Split Knockoffs for Controlling False Discoveries in Transformational Sparsity: A Robust Framework for High-Dimensional Data

Chris Junchi Li$^\diamond$

Department of Electrical Engineering and Computer Sciences$^\diamond$
University of California, Berkeley

Abstract

Variable selection plays a central role in high-dimensional statistical modeling, particularly in scenarios where the underlying parameter structure exhibits transformational sparsity. This paper presents a novel methodology, called the Split Knockoff method, designed to control the False Discovery Rate (FDR) in the selection of non-zero coefficients under linear transformations of parameters. The key idea is to relax the linear constraint into an unconstrained problem using variable splitting, leading to orthogonal designs that facilitate the construction of Split Knockoff copies. Our approach is robust even in high-dimensional settings where the number of features exceeds the number of samples. We demonstrate the effectiveness of the Split Knockoff method through simulation studies and an application to Alzheimer's Disease data, showing its power in identifying significant regions and connections in brain imaging. Theoretical guarantees for FDR control are provided under general conditions.

Keywords: False Discovery Rate, Transformational Sparsity, Split Knockoffs, High-Dimensional Statistics, Variable Selection

1 Introduction

The problem of variable selection is crucial in modern high-dimensional data analysis, where the goal is to identify a small set of relevant features from a large number of candidates. When the underlying model exhibits some form of transformational sparsity, such as linear transformations of parameters, traditional variable selection techniques can fall short. In many practical applications, the sparsity structure may be embedded in transformed parameters, requiring specialized methodologies for accurate selection and false discovery control.

Existing approaches to controlling the False Discovery Rate (FDR) include the original Knock-off filter for standard linear models and the Model-X Knockoff for more general settings. These methods provide rigorous FDR control, but they are often limited to direct sparsity scenarios where the sparsity occurs in the raw parameters rather than in their transformations. Generalizing these methods to handle transformational sparsity remains a significant challenge, particularly when the transformation matrix is non-trivial or when the problem dimension is high.

In this paper, we introduce a new method called *Split Knockoffs* to address the limitations of existing techniques in transformational sparsity settings. The core idea is to use a variable splitting approach to relax the original linear constraints, transforming the problem into an orthogonal design setting. By partitioning the data into two subsets, we construct knockoff copies that maintain crucial statistical properties for FDR control while allowing flexible handling of transformations. This enables the Split Knockoff approach to perform well even in scenarios where traditional methods would fail, such as when the transformation matrix has more rows than columns or exhibits strong correlations with the design matrix.

Backgrounds. Variable selection or sparse model selection is a fundamental problem in statistical research. Equipped with the wide spread of modern data acquisition facilities, one can simultaneously measure a large number of covariates or features and it is desired to discover a relatively small amount of dominant factors governing the variations. In many scenarios, such a sparsity constraint does not always rely on a small number of measured covariates or features, but is about some transformations, often linear, of parameters. For instance, in signal processing such as images, sparsity of edges or jumps lies in wavelet transforms [DJ95] or total variations [ROF92, CDOS12]; in genomic studies of human cancer, sparsity in 1-D fused LASSO [TSR+05] is associated with abnormality of copy numbers of genome orders in comparative genomic hybridization (CGH) data, a valuable way of understanding human cancer; in trend filtering [KKBG09], sparsity lies in the change point detection of piece-wise linear time series; in statistical ranking, sparsity of graph gradients disclose the partial orders or candidate groups in ties [HSXY16, HSXY20].

In this paper, consider the following *transformational sparsity* or *structural sparsity* problem in a linear regression where a linear transformation of parameters is sparse.

$$y = X\beta^* + \varepsilon, \ \gamma^* = D\beta^*, \tag{1}$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix, $\beta^* \in \mathbb{R}^p$ is the unknown coefficient vector, $D \in \mathbb{R}^{m \times p}$ is the linear transformer, $\gamma^* \in \mathbb{R}^m$ is the sparse vector, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ is Gaussian noise. Our purpose is to recover the support set of γ^*. For shorthand notations, we define the nonnull set $S_1 = \text{supp}(\gamma^*) = \{i : \gamma_i^* \neq 0\}$, and the null set $S_0 = \{i : \gamma_i^* = 0\}$. Note that if we take $m = p$ and $D = I_p$, this model is degenerated into the standard sparse linear regression. Hence model (1) can be viewed as a generalization of the traditional sparse regression problem.

Example. Considering brain imaging data analysis for Alzheimer's Disease, y represents the Alzheimer's Disease Assessment Scale (ADAS) of patients, $X_{i,j}$ measures the gray matter volume of brain region j in the cerebrum brain of subject i. Taking the identity matrix $D = I$ (where $m = p = 90$), one searches for highly atrophy brain regions for AD; taking D as the graph gradient operator on the brain region connectivity graph (where $m = 463 > p = 90$), abnormal connections of brain regions due to disease progression are discovered. Sparsity associated with various linear transformations above discloses both important lesion regions that undergo severe damages in disease progression and highly differential connections that link stable regions to Hippocampus, one of the most important regions accounting for Alzheimer's Disease [JLPS99]. Such discoveries based on the methodology in this paper are illustrated by Figure 1, whose details will be discussed in Section 7.

To evaluate the accuracy of an empirically discovered set $\widehat{S}$ of non-null parameters, the false discovery rate (FDR) is the goal of this paper in favor of the reproducibility of discoveries. Formally speaking, FDR of a selection set $\widehat{S}$ is defined as

$$\text{FDR} = \mathbb{E} \left[\frac{|\{i : i \in \widehat{S} \cap S_0\}|}{|\widehat{S}| \vee 1} \right]. \tag{2}$$

The problem of controlling the False Discovery Rate (FDR) has been widely studied since [BH95]. In particular, [BC+15] recently proposed the knockoff method for sparse linear regression when $D = I_p$, with a theoretical guarantee of an upper bound of FDR. This work has been further applied to group sparse models, multi-task regression models [DB16], Huber's robust regression

for outlier detection [XXCY16], and high-dimensional scenarios [BC$^+$19]. In particular, [CFJL18] proposed Model-X knockoff for random design, and [BCS$^+$20] showed that Model-X knockoff is robust to estimation error of random design parameters. Deep knockoff [RSC19] has been developed for nonparametric random designs, and side information is considered in [RC20]. Derandomized Knockoffs have been proposed in [RWC21]. However, for the general constraint $\gamma = D\beta$, it is not clear how to apply such knockoff methods to transformational sparsity, except for some special cases that can be reduced to sparse linear regression (see Section 5.1 for details).

To fill this gap, we propose a data-adaptive selection method for the transformational sparsity problem by developing a new method in the framework of knockoffs, called the *Split Knockoff* method. In our approach, the linear submanifold constraint $\gamma = D\beta$ is relaxed into its Euclidean neighborhood with a proximity parameter ν, which is known as the variable splitting technique in optimization. This leads to an orthogonal design as well as orthogonal knockoff copies. To overcome the challenge of exchangeability failure, the whole dataset is further split into two independent subsets, with β being estimated on one and sparse γ selected on the other with new designs of Split Knockoff statistics. Equipped with orthogonal Split Knockoff copies, the crucial new statistics design has signs as independent Bernoulli random variables, decoupled from magnitudes, that enables the inverse supermartingale inequalities to guarantee FDR control. It also enables us to handle some traditional types of knockoff statistics whose signs and magnitudes are dependent, via filtration refinement. The ν-relaxation above can be used for power optimization as a trade-off between incoherence improvement and loss of weak signals.

The methodology is validated by simulation experiments and applied to the study of Alzheimer's Disease, where atrophy brain regions and their abnormal connections are successfully discovered.

1.1 Contributions

The main contributions of this paper are summarized as follows:

- We propose a novel Split Knockoff method that generalizes the Knockoff framework to handle transformational sparsity by using a relaxation technique based on variable splitting.

- Theoretical guarantees for FDR control are established, leveraging new inverse supermartingale structures for robust analysis.

- We extend the methodology to high-dimensional settings by incorporating a screening procedure, allowing FDR control even when the number of features exceeds the number of samples.

- Extensive simulations demonstrate the effectiveness of the proposed method in various scenarios, including comparisons with standard Knockoffs. We also apply the methodology to real data from Alzheimer's Disease studies, showcasing its practical utility in medical imaging.

1.2 Organization of the Paper

The remainder of the paper is organized as follows:

- In Section 2, we present the methodology of Split Knockoffs, detailing the variable splitting technique and the construction of Split Knockoff statistics.

- Section 3 provides theoretical results on FDR control, including the development of new inverse supermartingale structures to address challenges related to heterogeneous noise.

3

- In Section 4, we extend the methodology to high-dimensional settings, discussing how to adapt the approach when the number of features exceeds the number of samples.

- Section 5 discusses several important topics: (i) the limitations of Knockoffs with generalized LASSO in transformational sparsity settings, (ii) the challenges of establishing theoretical FDR control due to the failure of exchangeability, and (iii) the impact of ν-relaxation on model selection consistency in Split LASSO regularization paths.

- In Section 6, we conduct simulation experiments to evaluate the performance of Split Knockoffs, comparing the method's FDR control and power with other approaches.

- Section 7 presents an application to Alzheimer's Disease using MRI data, where Split Knockoffs identify significant brain atrophy regions and abnormal inter-regional connections.

- Finally, Section 8 concludes the paper and outlines directions for future research.

Details of the proofs and additional results related to the Alzheimer's Disease application are provided in the supplementary material.

2 The Split Knockoff Method

Let's start with the generalized LASSO [TT$^+$11], the most popular method to handle the transformational sparsity. Recall that the generalized LASSO solves the following optimization problem with $\lambda > 0$,

$$\min_{\beta} \ \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\gamma\|_1, \quad \text{subject to } \gamma = D\beta. \tag{3}$$

However, the major hurdle to prohibit a knockoff design for adaptive variable selection lies in the linear constraint $\gamma = D\beta$ under general D, see Section 5.1 for detailed discussions.

To overcome this challenge, a natural treatment here is to relax the linear constraint $\gamma = D\beta$ to its Euclidean neighbourhood (proximity) in the lifted parameter space (β, γ), rendering an unconstrained optimization problem,

$$\min_{\beta, \gamma} \frac{1}{2n}\|y - X\beta\|_2^2 + \frac{1}{2\nu}\|D\beta - \gamma\|_2^2 + \lambda\|\gamma\|_1, \quad \lambda > 0, \tag{4}$$

where $\nu > 0$ is a parameter to control the Euclidean gap between $D\beta$ and γ. In other words, we shall allow model parameters to vary in the neighborhood or proximity of the linear subspace $D\beta = \gamma$, where the larger is ν, the larger is the relaxation proximity. Such a ν-relaxation renders an *orthogonal design* (identity matrix) for γ, leading to *orthogonal Split Knockoff* features that are crucial for FDR control, as well as improving the selection power to identify strong nonnull features in the presence of noise, as we shall see below. Such a treatment is known as the variable splitting in optimization, hence (4) is called *Split LASSO* in this paper.

Another critical treatment in our methodology is that, instead of using the whole dataset, we randomly split the data into two independent subsets, with one to estimate non-sparse intercept parameter β and the other to construct knockoffs for the selection of sparse parameter γ. Such a treatment is crucial to enable independent signs of Split Knockoff statistics, as Bernoulli random variables, recovering supermartingale structures for provable FDR control.

The detailed procedure goes as follows.

2.1 Data Splitting and Intercept Estimation on the First Dataset

Data Splitting. The dataset $\mathcal{D} = (X, y)$ is randomly split into two subsets as $\mathcal{D}_1 = (X_1, y_1)$ and $\mathcal{D}_2 = (X_2, y_2)$ with n_1 and n_2 samples respectively, where $n_1 + n_2 = n$, and $n_2 \geq m + p$.[1]

Estimation of Intercept $(\widehat{\beta}(\lambda))$ with $\mathcal{D}_1$. In the following, we present two typical choices of $\widehat{\beta}(\lambda)$ estimated with $\mathcal{D}_1$. Both choices are later shown to have theoretical FDR control (detailed in Theorem 1 in Section 3) and demonstrate the desired selection power empirically (see Sections 6 and 7 for more details).

1. Take $\widehat{\beta}(\lambda)$ from the Split LASSO regularization path (4), based on the first dataset $\mathcal{D}_1 = (X_1, y_1)$.

2. Alternatively, to maximize power, one can take $\widehat{\beta}(\lambda) = \widehat{\beta}_{\widehat{\lambda}, \widehat{\nu}}$ as an optimal estimator with minimal cross-validation loss (with respect to the parameters λ and ν) on the Split LASSO path.

In fact, the intercept $\widehat{\beta}(\lambda)$ determined by $\mathcal{D}_1$ can be any continuous function with respect to λ that satisfies[2] $\lim_{\lambda \to \infty} \frac{\widehat{\beta}(\lambda)}{\lambda} = 0$ to achieve theoretical FDR control, as stated in Theorem 1. However, we recommend learning $\widehat{\beta}(\lambda)$ from $\mathcal{D}_1$ to demonstrate the desired selection power in addition to FDR control. Examples of $\widehat{\beta}(\lambda)$ that satisfy the above conditions include constant estimators and the solution paths provided above.

2.2 Construction of Knockoff Matrix with the Second Dataset

Now we construct fake knockoff features based on the second dataset $\mathcal{D}_2 = (X_2, y_2)$. For this purpose, the transformational sparsity model (1) on the second dataset $\mathcal{D}_2 = (X_2, y_2)$ can be rewritten as the following linear regression in the lifted parameter (β^*, γ^*) space with heterogeneous noise:

$$\widetilde{y} = A_\beta \beta^* + A_\gamma \gamma^* + \widetilde{\varepsilon}, \tag{5}$$

where we denote ε_2 to be the Gaussian noise in y_2 and

$$\widetilde{y} = \begin{pmatrix} \frac{y_2}{\sqrt{n_2}} \\ 0_m \end{pmatrix}, \ A_\beta = \begin{pmatrix} \frac{X_2}{\sqrt{n_2}} \\ \frac{D}{\sqrt{\nu}} \end{pmatrix}, \ A_\gamma = \begin{pmatrix} 0_{n_2 \times m} \\ -\frac{I_m}{\sqrt{\nu}} \end{pmatrix}, \ \widetilde{\varepsilon} = \begin{pmatrix} \frac{\varepsilon_2}{\sqrt{n_2}} \\ 0_m \end{pmatrix}. \tag{6}$$

Knockoff Construction on $\mathcal{D}_2$. The split knockoff copy matrix $\widetilde{A}_\gamma$ satisfies

$$\widetilde{A}_\gamma^T \widetilde{A}_\gamma = A_\gamma^T A_\gamma, \ A_\beta^T \widetilde{A}_\gamma = A_\beta^T A_\gamma, \ A_\gamma^T \widetilde{A}_\gamma = A_\gamma^T A_\gamma - \mathrm{diag}(s), \tag{7}$$

where $s \in \mathbb{R}^m$ is some non-negative vector. Since the original features A_γ is an *orthogonal design*, the Split Knockoff matrix, $\widetilde{A}_\gamma$, is thus an *orthogonal matrix*, that imitates the inner product or angles of features in A_γ, but is different to the original features as much as possible. In particular, the existence of the nonsparse intercept features further requires that $\widetilde{A}_\gamma$ imitates the inner product or angles between A_γ and A_β. Explicit solutions of Equation (7) when $n_2 \geq m + p$ are discussed

---

[1] When this constraint is not satisfied, one can use the first dataset for variable screening to reduce the dimension as presented in Section 4.

[2] This condition is proposed to ensure (by Proposition 8) that Z defined in Section 2.3 won't be infinite.

in Section A. For convenience, $\widetilde{A}_\gamma$ is partitioned according to that of A_γ, $\widetilde{A}_\gamma^T = (\widetilde{A}_{\gamma,1}^T; \widetilde{A}_{\gamma,2}^T)^T$, i.e. $\widetilde{A}_{\gamma,1} \in \mathbb{R}^{n_2 \times m}$ is the submatrix consisting of the first n_2 rows of $\widetilde{A}_\gamma$ and $\widetilde{A}_{\gamma,2} \in \mathbb{R}^{m \times m}$ is the remaining submatrix.

2.3 Feature and Knockoff Significance

Given $(\widetilde{y}, A_\beta, A_\gamma, \widetilde{A}_\gamma)$ and $\widehat{\beta}(\lambda)$, we are now ready to compute feature and knockoff significance statistics.[3]

Feature Significance (Z) on $\mathcal{D}_2$.

1. Compute the Split LASSO regularization path for γ,

$$\gamma(\lambda) := \arg\min_\gamma \frac{1}{2}\|\widetilde{y} - A_\beta \widehat{\beta}(\lambda) - A_\gamma \gamma\|_2^2 + \lambda\|\gamma\|_1, \quad \lambda > 0. \tag{8}$$

2. Define the feature significance as the supremum of $\lambda > 0$ on the regularization path $\gamma_i(\lambda)$ such that $\gamma_i(\lambda)$ is nonzero for all $i \in \{1, 2, \cdots, m\}$:

$$Z_i = \sup\{\lambda > 0 : \gamma_i(\lambda) \neq 0\}, \tag{9}$$

or $Z_i = 0$ if the set $\{\lambda > 0 : \gamma_i(\lambda) \neq 0\}$ is empty.

Knockoff Significance ($\widetilde{Z}$) on $\mathcal{D}_2$.

1. Compute the Split LASSO regularization path for $\widetilde{\gamma}$,

$$\widetilde{\gamma}(\lambda) := \arg\min_{\widetilde{\gamma}} \frac{1}{2}\|\widetilde{y} - A_\beta \widehat{\beta}(\lambda) - \widetilde{A}_\gamma \widetilde{\gamma}\|_2^2 + \lambda\|\widetilde{\gamma}\|_1, \quad \lambda > 0. \tag{10}$$

2. Define the knockoff significance as the supremum of $\lambda > 0$ on the regularization path $\widetilde{\gamma}_i(\lambda)$ such that $\widetilde{\gamma}_i(\lambda)$ is nonzero, for all $i \in \{1, 2, \cdots, m\}$:

$$\widetilde{Z}_i = \sup\{\lambda > 0 : \widetilde{\gamma}_i(\lambda) \neq 0\}. \tag{11}$$

or $\widetilde{Z}_i = 0$ if the set $\{\lambda > 0 : \widetilde{\gamma}_i(\lambda) \neq 0\}$ is empty.

In summary, Equation (8) and (10) define regularization paths $\gamma(\lambda)$ and $\widetilde{\gamma}(\lambda)$ that typically transition from zero to nonzero as λ decreases from $+\infty$ to 0. The supremums of λ for which $\gamma(\lambda) \neq 0$ or $\widetilde{\gamma}(\lambda) \neq 0$ indicate the importance of γ and $\widetilde{\gamma}$, respectively. The larger these supremums, the more important the corresponding features are. These supremums are recorded as the feature significance Z or the knockoff significance $\widetilde{Z}$.

In addition to recording the supremums of λ for which $\gamma(\lambda) \neq 0$ or $\widetilde{\gamma}(\lambda) \neq 0$ as significance statistics, the signs of $\gamma(\lambda)$ and $\widetilde{\gamma}(\lambda)$ at such supremums can also be recorded. This allows for a *truncation of sign mismatch* on $\widetilde{Z}$ to define a new knockoff significance statistic.

---

[3]Here, we compute the regularization paths on A_γ and $\widetilde{A}_\gamma$ separately rather than jointly, which can lower the computational cost at lower dimensionality.

Formally speaking, for all i, define the reference signs r_i and $\widetilde{r}_i$ as the signs of $\gamma_i(\lambda)$ and $\widetilde{\gamma}_i(\lambda)$ "upon" becoming nonzero (i.e. at Z and $\widetilde{Z}$) when λ decreases from $+\infty$ to 0. That is,

$$r_i := \begin{cases} 1 & \text{if } Z_i > 0 \text{ and } \limsup_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = 1, \\ 0 & \text{if } Z_i = 0, \\ -1 & \text{if } Z_i > 0 \text{ and } \liminf_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = -1, \end{cases} \tag{12}$$

$$\widetilde{r}_i := \begin{cases} 1 & \text{if } \widetilde{Z}_i > 0 \text{ and } \limsup_{\lambda \to \widetilde{Z}_i^-} \operatorname{sign}(\widetilde{\gamma}_i(\lambda)) = 1, \\ 0 & \text{if } \widetilde{Z}_i = 0, \\ -1 & \text{if } \widetilde{Z}_i > 0 \text{ and } \liminf_{\lambda \to \widetilde{Z}_i^-} \operatorname{sign}(\widetilde{\gamma}_i(\lambda)) = -1. \end{cases} \tag{13}$$

Proposition 3 in Section 3 provides an equivalent definition of r and $\widetilde{r}$ through the Karush-Kuhn-Tucker (KKT) conditions of Equation (8) and Equation (10).

Using r and $\widetilde{r}$, the *sign mismatch truncation* of $\widetilde{Z}$ is defined by

$$\tau(\widetilde{Z}) := \widetilde{Z} \odot \mathbf{1}\{r = \widetilde{r}\} \tag{14}$$

where the symbol $\odot$ stands for Hadamard product. Such a truncation sets zeroes those knockoff significance values if their signs differ to their associated feature significance values. As we shall see later in this paper, the truncated knockoff significance may reduce the conservativeness in selections compared with the untruncated one.

2.4 Three Types of Split Knockoff Statistics

We introduce a family of Split Knockoff W statistics.

Split Knockoff Statistics.

1. $W^{\mathrm{S}} := Z \odot \operatorname{sign}(Z - \widetilde{Z})$[4], where S refers to "Split".

2. $W^{\mathrm{S}\tau} := Z \odot \operatorname{sign}(Z - \tau(\widetilde{Z}))$, where S$\tau$ refers to applying truncation (14) on W^{S}.

3. $W^{\mathrm{BC}} := (Z \vee \widetilde{Z}) \odot \operatorname{sign}(Z - \widetilde{Z})$, where BC refers to the original definition adopted by Barber-Candès in [BC+15].

All three versions will be handled in a unified framework in this paper. For shorthand notation, we use $W^\star$ to represent any one in the family, where $\star \in \{\mathrm{S}, \mathrm{S}\tau, \mathrm{BC}\}$. In all cases, as we wish to select i when $W_i^\star$ is large and positive. Let q be our target FDR, two data-dependent threshold rules on a pre-set nominal level q are defined as

$$\text{(Split Knockoff)} \quad T_q^\star = \min\left\{\lambda \in \mathcal{W}^\star : \frac{|\{i : W_i^\star \le -\lambda\}|}{1 \vee |\{i : W_i^\star \ge \lambda\}|} \le q\right\},$$

$$\text{(Split Knockoff+)} \quad T_q^\star = \min\left\{\lambda \in \mathcal{W}^\star : \frac{1 + |\{i : W_i^\star \le -\lambda\}|}{1 \vee |\{i : W_i^\star \ge \lambda\}|} \le q\right\},$$

or $T_q^\star = +\infty$ if this set is empty, where $\mathcal{W}^\star = \{|W_j^\star| : j = 1, 2, \cdots, m\} \backslash \{0\}$. In all cases, the selector is defined as

$$\widehat{S}^\star = \{i : W_i^\star \ge T_q^\star\}.$$

---

[4]It is shown in Proposition 8 that $\{\exists i : Z_i = \widetilde{Z}_i > 0\}$ is a zero probability event. For simplicity of notations, the event $\{\exists i : Z_i = \widetilde{Z}_i > 0\}$ will be omitted throughout this paper.

The following proposition summarizes an important inclusive relationship among the three selectors.

Proposition 1 (Inclusion Property of Selectors)**.** *For the selectors $\widehat{S}^{\mathrm{BC}}$, $\widehat{S}^{\mathrm{S}}$ and $\widehat{S}^{\mathrm{S}\tau}$, there holds*

$$\widehat{S}^{\mathrm{BC}} \subseteq \widehat{S}^{\mathrm{S}} \subseteq \widehat{S}^{\mathrm{S}\tau}.$$

In other words, $\widehat{S}^{\mathrm{S}\tau}$ achieves the highest selection power among these selectors while $\widehat{S}^{\mathrm{BC}}$ is the most conservative. This relationship is also validated by simulation experiments in Section 6. The proof of Proposition 1 is given in Section B.1. Next section will disclose that they all achieve the desired FDR control.

3   FDR Control of Split Knockoffs

In this section, we first show that the false discovery rates are under control for the Split Knockoff method, then present the key ideas on how to reach this analysis. In particular, we consider the basic setting with $n_2 \geq m + p$ here, leaving the high dimensional extension to Section 4.

Specifically, the following theorem of FDR control is established for all three selectors. For Split Knockoff, we control a "modified" FDR (mFDR) as [BC$^+$15] that adds q^{-1} in the denominator, which should have little effect if a large number of features are selected; for Split Knockoff+, we get the exact FDR control.

Theorem 1 (FDR Control of Split Knockoffs)**.** *For all $0 < q \leq 1$, $\star \in \{\mathrm{S}, \mathrm{S}\tau, \mathrm{BC}\}$, and all $\nu > 0$, there holds*

(a) *(mFDR of Split Knockoff)*
$$\mathbb{E}\left[\frac{\left|\left\{i : i \in \widehat{S}^{\star} \cap S_0\right\}\right|}{\left|\widehat{S}^{\star}\right| + q^{-1}}\right] \leq q.$$

(b) *(FDR of Split Knockoff+)*
$$\mathbb{E}\left[\frac{\left|\left\{i : i \in \widehat{S}^{\star} \cap S_0\right\}\right|}{\left|\widehat{S}^{\star}\right| \vee 1}\right] \leq q.$$

Since the FDR is uniformly under control for all $\nu > 0$, the hyperparameter ν from Split LASSO can be used to optimize the power of Split Knockoffs. The influence of ν on power is of two folds (see Section 5.3). On one hand, increasing ν may increase the power of discovering strong nonnull features whose magnitudes are large. It is because that enlarging ν will improve the incoherence condition of the Split LASSO such that strong nonnull features will appear earlier on the path than the weak ones and nulls. On the other hand, increasing ν may lose the power of discovering weak nonnull features whose magnitudes are below ν and the noise scale. Therefore a trade-off of these two aspects will lead to a good power by optimizing ν, which can be achieved empirically by cross-validation as what we describe in intercept estimation $\widehat{\beta}$. Performance of such a choice will be confirmed in simulation experiments in Section 6.

The main challenge in establishing the FDR control of Split Knockoffs lies in the failure of exchangeability (see Section 5.2 for more discussions). Such a failure leads to dependency between the magnitude and sign of the original type of W statistics (W^{BC}), which further fails the inverse

supermartingale argument in [BC+15, BC+19] for provable FDR control. To overcome this challenge, our key technical development is based on the orthogonal design and the rendered orthogonal Split Knockoff copies, which together with data splitting leads to independent Bernoulli processes for signs of W^{S} and $W^{\mathrm{S}\tau}$, decoupled or independent to their magnitudes. This enables us some new *inverse supermartingale structures* such that the FDR control can be proved, in a similar way to [BC+19]. Furthermore, W^{S} enjoys a particularly nice structure that its associated inverse supermartingale has a more refined filtration than that of W^{BC}, which enables us to reach an upper bound for W^{BC} via W^{S}. Below we will present the key ideas in detail for the FDR control.

First of all, following the standard steps in knockoffs as in [BC+15], we can transfer the problem of bounding FDR by q in Theorem 1 into the problem of bounding $\mathbb{E}\left[\mathcal{M}_{T_q^\star}(W^\star)\right]$ by one (see Section B.5 for details), where for any $T > 0$, $\mathcal{M}_T(W^\star)$ is defined as

$$\mathcal{M}_T(W^\star) = \frac{\sum_{i \in S_0} \mathbb{1}\{W_i^\star \geq T\}}{1 + \sum_{i \in S_0} \mathbb{1}\{W_i^\star \leq -T\}}. \tag{15}$$

The following proposition reveals the relationships among $\mathcal{M}_T(W^{\mathrm{S}\tau})$, $\mathcal{M}_T(W^{\mathrm{S}})$, and $\mathcal{M}_T(W^{\mathrm{BC}})$, in particular.

Proposition 2 (Inequalities of W Statistics). *The following holds for all $T > 0$.*

1. *For all i, $\{W_i^{\mathrm{S}\tau} \leq -T\} \subseteq \{W_i^{\mathrm{S}} \leq -T\} \subseteq \{W_i^{\mathrm{BC}} \leq -T\}$.*

2. *For all i, $\{W_i^{\mathrm{BC}} \geq T\} = \{W_i^{\mathrm{S}} \geq T\} \subseteq \{W_i^{\mathrm{S}\tau} \geq T\}$.*

3. *Combining the above two points together, there holds $\mathcal{M}_T(W^{\mathrm{S}\tau}) \geq \mathcal{M}_T(W^{\mathrm{S}}) \geq \mathcal{M}_T(W^{\mathrm{BC}})$.*

Proposition 2 suggests that the statistics $W^{\mathrm{S}\tau}$ are the least conservative in terms of FDR, since $\mathcal{M}_T(W^{\mathrm{S}\tau}) = \max\{\mathcal{M}_T(W^{\mathrm{S}\tau}), \mathcal{M}_T(W^{\mathrm{S}}), \mathcal{M}_T(W^{\mathrm{BC}})\}$, where $\mathcal{M}_T(W^\star)$ is introduced to upper bound the FDR for $\star \in \{\mathrm{S}, \mathrm{S}\tau, \mathrm{BC}\}$ respectively. The proof of Proposition 2 is provided in Section B.2.

Next, we will handle the FDR control of Split Knockoff statistics W^{S} and the Barber-Candès type statistics W^{BC} in two cases, respectively, while the analysis on the sign-mismatch truncated statistics $W^{\mathrm{S}\tau}$ can be done in a similar way as W^{S}. The section is organized as follows:

1. An inverse supermartingale structure on $\mathcal{M}_T(W^{\mathrm{S}})$ is introduced, which enables the optional stopping theorem to bound $\mathbb{E}\left[\mathcal{M}_{T_q^{\mathrm{S}}}(W^{\mathrm{S}})\right]$, in Section 3.1.

2. Barber-Candès type of knockoff statistics W^{BC} renders a coarser filtration than W^{S}, which enables $\mathbb{E}\left[\mathcal{M}_{T_q^{\mathrm{BC}}}(W^{\mathrm{S}})\right]$ providing an upper bound of $\mathbb{E}\left[\mathcal{M}_{T_q^{\mathrm{BC}}}(W^{\mathrm{BC}})\right]$, in Section 3.2.

3.1 Case I: Split Knockoff Statistics

The key property of Split Knockoff statistics W^{S} is that the magnitudes $|W^{\mathrm{S}}|$ are independent of the signs $\mathrm{sign}(W^{\mathrm{S}})$, and the signs $\mathrm{sign}(W^{\mathrm{S}})$ are independent Bernoulli random variables. This property allows us to construct an inverse supermartingale, which will be presented below..

We begin with the Karush-Kuhn-Tucker (KKT) conditions of optimization problems described in Equations (8) and (10). These equations imply that the optimal solutions $\gamma(\lambda)$, $\widetilde{\gamma}(\lambda)$ with respect to $\lambda > 0$ must satisfy the following KKT conditions:

$$\lambda\rho(\lambda) + \frac{\gamma(\lambda)}{\nu} = \frac{D\widehat{\beta}(\lambda)}{\nu}, \tag{16a}$$

$$\lambda \widetilde{\rho}(\lambda) + \frac{\widetilde{\gamma}(\lambda)}{\nu} = \frac{D\widehat{\beta}(\lambda)}{\nu} + \underbrace{\left\{ -\operatorname{diag}(s)\gamma^* + \frac{\widetilde{A}_{\gamma,1}^T}{\sqrt{n_2}}\varepsilon_2 \right\}}_{=:\zeta}, \tag{16b}$$

$$\rho(\lambda) \in \partial\|\gamma(\lambda)\|_1, \tag{16c}$$

$$\widetilde{\rho}(\lambda) \in \partial\|\widetilde{\gamma}(\lambda)\|_1, \tag{16d}$$

where ε_2 is the Gaussian noise in y_2, and $\partial\|\cdot\|_1$ is the set of subgradients of ℓ_1-norm $\|\cdot\|_1$. For shorthand notation, define $\zeta := -\operatorname{diag}(s)\gamma^* + \frac{\widetilde{A}_{\gamma,1}^T}{\sqrt{n_2}}\varepsilon_2$.

Proposition 3 below provides equivalent definitions of r and $\widetilde{r}$ through the functions $\rho(\lambda)$ and $\widetilde{\rho}(\lambda)$ by exploiting the KKT conditions (16). This suggests that the definitions of r and $\widetilde{r}$ given by Equations (12) and (13) are well-defined. The proof of Proposition 3 can be found in Section B.3.

Proposition 3. *The following definitions of r and $\widetilde{r}$ are equivalent with Equation* (12) *and* (13)*:*

$$r_i := \begin{cases} \operatorname{sign}(\rho_i(Z_i)) & \text{if } Z_i > 0, \\ 0 & \text{if } Z_i = 0, \end{cases} \quad \widetilde{r}_i := \begin{cases} \operatorname{sign}(\widetilde{\rho}_i(\widetilde{Z}_i)) & \text{if } \widetilde{Z}_i > 0, \\ 0 & \text{if } \widetilde{Z}_i = 0. \end{cases} \tag{17}$$

Now we are ready to show that the magnitudes $|W^{\mathrm{S}}|$ are independent to the signs $\{\operatorname{sign}(W^{\mathrm{S}})\}$. Note that $\widehat{\beta}(\lambda)$ is determined from the first dataset $\mathcal{D}_1 = (X_1, y_1)$, whence independent to the second dataset $\mathcal{D}_2 = (X_2, y_2)$ and particularly ζ. Therefore, it suffices to consider the random variable ζ conditional on a predetermined $\widehat{\beta}(\lambda)$. From Equation (16a), $\gamma(\lambda)$ is determined by $\widehat{\beta}(\lambda)$ and thus independent to ζ, so is $|W^{\mathrm{S}}| = Z$ as a function of $\gamma(\lambda)$. On the other hand, from Equation (16b), conditional on a determined $\widehat{\beta}(\lambda)$, $\widetilde{\gamma}(\lambda)$ as well as $\widetilde{Z}$, are determined by ζ from $\mathcal{D}_2 = (X_2, y_2)$. Therefore, $|W^{\mathrm{S}}|$, determined by $\widehat{\beta}(\lambda)$ on $\mathcal{D}_1 = (X_1, y_1)$, is independent to $\operatorname{sign}(W^{\mathrm{S}}) = \operatorname{sign}(Z - \widetilde{Z})$, depending on ζ with $\mathcal{D}_2 = (X_2, y_2)$.

To see that the signs $\{\operatorname{sign}(W^{\mathrm{S}})\}$ are independent Bernoulli random variables, ζ follows an independently joint Gaussian distribution as a result of the *orthogonal* Split Knockoff matrix satisfying (7) (see Equation (43) in Section B.6 for details), *i.e.*

$$\zeta \sim \mathcal{N}\left(-\operatorname{diag}(s)\gamma^*, \frac{1}{n_2}\operatorname{diag}(s)(2I_m - \operatorname{diag}(s)\nu)\sigma^2 \right). \tag{18}$$

Then we have the following lemma.

Lemma 1. *Given any determined $\widehat{\beta}(\lambda)$, $1\{W_i^{\mathrm{S}} < 0\}$ are some independent Bernoulli random variables. Furthermore, for $i \in S_0 \cap \{i : |W_i^{\mathrm{S}}| = Z_i > 0\}$, there holds*

$$\mathbb{P}[W_i^{\mathrm{S}} < 0] \geq \frac{1}{2}.$$

This lemma states that given $\widehat{\beta}(\lambda)$ which determines $|W^{\mathrm{S}}| = Z$, $\operatorname{sign}(W^{\mathrm{S}})$ consists of independent random variables with $\mathbb{P}[W_i^{\mathrm{S}} < 0] \geq 1/2$ on the nulls satisfying $|W_i^{\mathrm{S}}| = Z_i > 0$. With such a property, we can construct an inverse supermartingale whose optional stopping theorem gives an upper bound on $\mathbb{E}\left[\mathcal{M}_{T_q^{\mathrm{S}}}(W^{\mathrm{S}})\right]$.

For simplicity, we rearrange the index of W^{S}, such that $|W_{(1)}^{\mathrm{S}}| \geq |W_{(2)}^{\mathrm{S}}| \geq \cdots \geq |W_{(m^*)}^{\mathrm{S}}| > 0$ for $\{(1), (2), \cdots, (m^*)\} = S_0 \cap \{i : |W_i^{\mathrm{S}}| = Z_i > 0\}$. Further denote $B_{(i)} = 1\{W_{(i)}^{\mathrm{S}} < 0\}$, then there holds

$$\frac{\sum_{i \in S_0} 1\{W_i^{\mathrm{S}} \geq T_q^{\mathrm{S}}\}}{1 + \sum_{i \in S_0} 1\{W_i^{\mathrm{S}} \leq -T_q^{\mathrm{S}}\}} = \frac{1 + \sum_{i \in S_0} 1\{|W_i^{\mathrm{S}}| \geq T_q^{\mathrm{S}}\}}{1 + \sum_{i \in S_0} 1\{|W_i^{\mathrm{S}}| \geq T_q^{\mathrm{S}}, W_i^{\mathrm{S}} < 0\}} - 1,$$

10

$$= \frac{1 + J}{1 + B_{(1)} + B_{(2)} + \cdots + B_{(J)}} - 1, \tag{19}$$

where $J \leq m^*$ is defined to be the index satisfying

$$|W_{(1)}^{\mathrm{S}}| \geq |W_{(2)}^{\mathrm{S}}| \geq \cdots \geq |W_{(J)}^{\mathrm{S}}| \geq T_q^{\mathrm{S}} > |W_{(J+1)}^{\mathrm{S}}| \geq \cdots \geq |W_{(m^*)}^{\mathrm{S}}|,$$

in other words, $J = \mathrm{argmax}_{k \leq m^*}\{|W_{(k)}^{\mathrm{S}}| \geq T_q^{\mathrm{S}}\}$. The following lemma summarizes the *inverse supermartingale* inequality, which gives an upper bound for the Equation (19).

Lemma 2. *Given any determined $\widehat{\beta}(\lambda)$, let $\{G_i\}_{i=1}^m$ be some proper Borel sets such that $B_i = 1\{\zeta_i \in G_i\}$ with $\mathbb{P}[B_i = 1] = \rho_i$. Let $\rho > 0$ satisfy $\rho \leq \min_{i \in S_0}\{\rho_i\}$. Let J be a stopping time with respect to the filtration $\{\mathcal{F}_j\}_{j=1}^m$ in inverse time defined as*

$$\mathcal{F}_j = \sigma\left(\left\{\sum_{i=1}^j B_{(i)}, \zeta_{(j+1)}, \cdots, \zeta_{(m)}\right\}\right)$$

Then

$$\mathbb{E}\left[\frac{1 + J}{1 + B_{(1)} + B_{(2)} + \cdots + B_{(J)}}\right] \leq \rho^{-1}.$$

One can verify that $J = \mathrm{argmax}_{k \leq m^*}\{|W_{(k)}^{\mathrm{S}}| \geq T_q^{\mathrm{S}}\}$ is indeed a stopping time with respect to the filtration $\{\mathcal{F}_j\}_{j=1}^m$ in inverse time. Applying Lemma 2 to Equation (19) with the property that $\mathbb{P}[B_i = 1] \geq \rho = 1/2$ for $i \in S_0$ as shown in Lemma 1, we reach our desired result. The same procedure can be applied to the case of $W^{\mathrm{S}\tau}$.

3.2 Case II: Barber-Candès Type Statistics

The key lies in that W^{S} provides us a bridge to the analysis of W^{BC}, enabling an upper bound on $\mathcal{M}_{T_q^{\mathrm{BC}}}(W^{\mathrm{BC}})$ via an inverse martingale inequality for $\mathcal{M}_{T_q^{\mathrm{BC}}}(W^{\mathrm{S}})$ due to the following facts.

(a) $\mathcal{M}_T(W^{\mathrm{S}})$ upper bounds $\mathcal{M}_T(W^{\mathrm{BC}})$, that $\mathcal{M}_T(W^{\mathrm{BC}}) \leq \mathcal{M}_T(W^{\mathrm{S}})$ for all $T > 0$ as presented in Proposition 2;

(b) The inverse supermartingale associated with $\mathcal{M}_T(W^{\mathrm{S}})$ has a more refined filtration than that of $\mathcal{M}_T(W^{\mathrm{BC}})$, such that the stopping time T_q^{BC} is also a stopping time of the former.

Precisely, we have the following critical proposition.

Proposition 4 (Inclusion Property of Filtration)**.** *Let $\mathcal{F}^{\mathrm{BC}}(T)$ ($\mathcal{F}^{\mathrm{S}}(T)$) be the filtration induced by W^{BC} (W^{S}) respectively, i.e.*

$$\mathcal{F}^{\mathrm{BC}}(T) = \sigma\left(\#\{i : W_i^{\mathrm{BC}} \geq T\}, \#\{i : W_i^{\mathrm{BC}} \leq -T\}, \{\zeta_i : |W_i^{\mathrm{BC}}| < T\}\right),$$
$$\mathcal{F}^{\mathrm{S}}(T) = \sigma\left(\#\{i : W_i^{\mathrm{S}} \geq T\}, \#\{i : W_i^{\mathrm{S}} \leq -T\}, \{\zeta_i : |W_i^{\mathrm{S}}| < T\}\right).$$

Then there holds for any $T > 0$ that

$$\mathcal{F}^{\mathrm{BC}}(T) \subseteq \mathcal{F}^{\mathrm{S}}(T).$$

For any stopping time T_q^{BC} in reverse time adapted to the filtration $\mathcal{F}^{\mathrm{BC}}(T)$, Proposition 4 points out that $\mathcal{F}^{\mathrm{S}}(T)$ is a superset refinement of $\mathcal{F}^{\mathrm{BC}}(T)$, whence T_q^{BC} is also a stopping time in reverse time adapted to the filtration $\mathcal{F}^{\mathrm{S}}(T)$. Applying optional stopping theorem on $\mathcal{M}_{T_q^{\mathrm{BC}}}(W^{\mathrm{S}})$, there holds

$$\mathbb{E}\left[\mathcal{M}_{T_q^{\mathrm{BC}}}(W^{\mathrm{S}})\right] \leq 1.$$

Combining with the property that $\mathcal{M}_T(W^{\mathrm{BC}}) \leq \mathcal{M}_T(W^{\mathrm{S}})$ for any $T > 0$, there further holds

$$\mathbb{E}\left[\mathcal{M}_{T_q^{\mathrm{BC}}}(W^{\mathrm{BC}})\right] \leq \mathbb{E}\left[\mathcal{M}_{T_q^{\mathrm{BC}}}(W^{\mathrm{S}})\right] \leq 1,$$

which is our desired result.

The proof of Proposition 4 will be provided in Section B.4. The complete proof of Theorem 1 can be found in Section B.5, with the proofs of supporting lemmas in Section B.6.

4 A Generalization in High Dimensional Settings

In this section, we extend the Split Knockoffs into high dimensional settings, where the sample size is limited and the condition of $n_2 \geq m + p$ cannot be satisfied. In such settings, we will conduct feature screening in the first dataset $\mathcal{D}_1$ to deduct the number of features, and then perform Split Knockoffs on the screened subset of features in the second dataset $\mathcal{D}_2$.

In particular, we will use the first dataset $\mathcal{D}_1 = (X_1, y_1)$ to conduct initial screening and give estimated support sets $\widehat{S}_\beta$, $\widehat{S}_\gamma$ for β, γ respectively, such that $n_2 \geq |\widehat{S}_\beta| + |\widehat{S}_\gamma|$ is satisfied. The feature screening is a well-studied topic, and a list of the screening methods can be found in but not limited to [WR09, WDR$^+$10]. After the screening step, the rest steps of Split Knockoffs will be conducted on the estimated support sets $\widehat{S}_\beta$, $\widehat{S}_\gamma$.

For the rest steps, we first generate the intercept $\widehat{\beta}(\lambda)$ as a bounded continuous function ($\mathbb{R}_+ \to \mathbb{R}^{|\widehat{S}_\beta|}$) from $\mathcal{D}_1$. Then it remains to generate the feature and knockoff importance statistics Z and $\widetilde{Z}$ in order to perform Split Knockoffs.

Let $X_{\widehat{S}_\beta}$ be the submatrix of X_2, consisting of the columns indexed by $\widehat{S}_\beta$. Let $D_{\widehat{S}_\beta, \widehat{S}_\gamma}$ be the submatrix of D, consisting of the columns indexed by $\widehat{S}_\beta$ and rows indexed by $\widehat{S}_\gamma$. Further, let

$$\widetilde{y} = \begin{pmatrix} \frac{y_2}{\sqrt{n_2}} \\ 0_{|\widehat{S}_\gamma|} \end{pmatrix}, \ A_\beta = \begin{pmatrix} \frac{X_{\widehat{S}_\beta}}{\sqrt{n_2}} \\ \frac{D_{\widehat{S}_\beta, \widehat{S}_\gamma}}{\sqrt{\nu}} \end{pmatrix}, \ A_\gamma = \begin{pmatrix} 0_{n_2 \times |\widehat{S}_\gamma|} \\ -\frac{I_{|\widehat{S}_\gamma|}}{\sqrt{\nu}} \end{pmatrix}, \ \widetilde{\varepsilon} = \begin{pmatrix} \frac{\varepsilon_2}{\sqrt{n_2}} \\ 0_{|\widehat{S}_\gamma|} \end{pmatrix}. \tag{20}$$

For shorthand notations, we abuse the notations here and use the same notations as in Equation (6), while keep in mind that $\widetilde{y}$, A_β, A_γ and $\widetilde{\varepsilon}$ defined in Equation (20) are dependent on the estimated support sets $\widehat{S}_\beta$, $\widehat{S}_\gamma$. With Equation (20), we can construct the split knockoff copy matrix $\widetilde{A}_\gamma$ satisfying Equation (7) in high dimensional settings. Then the feature and knockoff statistics can be generated in the same way as in Equation (9) and Equation (11).

For the FDR control about the procedure above, the following Theorem 2 states that when the estimated support set $\widehat{S}_\beta$ includes the true support set of β, such a procedure above will not cause any loss in the FDR control. The proof of Theorem 2 is given in Section B.7.

Theorem 2. *Let S_β be the true support sets for β. Let Υ be the event that $S_\beta \subseteq \widehat{S}_\beta$, then there holds for all $0 < q \leq 1$, $\star \in \{\mathrm{S}, \mathrm{S}\tau, \mathrm{BC}\}$, and all $\nu > 0$,*

(a) (mFDR of Split Knockoff)

$$\mathbb{E}\left[\left.\frac{\left|\left\{i : i \in \widehat{S}^\star \cap S_0\right\}\right|}{\left|\widehat{S}^\star\right| + q^{-1}}\right| \Upsilon\right] \le q.$$

(b) (FDR of Split Knockoff+)

$$\mathbb{E}\left[\left.\frac{\left|\left\{i : i \in \widehat{S}^\star \cap S_0\right\}\right|}{\left|\widehat{S}^\star\right| \vee 1}\right| \Upsilon\right] \le q.$$

The event that $\Upsilon = \{S_\beta \subseteq \widehat{S}_\beta\}$ is often known as the sure screening event [FL08] in the literature. It is worth to mention that an immediate corollary follows from Theorem 2, that mFDR $\le q + \mathbb{P}\left[\Upsilon^C\right]$ for Split Knockoff and FDR $\le q + \mathbb{P}\left[\Upsilon^C\right]$ for Split Knockoff+. Theorem 2 achieves comparable results with Theorem 2 in [BC+19]. The simulation experiments that validate the effectiveness of this extension is presented in Section C.2.

5 Discussion

In this section, we first discuss a straightforward way of conducting Knockoffs on generalized LASSO under a special case of transformational sparsity when D is of full row-rank, and why such a procedure does not work in general. Next, we discuss how the exchangeability property fails for Split Knockoffs and the challenges for provable FDR control. Finally, we discuss how the model selection consistency of Split LASSO may affect the selection power of Split Knockoffs.

5.1 Generalized LASSO and Knockoffs

Within the large literature dealing with the transformational sparsity problem, generalized LASSO [TT+11] is the most popular one. In a special case when D is of full row-rank, i.e. a surjective linear map, generalized LASSO on transformational sparsity can be converted to LASSO on direct sparsity where one can design knockoffs; yet in general, it remains open how to do so.

Recall that the generalized LASSO solves the optimization problem below with $\lambda > 0$,

$$\min_\beta \ \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|D\beta\|_1. \tag{21}$$

In the special case that D is surjective, i.e. $\mathsf{rank}\, D = m \le p$, this problem can be equivalently represented by the following LASSO procedure with respect to $\lambda > 0$,

$$\min_\gamma \ \frac{1}{2n}\|y - XD^\dagger\gamma - XD_0\beta_0\|_2^2 + \lambda\|\gamma\|_1,$$

where $D^\dagger \in \mathbb{R}^{p\times m}$ is the pseudo inverse for D, $D_0 \in \mathbb{R}^{p\times(p-\mathsf{rank}\, D)}$ is a matrix whose columns spans the null space $\ker(D)$, and $\beta_0 \in \mathbb{R}^{p-\mathsf{rank}\, D}$ is the representation coefficient for the null space $\ker(D)$ without sparsity. Note that we can write $\beta^* = D^\dagger\gamma^* + D_0\beta_0^*$ for some $\beta_0^* \in \mathbb{R}^{p-\mathsf{rank}\, D}$, then Equation (1) becomes

$$y = XD^\dagger\gamma^* + XD_0\beta_0^* + \varepsilon.$$

13

In this regard, one can construct standard knockoffs in this special case that $\mathsf{rank}\,D = m \le p$. To see this, Taking $U \in \mathbb{R}^{n \times (n-p+\mathsf{rank}\,D)}$ as an orthogonal complement for the column space of $XD_0 \in \mathbb{R}^{n \times (p-\mathsf{rank}\,D)}$, we have

$$U^T y = U^T X D^\dagger \gamma^* + U^T \varepsilon. \tag{22}$$

Now one can treat $U^T X D^\dagger$ as a new design matrix, $U^T y$ as the response vector and $U^T \varepsilon \sim \mathcal{N}(0, I_{n-p+\mathsf{rank}\,D})$ as the Gaussian noise. Then the transformational sparsity problem is transferred to a sparse linear regression problem, where one can apply the standard knockoff method for the FDR control.

However, a shortcoming of this approach is that it may suffer poor selection power when D is nontrivial and/or X is highly co-related, such that the incoherence requirements for sparse recovery or model selection consistency [DH01, Tro04, ZY06, Wai09, ORX$^+$16, HSXY20] fail for the design matrix $U^T X D^\dagger$.

Furthermore, this conversion does not work in the general setting $m > p$. In the case $m > p$, γ lies in the column space of D, a proper subspace of $\mathbb{R}^m$, and it remains open how to design a knockoff method under such a constraint before our work.

5.2   Failure of Exchangeability in Split Knockoffs

The main challenge in establishing the FDR control of Split Knockoffs lies in the failure of exchangeability. In this section, we provide the details on how the exchangeability of standard Knockoffs fails in Split Knockoffs.

Specifically, [BC$^+$15] (Lemma 2-3), established pairwise exchangeability for both features and responses in standard knockoffs, the latter of which however fails in Split Knockoffs.

Proposition 5 (Failure of Pairwise Exchangeability)**.** *For any $S \subseteq S_0$, there holds*

$$[A_\gamma, \widetilde{A}_\gamma]_{\mathrm{swap}\{S\}}^T [A_\gamma, \widetilde{A}_\gamma]_{\mathrm{swap}\{S\}} = [A_\gamma, \widetilde{A}_\gamma]^T [A_\gamma, \widetilde{A}_\gamma], \tag{23a}$$

$$[A_\gamma, \widetilde{A}_\gamma]_{\mathrm{swap}\{S\}}^T \widetilde{y} \overset{d}{\neq} [A_\gamma, \widetilde{A}_\gamma]^T \widetilde{y}, \tag{23b}$$

where $[A_\gamma, \widetilde{A}_\gamma]_{\mathrm{swap}\{S\}}$ denotes a swap of the j-th column of A_γ and $\widetilde{A}_\gamma$ in $[A_\gamma, \widetilde{A}_\gamma]$ for all $j \in S$.

In other words, for Split Knockoffs, while pairwise exchangeability for the features holds in Equation (23a), that for the responses fails in Equation (23b). To see the details for Proposition 5, it can be calculated that

$$A_\gamma^T \widetilde{y} = 0, \quad \widetilde{A}_\gamma^T \widetilde{y} = \frac{\widetilde{A}_{\gamma,1}^T}{\sqrt{n_2}} X_2 \beta^* + \frac{\widetilde{A}_{\gamma,1}^T}{\sqrt{n}} \varepsilon_2, \tag{24}$$

where $\widetilde{A}_{\gamma,1}^T$ together with $\widetilde{A}_{\gamma,2}^T$ are defined in the end of Section 2.2. Moreover, by Equation (7), there holds

$$\widetilde{A}_{\gamma,1}^T \frac{X_2}{\sqrt{n_2}} + \widetilde{A}_{\gamma,2}^T \frac{D}{\sqrt{\nu}} = -\frac{D}{\nu}, \quad -\frac{\widetilde{A}_{\gamma,2}^T}{\sqrt{\nu}} = \frac{I}{\nu} - \mathrm{diag}(s). \tag{25}$$

Thus it can be solved that $\widetilde{A}_{\gamma,1}^T \frac{X_2}{\sqrt{n_2}} = \mathrm{diag}(s) \cdot D$. Plugging the solution into Equation (24), there holds

$$[A_\gamma, \widetilde{A}_\gamma]^T \widetilde{y} = \begin{bmatrix} 0 \\ -\mathrm{diag}(s)\gamma^* + \frac{\widetilde{A}_{\gamma,1}^T}{\sqrt{n}} \varepsilon \end{bmatrix}, \tag{26}$$

14

where swapping any $i \in S_0$ in the first and second block will lead to different distributions.

The failure of exchangeability may impose a theoretical challenge for knockoff-based methods, as the beautiful symmetry results [BC+15] can not be applicable. For random designs, [BCS+20] exploits KL divergence to measure the "distance" to exchangeability, and then gives approximate but not exact FDR control based on the "distance". It is also worth mentioning that this method no longer relies on the martingale arguments and uses direct analysis. However, it is not clear how to apply such methods to fixed designs in our scenario. Instead, our strategy in this paper is to exploit the orthogonal design in Split LASSO and the data splitting, which leaves us independent signs of Split Knockoff statistics, recovering the inverse supermartingale structure similar to [BC+19] without using the exchangeability.

5.3   Selection Power and Model Selection Consistency of Split LASSO

In this section, we discuss the model selection (sign) consistency of Split LASSO with respect to ν, to reveal how ν might affect the selection power of Split Knockoffs. As we shall see below, enlarging ν might increase the power of discovering strong nonnull features whose magnitudes are large, by improving the incoherence condition; on the other hand, doing so may lose the power of discovering weak nonnull features of small magnitudes.

This point is made precise by Proposition 6 on the model selection (sign) consistency of Split LASSO. Define $H_\nu := I_m - \frac{D[\Sigma_X + L_D]^{-1} D^T}{\nu}$, where $\Sigma_X = \frac{X^T X}{n}$ and $L_D = \frac{D^T D}{m}$. Further denote H_ν^{11} as the submatrix of index set S_1, H_ν^{00} for index S_0, and H_ν^{10}, H_ν^{01} to be the covariance matrices between S_1 and S_0. It is now ready to state Proposition 6, leaving its proof in Section B.8.

Proposition 6 (Model Selection Consistency of Split LASSO (4)). *Assume that the design matrix X and D satisfy*

- ***Restricted-Strongly-Convex:*** *there exists $C_{\min} > 0$, such that the smallest eigenvalue of H_ν^{11} is larger than $C_{\min}$;*

- ***ν-Incoherence Condition:*** *there exists a parameter $\chi_\nu \in (0, 1]$, such that $\|H_\nu^{01}[H_\nu^{11}]^{-1}\|_\infty \leq 1 - \chi_\nu$.*

Let the columns of X be normalized as $\max_{i \in [1:p]} \frac{\|x_i\|_2}{\sqrt{n}} \leq 1$. Then there exists some constant $C > 0$, such that for the sequence of $\{\lambda_n\}$ satisfying $\lambda_n > \frac{C}{\chi_\nu} \sqrt{\frac{\sigma^2 \log m}{n}}$, the following holds with probability larger than $1 - 4e^{-Cn\lambda_n^2}$.

1. *(No-false-positive) Split LASSO (4) has a unique solution $(\widehat{\beta}, \widehat{\gamma}) \in \mathbb{R}^p \times \mathbb{R}^m$ without false positives w.r.t. γ.*

2. *(Sign-consistency) In addition, $\widehat{\gamma}$ recovers the sign of γ^*, if there holds*

$$\min_{i \in S_1} \gamma_i^* > \lambda_n \nu \left[\frac{\sigma}{2 C_{\min}} + \|[H_\nu^{11}]^{-1}\|_\infty \right]. \tag{27}$$

For the ν-Incoherence Condition of Proposition 6, one can observe that as $\nu \to \infty$, $H_\nu = I_m - \frac{1}{\nu} D[\Sigma_X + L_D]^{-1} D^T \succeq I_m - \frac{1}{\nu} D[\Sigma_X]^{-1} D^T \to I_m$, therefore $H_\nu^{01} \to 0_{|S_0| \times |S_1|}$, while $H_\nu^{11} \to I_{|S_1|}$. In this situation, the term $\|H_\nu^{01}[H_\nu^{11}]^{-1}\|_\infty$ drops to zero, and the incoherence condition is satisfied with arbitrarily large $\chi_\nu \to 1$.

On the other hand, however, the increase of ν makes it harder to meet the condition on the minimal signal-noise-ratio (27). This may cause a potential loss in selecting weak nonnull features.

Hence a sufficiently large ν will ensure the incoherence condition for model selection consistency such that strong nonnull features will be selected earlier on the Split LASSO path than the nulls, at the cost of possibly losing weak nonnull features. A good power must rely on a proper choice of ν for the trade-off. In simulation Section 6, we indeed observe that the selection power of Split Knockoffs undergoes a first increase then decrease trend as ν grows. In practice, one may apply the cross validation over (ν, λ) on the Split LASSO path with subset data $\mathcal{D}_1$, to maximize the power for the optimal intercept estimator $\widehat{\beta}(\lambda) = \widehat{\beta}_{\widehat{\nu}, \widehat{\lambda}}$. Equipped with the FDR control for all $\nu > 0$ in Theorem 1, it maximizes the empirical power with a desired FDR. Such an empirical strategy is validated by simulation experiments in Section 6 and renders satisfied results in the study of Alzheimer's Disease in Section 7.

6 Simulation Experiment

In this section, we show by several simulation experiments that our proposed Split Knockoff method performs well with transformational sparsity. Particularly, under the cross-validation optimal choice of $\widehat{\beta}(\lambda) = \widehat{\beta}_{\widehat{\nu}, \widehat{\lambda}}$ and $\nu = \widehat{\nu}$, Split Knockoffs achieve both desired FDR control and high selection power in all the three choices of W statistics and respective selectors. The simulation experiments presented in this section are under the basic setting where $n_2 \geq m + p$, while the experiments under the high dimensional setting will be given in Section C.2.

6.1 Experimental Setting

In model (1), we generate $X \in \mathbb{R}^{n \times p}$ ($n = 500$ and $p = 100$) i.i.d. from $\mathcal{N}(0_p, \Sigma)$, where $\Sigma_{i,i} = 1$ and $\Sigma_{i,j} = c^{|i-j|}$ for $i \neq j$, with feature correlation $c = 0.5$. Define $\beta^* \in \mathbb{R}^p$ by

$$\beta_i^* := \begin{cases} 1 & i \leq 20, \ i \equiv 0, -1 (\text{mod } 3), \\ 0 & \text{otherwise.} \end{cases}$$

Then n linear measurements are generated by

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

where $\varepsilon \in \mathbb{R}^n$ is generated i.i.d. from $\mathcal{N}(0, 1)$.

For transformational sparsity, we need to specify the linear transformer D such that $\gamma^* = D\beta^*$, where γ^* is sparse. Our choice of β^* has two types of transformational sparsity that lead to the following three choices of D.

- β^* is sparse with many zero elements such that we can take $D_1 = I_p$ as the identity matrix where $m = p$.

- β^* is a uni-dimensional piecewise constant function such that we can take D_2 as the 1-D graph difference operator on a line, i.e. $D_2 \in \mathbb{R}^{(p-1) \times p}$, $D_2(i, i) = 1$, $D_2(i, i+1) = -1$ for $i = 1, \cdots, p-1$, and $D_2(i, j) = 0$ for other pairs of (i, j), where in this case $m = p - 1 < p$.

- Combining both cases, β^* is a sparse piecewise constant function such that we can take $D_3 = \begin{bmatrix} D_1 \\ D_2 \end{bmatrix} \in \mathbb{R}^{(2p-1) \times p}$, where in this case $m = 2p - 1 > p$.

In simulation experiments, we use `glmnet` package [FHT10, SFHT11] to compute regularization paths for Split LASSO, etc. For the data splitting, we randomly split the dataset $\mathcal{D} = (X, y)$ into two parts $\mathcal{D}_1 = (X_1, y_1)$ and $\mathcal{D}_2 = (X_2, y_2)$ with n_1 and n_2 samples respectively where $n_1 = 200$ and $n_2 = 300$. The performance of the selection power is presented together with the performance of the FDR control, where the selection power is defined as Power $= \frac{|S_1 \cap \widehat{S}|}{|S_1|}$.

For the first choice of $\widehat{\beta}(\lambda)$ in Section 2.1, we take $\widehat{\beta}(\lambda)$ as $\widehat{\beta}_\nu(\lambda)$, the solution path with respect to $\widehat{\beta}(\lambda)$ in the ν-Split LASSO regularization path with dataset $\mathcal{D}_1 = (X_1, y_1)$. For the regularization paths calculated in Split Knockoffs, we take $\log \lambda$ from a grid between 0 and -6 with a step size $h_\lambda = 0.01$.

For the second choice of $\widehat{\beta}(\lambda)$ in Section 2.1, we take $\widehat{\beta}(\lambda)$ as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$ in the Split LASSO path with dataset $\mathcal{D}_1 = (X_1, y_1)$, screening on $\log \nu$ from a grid between 0 and 2 with a step size 0.4, and $\log \lambda$ from a grid between 0 and -8 with a step size 0.4. For the regularization paths of feature and knockoff significance in Equation (8) and Equation (10), we take $\log \lambda$ from a grid between 0 and -6 with a step size $h_\lambda = 0.01$.

6.2 Performance of Split Knockoffs

In Figure 2, we plot the performance of Split Knockoffs on $\log(\nu)$ between 0 and 2 with a step size 0.2, where $\widehat{\beta}(\lambda)$ is taken as the solution path $\widehat{\beta}_\nu(\lambda)$ in the ν-Split LASSO path. In Figure 3, we plot the performance of Split Knockoffs on $\log(\nu)$ between 0 and 2 with a step size 0.2, where $\widehat{\beta}(\lambda)$ is taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$ on the Split LASSO paths.

In these cases, the FDR of Split Knockoffs are all under control; while the performance in the selection power differs from one to another. The cross-validation optimal estimator choice of $\widehat{\beta}(\lambda) = \widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$ shown in Figure 3, clearly improves the selection power of Split Knockoffs compared with the ν-Split LASSO solution path choice of $\widehat{\beta}(\lambda) = \widehat{\beta}_\nu(\lambda)$ shown in Figure 2.

In the family of W-statistics defined in Section 2.4, $W^{S\tau}$ shows a less conservative FDR and slightly better selection power compared with the others in Figure 2 in the tested regime of ν. Meanwhile, W^{BC} shows the most conservative performance compared with W^S and $W^{S\tau}$, which trades the selection power for a better empirical FDR control. Such a phenomenon is explained by Proposition 1 and 2.

In all cases, particularly for D_2 and D_3, the selection power of Split Knockoffs shows a first increasing then decreasing trend with respect to ν. This is discussed by Proposition 6 in Section 5.3, that increasing ν will help discover strong nonnull features at the cost of possibly losing weak ones. Thus a good power relies on a proper choice of ν for the trade-off. In practice, cross validation renders satisfactory results, which is validated in the next section.

6.3 Comparisons between Split Knockoffs and Knockoffs

In this section, we conduct more simulation experiments on the cross validation optimal estimator choice of $\widehat{\beta}(\lambda) = \widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$, which gives higher selection power compared with the ν-Split LASSO solution path choice of $\widehat{\beta}(\lambda) = \widehat{\beta}_\nu(\lambda)$, as shown in Section 6.2. Here, for the choice of ν in the calculation of the feature significance and knockoff significance by Equation (8) and Equation (10), we take the cross validation optimal choice of $\nu = \widehat{\nu}$ to maximize the selection power of Split Knockoffs.

We show in Table 1 the performance of Split Knockoffs with all the three versions of W statistics, under the above choice of $\widehat{\beta}(\lambda)$ and ν in the simulation settings described in Section 6.1. We also provide comparisons between Split Knockoffs and standard Knockoffs when both are applicable

Table 1. FDR and Power for Knockoffs and Split Knockoffs in simulation experiments ($q = 0.2$). The intercept $\widehat{\beta}(\lambda)$ for Split Knockoffs is taken as a fixed cross-validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$, with the ν for calculating the feature and knockoff significance taken as $\widehat{\nu}$. In this table, we present the average performance of Knockoffs and Split Knockoffs in FDR and Power, together with the standard deviations in 200 simulation instances. For shorthand notations, we use "SK(+)" to refer to "Split Knockoff(+)".

Performance	Knockoff	SK with W^{S}	SK with $W^{\mathrm{S}\tau}$	SK with W^{BC}
FDR in D_1	0.2233	0.2481	0.2517	0.0667
	±0.1584	±0.1465	±0.1465	±0.0706
Power in D_1	1.0000	1.0000	1.0000	1.0000
	±0.0000	±0.0000	±0.0000	±0.0000
FDR in D_2	0.2813	0.2206	0.2206	0.1492
	±0.1771	±0.1341	±0.1341	±0.1130
Power in D_2	0.5571	0.9886	0.9886	0.9886
	±0.3231	±0.0299	±0.0299	±0.0299
FDR in D_3	N/A	0.2374	0.2386	0.1529
	N/A	±0.1124	±0.1133	±0.0859
Power in D_3	N/A	0.9352	0.9352	0.9352
	N/A	±0.0509	±0.0509	±0.0509
Performance	Knockoff+	SK+ with W^{S}	SK+ with $W^{\mathrm{S}\tau}$	SK+ with W^{BC}
FDR in D_1	0.1787	0.1914	0.1929	0.0521
	±0.1487	±0.1444	±0.1447	±0.0649
Power in D_1	1.0000	1.0000	1.0000	1.0000
	±0.0000	±0.0000	±0.0000	±0.0000
FDR in D_2	0.1649	0.1709	0.1709	0.1085
	±0.1985	±0.1217	±0.1217	±0.0984
Power in D_2	0.2914	0.9886	0.9886	0.9886
	±0.3669	±0.0299	±0.0299	±0.0299
FDR in D_3	N/A	0.2100	0.2110	0.1347
	N/A	±0.1124	±0.1128	±0.0845
Power in D_3	N/A	0.9352	0.9352	0.9352
	N/A	±0.0509	±0.0509	±0.0509

(the case of D_1 and D_2). In particular, for the 1-D fused LASSO case that $D_2 \in \mathbb{R}^{(p-1) \times p}$ is the graph difference operator on a line, we make Knockoffs applicable by introducing the induced LASSO problem from the generalized LASSO problem (see Section 5.1 for details). We note that the Knockoffs are implemented in the whole dataset without data splitting.

As shown in Table 1, Split Knockoffs(+) with the cross-validation optimal choice $\widehat{\beta}(\lambda) = \widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$ achieve desired FDR in the family of W under all cases of transformational sparsity. In particular, Split Knockoffs with W^{BC} is the most conservative with a lower FDR and the same power. In all cases, (Split) Knockoff+ has a better control in the FDR compared with (Split) Knockoff, at the cost of a potential loss in selection power.

Compared with standard Knockoffs, Split Knockoffs exhibits higher power in the case of D_2. In this case, the correlated X and D_2 destroy the incoherence condition of induced the LASSO problem (see Equation (22) in Section 5.1 for discussions) such that standard Knockoffs suffer from losing the model selection consistency in the case of D_2. It hurts the selection power for standard

Knockoffs. In contrast, for Split Knockoffs, the improved ν-incoherence condition of Split LASSO as discussed in Section 5.3 leads to better model selection consistency on regularization paths which helps improve the selection power.

7 Application: Alzheimer's Disease

In this experiment, we apply the Split Knockoff method to study lesion regions of brains and their connections in Alzheimer's Disease (AD), which is the major cause of dementia and has attracted increasing attention in recent years.

7.1 Dataset

The data is obtained from ADNI (`http://adni.loni.ucla.edu`) dataset, acquired by structural Magnetic Resonance Imaging (MRI) scan. In total, the dataset contains $n = 752$ samples, with 126 AD, 433 Mild Cognitive Impairment (MCI), and 193 Normal Controls (NC). For each image, we implement the Dartel VBM [Ash07] for pre-processing, followed by the toolbox *Statistical Parametric Mapping* (SPM) for segmentation of gray matter (GM), white matter (WM), and cerebral spinal fluid (CSF). Then we use Automatic Anatomical Labeling (AAL) atlas [TMLP$^+$02] to partition the whole brain into $p = 90$ Cerebrum brain anatomical regions, with the volume of each region (summation of all GMs in the region) provided.

We use $X \in \mathbb{R}^{n \times p}$ to denote the design matrix, with each element $X_{i,j}$ representing the column-wise normalized volume of the region j in the subject i's brain. The response variable vector $y \in \mathbb{R}^n$ denotes the Alzheimer's Disease Assessment Scale (ADAS), which was originally designed to assess the severity of cognitive dysfunction [RMD84] and was later found to be able to clinically distinguish the diagnosed AD from normal controls [ZLV$^+$92]. We test two types of transformational sparsity:

(a) $D = I_p$, the identity matrix, for selecting the regions that account for the Alzheimer's disease, where in this case $m = p = 90$;

(b) D is the graph difference operator matrix on the brain region connectivity graph, for selecting the connections of regions with large activation changes accounting for the disease, where in this case $m = 463 > p = 90$;

In the experiments below, we choose the Split Knockoff statistics $W^{S\tau}$ to demonstrate the results as it exhibits the best selection power in simulations, where $\widehat{\beta}(\lambda)$ is taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$, screening over $\log \nu$ on a grid between 0 and 2 with a step size 0.4, and $\log \lambda$ on a grid between 0 and -8 with a step size 0.4. For the regularization paths in Equation (8) and (10), we take $\log \lambda$ from an arithmetic sequence between 0 and -6 with step size $h_\lambda = 0.01$. The dataset $\mathcal{D} = (X, y)$ is randomly split into two parts $\mathcal{D}_1 = (X_1, y_1)$ and $\mathcal{D}_2 = (X_2, y_2)$ with n_1 and n_2 samples respectively, where $n_1 = 150$ and $n_2 = n - n_1 = 602$. We show the experimental results on one random data split instance in this section, where for multiple instances of data splits the frequency plots on the most frequently selected regions and connections are provided as supplementary materials in Section D.

7.2 Region Selection

In this experiment, consider the selection of lesion regions, with $D = I_p$. The target FDR level is set at $q = 0.2$.

Table 2. Selected Regions by Split Knockoff on Alzheimer's Disease ($q = 0.2$). We choose the W statistics for Split Knockoff as $W^{S\tau}$, with $\widehat{\beta}(\lambda)$ taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$, where the minimal cross validation loss lies on the choice $\log\widehat{\nu} = 0$.

Region	Split Knockoff with $\log\nu$	
	[0, 1.8]	[1.9, 2]
Inferior frontal gyrus, opercular part (L)	$\checkmark$	$\checkmark$
Hippocampus (L)	$\checkmark$	$\checkmark$
Hippocampus (R)	$\checkmark$	$\checkmark$
Inferior parietal gyrus (R)	$\checkmark$	
Middle temporal gyrus (L)	$\checkmark$	

Our region selection results are summarized in Table 2 where $\log(\nu)$ is between 0 and 2 with a step size 0.1. At $\log\widehat{\nu} = 0$ where $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$ is optimal in terms of cross-validation loss, our algorithm selects five regions that are all related to AD and have been reported to suffer from degeneration during disease progression [VJ10, SWB$^+$09, KSR$^+$07, GKADNI$^+$10]. Specifically, it was reported in [VJ10] that the hippocampus in both sides are responsible for learning and memory; while the middle temporal gyrus participates in language and memory processing. Moreover, [TSP$^+$05] and [SNH$^+$18] found that the opercular part of the Inferior frontal gyrus may be associated with tone and inflectional morpheme processing. Finally, the inferior parietal gyrus which is associated with motor and sensory, was found to suffer from volume reduction after the stage from normal control (NC) to mild cognitive impairment [GKADNI$^+$10], *i.e.*, an intermediate stage between NC to Alzheimer's Disease.

7.3 Connection Selection

In this experiment, we consider the connection or edge detection, in which each connection is associated with two adjacent brain regions. Formally speaking, we set D as the graph gradient (difference) operator on the graph $G = (V, E)$ where V denotes the vertex set of brain regions and E denotes the (oriented) edge set of region pairs in neighbor, such that $D(\beta)(i, j) = \beta_i - \beta_j$ for $(i, j) \in E$. Thus D is the graph gradient operator that measures the differences in the impacts of Alzheimer's Disease between adjacent regions in brains. We shall expect a properly selected edge will connect regions of high contrast variations in the degree of atrophy during the disease progression. For example, a selected edge may connect an atrophy region in the brain significantly impaired by Alzheimer's Disease to another region which is less influenced by the disease.

For connection selection, our results are summarized in Table 3 where $\log(\nu)$ is between 0 and 2 with a step size 0.1. In total, there are eleven connections selected, six of which are selected at the cross-validation optimal $\widehat{\nu}$. Among all the eleven selected connections, eight of them are associated with the Hippocampus on the left side or the right side, where the Hippocampus is one of the most early affected regions during disease progression [JLPS99]. Similarly, the selected connections that involve the Amygdala echo the previous findings that the Amygdala is also affected early [Kna12], which can explain the neuropsychiatric symptoms that are commonly observed in mild stages of AD. These studies provide us with references to why these pairs of adjacent regions have different degrees of atrophy. Finally, the connection between Insula (L) and Middle frontal gyrus (L) may be a false discovery, as both regions were reported to be atrophied and this connection is not included in the cross-validation optimal selection. Although it was found that the atrophy happens in the whole region of the Insula while only the sub-region of the middle frontal gyrus [BDZ08], it is

Table 3. Selected connections by Split Knockoffs on Alzheimer's Disease ($q = 0.2$). We choose the W statistics for Split Knockoff as $W^{S\tau}$, with $\widehat{\beta}(\lambda)$ taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$, where the minimal cross-validation loss lies on the choice $\log\widehat{\nu} = 0.4$.

Connection		Split Knockoff with $\log\nu$			
Region 1	Region 2	[0, 0.1]	[0.2, 1.6]	1.7	[1.8, 2]
Hippocampus (L)	Posterior cingulate gyrus (L)	√	√	√	√
Hippocampus (L)	Lingual gyrus (L)	√	√	√	√
Hippocampus (L)	Fusiform gyrus (L)	√		√	
Hippocampus (L)	Precuneus (L)	√	√	√	√
Hippocampus (R)	Insula (R)	√			
Hippocampus (R)	Lingual gyrus (R)	√	√	√	√
Hippocampus (R)	Precuneus (R)	√	√	√	√
Hippocampus (R)	Superior temporal gyrus (R)	√			
Amygdala (R)	Putamen (R)	√	√	√	√
Amygdala (R)	Caudate nucleus (R)	√			
Middle frontal gyrus, orbital part (L)	Insula (L)	√			

still ambiguous to claim that the involved regions are significantly different in terms of degrees of degeneration. In summary, Figure 1 shows the selected regions and connections at cross-validated $\widehat{\nu}$. In the graph, each vertex represents a cerebrum brain region in Automatic Anatomical Labeling (AAL) atlas [TMLP+02], with abbreviations of each region marked in the vertex. A comparison table between the full region names and their abbreviations is now given in Table 5.

8   Conclusion

In this paper, we propose the *Split Knockoff* method as a data-adaptive approach for controlling the False Discovery Rate (FDR) in scenarios where sparsity is imposed on a linear transformation of parameters rather than directly on the parameters themselves. This transformational sparsity setting is common in applications such as total variations, wavelet transforms, fused LASSO, and trend filtering. By relaxing the original linear transformation constraint to its Euclidean neighborhood in a lifted parameter space, our method achieves FDR control while improving model selection consistency through enhanced conditions on Split LASSO regularization paths.

A key theoretical contribution is the development of new inverse supermartingale structures to ensure provable FDR control despite the failure of exchangeability caused by heterogeneous noise from the transformation. We extend the methodology to high-dimensional settings and demonstrate its effectiveness through simulation experiments, where the Split Knockoff method consistently achieves the desired FDR and power.

In a real-world application to Alzheimer's Disease, using structural Magnetic Resonance Imaging (MRI) data from the ADNI dataset, Split Knockoffs successfully identify important brain atrophy regions and abnormal inter-regional connections linked to disease progression. Future research directions include extending the methodology to directional FDR control and applying it to random design scenarios.

References

[Ash07]   John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007.

[BC+15]   Rina Foygel Barber, Emmanuel J Candès, et al. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[BC+19]   Rina Foygel Barber, Emmanuel J Candès, et al. A knockoff filter for high-dimensional selective inference. *Annals of Statistics*, 47(5):2504–2537, 2019.

[BCS+20]   Rina Foygel Barber, Emmanuel J Candès, Richard J Samworth, et al. Robust inference with knockoffs. *Annals of Statistics*, 48(3):1409–1431, 2020.

[BDZ08]   Geraldo F Busatto, Breno S Diniz, and Marcus V Zanetti. Voxel-based morphometry in Alzheimer's disease. *Expert review of neurotherapeutics*, 8(11):1691–1702, 2008.

[BH95]   Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, 57(1):289–300, 1995.

[CDOS12]   Jian-Feng Cai, Bin Dong, Stanley Osher, and Zuowei Shen. Image restoration: Total variation, wavelet frames, and beyond. *J. Amer. Math. Soc.*, 25(4):1033–1089, 2012.

[CFJL18]   Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: series B (Statistical Methodology)*, 80(3):551–577, 2018.

[DB16]   Ran Dai and Rina Foygel Barber. The knockoff filter for FDR control in group-sparse and multitask regression. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, 2016. PMLR 48:1851-1859. arXiv:1602.03589.

[DH01]   David L. Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

[DJ95]   David L Donoho and Iain M Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

[FHT10]   Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[FL08]   Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[GKADNI+10]   Sarah J Greene, Ronald J Killiany, Alzheimer's Disease Neuroimaging Initiative, et al. Subregions of the inferior parietal lobule are affected in the progression to Alzheimer's disease. *Neurobiology of aging*, 31(8):1304–1311, 2010.

[HSXY16]   Chendi Huang, Xinwei Sun, Jiechao Xiong, and Yuan Yao. Split LBI: An iterative regularization path with structural sparsity. In *Advances in Neural Information Processing Systems (NIPS) 29*, pages 3369–3377. 2016.

[HSXY20]   Chendi Huang, Xinwei Sun, Jiechao Xiong, and Yuan Yao. Boosting with structural sparsity: A differential inclusion approach. *Applied and Computational Harmonic Analysis*, 48(1):1–45, 2020.

[JLPS99]   Kirsi Juottonen, Mikko P Laakso, Kaarina Partanen, and Hilkka Soininen. Comparative MR analysis of the entorhinal cortex and hippocampus in diagnosing Alzheimer disease. *American Journal of Neuroradiology*, 20(1):139–144, 1999.

[KKBG09]   Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM review*, 51(2):339–360, 2009.

[Kna12]   Shira Knafo. Amygdala in Alzheimer's disease. *The Amygdala–A Discrete Multitasking Manager. IntechOpen*, pages 375–383, 2012.

[KSR+07]   Giorgos Karas, Philip Scheltens, Serge Rombouts, Ronald Van Schijndel, Martin Klein, Bethany Jones, Wiesje Van Der Flier, Hugo Vrenken, and Frederik Barkhof. Precuneus atrophy in early-onset Alzheimer's disease: a morphometric structural MRI study. *Neuroradiology*, 49(12):967–976, 2007.

[ORX+16]   Stanley Osher, Feng Ruan, Jiechao Xiong, Yuan Yao, and Wotao Yin. Sparse recovery via differential inclusions. *Applied and Computational Harmonic Analysis*, 41(2):436–469, 2016.

[RC20]   Zhimei Ren and Emmanuel Candès. Knockoffs with side information. 2020.

[RMD84]   W. G. Rosen, R. C. Mohs, and K. L. Davis. A new rating scale for Alzheimer's disease. *Am J Psychiatry*, 141(11):1356–64, 1984.

[ROF92]   Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[RSC19]   Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Deep knockoffs. *Journal of the American Statistical Association*, pages 1–12, 2019.

[RWC21]   Zhimei Ren, Yuting Wei, and Emmanuel Candès. Derandomizing knockoffs. *Journal of American Statistical Association*, 2021.

[SFHT11]   Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1, 2011.

[SNH+18]   Andrea Schremm, Mikael Novén, Merle Horne, Pelle Söderström, Danielle van Westen, and Mikael Roll. Cortical thickness of planum temporale and pars opercularis in native language tone processing. *Brain and Language*, 176:42–47, 2018.

[SWB+09]   N Schuff, N Woerner, L Boreta, T Kornfield, LM Shaw, JQ Trojanowski, PM Thompson, CR Jack Jr, MW Weiner, and Alzheimer's; Disease Neuroimaging Initiative. MRI of hippocampal volume loss in early Alzheimer's disease in relation to apoe genotype and biomarkers. *Brain*, 132(4):1067–1077, 2009.

[TMLP+02]   Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[Tro04]   Joel A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

[TSP+05]   Lorraine K Tyler, Emmanuel A Stamatakis, Brechtje Post, Billi Randall, and William Marslen-Wilson. Temporal and frontal systems in speech comprehension: An fMRI study of past tense processing. *Neuropsychologia*, 43(13):1963–1974, 2005.

[TSR+05]   Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[TT+11]   Ryan J Tibshirani, Jonathan Taylor, et al. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

[VJ10]   Prashanthi Vemuri and Clifford R Jack. Role of structural MRI in Alzheimer's disease. *Alzheimer's research & therapy*, 2(4):1–10, 2010.

[Wai09]   Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (LASSO). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

[WDR⁺10]   Jing Wu, Bernie Devlin, Steven Ringquist, Massimo Trucco, and Kathryn Roeder. Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 34(3):275–285, 2010.

[WR09]   Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

[XXCY16]   Qianqian Xu, Jiechao Xiong, Xiaochun Cao, and Yuan Yao. False discovery rate control and statistical quality assessment of annotators in crowdsourced ranking. In *International Conference on Machine Learning*, pages 1282–1291, 2016.

[ZLV⁺92]   Ronald F Zec, Edward S Landreth, Sandra K Vicari, Eleanor Feldman, James Belman, Alicia Andrise, Randall Robbs, Vinod Kumar, and Robert Becker. Alzheimer disease assessment scale: useful for both early detection and staging of dementia of the Alzheimer type. *Alzheimer Disease and Associated Disorders*, 1992.

[ZY06]   Peng Zhao and Bin Yu. On model selection consistency of Lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

A Construction of Split Knockoff Copies

In this section, we provide more details on the construction of Split Knockoff copy matrices. An explicit form of the construction is shown in the following proposition.

Proposition 7 (Split Knockoff Matrix). *If $n_2 \geq m + p$, for any vector s in Equation (7) satisfying*

$$\text{diag}(s) \succeq 0, \ 2C_\nu - \text{diag}(s) \succeq 0, \tag{28}$$

where $C_\nu := \Sigma_{\gamma,\gamma} - \Sigma_{\gamma,\beta}\Sigma_{\beta,\beta}^{-1}\Sigma_{\beta,\gamma}$, $\Sigma_{\beta,\beta} := A_\beta^T A_\beta$, $\Sigma_{\beta,\gamma} = \Sigma_{\gamma,\beta}^T := A_\beta^T A_\gamma$, and $\Sigma_{\gamma,\gamma} := A_\gamma^T A_\gamma$, there is a valid Split Knockoff matrix for $n_2 \geq m + p$,

$$\widetilde{A}_\gamma = A_\gamma(I_m - C_\nu^{-1}\,\text{diag}(s)) + A_\beta\Sigma_{\beta,\beta}^{-1}\Sigma_{\beta,\gamma}C_\nu^{-1}\,\text{diag}(s) + \widetilde{U}K, \tag{29}$$

where $\widetilde{U} \in \mathbb{R}^{(n_2+m)\times m}$ is the orthogonal complement of $[A_\beta, A_\gamma]$ and $K \in \mathbb{R}^{m\times m}$ satisfies $K^T K = 2\,\text{diag}(s) - \text{diag}(s)C_\nu^{-1}\,\text{diag}(s)$.

Remark. The requirement $n_2 \geq m + p$ is from the property that $\widetilde{U} \in \mathbb{R}^{(n_2+m)\times m}$ is an orthogonal complement of $[A_\beta, A_\gamma] \in \mathbb{R}^{(n_2+m)\times(m+p)}$.

In the following, we show that any s satisfying Equation (28) will ensure the existence of $\widetilde{A}_\gamma$. Note that a necessary and sufficient condition for the existence of $\widetilde{A}_\gamma$ satisfying Equation (7) is

$$G := \begin{bmatrix} \Sigma_{\beta,\beta} & \Sigma_{\beta,\gamma} & \Sigma_{\beta,\gamma} \\ \Sigma_{\beta,\gamma} & \Sigma_{\gamma,\gamma} & \Sigma_{\gamma,\gamma} - \text{diag}(s) \\ \Sigma_{\beta,\gamma} & \Sigma_{\gamma,\gamma} - \text{diag}(s) & \Sigma_{\gamma,\gamma} \end{bmatrix} \succeq 0.$$

This holds if and only if the Schur complement of $\Sigma_{\beta,\beta}$ is positive semi-definite, i.e.

$$\begin{bmatrix} C_\nu & C_\nu - \text{diag}(s) \\ C_\nu - \text{diag}(s) & C_\nu \end{bmatrix} \succeq 0,$$

which holds if and only if C_ν and its Schur complement are positive semi-definite, i.e.

$$C_\nu \succeq 0,$$
$$C_\nu - (C_\nu - \text{diag}(s))C_\nu^{-1}(C_\nu - \text{diag}(s)) = 2\,\text{diag}(s) - \text{diag}(s)C_\nu^{-1}\,\text{diag}(s) \succeq 0.$$

The equations above are equivalent with Equation (28). Then one can verify that the construction in Proposition 7 satisfies Equation (7) for any vector s satisfying Equation (28).

There are various choices for $s = (s_i)$ satisfying (28) for Split Knockoffs. Below we give two typical examples.

(a) (SDP for discrepancy maximization) One can maximize the discrepancy between Knockoffs and its corresponding features by solving the following SDP

$$\text{maximize} \quad \sum_i s_i,$$
$$\text{subject to} \quad 0 \leq s_i \leq \frac{1}{\nu} \text{ and } \frac{\text{diag}(s)}{2} \preceq C_\nu.$$

(b) (Equi-correlation) Take $s_i = 2\lambda_{\min}(C_\nu) \wedge \frac{1}{\nu}$ for all $i \in \{1, 2, \cdots, m\}$, where $\lambda_{\min}(C_\nu)$ represents the minimal eigenvalue of C_ν.

B Proofs

In this section, we will prove the propositions and theories in the main text in the following order. We will first prove Proposition 1, the inclusion property of selectors, then Proposition 4, a preliminary proposition for Theorem 1. After that, we will provide the proof of our main theorem on the FDR control — Theorem 1 — and some additional supporting lemmas. Finally, we will present the proof of Theorem 2, the FDR control theorem for Split Knockoffs in high dimensional settings, followed by the proof of Proposition 6, the model selection consistency of Split LASSO.

B.1 Proof of Proposition 1

Proof. We will prove Proposition 1 by showing the following two arguments one by one.

1. $\widehat{S}^{\mathrm{S}} \subseteq \widehat{S}^{\mathrm{S}\tau}$;

2. $\widehat{S}^{\mathrm{BC}} \subseteq \widehat{S}^{\mathrm{S}}$.

1) $\widehat{S}^{\mathrm{S}} \subseteq \widehat{S}^{\mathrm{S}\tau}$**.** First, we show in the following that $W_i^{\mathrm{S}} \leq W_i^{\mathrm{S}\tau}$ for all i. By definition, $\widetilde{Z}_i \geq \widetilde{Z}_i^{\mathcal{T}}$, therefore

$$W_i^{\mathrm{S}} = Z_i \cdot \mathrm{sign}(Z_i - \widetilde{Z}_i) \leq Z_i \cdot \mathrm{sign}(Z_i - \widetilde{Z}_i^{\mathcal{T}}) = W_i^{\mathrm{S}\tau}.$$

Our next goal is to show that for any determined $\mathcal{D}_1$, $\mathcal{D}_2$, there holds $T_q^{\mathrm{S}\tau} \leq T_q^{\mathrm{S}}$. Once such an argument is established, by the property $W_i^{\mathrm{S}} \leq W_i^{\mathrm{S}\tau}$ for all i, there holds

$$\widehat{S}^{\mathrm{S}} = \{i : W_i^{\mathrm{S}} \geq T_q^{\mathrm{S}}\} \subseteq \{i : W_i^{\mathrm{S}\tau} \geq T_q^{\mathrm{S}}\} \subseteq \{i : W_i^{\mathrm{S}\tau} \geq T_q^{\mathrm{S}\tau}\} = \widehat{S}^{\mathrm{S}\tau}.$$

Thus it will be sufficient to show $T_q^{\mathrm{S}\tau} \leq T_q^{\mathrm{S}}$ to prove $\widehat{S}^{\mathrm{S}} \subseteq \widehat{S}^{\mathrm{S}\tau}$.

By the property that $W^{\mathrm{S}} \leq W^{\mathrm{S}\tau}$ for all i, for all $\lambda > 0$, there holds

$$\{i : W_i^{\mathrm{S}\tau} \leq -\lambda\} \subseteq \{i : W_i^{\mathrm{S}} \leq -\lambda\}, \ \{i : W_i^{\mathrm{S}} \geq \lambda\} \subseteq \{i : W_i^{\mathrm{S}\tau} \geq \lambda\},$$

which further suggest that

$$\frac{|\{i : W_i^{\mathrm{S}\tau} \leq -\lambda\}|}{1 \vee |\{i : W_i^{\mathrm{S}\tau} \geq \lambda\}|} \leq \frac{|\{i : W_i^{\mathrm{S}} \leq -\lambda\}|}{1 \vee |\{i : W_i^{\mathrm{S}} \geq \lambda\}|}, \frac{1 + |\{i : W_i^{\mathrm{S}\tau} \leq -\lambda\}|}{1 \vee |\{i : W_i^{\mathrm{S}\tau} \geq \lambda\}|} \leq \frac{1 + |\{i : W_i^{\mathrm{S}} \leq -\lambda\}|}{1 \vee |\{i : W_i^{\mathrm{S}} \geq \lambda\}|}.$$

Thus by the definition of threshold T_q^{S} and $T_q^{\mathrm{S}\tau}$, there holds $T_q^{\mathrm{S}\tau} \leq T_q^{\mathrm{S}}$.

2) $\widehat{S}^{\mathrm{BC}} \subseteq \widehat{S}^{\mathrm{S}}$**.** The proof follows a similar stream as above. First, we show that $W_i^{\mathrm{BC}} \leq W_i^{\mathrm{S}}$ for all i, as

- if $Z_i > \widetilde{Z}_i$, $W_i^{\mathrm{BC}} = Z_i = W_i^{\mathrm{S}}$;

- if $Z_i < \widetilde{Z}_i$, $W_i^{\mathrm{BC}} = -\widetilde{Z}_i < -Z_i = W_i^{\mathrm{S}}$.

We will then show in the following that for any determined $\mathcal{D}_1$, $\mathcal{D}_2$, there holds $T_q^{\mathrm{S}} \leq T_q^{\mathrm{BC}}$. Once the argument is established, by the property $W_i^{\mathrm{BC}} \leq W_i^{\mathrm{S}}$ for all i, there holds

$$\widehat{S}^{\mathrm{BC}} = \{i : W_i^{\mathrm{BC}} \geq T_q^{\mathrm{BC}}\} \subseteq \{i : W_i^{\mathrm{S}} \geq T_q^{\mathrm{BC}}\} \subseteq \{i : W_i^{\mathrm{S}} \geq T_q^{\mathrm{S}}\} = \widehat{S}^{\mathrm{S}}.$$

Thus it will be sufficient to show $T_q^{\mathrm{S}} \leq T_q^{\mathrm{BC}}$ to prove $\widehat{S}^{\mathrm{BC}} \subseteq \widehat{S}^{\mathrm{S}}$.

By the property that $W^{\mathrm{BC}} \leq W^{\mathrm{S}}$ for all i, for all $\lambda > 0$, there holds

$$\{i : W_i^{\mathrm{S}} \leq -\lambda\} \subseteq \{i : W_i^{\mathrm{BC}} \leq -\lambda\}, \ \{i : W_i^{\mathrm{BC}} \geq \lambda\} \subseteq \{i : W_i^{\mathrm{S}} \geq \lambda\},$$

which further suggest that

$$\frac{|\{i : W_i^{\mathrm{S}} \leq -\lambda\}|}{1 \vee |\{i : W_i^{\mathrm{S}} \geq \lambda\}|} \leq \frac{|\{i : W_i^{\mathrm{BC}} \leq -\lambda\}|}{1 \vee |\{i : W_i^{\mathrm{BC}} \geq \lambda\}|}, \frac{1 + |\{i : W_i^{\mathrm{S}} \leq -\lambda\}|}{1 \vee |\{i : W_i^{\mathrm{S}} \geq \lambda\}|} \leq \frac{1 + |\{i : W_i^{\mathrm{BC}} \leq -\lambda\}|}{1 \vee |\{i : W_i^{\mathrm{BC}} \geq \lambda\}|}.$$

Thus by the definition of threshold T_q^{BC} and T_q^{S}, there holds $T_q^{\mathrm{S}} \leq T_q^{\mathrm{BC}}$. □

B.2 Proof of Proposition 2

Proof. We prove the properties in Proposition 2 one by one.

1) For all i, $\{W_i^{\mathrm{S}\tau} \leq -T\} \subseteq \{W_i^{\mathrm{S}} \leq -T\} \subseteq \{W_i^{\mathrm{BC}} \leq -T\}$.

We first show that there holds $\{W_i^{\mathrm{S}\tau} \leq -T\} \subseteq \{W_i^{\mathrm{S}} \leq -T\}$ for all i. For all i, note that $\{W_i^{\mathrm{S}} \leq -T\} = \{Z_i \geq T, \widetilde{Z}_i > Z_i\}$, while $\{W_i^{\mathrm{S}\tau} \leq -T\} = \{Z_i \geq T, \tau(\widetilde{Z})_i > Z_i\}$. Since by definition $\tau(\widetilde{Z})_i = \widetilde{Z}_i \cdot 1\{r_i = \widetilde{r}_i\} \leq \widetilde{Z}_i$, there holds $\{W_i^{\mathrm{S}\tau} \leq -T\} \subseteq \{W_i^{\mathrm{S}} \leq -T\}$.

Then we show that there holds $\{W_i^{\mathrm{S}} \leq -T\} \subseteq \{W_i^{\mathrm{BC}} \leq -T\}$ for all i. For all i, note that $\{W_i^{\mathrm{BC}} \leq -T\} = \{Z_i \vee \widetilde{Z}_i \geq T, \widetilde{Z}_i > Z_i\}$, while $\{W_i^{\mathrm{S}} \leq -T\} = \{Z_i \geq T, \widetilde{Z}_i > Z_i\}$. Since $Z_i \vee \widetilde{Z}_i \geq Z_i$, there holds $\{W_i^{\mathrm{S}} \leq -T\} \subseteq \{W_i^{\mathrm{BC}} \leq -T\}$.

2) For all i, $\{W_i^{\mathrm{BC}} \geq T\} = \{W_i^{\mathrm{S}} \geq T\} \subseteq \{W_i^{\mathrm{S}\tau} \geq T\}$.

We first show that $\{W_i^{\mathrm{S}} \geq T\} \subseteq \{W_i^{\mathrm{S}\tau} \geq T\}$ for all i. For all i, note that $\{W_i^{\mathrm{S}} \geq T\} = \{Z_i \geq T, \widetilde{Z}_i < Z_i\}$, while $\{W_i^{\mathrm{S}\tau} \geq T\} = \{Z_i \geq T, \tau(\widetilde{Z})_i < Z_i\}$. Since by definition $\tau(\widetilde{Z})_i = \widetilde{Z}_i \cdot 1\{r_i = \widetilde{r}_i\} \leq \widetilde{Z}_i$, there holds $\{W_i^{\mathrm{S}} \geq T\} \subseteq \{W_i^{\mathrm{S}\tau} \geq T\}$.

Then we show that $\{W_i^{\mathrm{BC}} \geq T\} = \{W_i^{\mathrm{S}} \geq T\}$ for all i. For all i, note that $\{W_i^{\mathrm{BC}} \geq T\} = \{Z_i \vee \widetilde{Z}_i \geq T, \widetilde{Z}_i < Z_i\}$, where $Z_i \vee \widetilde{Z}_i = Z_i$ for $Z_i > \widetilde{Z}_i$. Therefore, $\{W_i^{\mathrm{BC}} \geq T\} = \{Z_i \geq T, \widetilde{Z}_i < Z_i\} = \{W_i^{\mathrm{S}} \geq T\}$.

3) There holds $\mathcal{M}_T(W^{\mathrm{S}\tau}) \geq \mathcal{M}_T(W^{\mathrm{S}}) \geq \mathcal{M}_T(W^{\mathrm{BC}})$.

Since $\{W_i^{\mathrm{S}\tau} \leq -T\} \subseteq \{W_i^{\mathrm{S}} \leq -T\} \subseteq \{W^{\mathrm{BC}} \leq -T\}$ for all i, there holds

$$\sum_{i \in S_0} 1\{W_i^{\mathrm{BC}} \leq -T\} \geq \sum_{i \in S_0} 1\{W_i^{\mathrm{S}} \leq -T\} \geq \sum_{i \in S_0} 1\{W_i^{\mathrm{S}\tau} \leq -T\}. \tag{31}$$

Moreover, since $\{W_i^{\mathrm{BC}} \geq T\} = \{W_i^{\mathrm{S}} \geq T\} \subseteq \{W^{\mathrm{S}\tau} \geq T\}$ for all i, there holds

$$\sum_{i \in S_0} 1\{W_i^{\mathrm{S}\tau} \geq T\} \geq \sum_{i \in S_0} 1\{W_i^{\mathrm{S}} \geq T\} = \sum_{i \in S_0} 1\{W_i^{\mathrm{BC}} \geq T\}. \tag{32}$$

Combining Equation (31) with Equation (32), there holds $\mathcal{M}_T(W^{\mathrm{S}\tau}) \geq \mathcal{M}_T(W^{\mathrm{S}}) \geq \mathcal{M}_T(W^{\mathrm{BC}})$. □

B.3 Proof of Proposition 3

Proof. We first show that for $Z_i > 0$, $\{\limsup_{\lambda \to Z_i^-} \mathrm{sign}(\gamma_i(\lambda)) = 1\}$ is the complement of $\{\liminf_{\lambda \to Z_i^-} \mathrm{sign}(\gamma_i(\lambda)) = -1\}$, which is equivalent with the following two points

1. $\{\limsup_{\lambda \to Z_i^-} \mathrm{sign}(\gamma_i(\lambda)) \neq 1\} \cap \{\liminf_{\lambda \to Z_i^-} \mathrm{sign}(\gamma_i(\lambda)) \neq -1\} = \emptyset,$

27

2. $\{\limsup_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = 1\} \cap \{\liminf_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = -1\} = \emptyset$.

We prove the two points in the following.

1) We show by contradiction that for $Z_i > 0$, the statements $\limsup_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) \neq 1$ and $\liminf_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) \neq -1$ cannot both be true.

Suppose that both statements above are true, then since $\operatorname{sign}(x) \in \{-1, 0, 1\}$ for $x \in \mathbb{R}$, there holds:

- $\limsup_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) \neq 1$ suggests that there exists $\delta_1 > 0$, such that for $Z_i - \delta_1 < \lambda < Z_i$, there holds $\gamma_i(\lambda) \leq 0$;

- $\liminf_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) \neq -1$ suggests that there exists $\delta_2 > 0$, such that for $Z_i - \delta_2 < \lambda < Z_i$, there holds $\gamma_i(\lambda) \geq 0$.

Therefore, for $Z_i - \min\{\delta_1, \delta_2\} < \lambda < Z_i$, there holds $\gamma_i(\lambda) = 0$. Since $\widehat{\beta}(\lambda)$ is continuous, $\gamma(\lambda)$ satisfying Equation (16) is continuous. Therefore, $\gamma_i(Z_i) = \lim_{\lambda \to Z_i^+} \gamma_i(\lambda) = 0$. This further suggests that $\gamma_i(\lambda) = 0$ for $\lambda > Z_i - \min\{\delta_1, \delta_2\}$, which contradicts with the definition of Z in Equation (9).

2) We show by contradiction that for $Z_i > 0$, the statements $\limsup_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = 1$ and $\liminf_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = -1$ cannot both be true.

Suppose that both statements above are true, then since $\rho(\lambda) \in \partial \|\gamma(\lambda)\|_1$, there holds:

- $\limsup_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = 1$ suggests that $\limsup_{\lambda \to Z_i^-} \rho_i(\lambda) = 1$;

- $\liminf_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = -1$ suggests that $\liminf_{\lambda \to Z_i^-} \rho_i(\lambda) = -1$.

However, since $\widehat{\beta}(\lambda)$ is continuous, $\rho(\lambda)$ satisfying Equation (16) is continuous. Therefore, $1 = \limsup_{\lambda \to Z_i^-} \rho_i(\lambda) = \liminf_{\lambda \to Z_i^-} \rho_i(\lambda) = -1$, which leads to contradiction.

Then we show that Equation (12) is equivalent with the definition of r given in Equation (17). In particular, we show that for $Z_i > 0$, there holds

1. $\{\limsup_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = 1\} \subseteq \{\operatorname{sign}(\rho_i(Z_i)) = 1\}$,

2. $\{\liminf_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = -1\} \subseteq \{\operatorname{sign}(\rho_i(Z_i)) = -1\}$.

The above statements are sufficient because it is already shown above that for $Z_i > 0$, $\{\limsup_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = 1\}$ is the complement of $\{\liminf_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = -1\}$. We prove the two points in the following.

1) Suppose that $\limsup_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = 1$. Then since $\rho(\lambda) \in \partial \|\gamma(\lambda)\|_1$, there holds $\limsup_{\lambda \to Z_i^-} \rho_i(\lambda) = 1$. By the continuity of $\rho(\lambda)$, there holds $\rho_i(Z_i) = \limsup_{\lambda \to Z_i^-} \rho_i(\lambda) = 1$. Therefore $\operatorname{sign}(\rho_i(Z_i)) = 1$.

2) Suppose that $\liminf_{\lambda \to Z_i^-} \operatorname{sign}(\gamma_i(\lambda)) = -1$. Then since $\rho(\lambda) \in \partial \|\gamma(\lambda)\|_1$, there holds $\liminf_{\lambda \to Z_i^-} \rho_i(\lambda) = -1$. By the continuity of $\rho(\lambda)$, there holds $\rho_i(Z_i) = \liminf_{\lambda \to Z_i^-} \rho_i(\lambda) = -1$. Therefore $\operatorname{sign}(\rho_i(Z_i)) = -1$.

Equation (13) can be shown to be well-defined and equivalent with the definition of $\widetilde{r}$ given in Equation (17) in the exact same way as above. This ends the proof. $\qquad\square$

B.4 Proof of Proposition 4

Proof. We first present the following two useful properties on W^{S} and W^{BC}. Recall that by definitions of W^{BC} and W^{S}, there holds for all $i \in \{1, 2, \cdots, m\}$ that

(i) if $Z_i > \widetilde{Z}_i$, then $|W_i^{\mathrm{BC}}| = Z_i = |W_i^{\mathrm{S}}|$;

(ii) if $Z_i < \widetilde{Z}_i$, then $|W_i^{\mathrm{BC}}| = \widetilde{Z}_i > Z_i = |W_i^{\mathrm{S}}|$.

Now we proceed to prove the argument that $\mathcal{F}^{\mathrm{BC}}(T) \subseteq \mathcal{F}^{\mathrm{S}}(T)$. Below we show that, the filtration associated with $\mathcal{M}_T(W^{\mathrm{BC}})$ is a refinement of the filtration associated with $\mathcal{M}_T(W^{\mathrm{S}})$, hence a stopping time adapted to the latter is also adapted to the former.

To see this, recall that by definitions of W^{BC} and W^{S}, from (i) we see $\{i : W_i^{\mathrm{BC}} > 0\} = \{i : W_i^{\mathrm{S}} > 0\}$, on which $|W_i^{\mathrm{BC}}| = Z_i = |W_i^{\mathrm{S}}|$, hence

$$\#\{i : W_i^{\mathrm{BC}} \geq T\} = \#\{i : W_i^{\mathrm{S}} \geq T\};$$

from (i) and (ii) we see for all i, $|W_i^{\mathrm{S}}| \leq |W_i^{\mathrm{BC}}|$, hence

$$\{\zeta_i : |W_i^{\mathrm{BC}}| < T\} \subseteq \{\zeta_i : |W_i^{\mathrm{S}}| < T\}.$$

Therefore, $\{\zeta_i : |W_i^{\mathrm{BC}}| < T\}$ can be represented by $\{\zeta_i : |W_i^{\mathrm{S}}| < T\}$. Now it remains to consider $\#\{i : W_i^{\mathrm{BC}} \leq -T\}$. In fact,

$$
\begin{aligned}
\{i : W_i^{\mathrm{BC}} \leq -T\} &= \{i : W_i^{\mathrm{BC}} \leq -T, |W_i^{\mathrm{S}}| \geq T\} \cup \{i : W_i^{\mathrm{BC}} \leq -T, |W_i^{\mathrm{S}}| < T\}, \\
&= \{i : W_i^{\mathrm{BC}} \leq -T, W_i^{\mathrm{S}} \leq -T\} \cup \{i : W_i^{\mathrm{BC}} \leq -T, |W_i^{\mathrm{S}}| < T\}, \\
&= \{i : W_i^{\mathrm{S}} \leq -T\} \cup \{i : W_i^{\mathrm{BC}} \leq -T, |W_i^{\mathrm{S}}| < T\},
\end{aligned}
$$

which implies that $\#\{i : W_i^{\mathrm{BC}} \leq -T\}$ can be sufficiently determined by $\#\{i : W_i^{\mathrm{S}} \leq -T\}$ and $\{\zeta_i : |W_i^{\mathrm{S}}| < T\}$, both of which are already included in $\mathcal{F}^{\mathrm{S}}(T)$. This finally shows that $\mathcal{F}^{\mathrm{BC}}(T) \subseteq \mathcal{F}^{\mathrm{S}}(T)$. $\square$

B.5 Proof of Theorem 1

In this section, we will give the complete proof of Theorem 1. Our treatment includes the following four types of W-statistics:

1. $W^{\mathrm{S}} := Z \odot \mathrm{sign}(Z - \widetilde{Z})$, where S refers to "Split".

2. $W^{\mathrm{S}\tau} := Z \odot \mathrm{sign}(Z - \tau(\widetilde{Z}))$, where S$\tau$ refers to applying truncation (14) on W^{S}.

3. $W^{\mathrm{BC}} := (Z \vee \widetilde{Z}) \odot \mathrm{sign}(Z - \widetilde{Z})$, where BC refers to the original definition adopted by Barber-Candès in [BC$^+$15].

4. $W^{\mathrm{BC}\tau} := (Z \vee \tau(\widetilde{Z})) \odot \mathrm{sign}(Z - \tau(\widetilde{Z}))$, where BC$\tau$ refers to applying truncation (14) on W^{BC}.

The last one $W^{\mathrm{BC}\tau}$ is added here for completeness. For shorthand notation, we use $W^{\star}$ to represent any one of the four cases, where $\star \in \{\mathrm{S}, \mathrm{S}\tau, \mathrm{BC}, \mathrm{BC}\tau\}$. In addition, $(\mathrm{S}\star, \mathrm{BC}\star)$ is used to denote one of the pairs in $\{(\mathrm{S}, \mathrm{BC}), (\mathrm{S}\tau, \mathrm{BC}\tau)\}$ with truncation τ adopted or not.

We will first show that by a standard procedure in Knockoffs as in [BC$^+$15], the problem of bounding the FDR by q can be transferred into the problem of bounding $\mathbb{E}\left[\mathcal{M}_{T_q^\star}(W^{\star})\right]$ by one, where $\mathcal{M}_T(W^{\star})$ for any $T > 0$ is defined in Equation (15). Following that, we divide the proof into two parts:

1. In Section B.5.1, we will prove that $\mathbb{E}\left[\mathcal{M}_{T_q^{\mathrm{S}\star}}(W^{\mathrm{S}\star})\right] \leq 1$ by introducing an inverse super-martingale structure associated with $\mathcal{M}_T(W^{\mathrm{S}\star})$.

2. In Section B.5.2, we will show that $T_q^{\mathrm{BC}\star}$ is a stopping time with respect to a filtration associated with $\mathcal{M}_T(W^{\mathrm{S}\star})$, which enables us to show that $\mathbb{E}\left[\mathcal{M}_{T_q^{\mathrm{BC}\star}}(W^{\mathrm{BC}\star})\right] \leq \mathbb{E}\left[\mathcal{M}_{T_q^{\mathrm{BC}\star}}(W^{\mathrm{S}\star})\right] \leq 1$.

We begin the proof from the following common procedure of Knockoffs [BC+15], that the upper bound of FDR in the case of $W^\star$ is transferred into the upper bound of $\mathbb{E}\left[\mathcal{M}_{T_q^\star}(W^\star)\right]$. The procedure goes as the following for Split Knockoff(+).

(a) (Split Knockoff) The mFDR can be bounded by the following product:

$$\mathbb{E}\left[\frac{\sum_{i\in S_0}1\{W_i^\star \geq T_q^\star\}}{\sum_i 1\{W_i^\star \geq T_q^\star\} + q^{-1}}\right] \leq \mathbb{E}\left[\frac{1 + \sum_i 1\{W_i^\star \leq -T_q^\star\}}{\sum_i 1\{W_i^\star \geq T_q^\star\} + q^{-1}} \frac{\sum_{i\in S_0}1\{W_i^\star \geq T_q^\star\}}{1 + \sum_{i\in S_0}1\{W_i^\star \leq -T_q^\star\}}\right]. \quad (33)$$

By the definition of the Split Knockoff threshold, there holds

$$\frac{\sum_i 1\{W_i^\star \leq -T_q^\star\}}{1 \vee \sum_i 1\{W_i^\star \geq T_q^\star\}} \leq q \leq 1,$$

which implies

$$\sum_i 1\{W_i^\star \leq -T_q^\star\} \leq q \sum_i 1\{W_i^\star \geq T_q^\star\}.$$

Consequently, there holds

$$\frac{1 + \sum_i 1\{W_i^\star \leq -T_q^\star\}}{\sum_i 1\{W_i^\star \geq T_q^\star\} + q^{-1}} \leq \frac{1 + q[\sum_i 1\{W_i^\star \geq T_q^\star\}]}{\sum_i 1\{W_i^\star \geq T_q^\star\} + q^{-1}} = q.$$

Combined with Equation (33), there holds

$$\mathbb{E}\left[\frac{\sum_{i\in S_0}1\{W_i^\star \geq T_q^\star\}}{\sum_i 1\{W_i^\star \geq T_q^\star\} + q^{-1}}\right] \leq q\mathbb{E}\left[\frac{\sum_{i\in S_0}1\{W_i^\star \geq T_q^\star\}}{1 + \sum_{i\in S_0}1\{W_i^\star \leq -T_q^\star\}}\right] = q\mathbb{E}\left[\mathcal{M}_{T_q^\star}(W^\star)\right]. \quad (34)$$

(b) (Split Knockoff+) The following lines established the result,

$$\mathbb{E}\left[\frac{\sum_{i\in S_0}1\{W_i^\star \geq T_q^\star\}}{1 \vee \sum_i 1\{W_i^\star \geq T_q^\star\}}\right] \leq \mathbb{E}\left[\frac{1 + \sum_i 1\{W_i^\star \leq -T_q^\star\}}{1 \vee \sum_i 1\{W_i^\star \geq T_q^\star\}} \frac{\sum_{i\in S_0}1\{W_i^\star \geq T_q^\star\}}{1 + \sum_{i\in S_0}1\{W_i^\star \leq -T_q^\star\}}\right],$$

$$\leq q\mathbb{E}\left[\frac{\sum_{i\in S_0}1\{W_i^\star \geq T_q^\star\}}{1 + \sum_{i\in S_0}1\{W_i^\star \leq -T_q^\star\}}\right] = q\mathbb{E}\left[\mathcal{M}_{T_q^\star}(W^\star)\right]. \quad (35)$$

Then we transfer the problem of bounding (m)FDR by q in Theorem 1 into the problem of bounding $\mathbb{E}\left[\mathcal{M}_{T_q^\star}(W^\star)\right]$ by one using Equation (34) and Equation (35). Then we will prove the following two inequalities respectively in Section B.5.1 and Section B.5.2.

1. We will first prove that $\mathbb{E}\left[\mathcal{M}_{T_q^{\mathrm{S}\star}}(W^{\mathrm{S}\star})\right] \leq 1$ in Section B.5.1.

2. We will then show that $\mathbb{E}\left[\mathcal{M}_{T_q^{\mathrm{BC}\star}}(W^{\mathrm{BC}\star})\right] \leq \mathbb{E}\left[\mathcal{M}_{T_q^{\mathrm{BC}\star}}(W^{\mathrm{S}\star})\right] \leq 1$ in Section B.5.2.

B.5.1 Proof of Theorem 1: Case I

As a reminder for the notations, S$\star$ represents an arbitrary element from $(\mathrm{S}, \mathrm{S}\tau)$. In this section, We will target to formulate a supermartingale structure associated with $\mathcal{M}_T(W^{\mathrm{S}\star})$. In order to do such a thing, we will need to show that $W^{\mathrm{S}\star}$ is a statistics whose sign $\{\mathrm{sign}(W^{\mathrm{S}\star})\}$ and length $(|W^{\mathrm{S}\star}|)$ are independent from each other.

To show such a independence property, we will need to take a deeper look at the KKT conditions in Equation (16). From the definition, there holds

1. $\widehat{\beta}(\lambda)$ from $\mathcal{D}_1 = (X_1, y_1)$ and ζ from $\mathcal{D}_2 = (X_2, y_2)$ are independent from each other;

2. $\gamma(\lambda)$ is determined by $\widehat{\beta}(\lambda)$ from $\mathcal{D}_1 = (X_1, y_1)$, which is the same for Z and $r := \mathrm{sign}(\gamma(Z-))$ as functions of $\gamma(\lambda)$;

3. conditional on $\widehat{\beta}(\lambda)$, $\widetilde{\gamma}$ is determined by ζ from $\mathcal{D}_2 = (X_2, y_2)$, which is the same for $\widetilde{Z}$, $\widetilde{r} := \mathrm{sign}(\widetilde{\gamma}(\widetilde{Z}-))$, and $\tau(\widetilde{Z})$ as functions of $\widetilde{\gamma}$.

Therefore, conditional on $\widehat{\beta}(\lambda)$ from $\mathcal{D}_1 = (X_1, y_1)$ which determines $|W^{\mathrm{S}\star}| = Z$, the difference between the feature significance (16a) and knockoff significance (16b) lies on the random variable ζ from $\mathcal{D}_2 = (X_2, y_2)$. Thus the length $|W^{\mathrm{S}\star}|$ and sign $\{\mathrm{sign}(W^{\mathrm{S}\star})\}$ of $W^{\mathrm{S}\star}$ are independent from each other. Additional calculations on ζ shows that ζ are consist of independent Gaussian random variables, i.e. for the Split Knockoff matrix satisfying (7), the distribution of $\zeta + \mathrm{diag}(s)\gamma^*$ satisfies Equation (18). Then the following lemma can be given for $W^{\mathrm{S}\star} \in \{W^{\mathrm{S}}, W^{\mathrm{S}\tau}\}$, in addition to W^{S} only in Lemma 1.

Lemma 3. *Given any determined $\widehat{\beta}(\lambda)$, $1\{W_i^{\mathrm{S}\star} < 0\}$ are some independent Bernoulli random variables. Furthermore, for $i \in S_0 \cap \{i : |W_i^{\mathrm{S}\star}| = Z_i > 0\}$, there holds*

$$\mathbb{P}[W_i^{\mathrm{S}\star} < 0] \geq \frac{1}{2}.$$

For shorthand notations, we rearrange the index of $W^{\mathrm{S}\star}$, such that $|W_{(1)}^{\mathrm{S}\star}| \geq |W_{(2)}^{\mathrm{S}\star}| \geq \cdots \geq |W_{(m^*)}^{\mathrm{S}\star}| > 0$, and $\{(1), (2), \cdots, (m^*)\} = S_0 \cap \{i : |W_i^{\mathrm{S}}| = Z_i > 0\}$. Further denote $B_{(i)} = 1\{W_{(i)}^{\mathrm{S}\star} < 0\}$, then there holds

$$\frac{\sum_{i \in S_0} 1\{W_i^{\mathrm{S}\star} \geq T_q^{\mathrm{S}\star}\}}{1 + \sum_{i \in S_0} 1\{W_i^{\mathrm{S}\star} \leq -T_q^{\mathrm{S}\star}\}} = \frac{1 + \sum_{i \in S_0} 1\{|W_i^{\mathrm{S}\star}| \geq T_q^{\mathrm{S}\star}\}}{1 + \sum_{i \in S_0} 1\{|W_i^{\mathrm{S}\star}| \geq T_q^{\mathrm{S}\star}, W_i^{\mathrm{S}\star} < 0\}} - 1,$$

$$= \frac{1 + J}{1 + B_{(1)} + B_{(2)} + \cdots + B_{(J)}} - 1, \tag{36}$$

where $J \leq m^*$ is defined to be the index satisfying

$$|W_{(1)}^{\mathrm{S}\star}| \geq |W_{(2)}^{\mathrm{S}\star}| \geq \cdots \geq |W_{(J)}^{\mathrm{S}\star}| \geq T_q^{\mathrm{S}\star} > |W_{(J+1)}^{\mathrm{S}\star}| \geq \cdots \geq |W_{(m^*)}^{\mathrm{S}\star}|.$$

In other words, $J = \mathrm{argmax}_{k \leq m^*}\{|W_{(k)}^{\mathrm{S}\star}| \geq T_q^{\mathrm{S}\star}\}$.

Define the filtration $\{\mathcal{F}_j\}_{j=1}^m$ and $\{\mathcal{G}_j\}_{j=1}^m$ in inverse time as

$$\mathcal{F}_j = \sigma\left(\left\{\sum_{i=1}^j B_{(i)}, \zeta_{(j+1)}, \cdots, \zeta_{(m)}\right\}\right),$$

31

$$\mathcal{G}_j = \sigma\left(\left\{\sum_{i=1}^{j} B_{(i)}, B_{(j+1)}, \cdots, B_{(m)}\right\}\right).$$

Conditional on $\widehat{\beta}(\lambda)$, since ζ_i determines $\text{sign}(W_i^{\text{S}\star})$ and B_i, the filtration $\{\mathcal{F}_j\}_{j=1}^{m}$ is a refined filtration of $\{\mathcal{G}_j\}_{j=1}^{m}$ in inverse time, i.e. $\mathcal{G}_j \subseteq \mathcal{F}_j$. By [BC+15, BC+19], J is a stopping time on $\{\mathcal{G}_j\}_{j=1}^{m}$ in inverse time, thus J is also a stopping time on the refined filtration $\{\mathcal{F}_j\}_{j=1}^{m}$ in inverse time. Then it will be proper to apply Lemma 2 as a supermartingale inequality to give an upper bound on the expectation of Equation (36).

Applying Lemma 2 to Equation (36), with the estimation that $\mathbb{P}[B_i = 1] \geq \rho = \frac{1}{2}$ for $i \in S_0$ by Lemma 3, we will have the following inequality on $\mathcal{M}_{T_q^{\text{S}\star}}(W^{\text{S}\star})$ that

$$\mathbb{E}\left[\mathcal{M}_{T_q^{\text{S}\star}}(W^{\text{S}\star})\right] = \mathbb{E}\left[\frac{\sum_{i \in S_0} 1\{W_i^{\text{S}\star} \geq T_q^{\text{S}\star}\}}{1 + \sum_{i \in S_0} 1\{W_i^{\text{S}\star} \leq -T_q^{\text{S}\star}\}}\right] \leq \rho^{-1} - 1 = 1. \tag{37}$$

Combining such results with Equation (34) and Equation (35), and we will finish the proof.

B.5.2 Proof of Theorem 1: Case II

As a reminder of notations, $(\text{BC}\star, \text{S}\star)$ denotes an arbitrary element from $\{(\text{BC}, \text{S}), (\text{BC}\tau, \text{S}\tau)\}$. In the case of $\text{BC}\star$, due to the failure of exchangeability, $\mathcal{M}_T(W^{\text{BC}\star})$ is now no longer a supermartingale with stopping time $T_q^{\text{BC}\star}$. To address this challenge, we are going show that $\mathcal{M}_T(W^{\text{S}\star})$ gives an upper bound of $\mathcal{M}_T(W^{\text{BC}\star})$ and is associated with a supermartingale structure at the same time. Specifically, we will show the following properties in addition to Proposition 4:

(a) $\mathcal{M}_T(W^{\text{S}\star})$ provides an upper bound for $\mathcal{M}_T(W^{\text{BC}\star})$, that $\mathcal{M}_T(W^{\text{BC}\star}) \leq \mathcal{M}_T(W^{\text{S}\star})$;

(b) $T_q^{\text{BC}\star}$ induces a stopping time for the inverse martingale associated with $\mathcal{M}_T(W^{\text{S}\star})$ which enables the application of upper bounds in Case I to Case II.

a) Upper Bound Property

Comparing the definition of $W^{\text{BC}\star}$ and the definition of $W^{\text{S}\star}$, we have

1. for $W_i^{\text{S}\star} > 0$, $W_i^{\text{BC}\star} = Z_i = W_i^{\text{S}\star}$;

2. for $W_i^{\text{S}\star} < 0$, if $(\text{BC}\star, \text{S}\star) = (\text{BC}, \text{S})$, $W_i^{\text{BC}\star} = -\widetilde{Z}_i \leq -Z_i = W_i^{\text{S}\star}$;

3. for $W_i^{\text{S}\star} < 0$, if $(\text{BC}\star, \text{S}\star) = (\text{BC}\tau, \text{S}\tau)$, $W_i^{\text{BC}\star} = -\tau(\widetilde{Z})_i \leq -Z_i = W_i^{\text{S}\star}$;

Therefore, $\{i \in S_0 : W_i^{\text{BC}\star} \geq T\} = \{i \in S_0 : W_i^{\text{S}\star} \geq T\}$, while $\{i \in S_0 : W_i^{\text{S}\star} \leq -T\} \subseteq \{i \in S_0 : W_i^{\text{BC}\star} \leq -T\}$ for $T > 0$, which further indicates

$$\mathcal{M}_T(W^{\text{BC}\star}) = \frac{\sum_{i \in S_0} 1\{W_i^{\text{BC}\star} \geq T\}}{1 + \sum_{i \in S_0} 1\{W_i^{\text{BC}\star} \leq -T\}} \leq \frac{\sum_{i \in S_0} 1\{W_i^{\text{S}\star} \geq T\}}{1 + \sum_{i \in S_0} 1\{W_i^{\text{S}\star} \leq -T\}} =: \mathcal{M}_T(W^{\text{S}\star}). \tag{38}$$

Thus $\mathcal{M}_T(W^{\text{S}\star})$ offers an upper bound for $\mathcal{M}_T(W^{\text{BC}\star})$.

32

b) Stopping Time and Supermartingale Inequalities

To apply the supermartingale inequality in Lemma 2, we will need to check that $T_q^{\text{BC}\star}$ defined by $W^{\text{BC}\star}$ can be induced to a stopping time associated with $\mathcal{M}_T(W^{\text{S}\star})$ with respect to a proper filtration.

For shorthand notations, we rearrange the index on of $W^{\text{S}\star}$, such that $|W_{(1)}^{\text{S}\star}| \geq |W_{(2)}^{\text{S}\star}| \geq \cdots \geq |W_{(m^*)}^{\text{S}\star}| > 0$, and $\{(1), (2), \cdots, (m^*)\} = S_0 \cap \{i : |W_i^{\text{S}}| = Z_i > 0\}$. Further denote $B_{(i)} = 1\{W_{(i)}^{\text{S}\star} < 0\}$, then there holds

$$\frac{\sum_{i \in S_0} 1\{W_i^{\text{S}\star} \geq T_q^{\text{BC}\star}\}}{1 + \sum_{i \in S_0} 1\{W_i^{\text{S}\star} \leq -T_q^{\text{BC}\star}\}} = \frac{1 + \sum_{i \in S_0} 1\{|W_i^{\text{S}\star}| \geq T_q^{\text{BC}\star}\}}{1 + \sum_{i \in S_0} 1\{|W_i^{\text{S}\star}| \geq T_q^{\text{BC}\star}, W_i^{\text{S}\star} < 0\}} - 1,$$
$$= \frac{1 + J}{1 + B_{(1)} + B_{(2)} + \cdots + B_{(J)}} - 1, \tag{39}$$

where $J \leq m^*$ is defined to be the index satisfying

$$|W_{(1)}^{\text{S}\star}| \geq |W_{(2)}^{\text{S}\star}| \geq \cdots \geq |W_{(J)}^{\text{S}\star}| \geq T_q^{\text{BC}\star} > |W_{(J+1)}^{\text{S}\star}| \geq \cdots \geq |W_{(m^*)}^{\text{S}\star}|,$$

in other words, $J = \text{argmax}_{k \leq m^*}\{|W_{(k)}^{\text{S}\star}| \geq T_q^{\text{BC}\star}\}$. It can be shown $\mathcal{F}^{\text{S}\star}(T)$ is a refined filtration of $\mathcal{F}^{\text{BC}\star}(T)$, i.e. $\mathcal{F}^{\text{BC}\star}(T) \subseteq \mathcal{F}^{\text{BC}\star}(T)$ in a similar way as discussed in Proposition 4. Rigorously speaking, we will have the following lemma showing that J is also a stopping time in inverse time with respect to the filtration $\mathcal{F}$ associated with $W^{\text{S}\star}$.

Lemma 4. *For any determined $\widehat{\beta}(\lambda)$, $J = \max_{i \leq m^*}\{|W_{(i)}^{\text{S}\star}| \geq T_q^{\text{BC}\star}\}$ is a stopping time with respect to the filtration $\{\mathcal{F}_j\}_{j=1}^m$ in inverse time defined as*

$$\mathcal{F}_j = \sigma\left(\left\{\sum_{i=1}^j B_{(i)}, \zeta_{(j+1)}, \cdots, \zeta_{(m)}\right\}\right).$$

With this stopping time property, we can apply Lemma 2 again to get the desired FDR bound. Specifically we will have the following inequality on $\mathcal{M}_{T_q^{\text{BC}\star}}(W^{\text{BC}\star})$ and $\mathcal{M}_{T_q^{\text{BC}\star}}(W^{\text{S}\star})$:

$$\mathbb{E}\left[\mathcal{M}_{T_q^{\text{BC}\star}}(W^{\text{BC}\star})\right] \leq \mathbb{E}\left[\mathcal{M}_{T_q^{\text{BC}\star}}(W^{\text{S}\star})\right] = \mathbb{E}\left[\frac{\sum_{i \in S_0} 1\{W_i^{\text{S}\star} \geq T_q^{\text{BC}\star}\}}{1 + \sum_{i \in S_0} 1\{W_i^{\text{S}\star} \leq -T_q^{\text{BC}\star}\}}\right] \leq 1, \tag{40}$$

which ends the proof.

B.6 Proof of Supporting Lemmas

In this section, we will give the proof of Lemma 2, Lemma 3 (an extension of Lemma 1), and Lemma 4 respectively. As a reminder, $(\text{BC}\star, \text{S}\star)$ refers to any element from $\{(\text{BC}, \text{S}), (\text{BC}\tau, \text{S}\tau)\}$, where $W^{\text{BC}\tau}$ is defined in Section B.5 for completeness.

In the proof of Lemma 2, we explicitly construct a decomposition in B_i similar with that in proof of Lemma 1 in [BC$^+$19], based on the property that $B_i = 1\{\zeta_i \in G_i\}$ for some Borel set G_i. Such a decomposition enables us the desired result.

In the proof of Lemma 3, we give detailed analysis on the KKT conditions (16), which provides a specific mapping from the value of ζ to $\text{sign}(W^{\text{S}\star})$. Then $\mathbb{P}[W^{\text{S}\star} < 0]$ can be estimated based on the probability measure on ζ.

In the proof of Lemma 4, we apply the ideas of $\mathcal{F}^{\text{BC}}(T) \subseteq \mathcal{F}^{\text{S}}(T)$ introduced in Proposition 4 in Section 3 in a discrete form. We will show that such a result holds for any choice of $(\text{BC}\star, \text{S}\star)$ in the proof.

B.6.1 Proof of Lemma 2

Proof. We start from the following constructions. For each i, we divide the space $\mathbb{R}$ into 4 disjoint Borel sets, $\mathbb{R} = A_1^i \cup A_2^i \cup A_3^i \cup A_4^i$, with $G_i := A_2^i \cup A_3^i \cup A_4^i$, and

1. $\mathbb{P}[\zeta_i \in A_1^i] = 1 - \rho_i$;

2. $\mathbb{P}[\zeta_i \in A_2^i] = \rho \frac{1 - \rho_i}{1 - \rho}$;

3. $\mathbb{P}[\zeta_i \in A_3^i] = \rho \frac{\rho_i - \rho}{1 - \rho}$;

4. $\mathbb{P}[\zeta_i \in A_4^i] = \rho_i - \rho$.

The existence of such division is ensured by $1 - \rho_i + \rho \frac{1 - \rho_i}{1 - \rho} + \rho \frac{\rho_i - \rho}{1 - \rho} + \rho_i - \rho = 1$, and $\rho \frac{1 - \rho_i}{1 - \rho} + \rho \frac{\rho_i - \rho}{1 - \rho} + \rho_i - \rho = \rho_i$.

Define $U_i = A_1^i \cup A_2^i$, and $V_i = A_2^i \cup A_3^i$ for each i. Further define $Q_i = 1\{\zeta_i \in V_i\}$ for each i and a random set $A := \{i : \zeta_i \in U_i\}$. There holds from the definition that

$$
\begin{aligned}
Q_i \cdot 1\{i \in A\} + 1\{i \notin A\} &= 1\{\{\zeta_i \in V_i \cap U_i\} \cup \{\zeta_i \in U_i^C\}\}, \\
&= 1\{\{\zeta_i \in A_2^i\} \cup \{\zeta_i \in A_3^i \cup A_4^i\}\}, \\
&= 1\{\zeta_i \in G_i\} = B_i.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\frac{1 + J}{1 + B_{(1)} + B_{(2)} + \cdots + B_{(J)}} &= \frac{1 + |\{i \le J : (i) \in A\}| + |\{i \le J : (i) \notin A\}|}{1 + \sum_{i \le J, (i) \in A} Q_{(i)} + |\{i \le J : (i) \notin A\}|}, \\
&\le \frac{1 + |\{i \le J : (i) \in A\}|}{1 + \sum_{i \le J, (i) \in A} Q_{(i)}},
\end{aligned}
\tag{41}
$$

where the last step is by the inequality $\frac{a+c}{b+c} \le \frac{a}{b}$ whenever $0 < b \le a$ and $c \ge 0$.

Let $\widetilde{Q}_i = Q_i \cdot 1\{i \in A\}$, and define

$$
\mathcal{F}_j' = \sigma \left(\left\{ \sum_{i=1}^{j} \widetilde{Q}_{(i)}, \zeta_{(j+1)}, \cdots, \zeta_{(m)}, A \right\} \right),
$$

then clearly $\{\mathcal{F}_j'\}_{j=1}^m$ is a filtration in inverse time, satisfying $\mathcal{F}_j \subseteq \mathcal{F}_j'$. Therefore, J is also a stopping time on $\mathcal{F}_j'$. Moreover, note that by definition

$$
\mathbb{P}[Q_i = 1 | i \in A] = \mathbb{P}[\zeta_i \in V_i | \zeta_i \in U_i] = \frac{\mathbb{P}[A_2^i]}{\mathbb{P}[A_1^i \cup A_2^i]} = \rho = \mathbb{P}[Q_i = 1],
$$

$$
\mathbb{P}[Q_i = 1 | i \notin A] = \mathbb{P}[\zeta_i \in V_i | \zeta_i \notin U_i] = \frac{\mathbb{P}[A_3^i]}{\mathbb{P}[A_3^i \cup A_4^i]} = \rho = \mathbb{P}[Q_i = 1].
$$

Such an observation combining with the the independence on ζ_i means that, conditional on any fixed A, Q_i are i.i.d. Bernoulli random variables with parameter ρ. This further suggests that Q_i as i.i.d. Bernoulli random variables with parameter ρ, and are independent from A. This means that Q_i are exchangeable [BC$^+$19] conditional on any fixed A, then by Lemma 2 in the supplementary material of [BC$^+$19], there holds

$$
\mathbb{E}\left[\frac{1 + |\{i \le J : (i) \in A\}|}{1 + \sum_{i \le J, (i) \in A} Q_{(i)}} \bigg| A \right] \le \rho^{-1}.
$$

Taking expectation over A, and we will get our desired result. $\square$

B.6.2 Proof of Lemma 3

Note that by Equation (7), there holds

$$\widetilde{A}_\gamma^T A_\beta = \widetilde{A}_{\gamma,1}^T \frac{X_2}{\sqrt{n_2}} + \widetilde{A}_{\gamma,2}^T \frac{D}{\sqrt{\nu}} = A_\gamma^T A_\beta = -\frac{D}{\nu},$$

$$\widetilde{A}_\gamma^T A_\gamma = -\frac{\widetilde{A}_{\gamma,2}^T}{\sqrt{\nu}} = A_\gamma^T A_\gamma - \mathrm{diag}(s) = \frac{I_m}{\nu} - \mathrm{diag}(s),$$

where as a reminder, $\widetilde{A}_{\gamma,1} \in \mathbb{R}^{n_2 \times m}$ and $\widetilde{A}_{\gamma,2} \in \mathbb{R}^{m \times m}$ is defined in the end of Section 2.2. Therefore, there holds

$$\widetilde{A}_\gamma^T \widetilde{y} = \widetilde{A}_{\gamma,1}^T \frac{X_2 \beta^* + \varepsilon_2}{\sqrt{n_2}} = -\mathrm{diag}(s)\gamma^* + \frac{\widetilde{A}_{\gamma,1}^T}{\sqrt{n_2}} \varepsilon_2, \tag{42}$$

where ε_2 is the last n_2 rows of ε as defined in Section 3.1.

From Equation (16a), $\gamma(\lambda)$ is determined by $\widehat{\beta}(\lambda)$, suggesting that $\widehat{\beta}(\lambda)$ is the sufficient statistics for Z and r. From Equation (16b), $\widehat{\beta}(\lambda)$ and ζ_i are the sufficient statistics for $\widetilde{\gamma}_i(\lambda)$ and consequently $\widetilde{Z}_i$ and $\widetilde{r}_i$ for all i. Therefore, $\widehat{\beta}(\lambda)$ and ζ_i are the sufficient statistics for $W_i^{\mathrm{S}\star}$ for all i.

Now, we further present by the following lemma that $B_i := 1\{W_i^{\mathrm{S}\star} < 0\}$ are some independent random variables through the independence among ζ_i.

Lemma 5. *For any determined $\widehat{\beta}(\lambda)$, let $B_i := 1\{W_i^{\mathrm{S}\star} < 0\}$, then B_i are some independent random variables.*

Proof. Note that by Equation (7), there holds

$$\widetilde{A}_\gamma^T \widetilde{A}_\gamma = \widetilde{A}_{\gamma,1}^T \widetilde{A}_{\gamma,1} + \widetilde{A}_{\gamma,2}^T \widetilde{A}_{\gamma,2} = A_\gamma^T A_\gamma = \frac{I_m}{\nu},$$

$$\widetilde{A}_\gamma^T A_\gamma = -\frac{\widetilde{A}_{\gamma,2}^T}{\sqrt{\nu}} = A_\gamma^T A_\gamma - \mathrm{diag}(s) = \frac{I_m}{\nu} - \mathrm{diag}(s),$$

where $\widetilde{A}_{\gamma,2} \in \mathbb{R}^{m \times m}$ is defined to be the matrix that takes last m rows of $\widetilde{A}_\gamma$. Therefore, there holds

$$\widetilde{A}_{\gamma,1}^T \widetilde{A}_{\gamma,1} = \mathrm{diag}(s)(2I_m - \mathrm{diag}(s)\nu), \tag{43}$$

and $\zeta + \mathrm{diag}(s)\gamma^*$ follows the distribution presented in Equation (18). Therefore, ζ consists of some *independent* random variables. Combining with the fact that $\widehat{\beta}(\lambda)$ and ζ_i are the sufficient statistics for $W_i^{\mathrm{S}\star}$ for all i, we will get our desired result. □

We show by the following proposition that $Z_i < +\infty$ for all i, and the event $\{\exists i : Z_i = \widetilde{Z}_i > 0\}$ is a zero probability event, conditional on any determined $\widehat{\beta}(\lambda)$ satisfying $\lim_{\lambda \to \infty} \frac{\widehat{\beta}(\lambda)}{\lambda} = 0$. To avoid redundant notations, we assume that $Z_i \neq \widetilde{Z}_i$ for all i satisfying $Z_i > 0$ throughout this paper.

Proposition 8. *For any determined $\widehat{\beta}(\lambda)$, there holds for any i that*

- $Z_i < +\infty$;

- $\mathbb{P}[Z_i = \widetilde{Z}_i > 0] = 0$.

Proof. From Equation (16), there holds for all $\lambda > 0$ and $i \in \{1, 2, \cdots, m\}$ that

$$\rho_i(\lambda) + \frac{\gamma_i(\lambda)}{\lambda \nu} = \frac{[D\widehat{\beta}(\lambda)]_i}{\lambda \nu},$$

$$\widetilde{\rho}_i(\lambda) + \frac{\widetilde{\gamma}_i(\lambda)}{\lambda \nu} = \frac{[D\widehat{\beta}(\lambda)]_i}{\lambda \nu} + \frac{\zeta_i}{\lambda}.$$

Therefore, there holds for $Z_i > 0$ and $\widetilde{Z}_i > 0$ that

$$Z_i = \sup \left\{ \lambda : \left| \frac{[D\widehat{\beta}(\lambda)]_i}{\lambda \nu} \right| > 1 \right\}, \text{ and } \widetilde{Z}_i = \sup \left\{ \lambda : \left| \frac{[D\widehat{\beta}(\lambda)]_i}{\lambda \nu} + \frac{\zeta_i}{\lambda} \right| > 1 \right\}. \tag{44}$$

By the condition $\lim_{\lambda \to \infty} \frac{\widehat{\beta}(\lambda)}{\lambda} = 0$ in Section 2.1, there holds

$$\lim_{\lambda \to \infty} \left| \frac{[D\widehat{\beta}(\lambda)]_i}{\lambda \nu} \right| = 0.$$

Therefore, by Equation (44), there holds $Z_i < +\infty$.

Moreover, by the continuity of $\widehat{\beta}(\lambda)$, by Equation (44), there holds for $Z_i > 0$ and $\widetilde{Z}_i > 0$ that

$$\left| \frac{[D\widehat{\beta}(Z_i)]_i}{Z_i \nu} \right| = \lim_{\lambda \to Z_i^+} \left| \frac{[D\widehat{\beta}(\lambda)]_i}{\lambda \nu} \right| = 1, \text{ and } \left| \frac{[D\widehat{\beta}(\widetilde{Z}_i)]_i}{\widetilde{Z}_i \nu} + \frac{\zeta_i}{\widetilde{Z}_i} \right| = \lim_{\lambda \to \widetilde{Z}_i^+} \left| \frac{[D\widehat{\beta}(\lambda)]_i}{\lambda \nu} + \frac{\zeta_i}{\lambda} \right| = 1.$$

Therefore, if there further holds $Z_i = \widetilde{Z}_i > 0$, there holds

$$\left| \frac{[D\widehat{\beta}(Z_i)]_i}{Z_i \nu} \right| = \left| \frac{[D\widehat{\beta}(Z_i)]_i}{Z_i \nu} + \frac{\zeta_i}{Z_i} \right|,$$

which only have finite solutions of ζ_i for any $\widehat{\beta}(\lambda)$ (which determines Z_i). Therefore $\mathbb{P}[Z_i = \widetilde{Z}_i > 0] = 0$. □

We describe by the following lemma on how ζ determines $\text{sign}(W^{S\star})$ conditional on a determined $\widehat{\beta}(\lambda)$.

Lemma 6. *For any determined $\widehat{\beta}(\lambda)$, there holds for i satisfying $Z_i > 0$,*

$$\{r_i \zeta_i > 0\} \subseteq \{W_i^S < 0\}, \ \{r_i \zeta_i > 0\} \subseteq \{W_i^{S\tau} < 0\},$$

where r_i is defined in Equation (12) and is shown to be equal to $\text{sign}(\rho_i(Z_i))$ for $Z_i > 0$ in Proposition 3.

Proof. We prove the two properties one by one.

1) $\{r_i \zeta_i > 0\} \subseteq \{W_i^S < 0\}$. Suppose that $r_i \zeta_i > 0$. We are going to show below that $\widetilde{Z}_i > Z_i$ and $W_i^S < 0$.

36

By the continuity of $\widehat{\beta}(\lambda)$, $\gamma(\lambda)$ and $\rho(\lambda)$ solved from Equation (16a) are continuous. Therefore, by the definition of Z_i in Equation (9), for i satisfying $Z_i > 0$, there holds $\gamma_i(Z_i) = \lim_{\lambda \to Z_i^+} \gamma_i(\lambda) = 0$ and $|\rho_i(Z_i)| = \limsup_{\lambda \to Z_i^-} |\rho_i(\lambda)| = 1$. From Equation (16a), there holds

$$Z_i \rho_i(Z_i) = Z_i \rho_i(Z_i) + \frac{\gamma_i(Z_i)}{\nu} = \frac{[D\widehat{\beta}(Z_i)]_i}{\nu}.$$

Then from Equation (16b), there holds

$$Z_i \widetilde{\rho}_i(Z_i) + \frac{\widetilde{\gamma}_i(Z_i)}{\nu} = \frac{[D\widehat{\beta}(Z_i)]_i}{\nu} + \zeta_i = Z_i \rho_i(Z_i) + \zeta_i.$$

Multiple $r_i = \operatorname{sign}(\rho_i(Z_i))$ on both sides and combining with the fact that $|\rho_i(Z_i)| = 1$, there holds

$$r_i Z_i \widetilde{\rho}_i(Z_i) + \frac{r_i \widetilde{\gamma}_i(Z_i)}{\nu} = Z_i + r_i \zeta_i.$$

Therefore $r_i \widetilde{\gamma}_i(Z_i) = \nu r_i \zeta_i > 0$. By the continuity of $\widehat{\beta}(\lambda)$, $\widetilde{\gamma}(\lambda)$ and $\widetilde{\rho}(\lambda)$ solved from Equation (16b) are continuous. Therefore, there exists $\delta > 0$, such that $r_i \widetilde{\gamma}_i(Z_i + \delta) > \frac{\nu r_i \zeta_i}{2} > 0$. This further suggests that $\widetilde{Z}_i > Z_i + \delta > Z_i$ and $W_i^{\mathrm{S}} < 0$.

2) $\{r_i \zeta_i > 0\} \subseteq \{W_i^{\mathrm{S}\tau} < 0\}$. Suppose that $r_i \zeta_i > 0$. It is already shown above that $\widetilde{Z}_i > Z_i > 0$, and it remains to show that $r_i = \widetilde{r}_i = \operatorname{sign}(\widetilde{\rho}_i(\widetilde{Z}_i))$ for showing $W_i^{\mathrm{S}\tau} < 0$.

By Equation (16a) and the definition of Z_i in Equation (9), there holds for $\lambda > Z_i$ that

$$\lambda \rho_i(\lambda) = \lambda \rho_i(\lambda) + \frac{\gamma_i(\lambda)}{\nu} = \frac{[D\widehat{\beta}(\lambda)]_i}{\nu}.$$

Then from $r_i \times$Equation (16b), there holds for $\lambda > Z_i$ and $r_i \zeta_i > 0$ that

$$r_i \lambda \widetilde{\rho}_i(\lambda) + r_i \frac{\widetilde{\gamma}_i(\lambda)}{\nu} = \frac{r_i [D\widehat{\beta}(\lambda)]_i}{\nu} + r_i \zeta_i = r_i \lambda \rho_i(\lambda) + r_i \zeta_i > r_i \lambda \rho_i(\lambda) \geq -\lambda.$$

Therefore there holds $r_i \widetilde{\rho}_i(\lambda) > -1$ for $\lambda > Z_i$ and consequetly $r_i \widetilde{\rho}_i(\widetilde{Z}_i) > -1$ as $\widetilde{Z}_i > Z_i$.

However, since $\widetilde{Z}_i > Z_i > 0$, there holds $|\widetilde{\rho}_i(\widetilde{Z}_i)| = \limsup_{\lambda \to \widetilde{Z}_i^-} |\widetilde{\rho}_i(\lambda)| = 1$. Combining with $r_i \widetilde{\rho}_i(\widetilde{Z}_i) > -1$, there holds $r_i \widetilde{\rho}_i(\widetilde{Z}_i) = 1$, and consequently $\widetilde{r}_i = r_i$ for $\widetilde{r}_i = \operatorname{sign}(\widetilde{\rho}_i(\widetilde{Z}_i))$. Therefore, $W_i^{\mathrm{S}\tau} < 0$. $\qquad\square$

Then it remains to show that for $i \in S_0 \cap \{i : |W_i^{\mathrm{S}\star}| = Z_i > 0\}$, there holds $\mathbb{P}[B_i = 1] \geq \rho(\nu) \geq \frac{1}{2}$ to prove Lemma 3.

Proof. First, from Lemma 6, conditional on any $\widehat{\beta}(\lambda)$, there holds $\{r_i \zeta_i > 0\} \subseteq \{W_i^{\mathrm{S}\star} < 0\}$ for $i \in S_0 \cap \{i : |W_i^{\mathrm{S}\star}| = Z_i > 0\}$. While for $i \in S_0$, ζ_i is symmetrically distributed around 0 by Equation (18). Thus $\mathbb{P}[W_i^{\mathrm{S}\star} < 0] \geq \mathbb{P}[r_i \zeta_i > 0] = \frac{1}{2}$ conditional on $\widehat{\beta}(\lambda)$, for $i \in S_0 \cap \{i : |W_i^{\mathrm{S}\star}| = Z_i > 0\}$. This ends the proof. $\qquad\square$

B.6.3 Proof of Lemma 4

Proof. As a reminder, $(\mathrm{BC}\star, \mathrm{S}\star)$ represents an arbitrary element from $\{(\mathrm{BC}, \mathrm{S}), (\mathrm{BC}\tau, \mathrm{S}\tau)\}$, where $W^{\mathrm{BC}\tau}$ is defined in Section B.5 for completeness. By definition of stopping time in reverse time, we need to show that $\{J < k\} \in \mathcal{F}_k$ for $k \leq m^*$. We first validate that $\{\mathcal{F}_j\}_{j=1}^m$ is indeed a filtration in inverse time. By Equation (16), for any determined $\widehat{\beta}(\lambda)$, $\zeta_{(i)}$ will determine $W_{(i)}^{\mathrm{BC}\star}$ as well as $W_{(i)}^{\mathrm{S}\star}$ for all i, thus determine $B_{(i)}$. Such a fact validates our claim that $\{\mathcal{F}_j\}_{j=1}^m$ is indeed a filtration. Furthermore, by the fact that $W^{\mathrm{BC}\star}$ and $W^{\mathrm{S}\star}$ share the same sign, there holds:

- $\mathcal{F}_k$ includes $\{W_{(i)}^{\mathrm{BC}\star} : i > k\}$, $V_s^-(k) := \#\{W_{(i)}^{\mathrm{BC}\star} < 0, i \leq k\}$, and $V_s^+(k) := \#\{W_{(i)}^{\mathrm{BC}\star} > 0, i \leq k\} = k - V_s^-(k)$.

- the event $\{J < k\}$ is defined by, for all l such that $|W_{(l)}^{\mathrm{BC}\star}| \leq |W_{(k)}^{\mathrm{S}\star}|$:

$$\frac{|\{i : W_{(i)}^{\mathrm{BC}\star} \leq -|W_{(l)}^{\mathrm{BC}\star}|\}|}{1 \vee |\{i : W_{(i)}^{\mathrm{BC}\star} \geq |W_{(l)}^{\mathrm{BC}\star}|\}|} > q, \text{ for Split Knockoff;} \quad \frac{1 + |\{i : W_{(i)}^{\mathrm{BC}\star} \leq -|W_{(l)}^{\mathrm{BC}\star}|\}|}{1 \vee |\{i : W_{(i)}^{\mathrm{BC}\star} \geq |W_{(l)}^{\mathrm{BC}\star}|\}|} > q, \text{ for Split Knockoff+}$$

  i.e. is decided by $\#\{i : W_{(i)}^{\mathrm{BC}\star} \leq -|W_{(l)}^{\mathrm{BC}\star}|\}$ and $\#\{i : W_{(i)}^{\mathrm{BC}\star} \geq |W_{(l)}^{\mathrm{BC}\star}|\}$.

Hence it suffices to show that, for all l such that $|W_{(l)}^{\mathrm{BC}\star}| \leq |W_{(k)}^{\mathrm{S}\star}|$, $\mathcal{F}_k$ includes

$$\#\{i : W_{(i)}^{\mathrm{BC}\star} \leq -|W_{(l)}^{\mathrm{BC}\star}|\} \text{ and } \#\{i : W_{(i)}^{\mathrm{BC}\star} \geq |W_{(l)}^{\mathrm{BC}\star}|\}.$$

To see this, the first set is decomposed by

$$\begin{aligned}
&\#\{W_{(i)}^{\mathrm{BC}\star} \leq -|W_{(l)}^{\mathrm{BC}\star}|\} \\
=&\#\{W_{(i)}^{\mathrm{BC}\star} \leq -|W_{(l)}^{\mathrm{BC}\star}|, i \leq k\} + \#\{W_{(i)}^{\mathrm{BC}\star} \leq -|W_{(l)}^{\mathrm{BC}\star}|, i > k\}, \\
=&\#\{W_{(i)}^{\mathrm{BC}\star} < 0, i \leq k\} + \#\{W_{(i)}^{\mathrm{BC}\star} \leq -|W_{(l)}^{\mathrm{BC}\star}|, i > k\}.
\end{aligned}$$

The last step is by for $i \leq k$, $|W_{(i)}^{\mathrm{BC}\star}| \geq |W_{(i)}^{\mathrm{S}\star}| \geq |W_{(k)}^{\mathrm{S}\star}| \geq |W_{(l)}^{\mathrm{BC}\star}|$. Here the first part is equal to $V_s^-(k) := \#\{W_{(i)}^{\mathrm{BC}\star} < 0, i \leq k\}$, the second part is determined by $\{W_{(i)}^{\mathrm{BC}\star}, i > k\}$, both of which are in the σ-field of $\mathcal{F}_k$. This shows that $\mathcal{F}_k$ includes $\#\{i : W_{(i)}^{\mathrm{BC}\star} \leq -|W_{(l)}^{\mathrm{BC}\star}|\}$. Similarly, the second set is decomposed by

$$\begin{aligned}
&\#\{i : W_{(i)}^{\mathrm{BC}\star} \geq |W_{(l)}^{\mathrm{BC}\star}|\} \\
=& \#\{i : W_{(i)}^{\mathrm{BC}\star} \geq |W_{(l)}^{\mathrm{BC}\star}|, i \leq k\} + \#\{i : W_{(i)}^{\mathrm{BC}\star} \geq |W_{(l)}^{\mathrm{BC}\star}|, i > k\}, \\
=& \#\{i : W_{(i)}^{\mathrm{BC}\star} > 0, i \leq k\} + \#\{i : W_{(i)}^{\mathrm{BC}\star} \geq |W_{(l)}^{\mathrm{BC}\star}|, i > k\},
\end{aligned}$$

where last step is by for $i \leq k$, $|W_{(i)}^{\mathrm{BC}\star}| \geq |W_{(i)}^{\mathrm{S}\star}| \geq |W_{(k)}^{\mathrm{S}\star}| \geq |W_{(l)}^{\mathrm{BC}\star}|$. The first part is equal to $V_s^+(k) = \#\{W_{(i)}^{\mathrm{BC}\star} > 0, i \leq k\}$ and the second part is determined by $\{W_{(i)}^{\mathrm{BC}\star} : i > k\}$, both of which are in the σ-field of $\mathcal{F}_k$. Therefore $\mathcal{F}_k$ includes $\#\{i : W_{(i)}^{\mathrm{BC}\star} \geq |W_{(l)}^{\mathrm{BC}\star}|\}$. This finishes the proof. $\square$

B.7 Proof of Theorem 2

Proof. When the event Υ occurs, by definition, $\widetilde{y}$, A_β, A_γ and $\widetilde{\varepsilon}$ defined in Equation (20) satisfy

$$\widetilde{y} = A_\beta \beta^*_{\widehat{S}_\beta} + A_\gamma \gamma^*_{\widehat{S}_\gamma} + \widetilde{\varepsilon}, \tag{45}$$

where $\beta^*_{\widehat{S}_\beta}$ is a subvector of β^*, consisting of the rows in $\widehat{S}_\beta$, and $\gamma^*_{\widehat{S}_\gamma}$ is a subvector of γ^*, consisting of the rows in $\widehat{S}_\gamma$.

From Equation (8) and Equation (10), in the high dimensional setting, the KKT conditions that the solution $\gamma(\lambda)$, $\widetilde{\gamma}(\lambda)$ — both are $\mathbb{R}_+ \to \mathbb{R}^{|\widehat{S}_\gamma|}$ functions — should satisfy are

$$\lambda\rho(\lambda) + \frac{\gamma(\lambda)}{\nu} = \frac{D\widehat{\beta}(\lambda)}{\nu}, \tag{46a}$$

$$\lambda\widetilde{\rho}(\lambda) + \frac{\widetilde{\gamma}(\lambda)}{\nu} = \frac{D\widehat{\beta}(\lambda)}{\nu} + \underbrace{\left\{ -\operatorname{diag}(s)\gamma^*_{\widehat{S}_\gamma} + \frac{\widetilde{A}^T_{\gamma,1}}{\sqrt{n_2}}\varepsilon_2 \right\}}_{=:\zeta}, \tag{46b}$$

where $\rho(\lambda) \in \partial\|\gamma(\lambda)\|_1$, $\widetilde{\rho}(\lambda) \in \partial\|\widetilde{\gamma}(\lambda)\|_1$, and $\widetilde{A}_{\gamma,1}$ is defined to be the matrix that takes first n_2 rows of $\widetilde{A}_\gamma$. For shorthand notations, define $\zeta := -\operatorname{diag}(s)\gamma^*_{\widehat{S}_\gamma} + \frac{\widetilde{A}^T_{\gamma,1}}{\sqrt{n_2}}\varepsilon_2$. Follow the same steps as those in achieving Equation (43), it can be shown that the ζ should satisfy

$$\zeta \sim \mathcal{N}\left(-\operatorname{diag}(s)\gamma^*_{\widehat{S}_\gamma}, \frac{1}{n_2}\operatorname{diag}(s)(2I_{|\widehat{S}_\gamma|} - \operatorname{diag}(s)\nu)\sigma^2 \right). \tag{47}$$

Such a fact suggests that $\zeta + \operatorname{diag}(s)\gamma^*_{\widehat{S}_\gamma}$ consists of independent Gaussian random variables. With this property, follow the same steps as in Section B.5 (proof of Theorem 1), there holds for Split Knockoff

$$\mathbb{E}\left[\frac{\left|\left\{ i : i \in \widehat{S}^\star \cap S_0 \right\}\right|}{\left|\widehat{S}^\star\right| + q^{-1}} \middle| \Upsilon \right] \leq q,$$

and for Split Knockoff+

$$\mathbb{E}\left[\frac{\left|\left\{ i : i \in \widehat{S}^\star \cap S_0 \right\}\right|}{\left|\widehat{S}^\star\right| \vee 1} \middle| \Upsilon \right] \leq q.$$

This ends the proof. $\qquad\square$

B.8   Proof of Proposition 6

In this section, we approach the model selection consistency by constructing Primal-Dual Witness (PDW) of Split LASSO regularization paths, following the same treatment in the traditional LASSO problem [Wai09] and Split Linearized Bregman Iterations [HSXY16, HSXY20]. We first introduce the concept of the successful Primal-Dual Witness which has a unique solution of Split LASSO; then we introduce where the incoherence condition for Split LASSO comes from and establish the no-false-positive and sign consistency of Split LASSO regularization path, i.e. Proposition 6.

We first list the KKT conditions that an optimal solution of the Split LASSO problem (4) satisfies, as they will be commonly used throughout this section. The KKT conditions are

$$0 = -(\Sigma_X + L_D)\beta(\lambda) + \frac{D^T}{\nu}\gamma(\lambda) + \left\{ \Sigma_X\beta^* + \frac{X^T}{n}\varepsilon \right\}, \tag{48a}$$

$$\lambda\rho(\lambda) = \frac{D\beta(\lambda)}{\nu} - \frac{\gamma(\lambda)}{\nu}, \tag{48b}$$

where $\rho(\lambda) \in \partial\|\gamma(\lambda)\|_1$.

B.8.1   Primal-Dual Witness

In this section, we will introduce the lemma that ensures the uniqueness of the successful Primal-Dual Witness for Split LASSO problem. In the beginning, we will give the definition of the PDW. The set of witness $(\widehat{\beta}^\lambda, \widehat{\gamma}^\lambda, \widehat{\rho}^\lambda) \in \mathbb{R}^p \times \mathbb{R}^m \times \mathbb{R}^m$ is constructed in the following way:

1. First, we set $\widehat{\gamma}_{S_0}^\lambda = 0$, and obtain $(\widehat{\beta}^\lambda, \widehat{\gamma}_{S_1}^\lambda) \in \mathbb{R}^p \times \mathbb{R}^{|S_1|}$ by solving

$$(\widehat{\beta}^\lambda, \widehat{\gamma}_{S_1}^\lambda) = \underset{(\beta, \gamma_{S_1})}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \frac{1}{2\nu} \|D\beta - \gamma\|_2^2 + \lambda \|\gamma_{S_1}\|_1 \right\}. \tag{49}$$

2. Second, we choose $\widehat{\rho}_{S_1}^\lambda = \partial \|\widehat{\gamma}_{S_1}^\lambda\|_1$ as the subgradient of $\|\widehat{\gamma}_{S_1}^\lambda\|_1$.

3. Third, for no-false-positive, we solve for $\widehat{\rho}_{S_0}^\lambda \in \mathbb{R}^{S_0}$ satisfying the KKT condition (48), and check whether or not the dual feasibility condition $|\widehat{\rho}_j^\lambda| < 1$ for all $j \in S_0$ is satisfied.

4. Fourth, for model selection (sign) consistency, we check whether $\widehat{\rho}_{S_1}^\lambda = \operatorname{sign}(\beta_{S_1}^*)$ is satisfied.

Then we can give the following lemma.

Lemma 7. *When the PDW succeed, if the subproblem (49) is strictly convex, the solution $(\widehat{\beta}, \widehat{\gamma})$ is the unique optimal solution for split LASSO.*

Proof. When PDW succeed, we have $\|\widehat{\rho}_{S_0}\|_\infty < 1$, and therefore $(\widehat{\beta}, \widehat{\gamma})$ is a set of optimal solution, while $\widehat{\rho}$ is in the subgradient of $\|\widehat{\gamma}\|_1$. Let $(\widetilde{\beta}, \widetilde{\gamma})$ be any other optimal solution for Split LASSO. Denote

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

Then there holds

$$F(\widehat{\beta}, \widehat{\gamma}) + \lambda \langle \widehat{\rho}, \widehat{\gamma} \rangle = F(\widetilde{\beta}, \widetilde{\gamma}) + \lambda \|\widetilde{\gamma}\|_1,$$

which is

$$F(\widehat{\beta}, \widehat{\gamma}) - \lambda \langle \widehat{\rho}, \widetilde{\gamma} - \widehat{\gamma} \rangle - F(\widetilde{\beta}, \widetilde{\gamma}) = \lambda (\|\widetilde{\gamma}\|_1 - \langle \widehat{\rho}, \widetilde{\gamma} \rangle).$$

Also, by Equation (48), there holds $\frac{\partial F(\widehat{\beta}, \widehat{\gamma})}{\partial \widehat{\beta}} = 0$, and $\frac{\partial F(\widehat{\beta}, \widehat{\gamma})}{\partial \widehat{\gamma}} = -\lambda \widehat{\rho}$. Therefore

$$F(\widehat{\beta}, \widehat{\gamma}) + \langle \frac{\partial F(\widehat{\beta}, \widehat{\gamma})}{\partial \widehat{\gamma}}, \widetilde{\gamma} - \widehat{\gamma} \rangle + \langle \frac{\partial F(\widehat{\beta}, \widehat{\gamma})}{\partial \widehat{\beta}}, \widetilde{\beta} - \widehat{\beta} \rangle - F(\widetilde{\beta}, \widetilde{\gamma}) = \lambda (\|\widetilde{\gamma}\|_1 - \langle \widehat{\rho}, \widetilde{\gamma} \rangle). \tag{50}$$

Since F is convex, the left hand side of Equation (50) is non-positive, then there holds

$$\|\widetilde{\gamma}\|_1 \leq \langle \widehat{\rho}, \widetilde{\gamma} \rangle.$$

Since $|\widehat{\rho}_{S_0}| < 1$, there holds $\widetilde{\gamma}_{S_0} = 0$. Therefore $(\widetilde{\beta}, \widetilde{\gamma})$ is also an optimal solution for the subproblem (49). Thus if the subproblem (49) is strictly convex, then $(\widehat{\beta}, \widehat{\gamma})$ is the only solution for split LASSO. $\qquad \square$

B.8.2 Incoherence Condition and Path Consistency

In this section, We first show on how the incoherence condition in Proposition 6 is formalized and then we give the proof of Proposition 6.

From $D[\Sigma_X + L_D]^{-1} \times$ Equation (48a) $+ \nu \times$ Equation (48b), and the fact that $\gamma^* = D\beta^*$, there holds for the solution $(\widehat{\beta}, \widehat{\gamma})$ to Equation (4) that

$$
\begin{aligned}
\lambda\nu\widehat{\rho} = & -\widehat{\gamma} + \frac{D[\Sigma_X + L_D]^{-1}D^T\widehat{\gamma}}{\nu} + D[\Sigma_X + L_D]^{-1}(\Sigma_X + L_D - L_D)\beta^* + \dots \\
& \dots + D[\Sigma_X + L_D]^{-1}\frac{X^T}{n}\varepsilon, \\
= & -H_\nu(\widehat{\gamma} - \gamma^*) + \omega,
\end{aligned}
$$

where $\omega = D[\Sigma_X + L_D]^{-1}\frac{X^T}{n}\varepsilon$. From the definition of H_ν^{11}, H_ν^{00}, H_ν^{10}, and H_ν^{01}, there further holds

$$
\lambda\nu \begin{bmatrix} \widehat{\rho}_{S_1} \\ \widehat{\rho}_{S_0} \end{bmatrix} = - \begin{bmatrix} H_\nu^{11} & H_\nu^{10} \\ H_\nu^{01} & H_\nu^{00} \end{bmatrix} \begin{bmatrix} \widehat{\gamma}_{S_1} - \gamma_{S_1}^* \\ 0_{S_0} \end{bmatrix} + \begin{bmatrix} \omega_{S_1} \\ \omega_{S_0} \end{bmatrix},
$$

which means

$$
\lambda\nu\widehat{\rho}_{S_1} = -H_\nu^{11}(\widehat{\gamma}_{S_1} - \gamma_{S_1}^*) + \omega_{S_1}, \tag{51a}
$$
$$
\lambda\nu\widehat{\rho}_{S_0} = -H_\nu^{01}(\widehat{\gamma}_{S_1} - \gamma_{S_1}^*) + \omega_{S_0}. \tag{51b}
$$

Since H_ν^{11} is reversible, $\widehat{\gamma}_{S_1} - \gamma_{S_1}^*$ can be solved from Equation (51a), and there holds

$$
\widehat{\gamma}_{S_1} - \gamma_{S_1}^* = -\lambda\nu[H_\nu^{11}]^{-1}\widehat{\rho}_{S_1} + [H_\nu^{11}]^{-1}\omega_{S_1}. \tag{52}
$$

Then, plug Equation (52) into Equation (51b), there further holds

$$
\widehat{\rho}_{S_0} = H_\nu^{01}[H_\nu^{11}]^{-1}\widehat{\rho}_{S_1} + \frac{1}{\lambda\nu}\{\omega_{S_0} - H_\nu^{01}[H_\nu^{11}]^{-1}\omega_{S_1}\}. \tag{53}
$$

The ν-Incoherence Condition can be now formalized from the right hand side of Equation (53). Below, we are going to give the proof of Proposition 6.

Proof of Proposition 6.

From Equation (53) and ν-Incoherence Condition, there holds

$$
\begin{aligned}
\|\widehat{\rho}_{S_0}\|_\infty \leq & (1 - \chi_\nu) + \frac{1}{\lambda\nu}[\|\omega_{S_0}\|_\infty + (1 - \chi_\nu)\|\omega_{S_1}\|_\infty], \\
\leq & (1 - \chi_\nu) + \frac{2}{\lambda}\left\|\frac{\omega}{\nu}\right\|_\infty.
\end{aligned}
$$

By definition $\frac{\omega}{\nu} = D\left[\nu\frac{X^TX}{n} + D^TD\right]^{-1}\frac{X^T}{n}\varepsilon$, therefore

$$
\mathbb{P}\left(\left\|\frac{2\omega}{\lambda\nu}\right\|_\infty \geq \frac{\chi_\nu}{2}\right) \leq 2m\exp\left(-\frac{n}{2c\sigma^2}\frac{\lambda^2\chi_\nu^2}{16}\right),
$$

for some constant $c > 0$ related with D and X. Take $\lambda = \lambda_n > \frac{8}{\chi_\nu}\sqrt{\frac{c\sigma^2 \log m}{n}}$, there holds

$$\mathbb{P}\left(\|\widehat{\rho}_{S_0}\|_\infty > 1 - \frac{\chi_\nu}{2}\right) \leq 2\exp\left(-c_1 n\lambda_n^2\right). \tag{54}$$

for some constant $c_1 > 0$.

Take $\lambda = \lambda_n$ in Equation (52) and consider the infinity norm on both sides, there holds

$$\|\widehat{\gamma}_{S_1} - \gamma^*_{S_1}\|_\infty \leq \lambda_n \nu \|[H^{11}_\nu]^{-1}\|_\infty + \|[H^{11}_\nu]^{-1}\omega_{S_1}\|_\infty.$$

Note that the first term in the right hand side is a deterministic term, therefore we can only estimate the second term. By definition $\frac{\omega}{\nu} = D\left[\nu\frac{X^TX}{n} + D^TD\right]^{-1}\frac{X^T}{n}\varepsilon$, similarly like above, there holds

$$\mathbb{P}(\|[H^{11}_\nu]^{-1}\omega_{S_1}\|_\infty > \nu t) \leq 2|S_1|\exp\left(-\frac{n}{2c\sigma^2}t^2 C^2_{\min}\right).$$

for some constant $c > 0$. Take $t = \frac{\sigma\lambda_n}{2C_{\min}}$, there holds

$$\mathbb{P}(\|[H^{11}_\nu]^{-1}\omega_{S_1}\|_\infty > \lambda_n \nu \frac{\sigma}{2C_{\min}}) \leq 2\exp\left(-c_2 n\lambda_n^2\right),$$

for some constant $c_2 > 0$. After all, there holds

$$\|\widehat{\gamma}_{S_1} - \gamma^*_{S_1}\|_\infty \leq \lambda_n \nu \left[\frac{\sigma}{2C_{\min}} + \|[H^{11}_\nu]^{-1}\|_\infty\right],$$

with probability greater than $1 - 2\exp(-c_2\lambda_n^2 n)$.

Take $C = \max\{8\sqrt{c}, c_1, c_2\}$. Then from Equation (54), for $\lambda = \lambda_n > \frac{C}{\chi_\nu}\sqrt{\frac{\sigma^2 \log m}{n}}$, $\widehat{\gamma}$ will not have false discoveries with probability greater than $1 - 4\exp(-Cn\lambda_n^2)$. If there further holds $\min_{i\in S_1}\gamma^*_i > \lambda_n \nu\left[\frac{\sigma}{2C_{\min}} + \|[H^{11}_\nu]^{-1}\|_\infty\right]$, $\widehat{\gamma}$ recovers the support set of γ^*. $\qquad\square$

Remark. The influence of ν on the power can be understood from this theorem as follows: (a) for the early stage of the Split LASSO path characterized by $\lambda_n > \frac{C}{\chi_\nu}\sqrt{\frac{\sigma^2 \log m}{n}}$., there is no false positive and only nonnull features are selected here; (b) all the strong nonnull features whose magnitudes are larger than $O(\nu\sigma\chi_\nu^{-1}\sqrt{\log m/n})$ could be selected on the path with sign consistency. Hence a sufficiently large ν will ensure the incoherence condition for path consistency such that strong nonnull features will be selected earlier on the Split LASSO path than the nulls, at the cost of possibly losing weak nonnull features below $O(\nu\sigma\chi_\nu^{-1}\sqrt{\log m/n})$. Therefore, a good power must rely on a proper choice of ν for the trade-off.

Remark. There is a close relationship between Proposition 6 in this paper and the model selection consistency in [HSXY20]. First of all, both of them aim to address the regression with transformational sparsity problem, in which the most popular method is the generalized LASSO (3). However, the model selection consistency of generalized LASSO suffers from the incoherence condition which often fails in applications. To alleviate this issue, the variable splitting idea is adopted in both approaches, by relaxing the linear constraint to an Euclidean neighborhood controlled with a proximity parameter $\nu > 0$. In this paper, the Split LASSO in (4) adopts a new objective function to optimize; while [HSXY20] adopts a dynamical system called the inverse scale space which does not optimize an objective function but renders an iterative regularization path in

its discrete form. In terms of model selection consistency, Proposition 6 shows that Split LASSO achieves sign consistency under strong signal assumption and a family of ν-incoherence conditions which is weaker than that for generalized LASSO as ν increases, easier to meet in applications. Such a theoretical result is similar to the model selection consistency in [HSXY20] when a proper early stopping regularization is chosen instead. It is shown in [HSXY20] that on such iterative regularization paths, there exists an estimator with model selection consistency, provided that the signal is strong enough and the same family of ν-incoherence conditions (Equation (2.5), Page 12 in [HSXY20]) are satisfied.

However, the incoherence conditions above are unknown in practice since the true support set is unknown, hence they can not be used for data adaptive model selection. The departure of our current paper aims to develop a data adaptive model selection method towards FDR control. Here, Theorem 1 and its high dimensional extension Theorem 2, show that the FDR of Split Knockoffs can be uniformly controlled for all parameter $\nu > 0$.

On the other hand, the selection power of Split Knockoffs, is ν-dependent, which can be observed from the simulation experiments in Section 6, where the selection power of Split Knockoffs undergoes a first increase then decrease as ν grows. Such a phenomenon is possibly explained by Proposition 6. In Proposition 6, there are two folds of influence of ν on model selection consistency of Split LASSO. On the one hand, the ν-incoherence condition of Split LASSO — which is critical for discovering strong signals — will be easier to be satisfied with the increase of ν. On the other hand, the requirement on the signal noise ratio (27) in achieving the sign consistency becomes harder to satisfy and weak signals might get lost with the increase of ν. As a consequence, the selection power of Split Knockoffs shows a first increasing then decreasing trend, in the simulation experiments in Section 6. Therefore, a good choice of ν in terms of power will depend on such a trade-off. This suggests us to use cross-validation to select a good $\widehat{\nu}$ to reach a good selection power with a desired FDR control.

C   Supplementary Material on Simulation Experiments

In this section, we will provide various supplementary material on simulation experiments for Split Knockoffs. In particular, we will present the simulation experiments of Split Knockoffs in the cases where m is close to n and in high dimensional settings. Then, we will discuss the choice of the data splitting fraction in Split Knockoffs, as well as the robustness of the random data splitting. After that, simulation experiments are implemented to compare the performance of Split Knockoffs and Knockoffs when the signal strength varies. In the end, we will discuss the computational cost of Split Knockoffs.

C.1   Simulation Experiments where m is close to n

In this section, we conduct simulation experiments in a setting where m is closer to n compared with that of Section 6.1. In particular, we consider the following setting, whose choices of m, n and p are close to those in the connection selection of Alzheimer's Disease in Section 7.3 ($m_{\mathrm{AD}} = 463$, $n_{\mathrm{AD}} = 752$, $p_{\mathrm{AD}} = 90$).

In model (1), we generate $X \in \mathbb{R}^{n \times p}$ ($n = 750$ and $p = 100$) i.i.d. from $\mathcal{N}(0_p, \Sigma)$, where $\Sigma_{i,i} = 1$ and $\Sigma_{i,j} = c^{|i-j|}$ for $i \neq j$, with feature correlation $c = 0.5$. Define $\beta^* \in \mathbb{R}^p$ by

$$\beta_i^* := \begin{cases} 1 & i \leq 20, \ i \equiv 0, -1 (\mathrm{mod}\ 3), \\ 0 & \text{otherwise.} \end{cases}$$

Then n linear measurements are generated by

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

where $\varepsilon \in \mathbb{R}^n$ is generated i.i.d. from $\mathcal{N}(0,1)$.

The linear transformation $D \in \mathbb{R}^{m \times p}$ is specified in the following way. Consider the graph $G = (V, E)$, where V denotes the vertex set $V = \{1, 2, \cdots, p\}$, and E denotes the edge set. There is an edge connecting two vertices $i \neq j$ if and only if

$$|i - j| \leq 5 (\mathrm{mod}\ p).$$

Through this construction, each vertex in V is connected to 10 neighbouring vertices. Therefore, $|E| = 10 * p/2 = 500$. Take D to be the graph difference operator on G, then $m = 500 > p = 100$. Then we generate $\gamma^* = D\beta^*$.

For Split Knockoffs, we take $\widehat{\beta}(\lambda)$ as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu}, \widehat{\lambda}}$ in the Split LASSO path with dataset $\mathcal{D}_1 = (X_1, y_1)$. The dataset $\mathcal{D} = (X, y)$ is randomly split into two parts $\mathcal{D}_1 = (X_1, y_1)$ and $\mathcal{D}_2 = (X_2, y_2)$ with n_1 and n_2 samples respectively, where $n_1 = 150$ and $n_2 = n - n_1 = 600 = m + p$. The performance of Split Knockoffs is presented in Figure 4.

As presented in Figure 4, the performance of Split Knockoffs follows the same trend as that of the other simulation experiments in Section 6. The FDR is controlled universally for W^{S}, $W^{\mathrm{S}\tau}$ and W^{BC}, where $W^{\mathrm{S}\tau}$ exhibits the least conservative FDR control as suggested by Proposition 2. Meanwhile, the selection power of Split Knockoffs presents a first increase then decrease trend as suggested by Proposition 6, where $W^{\mathrm{S}\tau}$ exhibits the best selection power, as predicted by Proposition 1.

C.2 Simulation Experiments for Split Knockoffs in High Dimensional Settings

In this section, we will show the results of the simulation experiments on Split Knockoffs in the high dimensional setting in a similar setting as in Section 6.1. In model (1), we generate $X \in \mathbb{R}^{n \times p}$ ($n = 400$ and $p = 1000$) i.i.d. from $\mathcal{N}(0_p, \Sigma)$, where $\Sigma_{i,i} = 1$ and $\Sigma_{i,j} = c^{|i-j|}$ for $i \neq j$, with feature correlation $c = 0.5$. Define $\beta^* \in \mathbb{R}^p$ by

$$\beta_i^* := \begin{cases} 1 & i \leq 20,\ i \equiv 0, -1 (\mathrm{mod}\ 3), \\ 0 & \text{otherwise}, \end{cases}$$

in the same way as in Section 6.1. Then we generate n linear measurements, $y = X\beta^* + \varepsilon$, where $\varepsilon \in \mathbb{R}^n$ is generated i.i.d. from $\mathcal{N}(0,1)$. For transformational sparsity, we should specify the linear transformation D such that $\gamma^* = D\beta^*$, where γ^* is sparse. We choose three types of transformation D_1, D_2, D_3 in the same way as in Section 6.1.

In simulation experiments, we use `glmnet` package [FHT10, SFHT11] to compute regularization paths for Split LASSO, etc. For the data splitting, we randomly split the dataset $\mathcal{D} = (X, y)$ into two parts $\mathcal{D}_1 = (X_1, y_1)$ and $\mathcal{D}_2 = (X_2, y_2)$ with n_1 and n_2 samples respectively, where $n_1 = 100$ and $n_2 = 300$.

The estimated support sets $\widehat{S}_\beta$, $\widehat{S}_\gamma$ in this experiment are generated from $\mathcal{D}_1$ in the following two steps. First, to estimate a super support set $\widehat{S}_\beta$, we perform the standard LASSO on a proper $\lambda > 0$ (chosen from cross validation in the experiments),

$$\min_\beta \frac{1}{2n} \|y_1 - X_1\beta\|_2^2 + \lambda \|\beta\|_1.$$

44

Second, to estimate support set $\widehat{S}_\gamma$, we perform the Split LASSO on the reduced model with respect to a proper $\lambda > 0$ (chosen from cross validation in the experiments),

$$\min_{\beta,\gamma} \frac{1}{2n} \left\| y_1 - X_{\widehat{S}_\beta}\beta \right\|_2^2 + \frac{1}{2\nu}\|D\beta - \gamma\|_2^2 + \lambda\|\gamma\|_1,$$

where $X_{\widehat{S}_\beta}$ is the submatrix of X_1 consisting of the columns in $\widehat{S}_\beta$.

With the estimated support set $\widehat{S}_\beta$ and $\widehat{S}_\gamma$, we generate the intercept $\widehat{\beta}(\lambda) = \widehat{\beta}_{\widehat{\lambda},\widehat{\nu}}$ as an optimal estimator with minimal cross validation loss (with respect to λ and ν) on the following Split LASSO regularization path

$$\min_{\beta,\gamma} \frac{1}{2n} \left\| y_1 - X_{\widehat{S}_\beta}\beta \right\|_2^2 + \frac{1}{2\nu}\|D_{\widehat{S}_\beta,\widehat{S}_\gamma}\beta - \gamma\|_2^2 + \lambda\|\gamma\|_1, \quad \text{for } \lambda > 0,$$

where $D_{\widehat{S}_\beta,\widehat{S}_\gamma}$ is the submatrix of D, consisting of the columns in $\widehat{S}_\beta$ and rows in $\widehat{S}_\gamma$. With these choices of $\widehat{S}_\beta$, $\widehat{S}_\gamma$ and $\widehat{\beta}(\lambda)$, we proceed with the rest steps of Split Knockoffs following the instructions in Section 4.

As shown in Figure 5, all three versions of the Split Knockoffs achieve desired performance in FDR control in the high dimensional settings. It is worth to mention that similar with the performance of Split Knockoffs in the case $n \geq m + p$ in Section 6, Split Knockoffs with W^{BC} statistics exhibits stricter FDR control, at the cost of the selection power — especially when ν is large — compared with W^{S} and $W^{\text{S}\tau}$, while $W^{\text{S}\tau}$ exhibits the highest selection power with the most adaptive FDR control with respect to the target. Such an observation is explained by Proposition 1 that W^{BC} is the most conservative statistics, while $W^{\text{S}\tau}$ is the most aggressive statistics in terms of feature selection.

Moreover, we show in Table 4 that, for the calculation of the feature and knockoff significance, the cross validation optimal choice of ν — $\widehat{\nu}$ — can still achieve high power and desired FDR in Split Knockoffs in high dimensional settings. Specifically, we present the performance of Split Knockoffs with all three versions of W statistics, under the above choice of $\widehat{\beta}(\lambda)$ being taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$ on the Split LASSO paths, while ν is chosen to be $\widehat{\nu}$. For such a choice, W^{BC} operates slightly stricter FDR control at the cost of slightly lower selection power compared with W^{S} and $W^{\text{S}\tau}$.

C.3 Simulation Experiments on Data Splitting Fraction

In this section, we will use simulation experiments to study how the data splitting fraction can affect the performance of Split Knockoffs, and give an practical guideline on how to choose the data splitting fraction. We succeed all the simulation settings on the dataset (X, y) and transformational sparsity D_1, D_2, D_3 from Section 6.1, and take the regression parameter $\beta^* \in \mathbb{R}^p$ as

$$\beta_i^* := \begin{cases} 0.5 & i \leq 20, \ i \equiv 0, -1 (\text{mod } 3), \\ 0 & \text{otherwise.} \end{cases} \tag{55}$$

The signal strength used in this section was lowered compared with that of Section 6.1 in order to increase the variation and amplify the effects of data splitting fractions on the performance of Split Knockoffs. Then we conduct simulation experiments with respect to different data splitting fractions in the range from 0.1 to 0.8 with a step size 0.1. The performance of Split Knockoffs in the FDR and selection power with respect to different data splitting fractions is shown in Figure 6.

Table 4. FDR and Power for Split Knockoffs in high dimensional settings ($q = 0.2$). The intercept $\widehat{\beta}(\lambda)$ for Split Knockoffs is taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$, with the ν for calculating the feature and knockoff significance taken as $\widehat{\nu}$. In this table, we present the average performance of Split Knockoffs in FDR and Power, together with the standard deviations in 200 simulation instances. For shorthand notations, we use "SK(+)" to refer to "Split Knockoff(+)".

Performance	SK with W^{S}	SK with $W^{\mathrm{S}\tau}$	SK with W^{BC}	SK+ with W^{S}	SK+ with $W^{\mathrm{S}\tau}$	SK+ with W^{BC}
FDR in D_1	0.1003	0.1003	0.0945	0.0900	0.0900	0.0780
	±0.0874	±0.0874	±0.0849	±0.0840	±0.0840	±0.0807
Power in D_1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	±0.0000	±0.0000	±0.0000	±0.0000	±0.0000	±0.0000
FDR in D_2	0.2015	0.2015	0.1770	0.1683	0.1683	0.1414
	±0.1084	±0.1084	±0.1141	±0.1095	±0.1095	±0.1127
Power in D_2	0.9975	0.9975	0.9964	0.9961	0.9961	0.9939
	±0.0150	±0.0150	±0.0186	±0.0192	±0.0192	±0.0224
FDR in D_3	0.1749	0.1749	0.1637	0.1642	0.1642	0.1482
	±0.0838	±0.0838	±0.0866	±0.0860	±0.0860	±0.0876
Power in D_3	0.9991	0.9991	0.9991	0.9985	0.9985	0.9985
	±0.0069	±0.0069	±0.0069	±0.0082	±0.0082	±0.0082

In particular, since we extend Split Knockoffs to high dimensional settings in Section 4, our method is still applicable for the case where the data splitting fraction of $\mathcal{D}_1$ is large and $n_2 \geq m + p$ is dissatisfied. As one can observe in Figure 6, the FDR of Split Knockoffs is not sensitive to the data splitting fraction — in all cases, Split Knockoffs achieves desired FDR despite its potential theoretical inflation when the data splitting fraction of $\mathcal{D}_1$ is large. The main effect of the data splitting fraction lies on the power of Split Knockoffs. In Figure 6, the optimal selection power is commonly achieved when the data splitting fraction is balanced for $\mathcal{D}_1$ and $\mathcal{D}_2$. Therefore, a practical guideline for selecting the data splitting fraction should be: make the data splitting fraction balanced when the sample size permits, and favor the sample size of $\mathcal{D}_1$ a little bit for better selection power when the sample size is limited.

C.4 Simulation Experiments on Stability of Data Splitting

In order to handle the transformational sparsity, Split Knockoffs introduce a random data splitting scheme in Section 2. Consequently, the random data splitting will lead to the random estimated support sets. In this section, we will use simulation experiments to study the robustness on features in the estimated support sets selected by Split Knockoffs in random data splitting. We succeed all the simulation settings on the dataset (X, y) and transformational sparsity D_1, D_2, D_3 from Section 6.1. In the experiment, we randomly generate the dataset (X, y) for 20 times. For each generation, we randomly perform 100 different data splits and conduct Split Knockoffs with each data split. The estimated support sets of Split Knockoffs are recorded for the 20×100 tests.

In Figure 7, we present the selection frequencies of the features selected by Split Knockoffs in the 20×100 tests. For Split Knockoffs conducted in this section, we take $\widehat{\beta}(\lambda)$ as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$, and take the parameter ν for calculation of the feature and knockoff significance as $\widehat{\nu}$.

As shown in Figure 7, in all cases, the nonnull features will have relatively stable and high frequencies to be selected, while the null features occur randomly and have much lower selection frequencies. Such a result suggests that the nonnull features can be robustly selected in the selection set of Split Knockoffs with respect to random data splitting.

C.5 Simulation Experiments on the Signal Strength

In this section, we implement simulation experiments to compare the performance of Split Knockoffs and Knockoffs when the signal strength varies. We succeed all simulation settings in Section 6.1, except that we take $\beta^* \in \mathbb{R}^p$ as

$$\beta_i^* := \begin{cases} A & i \leq 20, \ i \equiv 0, -1 (\mathrm{mod} \ 3), \\ 0 & \text{otherwise}, \end{cases}$$

where $\log(A)$ varies from -0.5 to 0.5. The performance comparisons between Split Knockoffs and Knockoffs when the signal strength varies are presented in Figure 8.

As presented in Figure 8, both Split Knockoffs and Knockoffs achieve desired FDR control when applicable.

As for the selection power, in the case that the transformation is the identity matrix (D_1), Split Knockoffs exhibit lower selection power compared with Knockoffs when the signal strength is weak. In this case, the potential power improvement from the weaker ν-incoherence condition in Proposition 6 is overwhelmed by the dissatisfaction of the minimal signal strength condition (27). This leads to a loss in the selection power of Split Knockoffs.

On the other hand, when the linear transformation is nontrivial such as D_2, the power improvement brought by the ν-incoherence condition in Proposition 6 dominants the minimal signal strength condition (27). This enables Split Knockoffs to achieve higher selection power compared with Knockoffs.

C.6 Analysis on Computational Cost

In this section, we will give an analysis on the computational cost of Split Knockoffs. Among the procedures of Split Knockoffs, the computation of the two regularization paths in Equation (8) and (10) takes the most computational time. In particular, the computational cost on regularization paths varies for different choices of $\widehat{\beta}(\lambda)$. In the case that $\widehat{\beta}(\lambda)$ is a fixed point, the computational cost will be twice the cost of the generalized LASSO in the `glmnet` package [FHT10, SFHT11]. However, in the case that $\widehat{\beta}(\lambda)$ is a solution path that changes with respect to λ, an additional cost of the generalized LASSO is spent; in this regard, the computational cost will be three times the cost of the generalized LASSO.

D Supplementary Material for Alzheimer's Disease Experiments

In this section, we provide more details regarding the experiments on Alzheimer's Disease. In particular, we provide supplementary material for experiments in Alzheimer's Disease in Section 7, where we plot the frequencies of the most frequently selected regions by Split Knockoffs in multiple data splits. To be specific, we show the selection results of Split Knockoff under 100 different random sample splits of the dataset in Alzheimer's Disease. We take the W statistics for Split Knockoff as $W^{S\tau}$, with $\widehat{\beta}(\lambda)$ being taken as a cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$, and ν being taken as $\widehat{\nu}$. Throughout the random splitting, important regions and connections should be selected with high frequencies, while the false discoveries in each selection should have relatively low frequencies to be selected. We plot the top 10 most frequently selected regions/connections by Split Knockoff in 100 random data splits in Figure 9 and Figure 10, with abbreviations of each region marked in the figures. A comparison table between the full region names and their abbreviations can be found in Table 5.

Table 5: Names and Abbreviations for Cerebrum Brain Anatomical Regions

Region Name	Abbreviation
Precental gyrus	PreCG
Superior frontal gyrus, dorsolateral	SFGdor
Superior frontal gyrus, orbital part	ORBsup
Middle frontal gyrus	MFG
Middle frontal gyrus, orbital part	ORBmid
Inferior frontal gyrus, opercular part	IFGoperc
Inferior frontal gyrus, triangular part	IFGtriang
Inferior frontal gyrus, orbital part	ORBinf
Rolandic operculum	ROL
Supplementary motor area	SMA
Olfactory cortex	OLF
Superior frontal gyrus, medial	SFGmed
Superior frontal gyrus, medial orbital	ORBsupmed
Gyrus rectus	REC
Insula	INS
Anterior cingulate and paracingulate gyri	ACG
Median cingulate and paracingulate gyri	MCG
Posterior cingulate gyrus	PCG
Hippocampus	HIP
Parahippocampal gyrus	PHG
Amygdala	AMYG
Calcarine fissure and surrounding cortex	CAL
Cuneus	CUN
Lingual gyrus	LING
Superior occipital gyrus	SOG
Middle occipital gyrus	MOG
Inferior occipital gyrus	IOG
Fusiform gyrus	FFG
Postcentral gyrus	PoCG
Superior parietal gyrus	SPG
Inferior parietal, but supramarginal and angular gyri	IPL
Supramarginal gyrus	SMG
Angular gyrus	ANG
Precuneus	PCUN
Paracentral lobule	PCL
Caudate nucleus	CAU
Lenticular nucleus putamen	PUT
Lenticular nucleus, pallidum	PAL
Thalamus	THA
Heschl gyrus	HES
Superior temporal gyrus	STG
Temporal pole: superior temporal gyrus	TPOsup
Middle temporal gyrus	MTG
Temporal pole: middle temporal gyrus	TPOmid
Inferior temporal gyrus	ITG
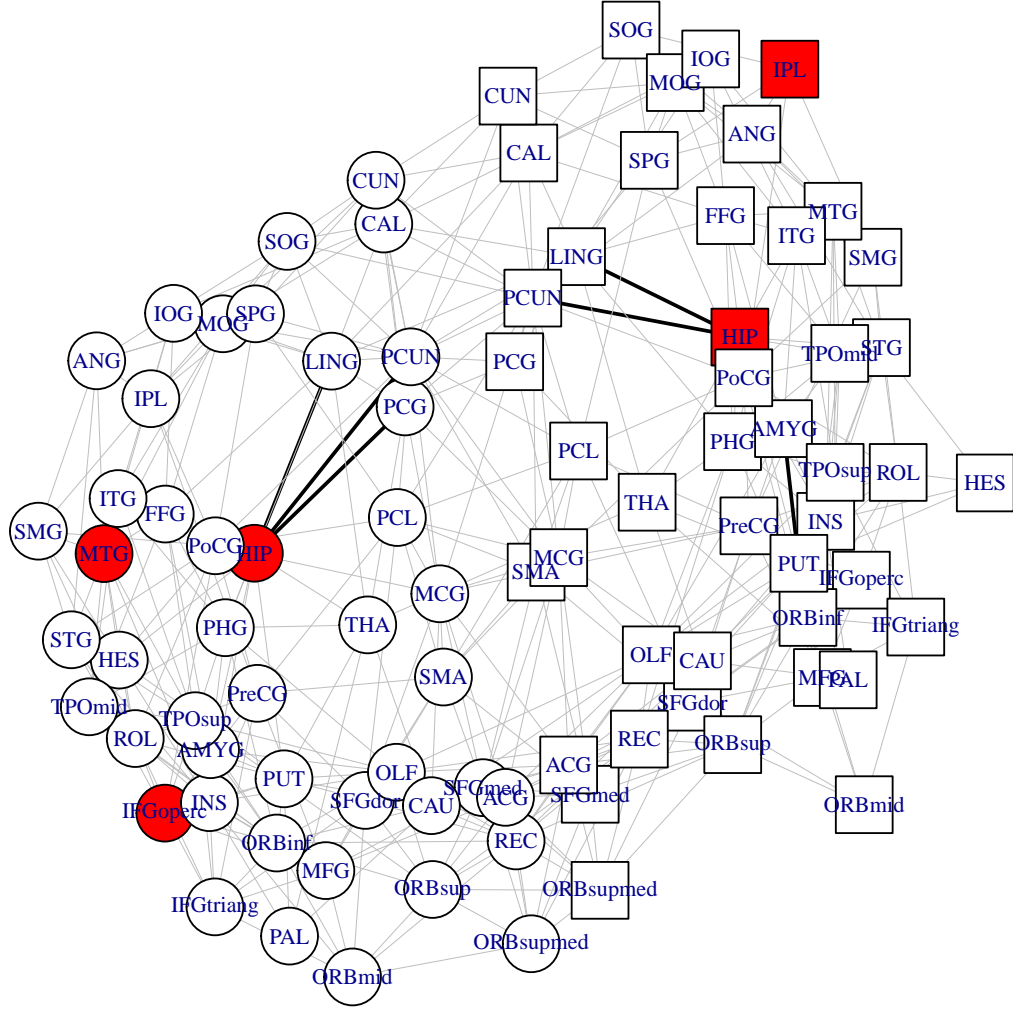
Figure 1. Selected brain regions and connections in Alzheimer's Disease. Each vertex represents a Cerebrum brain region in Automatic Anatomical Labeling (AAL) atlas [TMLP+02], whose abbreviations and full names are provided in Table 5. Here, vertices with a circle shape represent the left brain regions, while the ones with a square shape represent the right brain regions. An edge connects two vertices if and only if the associated two brain regions are adjacent. The method in this paper selects red vertices as significantly degenerate lesion regions and bold edges for highly differential connections.
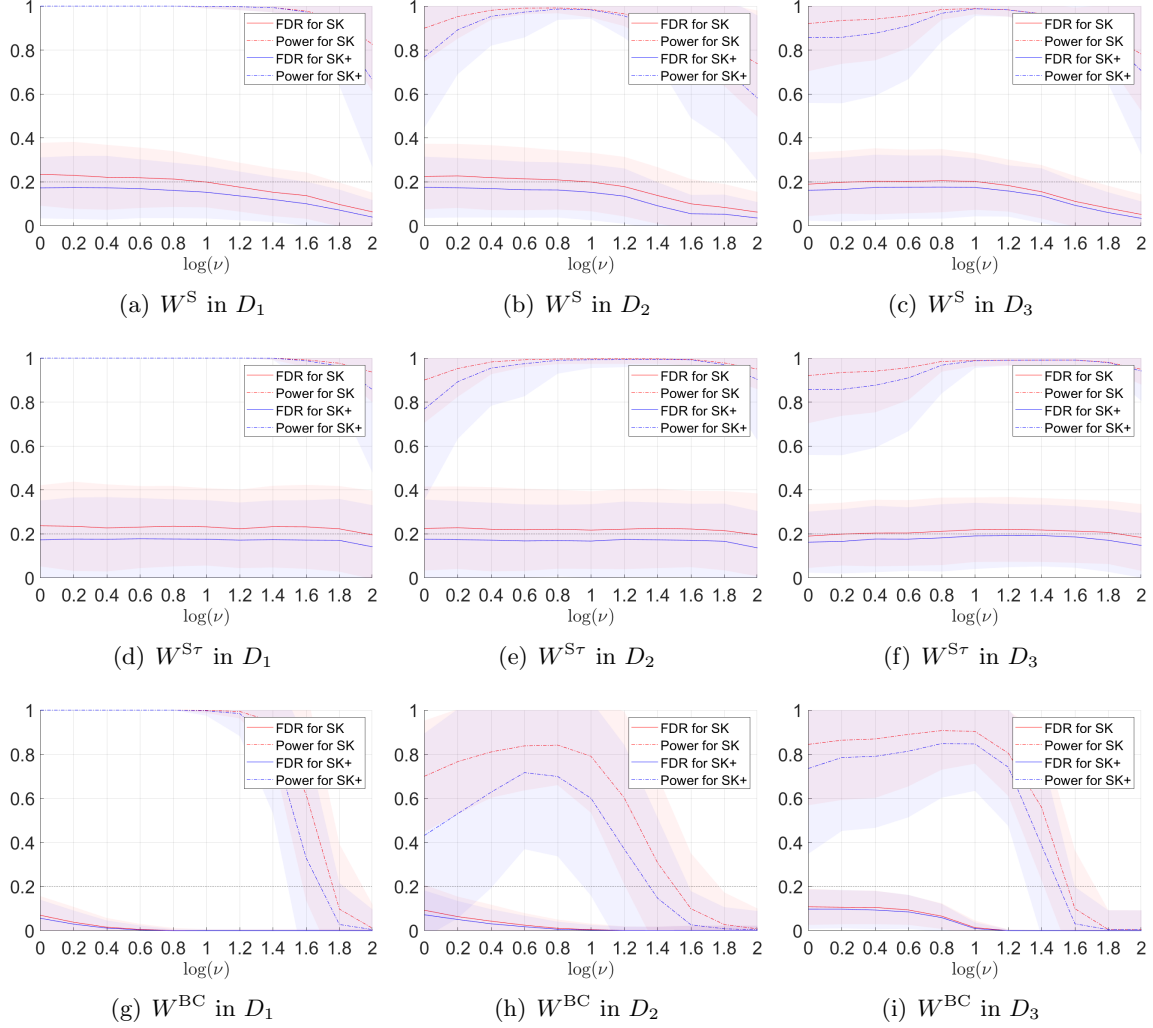
Figure 2. The performance of Split Knockoffs: FDR and Power for $q = 0.2$. $\widehat{\beta}(\lambda)$ is taken as $\widehat{\beta}_\nu(\lambda)$, the solution path of $\widehat{\beta}(\lambda)$ in the ν-Split LASSO path. The curves in the figures represent the average performance of Split Knockoffs in FDR and Power in 200 simulation instances, while the shaded areas represent the 80% confidence intervals truncated to the range $[0, 1]$.
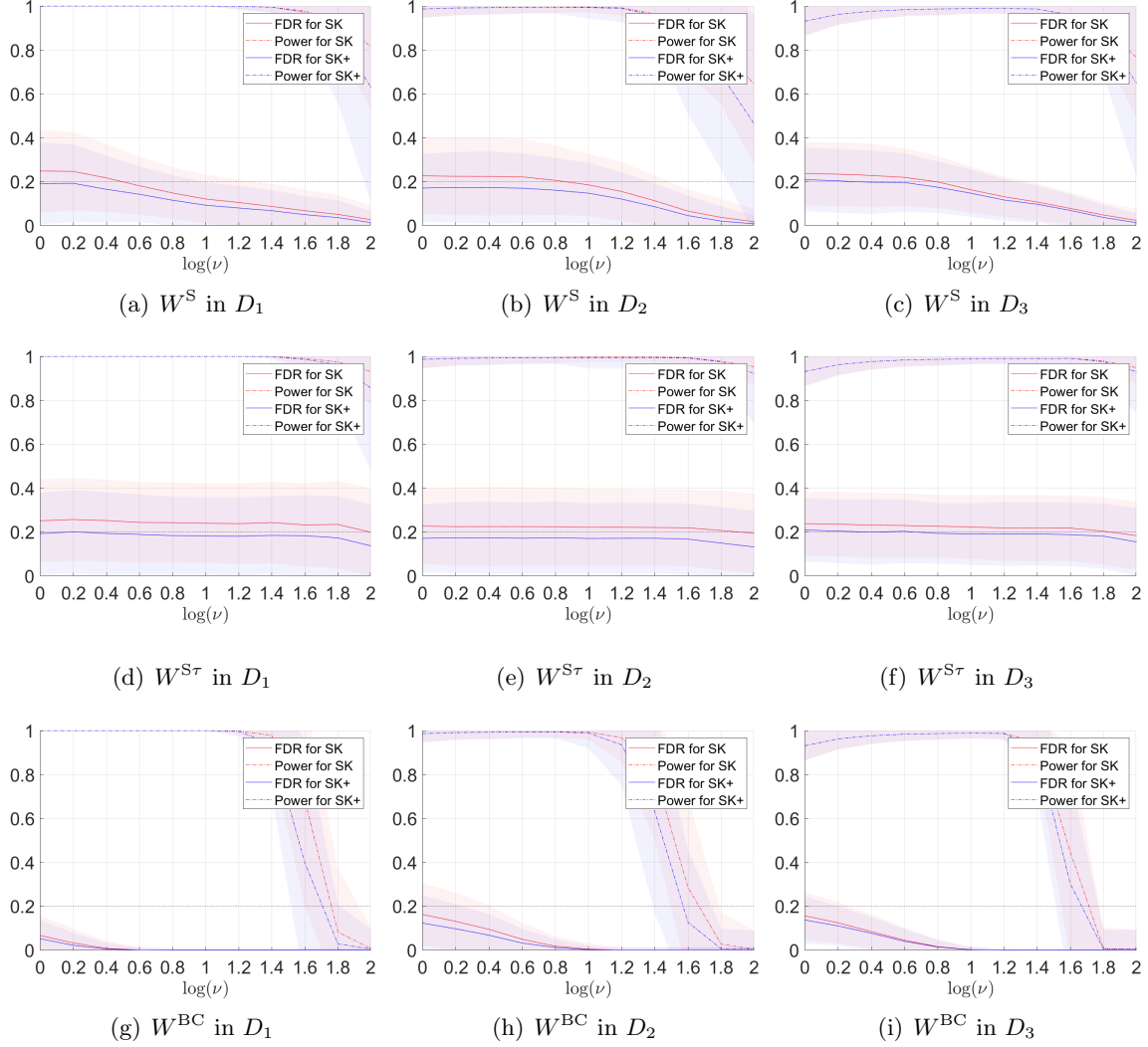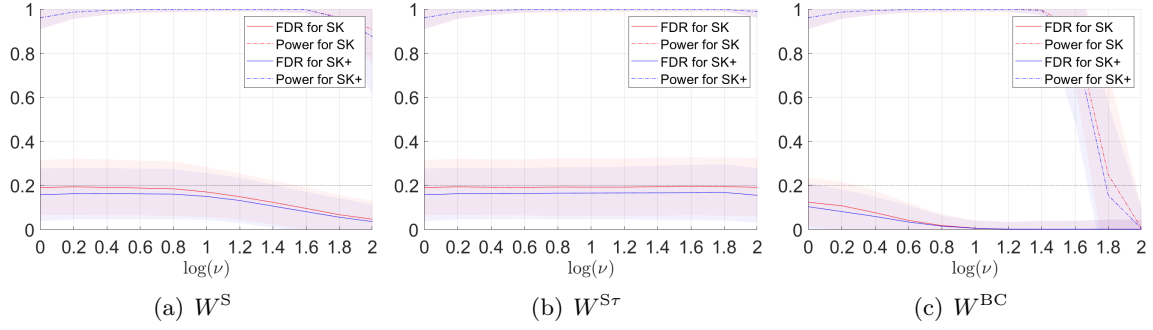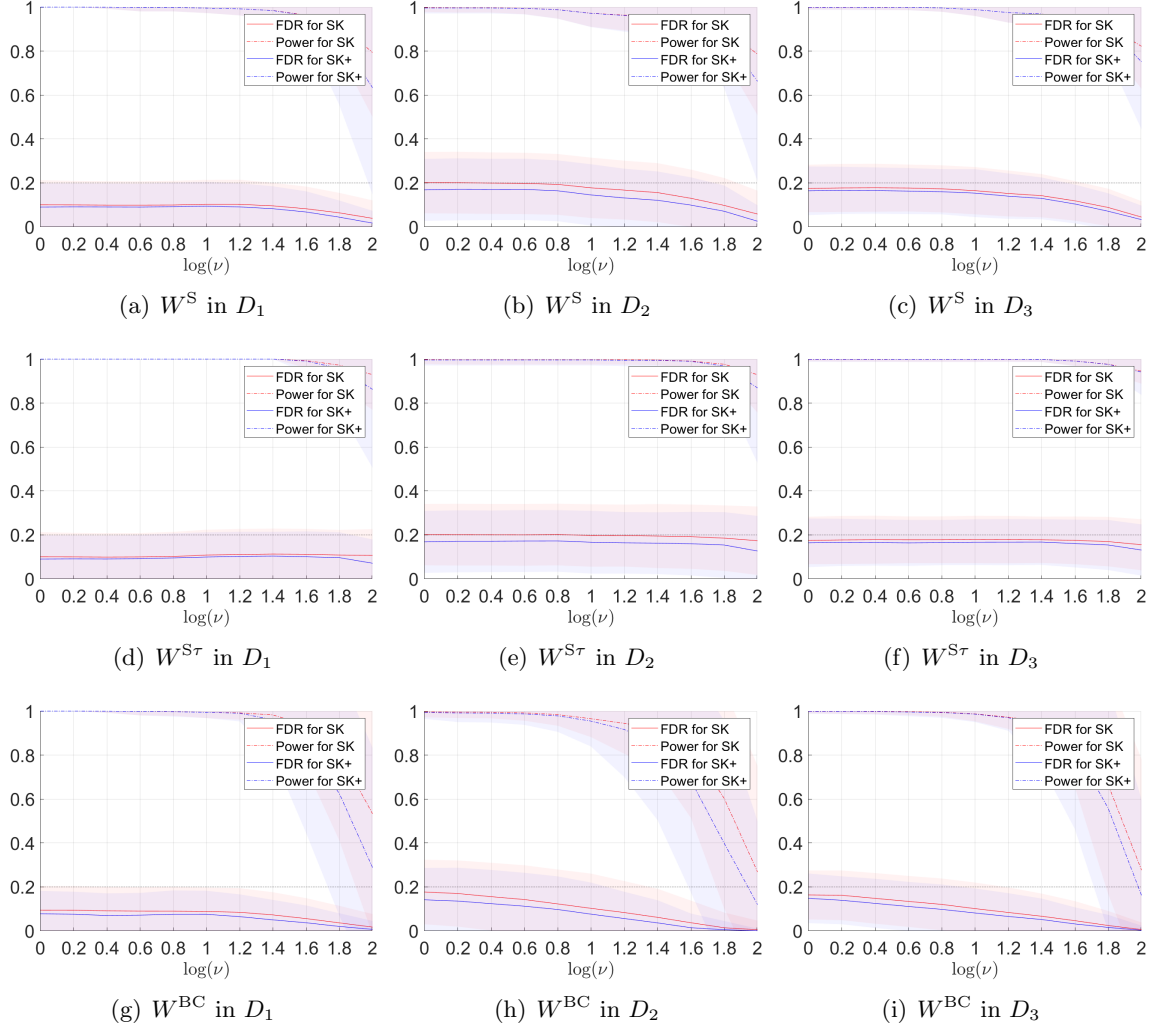
Figure 3. The performance of Split Knockoffs: FDR and Power for $q = 0.2$. $\widehat{\beta}(\lambda)$ is taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$. The curves in the figures represent the average performance of Split Knockoffs in FDR and Power in 200 simulation instances, while the shaded areas represent the 80% confidence intervals truncated to the range $[0, 1]$.
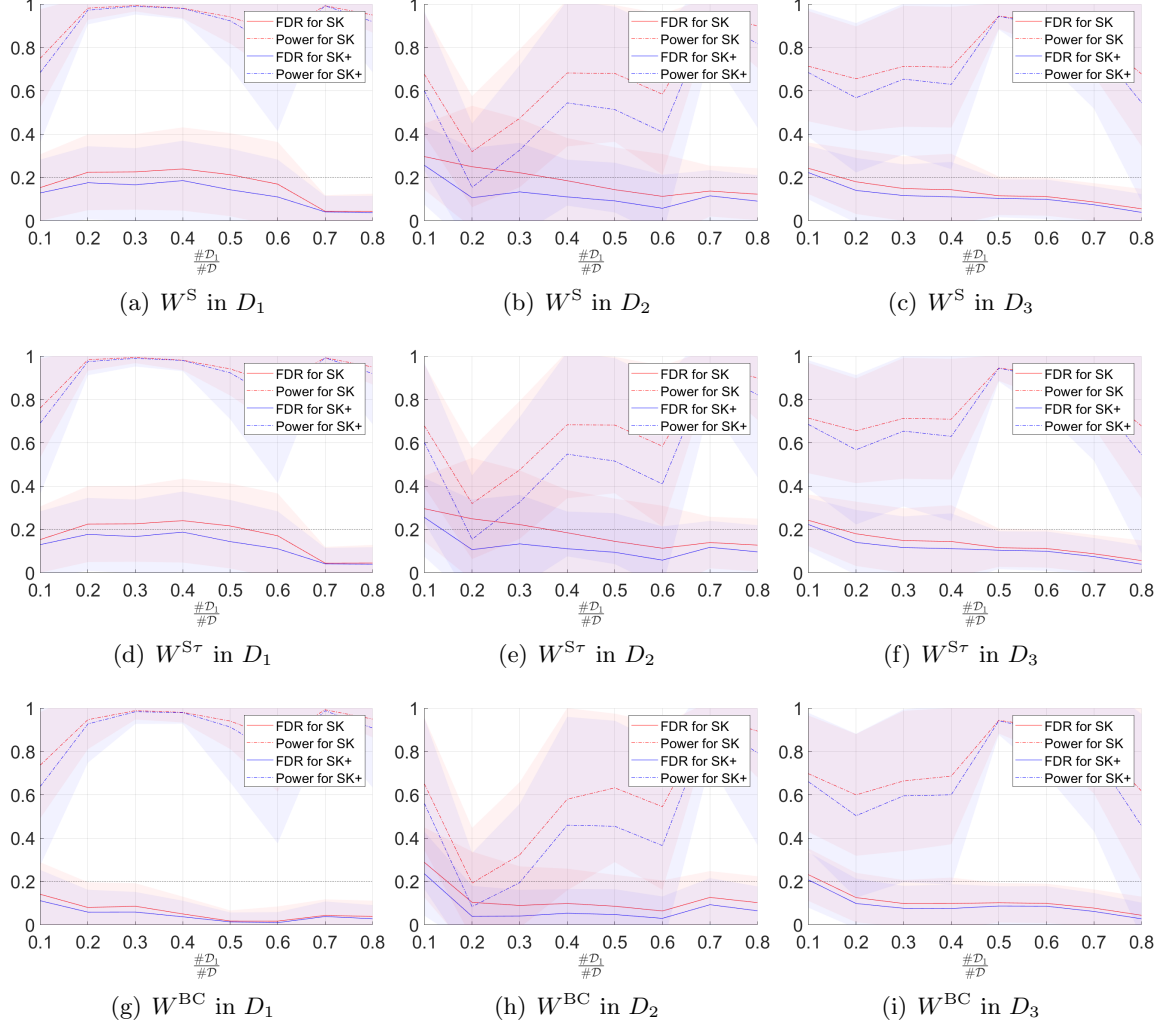
Figure 4. The performance of Split Knockoffs: FDR and Power for $q = 0.2$. $\widehat{\beta}(\lambda)$ is taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$. The curves in the figures represent the average performance of Split Knockoffs in FDR and Power in 200 simulation instances, while the shaded areas represent the 80% confidence intervals truncated to the range $[0, 1]$.

Figure 5. The performance of Split Knockoffs in high dimensional settings: FDR and Power for $q = 0.2$. $\widehat{\beta}(\lambda)$ is taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$. The curves in the figures represent the average performance of Split Knockoffs in FDR and Power in 200 simulation instances, while the shaded areas represent the 80% confidence intervals truncated to the range $[0, 1]$.

Figure 6. The performance of Split Knockoffs under different data splitting fractions: FDR and Power for $q = 0.2$. **The signal strength is taken to be 0.5 as in Equation** (55) **for this experiment instead of being 1 in Section 6.1.** The x-axis represents the percentage of samples in $\mathcal{D}_1$ among $\mathcal{D}$. The curves in the figures represent the average performance of Split Knockoffs in FDR and Power in 200 simulation instances, while the shaded areas represent the 80% confidence intervals truncated to the range $[0, 1]$. The intercept $\widehat{\beta}(\lambda)$ for Split Knockoffs is taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu}, \widehat{\lambda}}$, with the ν for calculating the feature and knockoff significance taken as $\widehat{\nu}$.
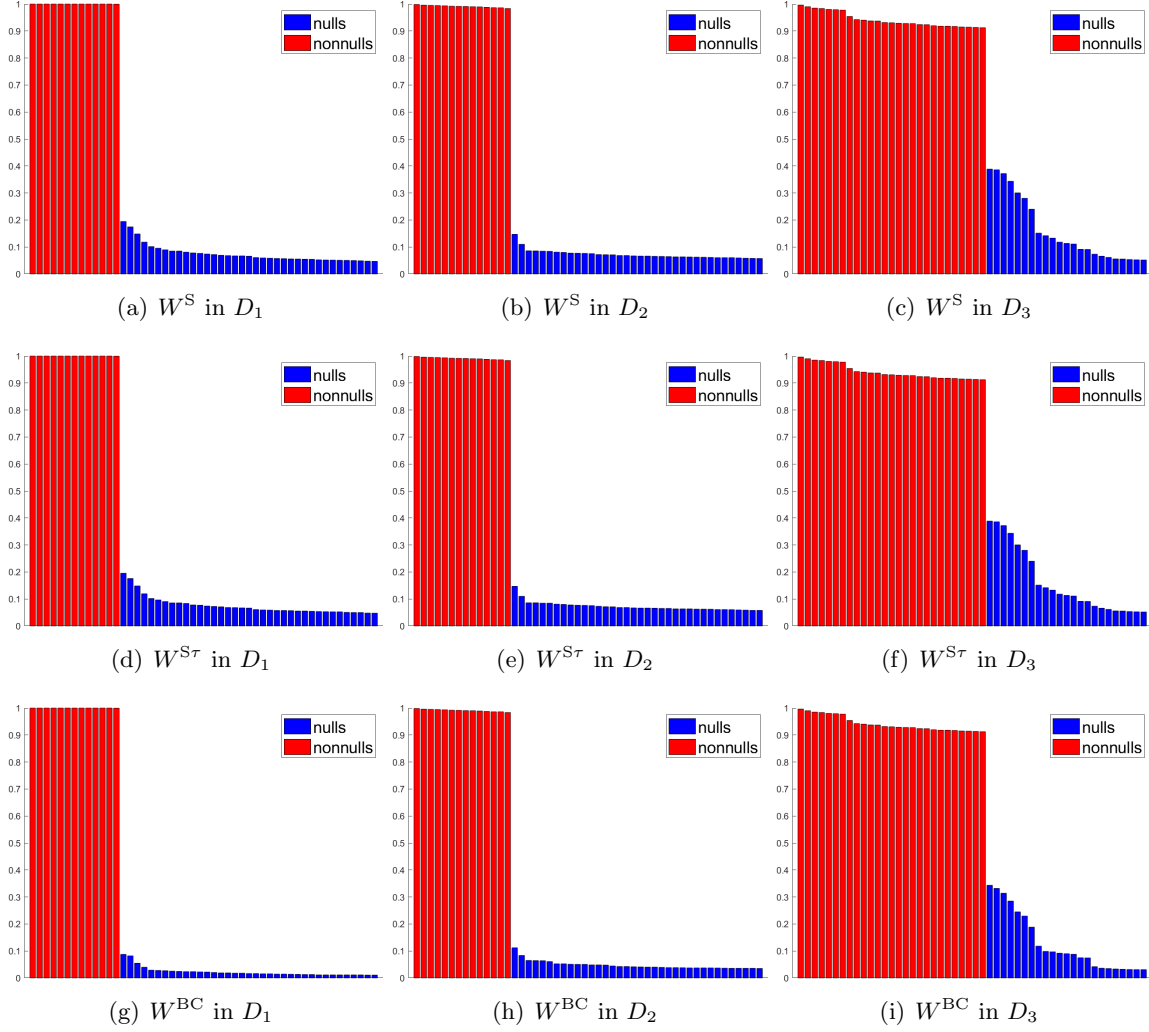
Figure 7. Top 50 most frequently selected features of Split Knockoffs in simulation experiments among 100 different data splits. Each bar in the figure represents the selection frequency of a particular feature. The intercept $\widehat{\beta}(\lambda)$ for Split Knockoffs is taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$, with the ν for calculating the feature and knockoff significance taken as $\widehat{\nu}$.
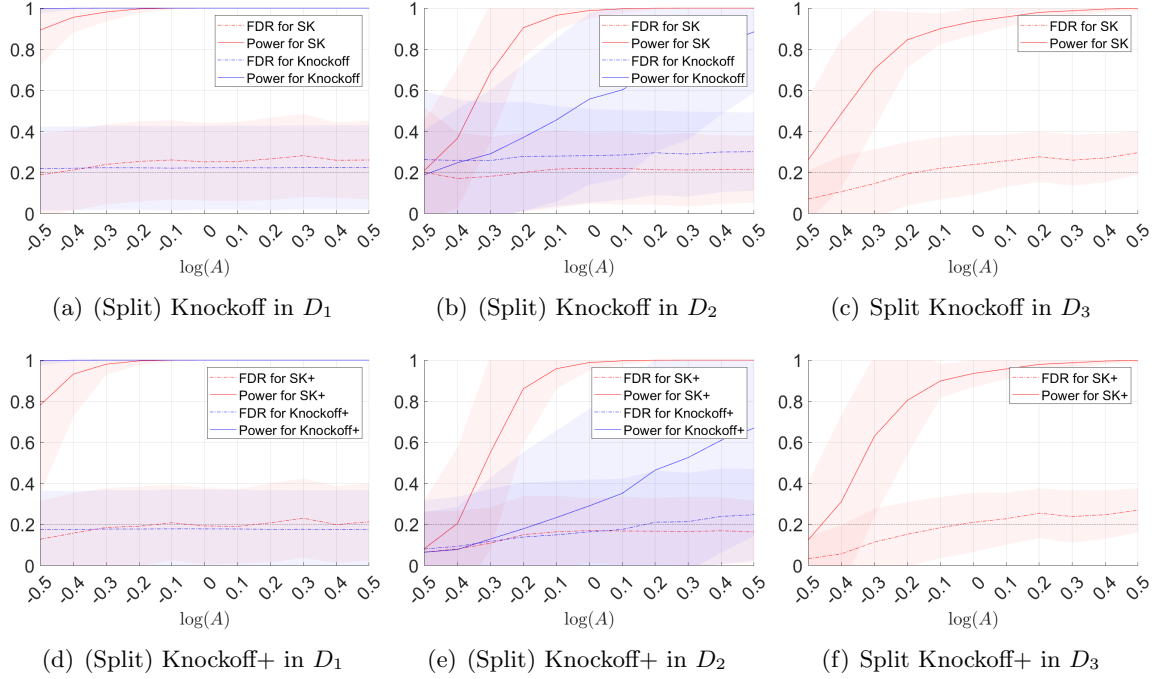
Figure 8. The performance of Split Knockoffs and Knockoffs under different signal noise ratios: FDR and Power for $q = 0.2$. The curves in the figures represent the average performance in FDR and Power in 200 simulation instances, while the shaded areas represent the 80% confidence intervals truncated to the range $[0, 1]$. The intercept $\widehat{\beta}(\lambda)$ for Split Knockoffs is taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu}, \widehat{\lambda}}$, with the ν for calculating the feature and knockoff significance taken as $\widehat{\nu}$. We take the W statistics of Split Knockoffs as $W^{S\tau}$ in this figure.
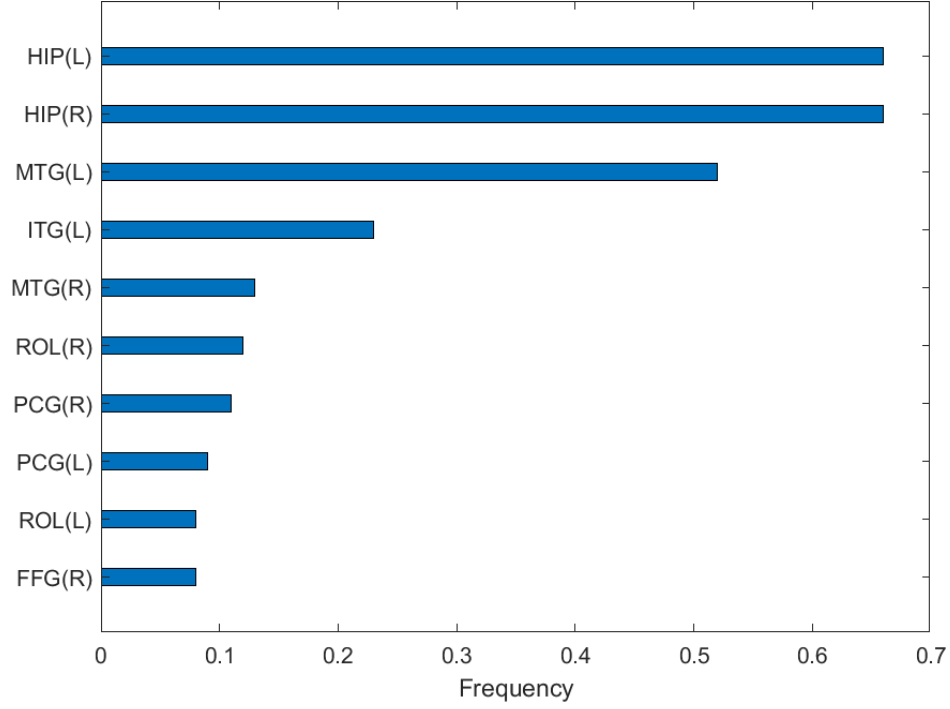
Figure 9. Top 10 most frequently selected lesion regions by Split Knockoff in 100 random data splits. We take the W statistics in Split Knockoff as $W^{S\tau}$, with $\widehat{\beta}(\lambda)$ being taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$, and ν being taken as $\widehat{\nu}$.
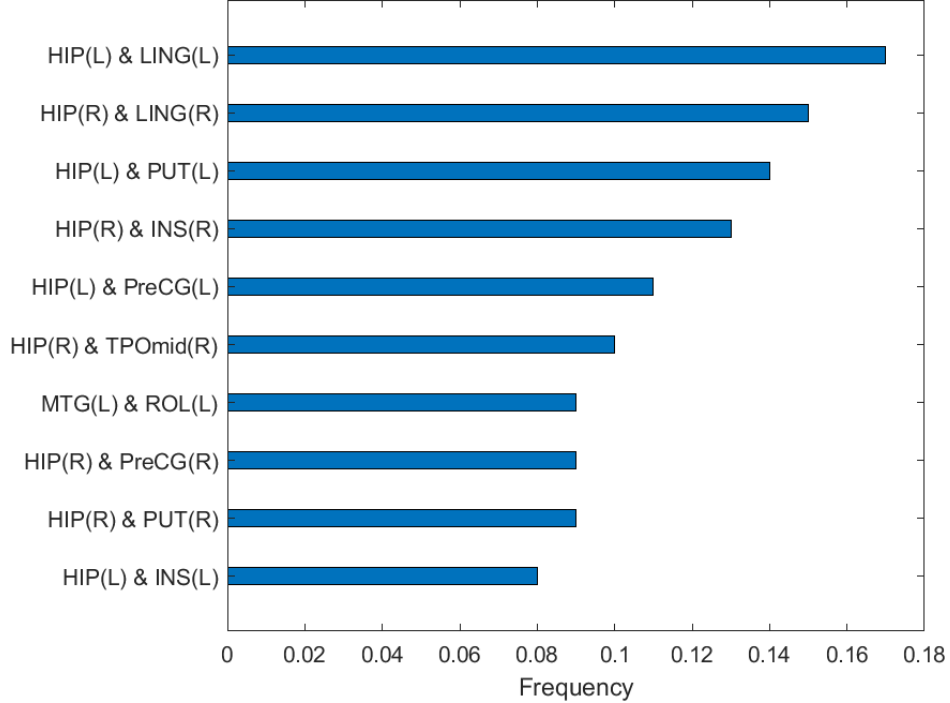

Figure 10. Top 10 most frequently selected connections of adjacent regions by Split Knockoff in 100 random data splits. We choose the W statistics for Split Knockoff as $W^{S\tau}$, with $\widehat{\beta}(\lambda)$ being taken as a fixed cross validation optimal estimator $\widehat{\beta}_{\widehat{\nu},\widehat{\lambda}}$, and ν being taken as $\widehat{\nu}$.