 2. As shown, the lesions are spatially clustered and thus satisfy spatial coherence while the procedural bias is dispersedly distributed and thus is implemented by the ℓ_1 sparsity. Besides, our method can reconstruct both types of features well (the 2nd and 4th columns) with minor residues (the 3rd and the 5th columns).

6 DISEASE PREDICTION OF ALZHEIMER'S DISEASE

In this section, we primarily validate our method on the Alzheimer's Disease Neuroimaging Initiative (**ADNI**) dataset (`http://adni.loni.usc.edu/`).

Data Description and Preprocessing. We incorporated ADNI data from ADNI1, ADNI2, and ADNI GO. Participants were selected based on the availability of baseline T1-weighted MRI scans and corresponding clinical assessments. Our dataset included three groups—Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and Normal Control (NC)—covering the full spectrum of cognitive health from normal aging, through early impairment, to clinically diagnosed AD. We follow Dartel VBM (Ashburner, 2007) to preprocess the ADNI data from two sMRI scan magnetic field strengths, namely 1.5 Tesla sMRI scan (*a.k.a*, 15T) and 3.0 Tesla sMRI scan (*a.k.a*, 30T). The 15T data contains 64 AD, 110 MCI, 90 NC; while the 30T data contains 66 AD, 247 MCI, and 90 NC. Similar to Sun et al. (2017) and Xin et al. (2015), we consider three classification tasks[1]: **i)** 15ADNC, *i.e.*, classify AD/NC on 15T; **ii)** 30ADNC, *i.e.*, classify AD/NC on 30T; **iii)** 15MCINC, *i.e.*, classify MCI/NC on 15T. We use $y = -1$ and $y = 1$ to respectively represent the diseased and non-diseased status.

Compared Baselines. We compare with **i) Support Vector Machine (SVM)**, **ii) Random Forest**, **iii) Lasso** (Tibshirani, 1996; Liu et al., 2012), **iv) Elastic Net** (Zou and Hastie, 2005; Xiao et al., 2021), **v) GenLasso** (Xin et al., 2014) with $D = \left(I_p^\top, \rho D_G^\top\right)^\top$ where D_G corresponds to the Total Variation matrix and $\rho > 0$ balances the spatial smoothness and ℓ_1 sparsity, and **vi) GSplit LBI** (Sun et al., 2017) that introduces a dual parameter such that the sparse parameter

[1]We also predict Alzheimer's Disease Assessment Scale (ADAS) as a regression task in Appx. 8.2.

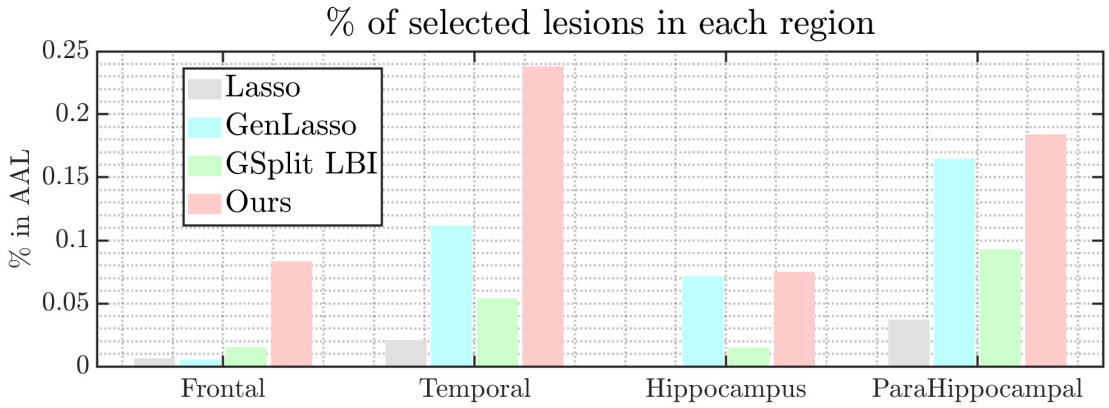


Figure 3: Ratio % of selected lesions in each region R: $\frac{|\text{supp}(\tilde{\beta})\cap R|}{|R|}$. Here, we choose R as the Frontal Lobe, Temporal Lobe, Hippocampus and ParaHippocampal.

Table 2: Classification accuracy of our method and other baselines.

Method \ Setting	15ADNC	30ADNC	15MCINC
SVM	83.10%	87.58%	73.50%
Random Forest	84.00%	84.00%	66.00%
Lasso	88.31%	87.50%	58.00%
Elastic Net	88.31%	87.50%	68.00%
GenLasso	85.06%	86.93%	63.00%
GSplit LBI	88.31%	88.64%	71.50%
Ours	**88.96%**	**91.48%**	**74.00%**

is enforced with spatial smoothness and non-negativity constraint to learn the lesion features and a dense parameter to additionally learn the procedural bias.

Implementation Details. We set $\ell(\beta)$ as the logistic regression loss, and implement Heter-LBI with $\kappa = 5$, $\alpha = \frac{1}{\kappa\|H\|_2}$ and $\nu = 0.1$. The stopping time t is selected according to the cross-validation results for our method and GSplit LBI. For Lasso, Elastic Net, and GenLasso, λ is optimized from $\{0.05, 0.05, ..., 1, 10, 10^2\}$; for Elastic Net, the mixture parameter α is optimized from $\{0, 0.05, 0.1, ..., 1\}$. For GenLasso, GSplit LBI, and our method, the smoothing hyperparameter ρ is optimized from $\{0.5, 1, ..., 10\}$. The SVM parameters are set with C and γ taking values from $\{10^{-5}, 10^{-4}, \ldots, 10^5\}$. In Random Forest, the number of trees ranged from 50 to 500 in steps of 100, and the minimum leaf size ranged from 1 to 5.

Results Analysis. We report the 10-fold cross-validation result in Tab. 2. As shown, our method can outperform others in all tasks. The advantage of our method over Lasso, Elastic Net, and GenLasso may be due to the enforcement of heterogeneous sparsity on the lesions and procedural bias. Although the dense parameter of GSplit LBI can also capture the procedural bias and thus perform better than other baselines, it can also include noise information, which may cause over-fitting of the trained model.

To further evaluate our method under the same parameter settings, we conducted validation on the OASIS dataset (`https://sites.wustl.edu/oasisbrains`), another AD diagnostic dataset. In our experiment, we selected a subset consisting of 82 AD cases and 81 NC cases for the classification task. Our method achieved a classification accuracy of 76.67%, surpassing the Lasso method's accuracy of 75%.

Visualization of Selected Features. To further validate the effectiveness of our method in identifying heterogeneous features, we visualize the selected lesions (marked by orange) and procedural bias (marked by blue) on the task of ADNI in Fig. 1 (b)-(d). Additional visualizations and details can be found in Appx. 8.1. As shown, the lesion features selected by our method are located around the two-side hippocampus, parahippocampal, and medial temporal lobe, which are believed to be the early damaged regions (Mu and Gage, 2011; van Hoesen et al., 2000; Visser et al., 2002); while the procedural bias is located at the edges of the gyrus and the lateral ventricular, which are believed to be enlarged (Ashburner and Friston, 2001). In contrast, the Lasso (*resp.* GenLasso) is too under-smoothed (*resp.* over-smoothed) to capture the lesions (*resp.* procedural bias) well.

To formally measure the effectiveness of selecting lesion features, we calculate the ratio of selected lesions in each region R, $\frac{|\text{supp}(\tilde{\beta})\cap R|}{|R|}$. As shown in Fig. 3, in regions Frontal Lobe (Cajanus et al., 2019), Temporal Lobe, Hippocampus, and ParaHippocampal that have been well-established as the early degenerated regions in AD, the selected lesions by our method account for a larger proportion than others.

Stability of Feature Selection[2]**.** In addition to better interpretability, Fig. 4 - 5 (Appx. 8.1) also shows that our method is more stable in terms of feature selection. To formally assess the stability, we adopt the multi-set Dice Coefficient (mDC) (Xin et al., 2015) of selected features across 10 folds: $\text{mDC} = \frac{10|\cap S_k|}{\sum_{k=1}^{10}|S_k|}$, where S_k denotes the set of selected features in the k-th fold. To measure the stability of the whole feature selection set ($S^\beta := \text{supp}(\beta^*)$), lesion features ($S^{+,I}$), and procedural bias ($S^-$), we respectively report the mDC, mDC^l ("l" stands for lesion), and mDC^p ("p" stands for procedural bias) in Tab. 3. As shown, our method is constantly more stable than other methods.

Table 3: mDC, mDC^l and mDC^p of selected features.

	15ADNC			30ADNC		
	mDC	mDC^l	mDC^p	mDC	mDC^l	mDC^p
Lasso	0.2038	0.1709	0.2941	0.1811	0.2283	0.1071
Elastic Net	0.2038	0.1709	0.2941	4399	0.5073	0.3605
GenLasso	0.3362	0.6728	0.0035	0.5221	0.5725	0.4925
GSplit LBI	0.3240	0.3240	N/A	0.4267	0.4267	N/A
Ours	**0.4542**	**0.5084**	**0.4294**	**0.5757**	**0.5824**	**0.5349**

7 CONCLUSIONS

In this paper, we propose a new method to select lesion voxels and procedural bias with heterogeneous sparsity in neuroimaging analysis. Our method is guided by a differential inclusion whose solution existence and uniqueness are guaranteed. We theoretically show the model selection consistency and ℓ_2 consistency of the solution path in estimating lesion features and the procedural bias. Our method achieves better prediction results for Alzheimer's Disease. Moreover, the identified lesion features are spatially clustered into early damaged regions; while the features belonging to procedural bias are dispersedly distributed in the enlarged space such as lateral ventricles.

Limitations and Future work. Although splitting the original parameter into two parts allows us to identify heterogeneous sparsity, it increases computational cost. Additionally, while our method provides a general framework beyond Alzheimer's Disease, it would be interesting to apply it to other conditions, such as Frontotemporal dementia and Parkinson's disease. Such an exploration and more efficient implementation of our algorithm will be explored in the future.

[2]We only calculate the mDC, mDC^l for the sparse estimator of GSplt LBI, since it does not select procedural bias.

Impact Statements

This paper presents work whose goal is to advance the fields of Sparse Learning and neuroimaging analysis. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (KRH2305058) and the R&D Program of the Beijing Municipal Education Commission (KM202310025015).

References

A Marshall, G., S Zoller, A., Lorius, N., E Amariglio, R., J Locascio, J., A Johnson, K., A Sperling, R., M Rentz, D., Initiative, A. D. N., et al. (2015). Functional activities questionnaire items that best discriminate and predict progression from clinically normal to mild cognitive impairment. *Current Alzheimer Research*, 12(5):493–502.

Aggleton, J. P., Pralus, A., Nelson, A. J., and Hornberger, M. (2016). Thalamic pathology and memory loss in early alzheimer's disease: moving the focus from the medial temporal lobe to papez circuit. *Brain*, 139(7):1877–1890.

Almuqhim, F. and Saeed, F. (2021). Asd-saenet: a sparse autoencoder, and deep-neural network model for detecting autism spectrum disorder (asd) using fmri data. *Frontiers in Computational Neuroscience*, 15:654315.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113.

Ashburner, J. and Friston, K. J. (2001). Why voxel-based morphometry should be used. *Neuroimage*, 14(6):1238–1243.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

Burger, M., Möller, M., Benning, M., and Osher, S. (2013). An adaptive inverse scale space method for compressed sensing. *Mathematics of Computation*, 82(281):269–299.

Cajanus, A., Solje, E., Koikkalainen, J., Lötjönen, J., Suhonen, N.-M., Hallikainen, I., Vanninen, R., Hartikainen, P., de Marco, M., Venneri, A., et al. (2019). The association between distinct frontal brain volumes and behavioral symptoms in mild cognitive

impairment, alzheimer's disease, and frontotemporal dementia. *Frontiers in neurology*, 10:1059.

Chouliaras, L. and O'Brien, J. T. (2023). The use of neuroimaging techniques in the early and differential diagnosis of dementia. *Molecular Psychiatry*, 28(10):4084–4097.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., and Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns. *Neuroimage*, 43(1):44–58.

Huang, C., Sun, X., Xiong, J., and Yao, Y. (2016). Split lbi: An iterative regularization path with structural sparsity. *Advances In Neural Information Processing Systems*, 29.

Huang, C., Sun, X., Xiong, J., and Yao, Y. (2020). Boosting with structural sparsity: A differential inclusion approach. *Applied and Computational Harmonic Analysis*, 48(1):1–45.

Ke, H., Chen, D., Yao, Q., Tang, Y., Wu, J., Monaghan, J., Sowman, P., and McAlpine, D. (2023). Deep factor learning for accurate brain neuroimaging data analysis on discrimination for structural mri and functional mri. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Kueper, J. K., Speechley, M., and Montero-Odasso, M. (2018). The alzheimer's disease assessment scale–cognitive subscale (adas-cog): modifications and responsiveness in pre-dementia populations. a narrative review. *Journal of Alzheimer's Disease*, 63(2):423–444.

Liu, M., Zhang, D., Shen, D., Initiative, A. D. N., et al. (2012). Ensemble sparse classification of alzheimer's disease. *NeuroImage*, 60(2):1106–1116.

Mohs, R. C. (1996). The alzheimer's disease assessment scale. *International Psychogeriatrics*, 8(2):195–203.

Mu, Y. and Gage, F. H. (2011). Adult hippocampal neurogenesis and its role in alzheimer's disease. *Molecular neurodegeneration*, 6(1):1–9.

Osher, S., Burger, M., Goldfarb, D., Xu, J., and Yin, W. (2005). An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489.

Osher, S., Ruan, F., Xiong, J., Yao, Y., and Yin, W. (2016). Sparse recovery via differential inclusions. *Applied and Computational Harmonic Analysis*, 41(2):436–469.

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., Jack, C. R., Jagust, W. J., Shaw, L. M., Toga, A. W., et al. (2010). Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209.

Sun, X., Hu, L., Yao, Y., and Wang, Y. (2017). Gsplit lbi: Taming the procedural bias in neuroimaging for disease prediction. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 107–115. Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso.

Vaiter, S., Peyré, G., Dossal, C., and Fadili, J. (2012). Robust sparse analysis regularization. *IEEE Transactions on information theory*, 59(4):2001–2016.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

van Hoesen, G. W., Augustinack, J. C., Dierking, J., Redman, S. J., and Thangavel, R. (2000). The parahippocampal gyrus in alzheimer's disease: clinical and preclinical neuroanatomical correlates. *Annals of the New York Academy of Sciences*, 911(1):254–274.

Visser, P., Verhey, F., Hofman, P., Scheltens, P., and Jolles, J. (2002). Medial temporal lobe atrophy predicts alzheimer's disease in patients with minor cognitive impairment. *Journal of Neurology, Neurosurgery & Psychiatry*, 72(4):491–497.

Xiao, R., Cui, X., Qiao, H., Zheng, X., Zhang, Y., Zhang, C., and Liu, X. (2021). Early diagnosis model of alzheimer's disease based on sparse logistic regression with the generalized elastic net. *Biomedical Signal Processing and Control*, 66:102362.

Xin, B., Hu, L., Wang, Y., and Gao, W. (2015). Stable feature selection from brain smri. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Xin, B., Kawahara, Y., Wang, Y., and Gao, W. (2014). Efficient generalized fused lasso and its application to the diagnosis of alzheimer's disease. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Ye, G.-B. and Xie, X. (2011). Split bregman method for large scale fused lasso. *Computational Statistics & Data Analysis*, 55(4):1552–1569.

Yin, W., Osher, S., Goldfarb, D., and Darbon, J. (2008). Bregman iterative algorithms for \ell_1-minimization with applications to compressed sens-

ing. *SIAM Journal on Imaging sciences*, 1(1):143–168.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Differential Inclusion approach for Learning Heterogeneous Sparsity in Neuroimaging Analysis: Supplementary Materials

1 SUPPORTING LEMMAS

Lemma 1.1 (Lemma 1 in Huang et al. (2016)). *Suppose that $\epsilon \sim N(0_n, \sigma^2 I_n)$, then*

$$\mathbb{P}\left(\frac{\|B\epsilon\|_\infty}{\sigma} \geq z\right) \leq 2q \exp\left(-\frac{z^2}{2\|B\|_2^2}\right) \left(B \in \mathbb{R}^{q\times n}, z \geq 0\right), \tag{14}$$

$$\mathbb{P}\left(\frac{\|\epsilon\|_2^2}{n\sigma^2} \geq 1+z\right) \leq \exp\left(-\frac{n(z-\log(1+z))}{2}\right) (z \geq 0). \tag{15}$$

By Eq. (14)*, we have that for $B \in \mathbb{R}^{q\times n}$, with probability not less than $1-2q/m^2$,*

$$\|B\epsilon\|_\infty \leq 2\sigma \cdot \|B\|_2 \sqrt{\log m}. \tag{16}$$

By Eq. (15) *we have that with probability not less than $1-\exp(-4n/5)$,*

$$\|\epsilon\|_2 \leq 2\sigma\sqrt{n}. \tag{17}$$

Lemma 1.2 (Lemma 6 in Huang et al. (2016)). *Denote $\Sigma := \frac{I-\tilde{D}A^{-1}\tilde{D}^\top}{\nu}$ with $A := \tilde{X}^*\tilde{X} + \frac{\tilde{D}^\top\tilde{D}}{\nu}$. Then the RSC condition is equivalent to $\Sigma_{S,S} \succeq \lambda_\Sigma I_s$ for some $\lambda_\Sigma > 0$. Besides, we have $\lambda_H \leq \lambda_\Sigma$.*

Lemma 1.3 (Lemma 2 in Huang et al. (2016)). *$\Sigma_{S,S} \succ 0$ if and only if $\ker(D_{S^c}) \cap \ker(X) \subseteq \ker(D_S)$.*

2 PROOF OF PROP. 4.4

Proposition 2.1. *The RSC condition in assum. 4.3 holds if there exists a constant $\lambda_X > 0$ such that $X_{S^\beta}^\top X_{S^\beta} \succeq \lambda_X$, with $S^\beta := \mathrm{supp}(\beta^*)$.*

Proof. According to Lemma 1.3, we have that the RSC condition holds if

$$\ker(\tilde{D}_{S^c}) \cap \ker(\tilde{X}) \subset \ker(\tilde{D}_S). \tag{18}$$

If $X_{S^\beta}^\top X_{S^\beta}$ is positive, then we have $\ker(\tilde{X}_{S^{+,I}\cup(p+S^-)}) = 0$ since $S^{+,I} \cap S^- = \emptyset$. For any $\beta \in \ker(\tilde{D}_{S^c}) \cap \ker(\tilde{X})$, we have $\beta_{(S^{+,I}\cup(p+S^-))^c} = 0$. Together with $\tilde{X}\beta = 0$, we have $\tilde{X}_{S^{+,I}\cup(p+S^-)}\beta_{S^{+,I}\cup(p+S^-)} = 0$. We then have $\beta = 0$, which means $\ker(\tilde{D}_{S^c}) \cap \ker(\tilde{X}) = 0 \subset \ker(\tilde{D}_S)$. Indeed, we have that

$$\ker(\tilde{D}_{(\mathcal{G}^{+,I}\setminus S^{+,I})\cup(\mathcal{G}^-\setminus S^-)}) \cap \ker(\tilde{X}) \subset \ker(\tilde{D}_S) \tag{19}$$

holds if and only if $X_{S^\beta}^\top X_{S^\beta}$ is positive. The sufficiency can be obtained similarly. For the necessity, note that if Eq. (18) holds, we have that for any $\beta \in \ker(\tilde{D}_{S^c}) \cap \ker(\tilde{X})$, $\beta_{S^{+,I}\cup(p+S^-)} = 0$. If this does not hold, then there exists a β' with $\beta'_{S^{+,I}\cup(p+S^-)} \neq 0$, which means $\beta' \notin \ker(\tilde{D}_S)$. This violates the condition. Therefore, if $X_{S^\beta}^\top X_{S^\beta}$ is positive, we have that there exists a β' with $\beta'_{S^{+,I}\cup(p+S^-)} \neq 0$, such that $\beta' \in \ker(\tilde{X}_{S^{+,I}\cup(p+S^-)})$. We set $\beta'_{(S^{+,I}\cup(p+S^-))^c} = 0$, then we have $\beta' \in (\tilde{D}_{(\mathcal{G}^{+,I}\setminus S^{+,I})\cup(\mathcal{G}^-\setminus S^-)}) \cap \ker(\tilde{X})$ however does not belong to $\ker(\tilde{D}_S)$. This violates the condition. □

3 UNIQUENESS OF SOLUTION PATH

In this section, we discuss the existence and uniqueness of the solution of Heter-LBISS in Eq. (6) and Heter-ISS in Eq. (3).

Proposition 3.1 (Solution existence and uniqueness for Heter (LB) ISS)**.** *With $v(t), \rho(t), g(t)$ defined in Eq. (4), we have*

1. *For Heter-ISS in Eq. (3), assume that $\rho(t), g(t)$ are right continuously differentiable; and $\beta(t), \gamma(t)$ are right continuous. Then the solution exists. Besides, the $v(t)$ is unique. If additionally $H_{(\beta,S(t)),(\beta,S(t))} \succ 0$ with $S(t) := \mathrm{supp}(\gamma(t))$ for some $0 \le t \le \tau$, we have $(\beta(t), \gamma(t))$ are unique when $0 \le t \le \tau$.*
2. *For Heter-LBISS in Eq. (6), assume that $v(t), \beta(t)$ are right continuously differentiable, then there is a unique solution for $t \ge 0$.*

Proof. We first prove the case for Heter ISS in Eq. (3). Denote $A := \tilde{X}^*\tilde{X} + \tilde{D}^\top\tilde{D}/\nu$, where $\tilde{D} := [D; I_{p\times p}]$. With $\Sigma = \left(I - DA^\dagger D^T\right)/\nu$, according to Eq. (3a), we have

$$\beta(t) := A^{-1}(\tilde{X}^* y + \tilde{D}^\top\gamma(t)/\nu). \tag{20}$$

Besides, according to Eq. (3b), *i.e.*,

$$\dot{v}(t) := -\nabla_\gamma \ell(\beta(t), \gamma(t)) = \frac{\tilde{D}\beta(t) - \gamma(t)}{\nu}, \tag{21}$$

we substitute Eq. (20) into Eq. (21) and obtain:

$$\dot{v}(t) := \tilde{D}A^{-1}\tilde{X}^* y - \Sigma\gamma(t) := -\Sigma^{\frac{1}{2}}\left(\Sigma^{\frac{1}{2}}\gamma(t) - \Sigma^{\dagger\frac{1}{2}}A^{-1}\tilde{X}^* y\right).$$

Note that $v(t) \in \partial f(\gamma) := \left(\|\gamma\|_1 + \mathbb{1}(\gamma^{l,I} \ge 0) + \mathbb{1}(\gamma^p \le 0)\right)$; therefore, if we denote $g(\gamma(t)) := \frac{1}{2}\|\tilde{y} - \Sigma^{\frac{1}{2}}\gamma(t)\|_2^2$ with $\tilde{y} := \Sigma^{\dagger\frac{1}{2}}A^{-1}\tilde{X}^* y$. Then the Heter ISS is equivalent to the following differential inclusion:

$$\dot{v}(t) = -\nabla_\gamma g(\gamma(t)), \tag{22a}$$

$$v(t) \in \partial\left(\|\gamma\|_1 + \mathbb{1}_{\mathcal{C}}\right), \tag{22b}$$

where $\mathcal{C} := \{\gamma : \gamma^{l,I} \ge 0,\ \gamma^p \le 0\}$. The existence of a slightly different version (*i.e.*, $v(t) \in \partial\|\gamma(t)\|_1$) of Eq. (22) has been studied in Burger et al. (2013) and Osher et al. (2016). Similar to the [Theorem 1 in Burger et al. (2013)], we can obtain the solution by defining a sequence of changing points of t, *i.e.*, $0 = t_0 < t_1 < t_2 <$ such that $v(t_0) = \gamma(t_0) = 0$ and for each k, $t_{k+1} := \sup\{v(t_k) - (t - t_k)\nabla_{\gamma(t)} g(\gamma(t_k)) \in \partial\left(\|\gamma(t_k)\|_1 + \mathbb{1}_{\mathcal{C}}(\gamma(t_k))\right)\}$ and it is allowed that $t_{k+1} = \infty$. Simply speaking, t_{k+1} is the earliest time the sub-gradient of the penalty function changes. Writing $v := [(v^{l,I})^\top, (v^{l,D})^\top, (v^p)^\top]^\top$, for each t, we define $S^{l,I}(t) := \{i : v_i^{l,I}(t) = 1\}$, $S^{l,D}(t) := \{i : |v_i^{l,D}(t)| = 1\}$, $S^p(t) := \{i : v_i^p(t) = 1\}$, $S(t) := \mathrm{supp}(\gamma(t))$. With these changing points, we define

$$v(t) := v(t_k) - (t - t_k)\nabla g(\gamma(t_k)), \tag{23a}$$

$$\gamma(t) := \arg\min_\gamma g(\gamma),\ \text{subject to:}$$

$$\gamma_i^{l,I} \ge 0,\ \forall i \in S^{l,I}(t_k) \cup S^p(t_k); v_i^{l,D}(t_k)\gamma_i^{l,D} \ge 0,\ \forall i \in S^{l,D}(t_k); \gamma_i = 0,\ \forall i \in S^p(t_k), \tag{23b}$$

for $t \in [t_k, t_{k+1})$ and for each k. We show that the $(v(t), \gamma(t))_t$ given by Eq. (23) is the solution of Eq. (22). We first show that $t_{k+1} \ne t_k$. We discuss two cases, $\nabla_i g(\gamma(t_k)) = 0$ and $\nabla_i g(\gamma(t_k)) \ne 0$, respectively. For $\nabla_i g(\gamma(t_k)) = 0$, we have $v_i(t) = v_i(t_k) - (t - t_k)\nabla_i g(\gamma(t)) = v_i(t_k) \in \partial\left(\|\gamma(t_k)\|_1 + \mathbb{1}_{\mathcal{C}}(\gamma(t_k))\right)$. For $\nabla_i g(\gamma(t_k)) \ne 0$, which can only happen when $\gamma_i(t_k) = 0$ according to the KKT optimality conditions of Eq. (23b). Further, by inspecting the same KKT optimality conditions, we have that among the set of elements of $\gamma(t_k)$ with $\nabla_i g(\gamma(t_k)) \ne 0$, $\nabla_i g(\gamma(t_k)) < 0$ for $v_i(t_k) = 1$ and $\gamma_i(t_k) = 0$; $\nabla_i g(\gamma(t_k)) > 0$ for $v_i(t_k) = -1$ and $\gamma_i(t_k) = 0$. Therefore, as long as t satisfies that $|(t - t_k)\nabla_i g(\gamma(t_k))| \le 2$, we have $v_i(t) \in \partial\left(\|\gamma(t_k)\|_1 + \mathbb{1}_{\mathcal{C}}(\gamma(t_k))\right)$. This implies $t_{k+1} - t_k > 0$. Taking derivatives with respect to t in Eq. (23a), we obtain that $(v(t), \gamma(t))$ satisfies Eq. (22a). The constraints on $\gamma(t)$ in Eq. (23b) further makes the $(v(t), \gamma(t))$ satisfies Eq. (22b). Therefore, $(v(t), \gamma(t))$ is the solution of Eq. (23).

Next, we prove the uniqueness. We first show the uniqueness of $\nabla f(\gamma(t))$ and $\Sigma^{\frac{1}{2}}\gamma(t)$, which ensures the uniqueness of $v(t)$ due to $v(0) = 0$ and Eq. (22a). We discuss two cases, $S(t) = \emptyset$ and $S(t) \neq \emptyset$, respectively. When $S(t) = \emptyset$, we have that $\gamma(t) = 0$; hence $\nabla f(\gamma(t)) = \nabla f(0)$ and $\Sigma^{\frac{1}{2}}\gamma(t) = 0$. Therefore $\nabla f(\gamma(t))$ and $\Sigma^{\frac{1}{2}}\gamma(t)$ are unique. When $S(t) \neq \emptyset$, note that Eq. (22) implies:

$$\begin{cases} \gamma_i(t) \geq 0 \text{ and } \nabla_i g(\gamma(t)) \geq 0, & \forall v_i(t) = 1 \text{ and } i \in S(t) \\ \gamma_i(t) \leq 0 \text{ and } \nabla_i g(\gamma(t)) \leq 0, & \forall v_i(t) = -1 \text{ and } i \in S(t) \\ \gamma_i(t) = 0, & \forall i \in (S(t))^c. \end{cases} \tag{24}$$

This is because when $v_i(t) = 1$ (resp. $v_i(t) = -1$) for $i \in S(t)$, $\dot{v}_i(t) \leq 0$ (resp. $\dot{v}_i(t) \geq 0$) since $v_i(t) = 1$ (resp. $v_i(t) = -1$) reaches the maximality (resp. minimality). According to Eq. (22a), we have $\nabla_i f(\gamma(t)) \geq 0$ (resp. $\nabla_i g(\gamma(t)) \leq 0$) for $v_i(t) = 1$ (resp. $v_i(t) = -1$) for $i \in S(t)$. According to the definition of $v(t)$ in Eq. (22b), we have that $\gamma_i(t) \geq 0$, $\gamma_i(t) \leq 0$ and $\gamma_i(t) = 0$ when $v_i(t) = 1, i \in S(t)$, $v_i(t) = -1, i \in S(t)$ and $i \in (S(t))^c$, respectively. Besides, we have

$$\gamma_i(t)\nabla_i g(\gamma(t)) = 0. \tag{25}$$

This is obvious when $\gamma_i(t) = 0$. When $\gamma_i(t) \neq 0$, due to the right continuity of $\gamma(t)$, there exists $\delta > 0$, such that $\gamma_i(s) \neq 0$ and does not change the sign on $[t, t + \delta)$. Besides, we have $\gamma(s) \in \mathcal{C}$ on $(t, t + \delta)$ as long as $\gamma(t) \in \mathcal{C}$. Therefore, we have $\dot{v}_i(t) = 0$, which means $\nabla_i f(\gamma(t)) = 0$. Eq. (24), (25) are the KKT conditions of Eq. (23b), which is a constrained convex optimization problem. Since $g(\gamma(t)) := h(\Sigma^{\frac{1}{2}}\gamma(t))$ with $h(a) := \frac{1}{2}\|\tilde{y} - a\|_2^2$, therefore the h is a strict convex function. Therefore, we obtain the uniqueness of $\Sigma^{\frac{1}{2}}\gamma(t)$ and hence $\nabla g(\gamma(t))$. Therefore, we have that $v(t)$ is unique. If further we have $\Sigma_{S(t),S(t)} \succ 0$ for $0 \leq t \leq \tau$, we have $\gamma(t)$ is unique. Particularly, if we can recover the true signal set $S(t) = S$, the solution is unique.

For Heter LBISS in Eq. (6), note that it is equivalent to

$$\begin{pmatrix} \dot{\beta}(t) \\ \dot{z}(t) \end{pmatrix} = -\begin{pmatrix} -\kappa\tilde{X}^*(\tilde{X}\beta(t) - y) - \kappa\tilde{D}^T(\tilde{D}\beta(t) - \kappa\mathcal{S}_{\text{combine}}(z(t), 1))/\nu \\ -(\kappa\mathcal{S}_{\text{combine}}(z(t), 1) - D\beta(t))/\nu \end{pmatrix} := h(\beta(t), z(t)),$$

where $\mathcal{S}_{\text{combine}}(z(t), 1) := [\mathcal{S}_+(z^{l,I}(t), 1); \mathcal{S}(z^{l,D}(t), 1); \mathcal{S}_-(z^p(t), 1)]$ with $\mathcal{S}_+(z, 1) := \max(z - 1, 0)$, $\mathcal{S}(z, 1) := \max(|z| - 1, 0)$, and $\mathcal{S}_(z, 1) := \max(-z + 1, 0)$. Since $h(\beta(t), z(t))$ is Lipschitz continuous, the Picard-Lindelöf Theorem implies that there exists a unique solution to this ODE. □

4 NON-INCREASING PROPERTIES

In this section, we show that $\ell(\beta(t), \gamma(t))$ is non-increasing w.r.t. t for the path given by Heter-ISS in Eq. (3) and Heter-LBISS in Eq. (6).

Proposition 4.1 (Non-increasing of $\ell(\beta(t), \gamma(t))$ along the paths given by Heter (LB)ISS)**.** *For a solution $(\rho(t), \beta(t), \gamma(t))$ of Heter ISS in Eq.* (3) *and Heter LBISS in Eq.* (6)*, $\ell(\beta(t), \gamma(t))$ is non-increasing w.r.t. t.*

Proof. This proof is similar to that in Huang et al. (2016). Denote $v(t), \beta(t), \gamma(t), S^{l,I}(t), S^{l,D}(t), S^p(t), S(t)$ in Prop. 3.1. According to the proof in Prop. 3.1, the $(\beta(t), v(t), \gamma(t))$ in Heter ISS (3) is the solution of the following constrained optimization:

$$\min_{\beta,\gamma} \ell(\beta(t), \gamma(t))$$
$$\text{Subject to } \begin{cases} \gamma_i(t) \geq 0, & v_i(t) = 1,\ i \in S(t), \\ \gamma_i(t) \leq 0, & v_i(t) = -1,\ i \in S(t), \\ \gamma_i(t) = 0, & i \in (S(t))^c. \end{cases}$$

For any t, due to the continuity of $v(t)$ (since $v(t)$ is piece-wise linear), there exists $\delta > 0$, such that for any $t' \in (t - \delta, t + \delta)$, $-1 < v_i(t') < 1$ for $i \in S(t)$ (in fact, we only need $-1 < v_i(t') < 1$ for $i \in S^{l,D}(t)$), then we have $\gamma_i(t) \geq 0$. For $v_i(t) = -1$, $i \in S(t)$, we have $v_i(t') < 1$ for $i \in S(t)$; then we have $\gamma_i(t') \leq 0$. Finally, for $i \in S(t)$, *i.e.*, $v_i^{l,I}(t) < 1$ for $i \in S^{l,I}(t)$ and $v_i^p(t) < 1$ for $i \in S^p(t)$; and $-1 < v_i^{l,D}(t) < 1$ for $i \in S^{l,D}(t)$, we respectively have $v_i^{l,I}(t') < 1$ for $i \in S^{l,I}(t')$ and $v_i^p(t') < 1$ for $i \in S^p(t)$; and $-1 < v_i^{l,D}(t') < 1$ for $i \in S^{l,D}(t)$. In

these three cases, we have $\gamma_i(t) = 0$. Therefore, we have $\gamma_i(t) = 0$ for $i \in (S(t))^c$. Since $(\beta(t), v(t), \gamma(t)$ are right continuous, the $\ell(\beta(t), \gamma(t))$ is also right continuous since ℓ is continuously differentiable. Then the $\gamma(t')$ satisfies the constraints, which implies $\ell(\beta(t), \gamma(t)) \geq \ell(\beta(t'), \gamma(t'))$ for any $t' \in (t-\delta, t+\delta)$. This means $\ell(\beta(t), \gamma(t))$ is the local minimum for every t.

Suppose $\ell(\beta(t), \gamma(t))$ is not non-increasing, then there exists $\tilde{t}_1 < \tilde{t}_2$ such that $\ell(\beta(\tilde{t}_1), \gamma(\tilde{t}_1)) < \ell(\beta(\tilde{t}_2), \gamma(\tilde{t}_2))$. Since there are at most countable changing points in $(\tilde{t}_1, \tilde{t}_2)$, we first assume there exists one changing point $t_k \in (\tilde{t}_1, \tilde{t}_2)$ such that $\ell(\beta(t_k), \gamma(t_k)) > \lim_{t \to t_k -} \ell(\beta(t), \gamma(t))$, which violates the local minimum property of t_k. If there is no such changing point, then there exists at least one piece, say $[t_k, t_{k+1}) \subset (\tilde{t}_1, \tilde{t}_2)$ such that $\ell(\beta(t_k), \gamma(t_k)) < \ell(\beta(t_k), \gamma(t_k)) < \ell(\beta(t_{k+1}), \gamma(t_{k+1}))$. Since $\ell(\beta(t), \gamma(t))$ is continuous in $[t_k, t_{k+1}]$, then there exists at least one $t' \in [t_k, t_{k+1}]$ such that $(\beta(t'), \gamma(t'))$ is not the local minimum of $\ell(\beta(t'), \gamma(t'))$.

For Heter-LBISS in Eq. (6), note that $\langle \dot{v}(t), \gamma(t) \rangle = 0$ according to the proof in Prop. 3.1. Since

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\ell(\beta(t), \gamma(t)) &= \left\langle \begin{pmatrix} \dot{\beta}(t) \\ \dot{\gamma}(t) \end{pmatrix}, \begin{pmatrix} \nabla_\beta \ell(\beta(t), \gamma(t)) \\ \nabla_\gamma \ell(\beta(t), \gamma(t)) \end{pmatrix} \right\rangle \\
&= \left\langle \begin{pmatrix} \dot{\beta}(t) \\ \dot{\gamma}(t) \end{pmatrix}, \begin{pmatrix} -\dot{\beta}(t)/\kappa \\ -\dot{v}(t) - \dot{\gamma}(t)/\kappa \end{pmatrix} \right\rangle = \frac{1}{\kappa} \left\| \begin{pmatrix} \dot{\beta}(t) \\ \dot{\gamma}(t) \end{pmatrix} \right\|_2^2 \leq 0,
\end{aligned}$$

we have that the $\ell(\beta(t), \gamma(t))$ is non-increasing. □

Proposition 4.2 (Non-increasing of $\ell(\beta(k), \gamma(k))$ w.r.t. k in Heter LBI (8))**.** *For the solution $(v(k), \beta(k), \gamma(k))$ of Heter LBI in Eq.* (8)*, the $\ell(\beta(k), \gamma(k))$ is non-increasing w.r.t. k, if $\kappa\alpha\|H\|_2 < 2$.*

Proof. We have

$$\begin{aligned}
&\ell(\beta(k+1), \gamma(k+1)) - \ell(\beta(k), \gamma(k)) = \\
&\left\langle \nabla \ell(\beta(k), \gamma(k)), \begin{pmatrix} \beta(k+1) - \beta(k) \\ \gamma(k+1) - \gamma(k) \end{pmatrix} \right\rangle + \\
&\qquad \frac{1}{2}(\beta^\top(k+1) - \beta^\top(k), \gamma^\top(k+1) - \gamma^\top(k)) H \begin{pmatrix} \beta(k+1) - \beta(k) \\ \gamma(k+1) - \gamma(k) \end{pmatrix} \\
&\leq -\frac{1}{\alpha} \left\langle -\alpha \nabla \ell(\beta(k), \gamma(k)), \begin{pmatrix} \beta(k+1) - \beta(k) \\ \gamma(k+1) - \gamma(k) \end{pmatrix} \right\rangle + \frac{\|H\|_2}{2} \left\| \begin{pmatrix} \beta(k+1) - \beta(k) \\ \gamma(k+1) - \gamma(k) \end{pmatrix} \right\|_2^2 \\
&\leq -\frac{1}{\alpha} \left\langle \begin{pmatrix} \frac{\beta(k+1) - \beta(k)}{\kappa} \\ v(k+1) - v(k) + \frac{\gamma(k+1) - \gamma(k)}{\kappa} \end{pmatrix}, \begin{pmatrix} \beta(k+1) - \beta(k) \\ \gamma(k+1) - \gamma(k) \end{pmatrix} \right\rangle + \\
&\qquad \frac{\|H\|_2}{2} \left\| \begin{pmatrix} \beta(k+1) - \beta(k) \\ \gamma(k+1) - \gamma(k) \end{pmatrix} \right\|_2^2 \\
&\leq -\frac{1}{\alpha} \langle v(k+1) - v(k), \gamma(k+1) - \gamma(k) \rangle + \left(\frac{\|H\|_2}{2} - \frac{1}{\kappa\alpha} \right) \left\| \begin{pmatrix} \beta(k+1) - \beta(k) \\ \gamma(k+1) - \gamma(k) \end{pmatrix} \right\|_2^2
\end{aligned}$$

Since

$$\begin{aligned}
&\langle v(k+1) - v(k), \gamma(k+1) - \gamma(k) \rangle \\
&= \|\gamma(k+1)\|_1 + \not{\kappa}_{\mathcal{C}}(\gamma_{k+1}) + \|\gamma(k)\|_1 + \not{\kappa}_{\mathcal{C}}(\gamma_k) - \langle v(k), \gamma(k+1) \rangle - \langle v(k+1), \gamma(k) \rangle \geq 0,
\end{aligned}$$

we have $\ell(\beta(k+1), \gamma(k+1)) \leq \ell(\beta(k), \gamma(k))$ as long as $\kappa\alpha\|H\|_2 \leq 2$. □

5 ORACLE DYNAMICS

In this section, we introduce Oracle dynamics, *i.e.*, the dynamics on the true signal set S. We first show that the solution of the Oracle dynamics is sign-consistent to the Oracle Solution as long as $t > \tau_\infty < \bar{\tau}$ where $\bar{\tau}$ denotes the time such that before $\bar{\tau}$ the no-false-positive property holds. In this regard, the Heter LBISS will be sign-consistent to the Oracle Solution, which will be shown to be sign-consistent to the (β^*, γ^*).

We consider the following Oracle dynamics of Heter LBISS:

$$\gamma'_{S^c}(t) = 0, \tag{26a}$$

$$\frac{\dot{\beta}'}{\kappa} = -\tilde{X}^*(\tilde{X}\beta'(t) - y) + \frac{\tilde{D}^\top(\tilde{D}\beta'(t) - \gamma'(t))}{\nu}, \tag{26b}$$

$$\dot{v}_S(t) = -\frac{\gamma'_S(t) - \tilde{D}_S\beta'(t)}{\nu}, \tag{26c}$$

$$v_S(t) \in \partial\left(\|\gamma_S(t)\|_1 + \mathbb{1}(\gamma^{l,I}_{S^{+,I}}(t) \geq 0) + \mathbb{1}(\gamma^p_{S^p}(t) \leq 0)\right). \tag{26d}$$

We call the dynamics in Eq. (26) as *Oracle dynamics* in terms of the support set since the dynamics constantly keep the value of γ on S^c as 0 in Eq. (26a). Denote $(\beta^{o,+}, \beta^{o,-}, \gamma^{o,+}, \gamma^{o,-})$ as the Oracle point:

$$(\beta^{o,+}, \beta^{o,-}, \gamma^{o,+}, \gamma^{o,-}) \in \arg \min_{(\gamma^{o,+,I})_{S^+,I} \geq 0, (\gamma^{o,-})_{S^-} \leq 0, \gamma_{S^c}=0} \ell(\beta, \gamma).$$

We then expect the Oracle dynamics of $(\beta'^{,l}(t), \beta'^{,p}(t), \gamma'(t))$ converge to the Oracle point $(\beta^{o,+}, \beta^{o,-}, \gamma^o)$ in terms of ℓ_2 distance:

$$d(t) = \sqrt{\|d_\beta(t)\|_2^2 + \|d_\gamma(t)\|_2^2},$$

where $d_\beta(t) = \beta'(t) - \beta^o$, $d_{\gamma,S}(t) = \gamma'_S(t) - \gamma^o_S$. For simplicity, we rewrite Eq. (26) as:

$$\gamma'_S(t) = 0, \tag{27a}$$

$$\begin{pmatrix} 0 \\ \dot{v}'_S(t) \end{pmatrix} + \frac{1}{\kappa}\begin{pmatrix} \dot{\beta}'(t) \\ \dot{\gamma}'_S(t) \end{pmatrix} = -H_{(\beta,S),(\beta,S)}\begin{pmatrix} d_\beta(t) \\ d_{\gamma,S}(t) \end{pmatrix}, \tag{27b}$$

Define the *potential function* $\Psi(t)$ as:

$$\Psi(t) := \|\gamma^o_S\|_1 + \mathbb{1}(\gamma^o_S \in \mathcal{C}) - \|\gamma'_S\|_1 - \mathbb{1}(\gamma'_S \in \mathcal{C}) + \frac{\|\gamma'_S(t) - \gamma^o_S\|_2^2}{2\kappa} + \frac{\|\beta'_S(t) - \beta^o\|_2^2}{2\kappa}.$$

The following property shows that the potential function decreases fast enough as t grows.

Lemma 5.1 (Generalized Bihari's inequality)**.** *Denote $\gamma^o_{\min} := \min_{i\in S}|\gamma^o_i|$, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\Psi(t) \leq -\lambda_H F^{-1}(\Psi(t)),$$

where

$$F(x) := \frac{x}{2\kappa} + \begin{cases} 0, & 0 \leq x < (\gamma^o_{\min})^2, \\ Mx/\gamma^o_{\min}, & (\gamma^o_{\min})^2 \leq x < s(\gamma^o_{\min})^2, \\ M\sqrt{sx}, & x \geq s(\gamma^o_{\min})^2, \end{cases}$$
$$F^{-1}(x) := \inf(y : F(y) \geq x)(y \geq 0),$$

where $M := \max_{i\in S}\max_{t\geq 0}|v'_i(t)|$.

Proof. We rewrite the $\ell(\beta'(t), \gamma'_S(t)) := \frac{1}{2n}\left\|\begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}\begin{pmatrix} \beta'(t) \\ \gamma'_S(t) \end{pmatrix}\right\|_2^2$. According to the definition of (β^o, γ^o_S) and the KKT optimality conditions, we have

$$\frac{1}{n}\begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}^\top \left(\begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}\begin{pmatrix} \beta^o \\ \gamma^o_S \end{pmatrix} - \begin{pmatrix} y \\ 0 \end{pmatrix}\right) + \begin{pmatrix} 0 \\ g^o_S \end{pmatrix} = 0, \tag{28}$$

$$\langle \gamma^o_S, g^o_S \rangle = 0,\ g^o_i \leq 0, \forall i \in S^{+,I};\ g^o_i = 0, \forall i \in S^{+,D};\ g^o_i \leq 0, \forall i \in S^-. \tag{29}$$

According to the definition of $\Psi(t)$, we have

$$\begin{aligned}
\frac{d\Psi(t)}{dt} &= -\langle \dot{\gamma}'_S(t), v'_S(t)\rangle - \langle \gamma_i^o - \gamma'_S(t), \dot{v}'_S(t)\rangle + \langle \dot{\gamma}'_S(t), v'_S(t)\rangle + \left\langle \begin{pmatrix} \beta'(t)-\beta^o \\ \gamma'_S(t)-\gamma_S^o \end{pmatrix}, \begin{pmatrix} \frac{\dot{\beta}'(t)}{\kappa} \\ \frac{\dot{\gamma}'_S(t)}{\kappa} \end{pmatrix} \right\rangle \\
&= \left\langle \begin{pmatrix} \beta'(t)-\beta^o \\ \gamma'_S(t)-\gamma_S^o \end{pmatrix}, \begin{pmatrix} \frac{\dot{\beta}'(t)}{\kappa} \\ \dot{v}'_S(t) + \frac{\dot{\gamma}'_S(t)}{\kappa} \end{pmatrix} \right\rangle \\
&= -\left(d_\beta^\top(t), d_{\gamma,S}^\top(t)\right) \frac{1}{n} \begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}^\top \left(\begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix} \begin{pmatrix} \beta'(t) \\ \gamma'_S(t) \end{pmatrix} - \begin{pmatrix} y \\ 0 \end{pmatrix} \right) \\
&= -\left(d_\beta^\top(t), d_{\gamma,S}^\top(t)\right) H_{(\beta,S),(\beta,S)} \begin{pmatrix} d_\beta(t) \\ d_{\gamma,S}(t) \end{pmatrix} + \left(d_\beta^\top(t), d_{\gamma,S}^\top(t)\right) \begin{pmatrix} 0 \\ g_S^o \end{pmatrix} \\
&\le -\left(d_\beta^\top(t), d_{\gamma,S}^\top(t)\right) H_{(\beta,S),(\beta,S)} \begin{pmatrix} d_\beta(t) \\ d_{\gamma,S}(t) \end{pmatrix} := -2L(t),
\end{aligned}$$

where the last inequality is due to $\langle \gamma'_S(t) - \gamma_S^o, g_S^o\rangle = \langle \gamma'_S(t), g_S^o\rangle \le 0$. Therefore, it is sufficient to show that $\frac{2L(t)}{\lambda_H} \ge F^{-1}(\Psi(t)) \implies F\left(\frac{2L(t)}{\lambda_H}\right) \ge \Psi(t)$. We first prove that M is finite. Since

$$\frac{d\Psi(t)}{dt} \le -\lambda_H d^2(t),$$

$\Psi(t)$ is non-increasing. We show that there exists t_0 such that $\forall t > t_0$, $v_i(t) = 1$ for all $i \in S$. Since we know that $d^2(t) < (\gamma_{\min}^o)^2$ means $\|\gamma_S^o\|_1 - \langle \gamma_S^o, v_S^o\rangle = 0$, we denote $t_0 := \inf\{t : d^2(t) \le (\gamma_{\min}^o)^2/2\}$ and have that:

$$\Psi(t_0) \le \Psi(0) - \lambda_H * t_0 * (\gamma_{\min}^o)^2/2.$$

We then have $t_0 \le \frac{2\Psi(0)}{\lambda_H(\gamma_{\min}^o)^2}$. Due to the right continuity of $(\beta(t), \gamma(t))$, we then have $d^2(t_0) \le (\gamma_{\min}^o)^2/2$, which means $\Psi(t_0) \le \frac{(\gamma_{\min}^o)^2}{4\kappa}$. If not there exists $i \in S$ such that $v_i'(t_0) < 1$, implying a contradiction that $\Psi(t_0) > \frac{d^2(t_0)}{2\kappa} \ge \frac{(\gamma_{\min}^o)^2}{2\kappa}$. Due to the non-increasing property of $\Psi(t)$, we have $\Psi(t) \le \Psi(t_0) \le \frac{(\gamma_{\min}^o)^2}{4\kappa}$. This means $\forall t \ge t_0$, we have $v_i'(t) = 1$. Then we have $M := \max_{i\in S} \max(\max_{0\le t\le t_0} |v_i'(t)|, 1)$. Due to the continuity of $v'(t)$, we have that M is finite.

Now we prove $\frac{2L(t)}{\lambda_H} \ge F^{-1}(\Psi(t))$. Note first that we have $\|\gamma_S^o\|_1 - \langle \gamma_S^o, v'_S(t)\rangle = 0$ if $\|\gamma'_S(t) - \gamma_S^o\|_2^2 < (\gamma_{\min}^o)^2$. Besides, we have $\|\gamma_S^o\|_1 - \langle \gamma_S^o, v'_S(t)\rangle \le M \sum_{v_i'(t)<1} \|\gamma_S^o\|_1$, which simultaneously less and equal than $\frac{M}{\gamma_{\min}^o}\|\gamma'_S(t) - \gamma_S^o\|_2^2$ and $M\sqrt{s\|\gamma'_S(t) - \gamma_S^o\|_2^2}$. Therefore, we have $\Psi(t) - \frac{1}{2\kappa}\left(\|d_{\gamma,S}(t)\|_2^2 + \|d_\beta(t)\|_2^2\right) \le F\left(\|d_{\gamma,S}(t)\|_2^2\right) - \frac{1}{2\kappa}\|d_{\gamma,S}(t)\|_2^2$. Since $F(y+x) \ge F(y) + \frac{x}{2\kappa}$ for any $x \ge 0, y \ge 0$ and $L(t) \ge \lambda_H d^2(t)$, we have

$$F\left(\frac{2L(t)}{\lambda_H}\right) \ge F(d^2(t)) = F\left(\|d_\beta(t)\|_2^2 + \|d_{\gamma,S}(t)\|_2^2\right) \ge F\left(\|d_{\gamma,S}(t)\|_2^2\right) + \frac{1}{2\kappa}\|d_\beta(t)\|_2^2 \ge \Psi(t).$$

The proof is completed. □

Lemma 5.2. *We have*

$$d(t) < c\gamma_{\min}^o, \forall\ c < 1, \tag{30}$$

if $t \ge \tau_\infty(c) := \frac{1}{\kappa\lambda_H}\log\frac{1}{c} + \frac{M\log s + 2M + d(0)/\kappa}{\lambda_H \gamma_{\min}^o}(0 < c < 1)$. *This means we have* $\operatorname{sign}(\gamma_i'(t)) = \operatorname{sign}(\gamma_i^o)$ *for* $i \in S$. *Besides, we have*

$$d(t) \le \min\left(\frac{2M\sqrt{s} + d(0)/\kappa}{\lambda_H t}, \sqrt{\frac{2\left(1 + \nu\Lambda_X^2 + \Lambda_{\tilde{D}_S}^2\right)}{\lambda_H \nu}} \cdot d(0)\right). \tag{31}$$

Proof. Suppose there exists $t' \le \tau_\infty(c)$ such that $d(t') < c\gamma_{\min}^o$, then we have

$$\Psi(t) \le \Psi(t') \le c\gamma_{\min}^o, \forall t \ge \tau_\infty(c),$$

due to the non-increasing property of $\Psi(t)$ and that $d(t') < c\gamma^o_{\min}$ with $c < 1$ implies that $v_i'(t) = 1$ thus $\|\gamma^o_S\|_1 - \langle \gamma^o_S, v'_S\rangle = 0$. Therefore, we have $d(t) < c\gamma^o_{\min}$ for $t \geq \tau_\infty(c)$. Suppose there does not exist t with $0 \leq t \leq \tau_\infty(c)$ such that $d(t) < c\gamma^o_{\min}$, then we have

$$\Psi(t) \geq \frac{d^2(t)}{2\kappa} \geq \frac{c^2(\gamma^o_{\min})^2}{2\kappa},$$

which means $F^{-1}(\Psi(t)) > 0$. According to lemma 5.1, we have

$$\begin{aligned}
&\lambda_H \tau_\infty(c) \leq \int_0^{\tau_\infty(c)} \frac{-\frac{\mathrm{d}}{\mathrm{d}t}\Psi(t)}{F^{-1}(\Psi(t))}\mathrm{d}t = \int_{\Psi(\tau_\infty(c))}^{\Psi(0)} \frac{\mathrm{d}x}{F^{-1}(x)} \\
\leq& \left(\int_{c^2(\gamma^o_{\min})^2/(2\kappa)}^{(\gamma^o_{\min})^2/(2\kappa)} + \int_{(\gamma^o_{\min})^2/(2\kappa)}^{F\left((\gamma^o_{\min})^2\right)} + \int_{F\left((\gamma^o_{\min})^2\right)}^{F\left(s(\gamma^o_{\min})^2\right)} + \int_{F\left(s(\gamma^o_{\min})^2\right)}^{F\left(d(0)^2\right)}\right) \frac{\mathrm{d}x}{F^{-1}(x)} \\
\leq& \int_{c^2(\gamma^o_{\min})^2/(2\kappa)}^{(\gamma_{\min})^2/(2\kappa)} \frac{\mathrm{d}x}{2\kappa x} + \int_{(\gamma^o_{\min})^2/(2\kappa)}^{F\left((\gamma^o_{\min})^2\right)} \frac{1}{(\gamma^o_{\min})^2}\,\mathrm{d}x + \int_{(\gamma^o_{\min})^2}^{s(\gamma^o_{\min})^2} \frac{\mathrm{d}F(x)}{x} + \int_{s(\gamma^o_{\min})^2}^{d(0)^2} \frac{\mathrm{d}F(x)}{x} \\
=& \frac{1}{2\kappa}\log\frac{1}{c^2} + \frac{2}{\gamma^o_{\min}} + \int_{(\gamma^o_{\min})^2}^{s(\gamma^o_{\min})^2} \left(\frac{1}{2\kappa x} + \frac{M}{\gamma^o_{\min} x}\right)\mathrm{d}x + \int_{s(\gamma^o_{\min})^2}^{d(0)^2} \left(\frac{1}{M\kappa x} + \frac{M\sqrt{s}}{2x\sqrt{x}}\right)\mathrm{d}x \\
<& \frac{1}{2\kappa}\log\frac{1}{c^2} + \frac{M}{\gamma^o_{\min}} + \frac{1}{2\kappa}\log\frac{d(0)^2}{(\gamma^o_{\min})^2} + \frac{M\log s}{\gamma^o_{\min}} + \frac{M}{\gamma^o_{\min}} \leq \frac{1}{\kappa}\log\frac{1}{c} + \frac{M\log s + 2M + d(0)/\kappa}{\gamma^o_{\min}}.
\end{aligned}$$

Therefore, we have $\tau_\infty(c) < \frac{1}{\kappa\lambda_H}\log\frac{1}{c} + \frac{M\log s + 2M + d(0)/\kappa}{\lambda_H \gamma^o_{\min}}$, which contradicts to the definition of $\tau_\infty(c)$. This means Eq. (30) holds.

To prove Eq. (31), first note that

$$\begin{aligned}
&\ell(\beta'(t), \gamma'_S(t)) := \frac{1}{2n}\left\|\begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}\begin{pmatrix} \beta'(t) \\ \gamma'_S(t) \end{pmatrix}\right\|_2^2 \\
=& \frac{1}{2n}\left\|\begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}\begin{pmatrix} \beta^o \\ \gamma^o_S \end{pmatrix} - \begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}\begin{pmatrix} d_\beta(t) \\ d_\gamma(t) \end{pmatrix}\right\|_2^2 \\
=& \frac{1}{2n}\left\|\begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}\begin{pmatrix} \beta^o \\ \gamma^o_S \end{pmatrix}\right\|_2^2 + \frac{1}{2n}\left\|\begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}\begin{pmatrix} d_\beta(t) \\ d_\gamma(t) \end{pmatrix}\right\|_2^2 \\
&\left\langle \begin{pmatrix} d_\beta(t) \\ d_\gamma(t) \end{pmatrix}, \frac{1}{n}\begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}^\top \left(\begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}\begin{pmatrix} \beta^o \\ \gamma^o_S \end{pmatrix} - \begin{pmatrix} y \\ 0 \end{pmatrix}\right)\right\rangle \\
=& \frac{1}{2n}\left\|\begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}\begin{pmatrix} \beta^o \\ \gamma^o_S \end{pmatrix}\right\|_2^2 + \\
&\frac{1}{2n}\left\|\begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}\begin{pmatrix} d_\beta(t) \\ d_\gamma(t) \end{pmatrix}\right\|_2^2 - \left\langle \begin{pmatrix} d_\beta(t) \\ d_\gamma(t) \end{pmatrix}, \begin{pmatrix} 0 \\ g^o_S \end{pmatrix}\right\rangle, \\
=& \frac{1}{2}\left(d^\top_\beta(t), d^\top_{\gamma,S}(t)\right) H_{(\beta,S),(\beta,S)}\begin{pmatrix} d_\beta(t) \\ d_\gamma(t) \end{pmatrix} - d^\top_{\gamma,S}(t) g^o_S + \frac{1}{2n}\left\|\begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}\begin{pmatrix} \beta^o \\ \gamma^o_S \end{pmatrix}\right\|_2^2 \\
:=& \underbrace{L(t) + d^\top_{\gamma,S}(t) g^o_S}_{\overline{L}(t)} + \frac{1}{2n}\left\|\begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} \tilde{X} & 0 \\ -\sqrt{n/\nu}\tilde{D}_S & I_s \end{pmatrix}\begin{pmatrix} \beta^o \\ \gamma^o_S \end{pmatrix}\right\|_2^2,
\end{aligned}$$

where the third equation is due to Eq. (28). Besides, we have $\overline{L}(t) \geq L(t)$. Using the same technique in Prop. 4.1, we can derive that the $\ell(\beta'(t), \gamma'_S(t))$ along the Oracle dynamics is also non-increasing. Therefore, the $\overline{L}(t)$ is also non-increasing. If $L(t) = 0$ at some t, then Eq. (31) naturally holds since $d^2(t) \leq \frac{2L(t)}{\lambda_H} = 0$. If $L(t) > 0$, denote $C = \sqrt{2L(t)/\lambda_H} > 0$, then for any $0 \leq t' \leq t$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t'}\Psi\left(t'\right) = -\overline{L}(t') - L(t') \leq -\overline{L}(t) - L(t) \leq -2L(t) = -\lambda_H C^2.$$

Besides, we denote $\tilde{F}(x) = x/(2\kappa) + M\sqrt{sx} \geq F(x)$ (where $F(x)$ is defined in lemma (5.1)), we have

$$\frac{\mathrm{d}}{\mathrm{d}t'}\Psi\left(t'\right) \leq -\lambda_H F^{-1}\left(\Psi\left(t'\right)\right) \leq -\lambda_H \tilde{F}^{-1}\left(\Psi\left(t'\right)\right)$$

Since

$$\tilde{F}\left(d(0)^2\right) \geq \tilde{F}\left(\|\gamma_S^o\|_2^2\right) + \|\beta^o\|_2^2/(2\kappa) \geq \Psi(0).$$

Next, we prove $\lambda_H t \leq \frac{2M\sqrt{s}}{C} + \frac{1}{2\kappa}\left(1 + \log\frac{d(0)^2}{C^2}\right) \leq \frac{2M\sqrt{s}+d(0)/\kappa}{C}$. We first consider the case when $d(0) > C$, we have:

$$\begin{aligned}
\lambda_H t &\leq \int_0^t \frac{-\frac{\mathrm{d}}{\mathrm{d}t'}\Psi\left(t'\right)}{\max\left(C^2, \tilde{F}^{-1}\left(\Psi\left(t'\right)\right)\right)}\mathrm{d}t' = \int_{\Psi(t)}^{\Psi(0)} \frac{\mathrm{d}x}{\max\left(C^2, \tilde{F}^{-1}(x)\right)} \\
&\leq \int_{\tilde{F}(0)}^{\tilde{F}\left(d(0)^2\right)} = \int_{\tilde{F}(0)}^{\tilde{F}\left(C^2\right)} \frac{\mathrm{d}x}{C^2} + \int_{C^2}^{d(0)^2} \frac{\mathrm{d}\tilde{F}(x)}{x} \\
&= \frac{C^2/(2\kappa) + 2\sqrt{s}C}{C^2} + \int_{C^2}^{d(0)^2}\left(\frac{1}{2\kappa x} + \frac{M\sqrt{s}}{2x\sqrt{x}}\right)\mathrm{d}x \\
&\leq \frac{2M\sqrt{s}}{C} + \frac{1}{2\kappa}\left(1 + \log\frac{d(0)^2}{C^2}\right) \leq \frac{2M\sqrt{s} + d(0)/\kappa}{C}
\end{aligned}$$

If otherwise $d(0) \leq C$,

$$\begin{aligned}
\lambda_H t &\leq \int_{\tilde{F}(0)}^{\tilde{F}\left(d(0)^2\right)} \frac{\mathrm{d}x}{\max\left(C^2, \tilde{F}^{-1}(x)\right)} \leq \int_{\tilde{F}(0)}^{\tilde{F}\left(d(0)^2\right)} \frac{\mathrm{d}x}{C^2} \\
&= \frac{d(0)^2/(2\kappa) + 2\sqrt{s}\cdot d(0)}{C^2} \leq \frac{2M\sqrt{s} + d(0)/\kappa}{C}.
\end{aligned}$$

Then we have that

$$d(t)^2 \leq \frac{2}{\lambda_H}L(t) = \frac{2}{\lambda_H}\cdot\frac{\lambda_H C^2}{2} \leq \left(\frac{2M\sqrt{s} + \cdot d(0)/\kappa}{\lambda_H t}\right)^2.$$

To prove $d(t) \leq \sqrt{\frac{2\left(1+\nu\Lambda_X^2+\Lambda_{\tilde{D}_S}^2\right)}{\lambda_H \nu}}\cdot d(0)$, note that $\overline{L}(0) := L(0)$, we then have

$$d(t)^2 \leq \frac{2}{\lambda_H}L(t) \leq \frac{2}{\lambda_H}\overline{L}(t) \leq \frac{2}{\lambda_H}\overline{L}(0) = \frac{2}{\lambda_H}L(0).$$

Since

$$\begin{aligned}
2L(0) &:= ((\beta^o)^\top, (\gamma_S^o)^\top) H_{(\beta,S),(\beta,S)} \begin{pmatrix} \beta^o \\ \gamma_S^o \end{pmatrix} := \frac{1}{n}\|\tilde{X}\beta\|_2^2 + \frac{1}{\nu}\|\tilde{D}_S\beta^o - \gamma_S^o\|_2^2 \\
&\leq \frac{1}{n}\|\tilde{X}\beta\|_2^2 + \frac{2}{\nu}\|\tilde{D}_S\beta^o\|_2^2 + \frac{2}{\nu}\|\gamma_S^o\|_2^2 \leq \frac{2(\nu\Lambda_X + \Lambda_{\tilde{D}_S}^2 + 1)}{\nu}d^2(0).
\end{aligned}$$

The proof is completed. □

6 SPARSE PARAMETER $\tilde{\beta}$ VIA PROJECTION

In this section, we introduce an efficient projection method to obtain sparse estimators $\tilde{\beta}_k$ at each k via

$$\tilde{\beta}_k^l := \underset{x\geq 0, x(\mathcal{G}^{+,I}\setminus S_k^{l,I})=0, D_G(\mathcal{G}^{+,D}\setminus S_k^{l,D},:)x=0}{\operatorname{argmin}} \|x - \beta_k^l\|_2,\ \tilde{\beta}_k^p := \underset{x\geq 0, x(\mathcal{G}^-\setminus S_k^p)=0}{\operatorname{argmin}} \|x - \beta^p(t)\|_2. \tag{32}$$

The $\tilde{\beta}_k^p$ can be directly obtained as: $\tilde{\beta}_k^p(i) = 0$ for all $i \in \mathcal{G}^-\setminus S_k^p \cap \{i : \beta_k^p(i) \geq 0\}$; $= \beta_k^p(i)$ otherwise. For $\tilde{\beta}_k^l$, the following proposition provides an efficient method by exploiting the connected components of the graph $G := (V, E_{\mathcal{G}^{+,D}\setminus S_k^{l,D}})$:

Proposition 6.1. *Denote* $G := (V, E_{\mathcal{G}^{+,D}\backslash S_k^{l,D}})$ *where* $V := \{1, ..., p\}$ *and* $E_{\mathcal{G}^{+,D}\backslash S_k^{l,D}}$ *denotes the edge set induced by* $\mathcal{G}^{+,D}\backslash S_k^{l,D}$*:* $(i,j) \in \mathcal{G}^{+,D}\backslash S_k^{l,D} \implies \tilde{\beta}_k^l(i) = \tilde{\beta}_k^l(j)$*. If* $G := G_1 \cup G_2 \cup ... \cup G_C$ *contains* C *connected components with* $G_k := (V_c, E_c)$*, we have*

$$\tilde{\beta}_k^l(j) = \begin{cases} 0 & V_c \cap \left(\mathcal{G}^{+,I}\backslash S_k^{l,I}\right) \neq \emptyset \\ \overline{\beta_k^l(V_c)} \mathbb{1}\left(\overline{\beta_k^l(V_c)} \geq 0\right) & V_c \cap \left(\mathcal{G}^{+,I}\backslash S_k^{l,I}\right) = \emptyset \end{cases}, \text{ for all } V_c \text{ and } j \in V_c. \tag{33}$$

Moreover, the complexity of calculating $\tilde{\beta}^l(t)$ *is* $O(p^2 + m)$*, which is less than the complexity of gradient descent w.r.t.* β*,* i.e.*,* $O(n^2p^2 + m)$*.*

Proof. Since $D_G(\mathcal{G}^{+,D}\backslash S_k^{l,D}, :)$ is the graph-difference matrix, then if two nodes are connected by an edge in, they share the same value. We can thus obtain the connected components of $G := G_1 \cup G_2 \cup ... \cup G_C$ via DFS (Depth First Search) with complexity $O(p + m)$, in which nodes in each component share the same value. If $V_c \cap (\mathcal{G}^{+,D}\backslash S_k^{l,D}) \neq \emptyset$, then we have $\tilde{\beta}_k^l(V_k) = 0$ the $\tilde{\beta}_k^l$ on V_c shares the same value and each element $\mathcal{G}^{+,I}\backslash S_k^{l,I}$ should be 0 due to the constraint. If otherwise $V_c \cap \mathcal{G}^{+,I}\backslash S_k^{l,I} = \emptyset$, then we optimize over $\tilde{\beta}_k^l(V_c)$ to minimize $\|\tilde{\beta}_k^l(V_c) - \beta_k^l(V_c)\|_2$. Since each elements $\tilde{\beta}_k^k(V_c)$ are the same (denoted by a), then we optimize a to minimize $\sum_{j \in V_c}(a - \beta_k^l(j)^2$. The minimizer is thus the average of the vector $\beta_k^l(V_c)$, *i.e.*, $\overline{\beta_k^l(V_c)}$ if $\overline{\beta_k^l(V_c)} \geq 0$; otherwise, due to the non-negativity constraint we have $\tilde{\beta}_k^l(V_c) = 0$. The computation of the average for each component is $O(p)$. Since there are at most p components, then the overall complexity is $O(p^2 + m)$. □

7 NUMERICAL EXPERIMENTS

In this section, we apply our method to synthetic data. To measure the model selection consistency of Heter-LBI, we calculate the Area Under Curve (AUC).

Data Generation. We set $n = 500$, $p = h \times w = 1600$ with $h = w = 40$. We first generate the matrix form of the signal, *i.e.*, $B^* \in \mathbb{R}^{h \times w}$ such that $B^*(i,j) = 1$ if $21 \leq i \leq 27$, $21 \leq j \leq 27$; $= -1$ if i, j are odd numbers less than 15; otherwise $= 0$. For D_G with $G := (\mathcal{V}, \mathcal{E})$, we set $\mathcal{E} := \{((i_1, j_1), (i_2, j_2)) : i_1, i_2 \leq h, j_1, j_2 \leq w, |i_1 - i_2| + |j_1 - j_2| = 1\}$. We then vectorize B^* as $\beta^* \in \mathbb{R}^{p \times 1}$, with $\beta^{*,+} := \max(\beta^*, 0)$ and $\beta^{*,-} := \max(-\beta^*, 0)$. Then we set $\gamma^{*,+} := [I, D_G^\top]^\top \beta^{*,+}$ and $\gamma^{*,-} := \beta^{*,-}$. We generate $X \in \mathbb{R}^{n \times p}$ where $X_{i,j} \sim_{i.i.d} \mathcal{N}(0,1)$. We fix X and generate $y = X\beta^* + \varepsilon$, where $\varepsilon_i \sim_{i.i.d} \mathcal{N}(0,1)$.

Implementation Details. We run Heter-LBI with $\kappa = 10$, $\nu = 0.2$ and $\alpha = \frac{1}{\kappa \|H\|_2}$. To calculate AUC, we record the selection time $t_\gamma := [t_1, ..., t_m]$ with $t_i := \min_{t>0}\{\gamma_i(t) \neq 0\}$ and the selection order of each element according to t_γ, *i.e.*, $\pi_\gamma := [i_1, ..., i_m]$ such that $t_{i_1} \leq t_{i_2} \leq ... \leq t_{i_m}$. Then the AUC of π_γ and the Oracle selection vector $[\mathbb{1}(i_1 \in S), ..., \mathbb{1}(i_m \in S)]$ defined by S can measure the goodness of variable selection in the solution path. That is, the higher the AUC implies that variables in S are selected earlier than variables in S^c.

Table 4: Mean and the std of AUC over 20 times for Heter-LBI (Transposed).

	S^+	S^-	S	S_β
mean	0.9998	0.9849	0.9936	0.9608
std	0.0000	0.0005	0.0002	0.0016

Results Analysis. To remove the randomness effect, we repeat for 20 times. We report the mean and the standard deviation (std) of AUC for $S^+ := \text{supp}(\gamma^{*,+})$, $S^- := \text{supp}(\gamma^{*,-})$, S and $S^\beta := \text{supp}(\beta^*)$ in Tab. 4. As shown, our method can select the true signals for both γ and β.

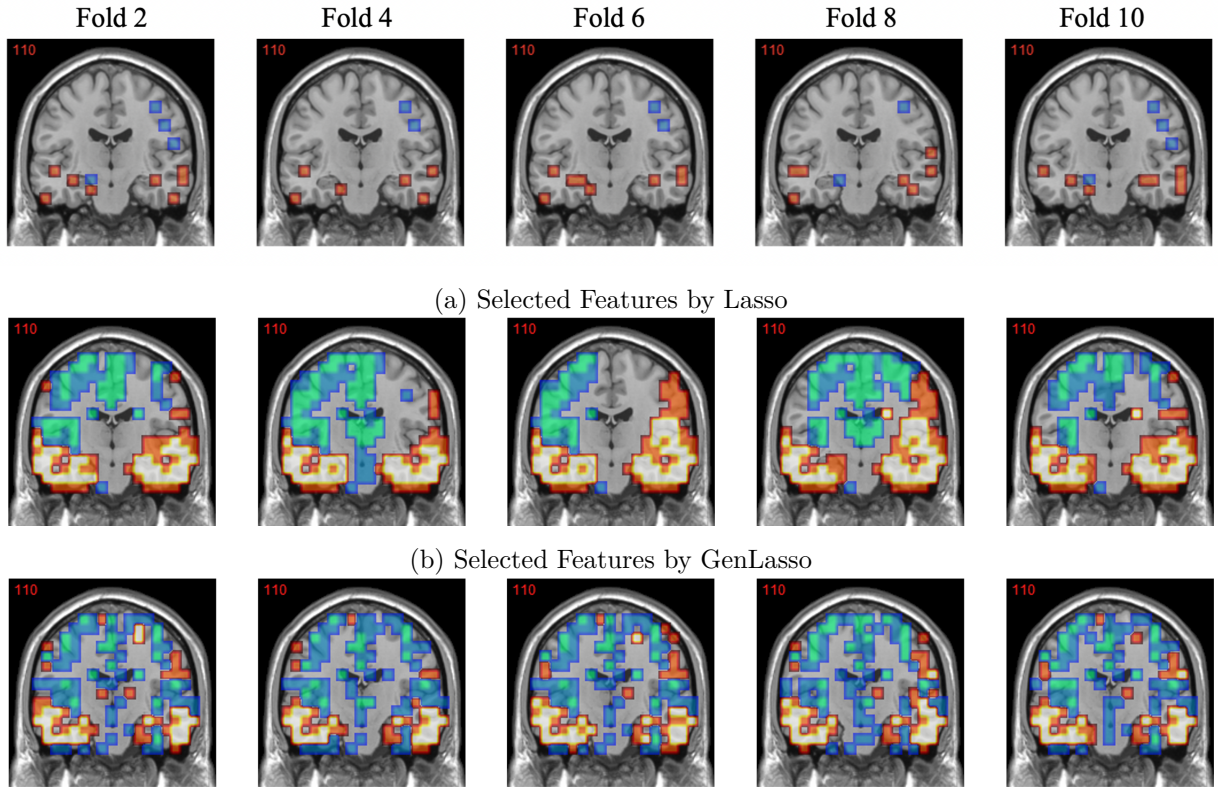
Visualization of Reconstruction. To illustrate the estimation error, we also visualize original "lesion features" $\beta^{*,+}$ (marked by red) and "procedural bias" $-\beta^{*,-}$ (marked by blue) in Fig. 2. As shown, the lesions are spatially clustered and thus satisfy the spatial coherence while the procedural bias is dispersedly distributed and thus is implemented by the ℓ_1 sparsity. Besides, our method can reconstruct both types of features well (the 2nd and 4th columns) with minor residues (the 3rd and the 5th columns).

8 MORE EXPERIMENTAL RESULTS ON AD

More about ADNI dataset. We consider $p = 2,527$ voxels, which have an average value in the GM population template that is greater than 0.1.

8.1 Visualization of ADNI

We visualize the selected lesions (marked by orange) and procedural bias (marked by blue) on the task of 30ADNC (Fig. 4) and 15ADNC (Fig. 5). The results are similar for both tasks. They show that some selected lesions by our method and GenLasso are spatially clustered into the Thalamus, which has been found to be an early degenerating region among patients (Aggleton et al., 2016).



(a) Selected Features by Lasso

(b) Selected Features by GenLasso

(c) Selected Features by our method

Figure 4: Visualization of (selected) lesion features (marked by orange) and the procedural bias (marked by blue) at the 2nd, 4th, 6th, 8th, and 10th folds on the 30ADNC. Top, middle, and bottom present the selected features by Lasso, GenLasso, and our method.

8.2 Prediction of Alzheimer's Disease Assessment Scale (ADAS) score

Data Description. In this experiment, we predict Alzheimer's Disease Assessment Scale (ADAS) score (Kueper et al., 2018), which is a regression task and hence we consider the linear model, *i.e.*, $\ell(\beta) := \frac{1}{2n}\|y - \tilde{X}\beta\|_2^2$. Our data is divided into 15T and 30T, in which 15T contains 64 AD, 229 MCI, and 103 NC; and 30T contains 62 AD, 204 MCI, and 90 NC.

Results Analysis. We report the 5-fold validation relative mean squared error (MSE) in Tab. 5. As shown, our method achieves comparable results to the best among all baseline methods. However, almost all methods suffer from large errors, which may be due to that ADAS is only one of the measurements (Mohs, 1996) that also include the Mini-Mental State Examination (MMSE), Clinical Dementia Rating (CDR), The Functional Activities Questionnaire (FAQ) (A Marshall et al., 2015). Therefore, it may be less reflective of the disease status compared to the disease label y.

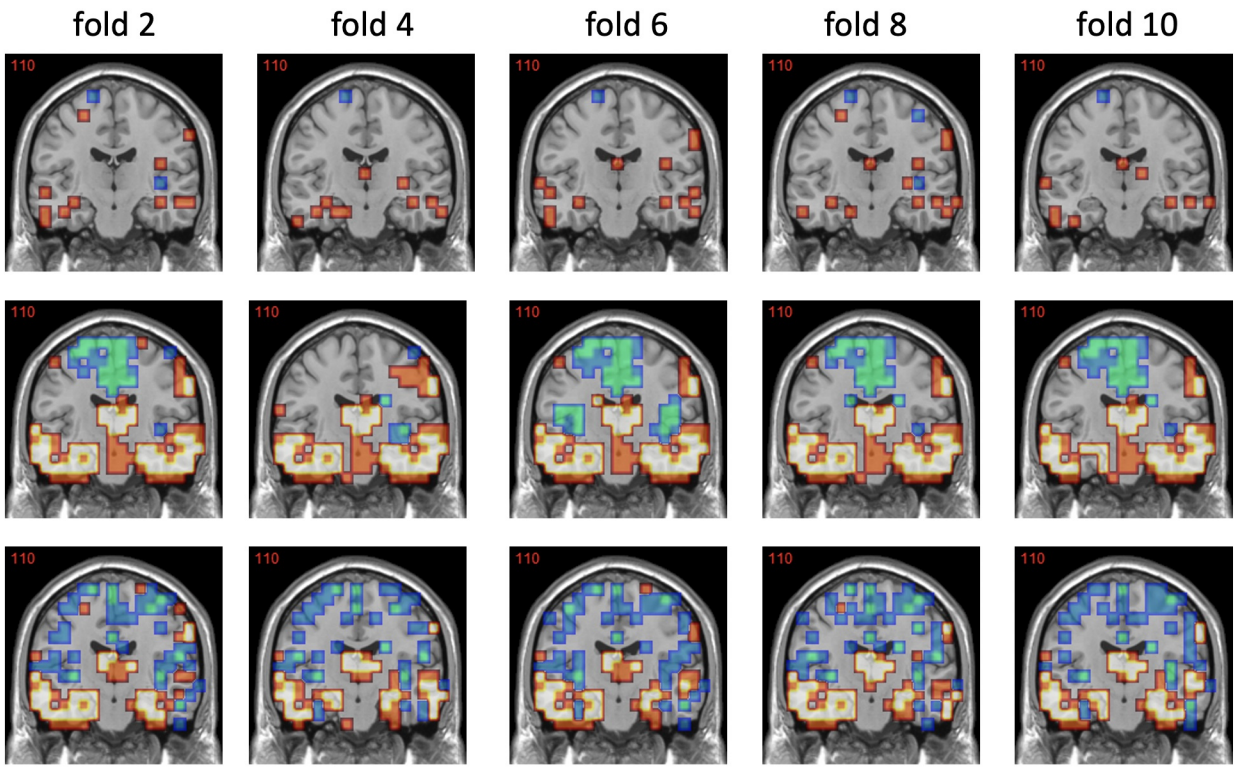


Figure 5: Visualization of (selected) lesion features (marked by orange) and the procedural bias (marked by blue) at the 2nd, 4th, 6th, 8th, and 10th folds on the 15ADNC. Top, middle, and bottom present the selected features by Lasso, GenLasso, and our method.

Table 5: Relative Mean Square Error (MSE) when y denotes ADAS.

Method	Lasso	Elastic Net	TV + ℓ_1	GSplit LBI	Ours
15T	0.1446	0.1317	0.1394	**0.1290**	0.1304
30T	0.2197	0.2048	0.2156	0.2047	**0.2020**

Stability Analysis. We in Tab. 6 report the mDC, mDC^l and mDC^p of Selected Features in 15T and 30T tasks. Our method achieves comparable mDC to the one of GSplit LBI on 30T; while on 15T our method achieves the second stable result.

Table 6: mDC, mDC^l and mDC^p of Selected Features in Regression.

	Method	Lasso	Elastic Net	GenLasso	GSplit LBI	Ours
15T	mDC	0.1408	**0.4943**	0.2155	0.2521	0.3719
	mDC^l	0.1610	**0.5318**	0.2132	0.2521	0.3260
	mDC^p	0.0450	**0.3911**	0.2336	N/A	0.3775
30T	mDC	0.1270	0.2462	0.2638	**0.4543**	0.4154
	mDC^l	0.1509	0.2642	0.3287	**0.4543**	0.3768
	mDC^p	0.0904	0.2105	0.0926	N/A	**0.4129**

Visualization of Selected Features. We visualize the selected features on 15T and 30T in Fig. 6 and Fig. 7, respectively. Although Lasso can well select procedural bias, their selected lesions are dispersedly distributed. Besides, some features that are selected as procedural bias are taken as lesions by GenLasso, which may be due to the over-smoothing of GenLasso in feature selection.

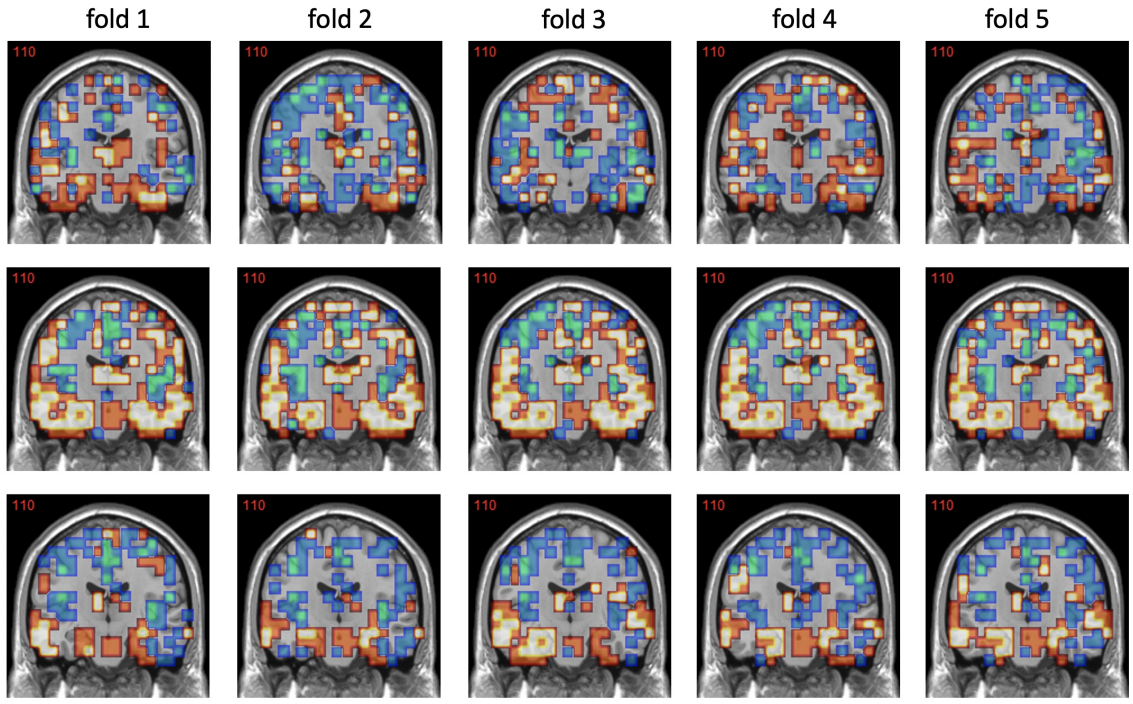


Figure 6: Visualization of (selected) lesion features (marked by orange) and the procedural bias (marked by blue) at the 1-5 folds on the 15T. Top, middle, and bottom present the selected features by Lasso, GenLasso, and our method.

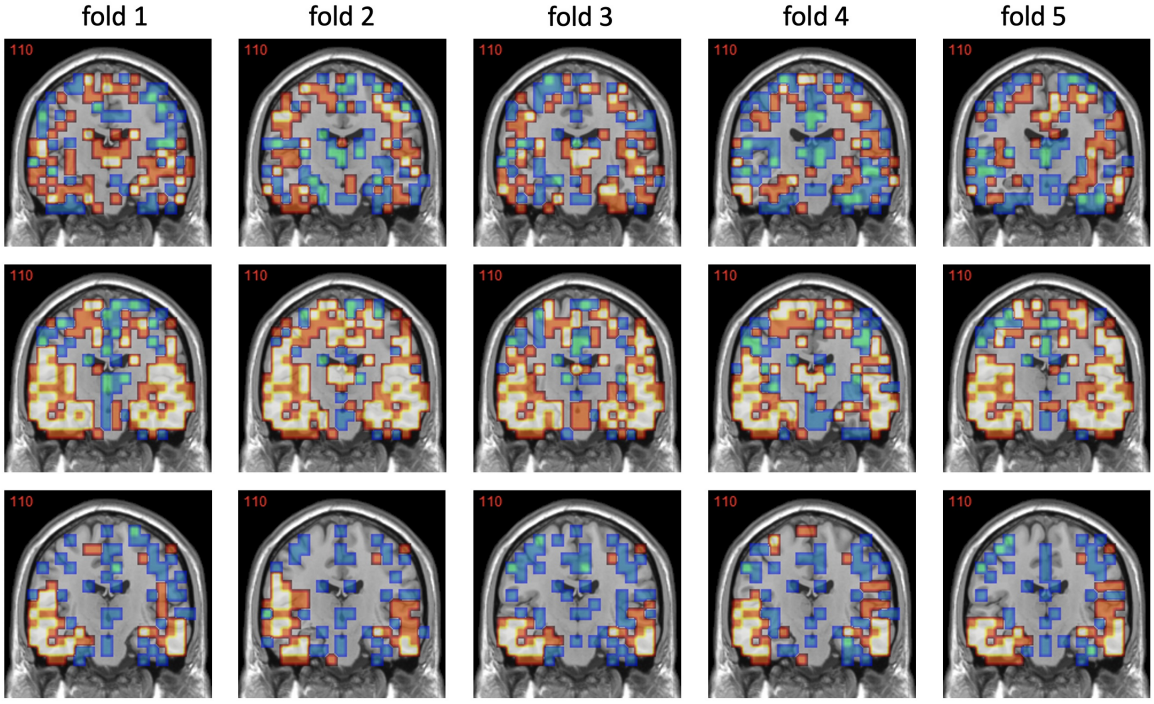


Figure 7: Visualization of (selected) lesion features (marked by orange) and the procedural bias (marked by blue) at the 1-5 folds on the 30T. Top, middle, and bottom present the selected features by Lasso, GenLasso, and our method.