Provable FDR Control for Deep Feature Selection: Deep MLPs and Beyond

Kazuma Sawaya
The University of Tokyo

Abstract

We develop a flexible feature selection framework based on deep neural networks that approximately controls the false discovery rate (FDR), a measure of Type-I error. The method applies to architectures whose first layer is fully connected. From the second layer onward, it accommodates multilayer perceptrons (MLPs) of arbitrary width and depth, convolutional and recurrent networks, attention mechanisms, residual connections, and dropout. The procedure also accommodates stochastic gradient descent with data-independent initializations and learning rates. To the best of our knowledge, this is the first work to provide a theoretical guarantee of FDR control for feature selection within such a general deep learning setting.

Our analysis is built upon a multi-index data-generating model and an asymptotic regime in which the feature dimension n diverges faster than the latent dimension q^*, while the sample size, the number of training iterations, the network depth, and hidden layer widths are left unrestricted. Under this setting, we show that each coordinate of the gradient-based feature-importance vector admits a marginal normal approximation, thereby supporting the validity of asymptotic FDR control. As a theoretical limitation, we assume $\boldsymbol{B}$-right orthogonal invariance of the design matrix, and we discuss broader generalizations. We also present numerical experiments that underscore the theoretical findings.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

Feature selection is the task of identifying features that are truly relevant to the response $\boldsymbol{y}$. It plays a dual role in modern machine learning: it underpins scientific discovery, such as identifying genes associated with Alzheimer's disease, and it enhances the interpretability of predictive models. Two largely separate research threads have pursued this goal, namely *high-dimensional statistics* and *explainable AI (XAI)*.

On the statistics side, a central objective has been to provide theoretical guarantees for false discovery rate (FDR) control (Abramovich et al., 2006; Meinshausen and Bühlmann, 2006; Barber and Candès, 2015; Candès et al., 2018; Xing et al., 2023; Dai et al., 2023a,b; Du et al., 2023), often by debiasing sparse estimators such as the LASSO (Tibshirani, 1996). These guarantees, albeit rigorous, typically come at the expense of strong modeling assumptions (e.g., the true model follows a generalized linear model, or the feature distribution is known), which can diverge from complex real-world phenomena and thereby cast doubt on the reliability of the selected features.

On the XAI side, a wealth of attribution methods, e.g., LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), random forest feature importance (Breiman, 2001), and saliency maps (Simonyan et al., 2013), quantify the contribution of individual features in black-box models. Thresholding such scores yields a practical selection heuristic, yet without guarantees on Type-I error; thus, error control has remained elusive.

This paper bridges these threads. We propose a feature-selection procedure for deep neural networks that approximately controls the FDR while retaining the modeling flexibility of modern architectures. Our approach is built upon *input sensitivity* $\boldsymbol{\xi}^{(t)} \in \mathbb{R}^n$ defined by the gradient of the trained neural network's output with respect to input, and a simple data-splitting aggregation scheme. The analysis sheds light on when input sensitivity admits a normal approximation and how this leads to valid error control,

irrespective of architecture details. Our **main contributions** are:

- *Flexible scope across architectures.* We develop a feature-selection method applicable to multilayer perceptrons (MLPs) with arbitrary width and depth, as well as convolutional and recurrent networks, attention mechanisms, residual connections, and dropout.
- *Agnostic to the training protocol.* Our guarantees accommodate stochastic gradient descent with arbitrary, data-independent initialization schemes and learning rates.
- *Normal approximation of input sensitivity.* We show that input sensitivity $\boldsymbol{\xi}^{(t)}$ is asymptotically normal when the feature dimension n is sufficiently larger than the latent dimension q^*. This holds regardless of the sample size m and all the network parameters (e.g., width and depth). Also, this result holds at each training iteration t, enabling early stopping.
- *Asymptotic FDR control via sample splitting.* By aggregating input sensitivity across splits, our procedure achieves asymptotic FDR control with a simple and implementable pipeline.
- *General data-generating process.* The theory is established under a multi-index model as the data-generating process with unknown nonlinearity. This is a flexible framework that captures rich latent structures beyond generalized linear models.
- *New proof technique.* At the crux of the analysis is a technique relying on the recursive inheritance of orthogonal invariance of the input sensitivity, which may be of independent interest.
- *Empirical support.* Numerical experiments underscore that the normal approximation and FDR control hold under the stated conditions, aligning with the theory.

On the other hand, a major limitation of our framework lies in the assumption that the design matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ is $\boldsymbol{B}$-right orthogonally invariant (See Assumption 1 (ii) for the definition). While this assumption accommodates time dependence, heavy-tailed distributions, and low-rank structures, it does not allow us to account for specific forms of feature correlations (See Appendix B for details). To address this limitation, we discuss potential extensions to the general correlation structures in Appendix C.

It should be noted that FDR control represents only the minimal requirement of avoiding excessive inclusion of irrelevant variables in feature selection. With respect to the complementary criterion of Type-II error (the ability to correctly identify truly relevant features), our analysis, like much of the existing literature, provides no theoretical guarantees, although numerical evaluations are reported. For instance, a procedure that selects nothing trivially attains an FDR of zero, yet suffers a Type-II error of one. A comprehensive assessment of feature selection methods therefore requires attention to both criteria.

1.1 Related works

FDR control in feature selection. There are two main research streams of FDR control in feature selection: estimator-based FDR control and knockoff filters. In the former, e.g., for linear regression, whether coefficients are zero or nonzero determines which features should be selected, and FDR control can be achieved by invoking asymptotic normality of the estimators. Representative examples include the Gaussian mirror (Xing et al., 2023) and data splitting (Dai et al., 2023a). A key advantage of these approaches is scale-free property, since they avoid estimating the asymptotic variance, and they have been extended to generalized linear models (Dai et al., 2023b). Similar techniques have also been applied to sliced inverse regression under multi-index models (Zhao and Xing, 2022), although the theoretical assumptions on the feature dimension and the sparsity level are rather stringent.

As the second line of research, the model-X knockoff (Candès et al., 2018) is groundbreaking in that it can control the FDR without assuming the structure of $\boldsymbol{y} \mid \boldsymbol{X}$ where $\boldsymbol{y} \in \mathbb{R}^m$ is a response vector. However, it requires knowledge of the joint distribution of $\boldsymbol{X}$, which is restrictive. To address this limitation, methods that estimate the distribution of $\boldsymbol{X}$ using generative networks have been proposed (Jordon et al., 2019; Romano et al., 2020). In addition, a sequential knockoff sampler for a given feature distribution has been proposed (Bates et al., 2021).

FDR control via neural networks. Several studies have proposed FDR control methods that employ neural networks, although without theoretical guarantees and typically within restricted classes of architectures. The Neural Gaussian Mirror (Xing et al., 2020) defines a kernel-based conditional dependence measure and performs feature selection with MLPs. DeepPINK (Lu et al., 2018) is a knockoff framework with a specially designed network architecture in which the features are assumed to be jointly Gaussian, and DeepLINK (Zhu et al., 2021) relaxes this distributional restriction. These approaches have been shown empirically to achieve FDR control. Nevertheless, because theoretical guarantees are not provided, it remains unclear

under what conditions FDR control is achievable, and the range of supported network architectures is limited.

1.2 Notations

Vectors and matrices are typeset in boldface (e.g., $\boldsymbol{x}, \boldsymbol{B}$). For $n \in \mathbb{N}$, $[n] = \{1, \ldots, n\}$. For $S \subset [n]$, $S^{\mathsf{c}} = [n] \setminus S$. For $\boldsymbol{a} \in \mathbb{R}^n$ and $S \subset [n]$, we denote by $\boldsymbol{a}_S$ the subvector of $\boldsymbol{a}$ consisting of the entries indexed by S. For a matrix $\boldsymbol{A}$, let $\boldsymbol{A}^+$ be the Moore-Penrose pseudo-inverse of $\boldsymbol{A}$. $\Phi : \mathbb{R} \to \mathbb{R}$ and $\phi : \mathbb{R} \to \mathbb{R}$ are the cumulative distribution function and the density function of the standard Gaussian distribution, respectively.

2 PROBLEM FORMULATION

Suppose that we observe a response vector $\boldsymbol{y} = (y_1, \ldots, y_m)^\top \in \mathbb{R}^m$ together with a design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)^\top \in \mathbb{R}^{m \times n}$, where m is the sample size and n is the number of features. Our goal is to select a relevant subset of feature indices from $[n]$ that are associated with $\boldsymbol{y}$. A desirable feature selection procedure controls the false discovery rate (FDR) (Benjamini and Hochberg, 1995), defined by

$$\mathsf{FDR} = \mathbb{E}[\mathsf{FDP}], \quad \mathsf{FDP} = \frac{\#\{j \notin S : j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} \vee 1},$$

where S denotes the set of indices corresponding to the relevant features, and $\hat{S}$ is the set of selected indices.

We consider the situation where $(\boldsymbol{y}, \boldsymbol{X})$ follows the multi-index model defined below.

Definition 1 (Multi-index model)**.** *We say that a pair of the response vector $\boldsymbol{y} \in \mathbb{R}^m$ and the design matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ follows a* multi-index model *if there exists a weight matrix $\boldsymbol{B} \in \mathbb{R}^{n \times q^*}$, a deterministic function $g : \mathbb{R}^{q^*} \times \mathbb{R} \to \mathbb{R}$, and noise variables $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)^\top$ independent of $\boldsymbol{X}$ such that, for each $i \in [m]$,*

$$y_i = g(\boldsymbol{B}^\top \boldsymbol{x}_i, \varepsilon_i). \tag{1}$$

Let $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)^\top$. This formulation (1) encompasses linear regression, logistic regression, and certain neural network models with q^* hidden units in the first layer. The column space of $\boldsymbol{B}$ is often referred to as the central subspace of $\boldsymbol{X}$.

We now consider fitting a neural network to the observations. Let $f_{\mathcal{W}} : \mathbb{R}^n \to \mathbb{R}$ denote a neural network parameterized by the set $\mathcal{W}$, e.g., including weight matrices $(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L)$, where L is the number of layers. Given a loss function $\mathcal{L} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the empirical risk minimization problem is

$$\min_{\mathcal{W}} \sum_{i=1}^{m} \mathcal{L}(y_i, f_{\mathcal{W}}(\boldsymbol{x}_i)).$$

For example, we may use the quadratic loss $\mathcal{L}(u, v) = (u - v)^2$ for regression and the cross-entropy loss $\mathcal{L}(u, v) = \log(1 + \exp(v)) - uv$ for binary classification. The optimization is performed by (stochastic) gradient descent, starting from initial parameters $\mathcal{W}^{(0)}$ and yielding updated parameters $\mathcal{W}^{(t)}$ after t iterations.

After training, we evaluate feature importance by the partial derivative of the fitted network with respect to each input feature. Specifically, for $t \in \mathbb{N}$ and $j \in [n]$, define

$$\xi_j^{(t)} \equiv \sum_{i=1}^{m} \frac{\partial}{\partial x_{ij}} f_{\mathcal{W}^{(t)}}(\boldsymbol{x}_i). \tag{2}$$

If the fitted network is differentiable almost everywhere, the input sensitivity $\xi_j^{(t)}$ can serve as a measure of the contribution of the j-th feature to the response $\boldsymbol{y}$. After computing $\boldsymbol{\xi}^{(t)} = (\xi_1^{(t)}, \ldots, \xi_n^{(t)})^\top$, we then determine an appropriate cutoff to control the FDR.

3 THEORETICAL BACKGROUND

To control the FDR, we need marginal distributional characterizations of the input sensitivity $\boldsymbol{\xi}^{(t)}$ under the null. In this section, we characterize the distribution of suitably transformed $\boldsymbol{\xi}^{(t)}$ for each $n, m, q^*, t \in \mathbb{N}$. Based on this, we establish the marginal asymptotic normality of the feature importance uniformly under the null as $n \to \infty$ with $q^* = o(n)$, for arbitrary m and t.

In what follows, we formally define the index set S^{c} of null features.

Definition 2. *We say that x_j for $j \in [n]$ is a null feature if,*

$$y \perp\!\!\!\perp x_j \mid \boldsymbol{x}_{-j}. \tag{3}$$

We then define S^{c} as the index set of all null features.

Assumption 1 (Multi-index model and $\boldsymbol{B}$-ROI design)**.** *(i) The observation $(\boldsymbol{y}, \boldsymbol{X}) \in \mathbb{R}^m \times \mathbb{R}^{m \times n}$ follows the multi-index model in Definition 1 with a full rank $\boldsymbol{B} \in \mathbb{R}^{n \times q^*}$.*
(ii) The design matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ satisfies $\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{X}\boldsymbol{U}$ for any orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ such that $\boldsymbol{U}\boldsymbol{B} = \boldsymbol{B}$. We shall call this property $\boldsymbol{B}$-ROI in the sequel.

The design assumption (ii) permits row-wise dependence, heavy-tailed marginals, and low-rank structures. Still, it rules out certain forms of column dependence, discrete-valued entries, and multi-modal distributions. See Appendix B for further discussion. We

also discuss the robustness to elliptical designs in Section C. Additional mild regularity conditions under which (3) holds if and only if $\boldsymbol{b}_j = \boldsymbol{0}_{q^*}$ are provided in Appendix A.4.

Assumption 2 (Loss function)**.** *Suppose that the loss function $\mathcal{L} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ has a finite partial derivative with respect to the second argument $\partial_2 \mathcal{L}(u, v)$ for almost every $v \in \mathbb{R}$.*

This covers most losses encountered in practice.

Assumption 3 (Architecture of the neural network)**.** *The first layer of the network is taken to be dense and fully connected, and we denote its weight matrix by $\boldsymbol{W}_1 \in \mathbb{R}^{n \times q}$. The dependence of the entire network $f_{\mathcal{W}}(\boldsymbol{x})$ on any input $\boldsymbol{x} \in \mathbb{R}^n$ arises solely through the transformed representation $\boldsymbol{W}_1^\top \boldsymbol{x}$.*

This assumption still encompasses multilayer perceptrons with arbitrary width, depth, and activation functions. From the second layer onward, we allow any structures, including residual connections and dropout, which is a pragmatic modeling choice.

Intuitively, the linear representation $\boldsymbol{W}_1^\top \boldsymbol{x}$ serves as a surrogate for $\boldsymbol{B}^\top \boldsymbol{x}$ in the multi-index model, while the network's subsequent nonlinearity approximates $g(\cdot)$. The matrix sizes of $\boldsymbol{W}_1$ and $\boldsymbol{B}$ need not match.

Assumption 4 (SGD options)**.** *(i) Every element of initial parameters $\mathcal{W}^{(0)}$ is independent of $(\boldsymbol{y}, \boldsymbol{X})$, $\boldsymbol{W}_1^{(0)}$ and $\mathcal{W}_{\setminus 1}^{(0)}$ are independent, and $\boldsymbol{W}_1^{(0)}$ satisfies $\tilde{\boldsymbol{U}} \boldsymbol{W}_1^{(0)} \overset{\mathrm{d}}{=} \boldsymbol{W}_1^{(0)}$ for any orthogonal matrix $\tilde{\boldsymbol{U}} \in \mathbb{R}^{n \times n}$. (ii) Let the mini-batch indices $I_t \subseteq [m]$ and the learning rate $\eta_t > 0$ be independent of $(\boldsymbol{y}, \boldsymbol{X}, \mathcal{W}^{(0)})$ for all $t \in \mathbb{N}$.*

For instance, the entrywise i.i.d. Gaussian $\boldsymbol{W}_1^{(0)}$ with any common variance, including He- and Xavier-initializations, satisfies the assumption.

Under these assumptions, we obtain the following.

Proposition 1. *Under Assumptions 1–4, for each $m, n, q^* \in \mathbb{N}$ and iteration $t \in \mathbb{N}$ of the SGD with/without replacement, conditioning on the learning-rate and mini-batch schedule,*

$$\frac{\boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{\xi}^{(t)}}{\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{\xi}^{(t)}\|}$$

is uniformly distributed on the unit sphere lying in $\mathrm{Col}(\boldsymbol{B})^\perp$. Here, $\mathrm{Col}(\boldsymbol{B})^\perp$ is the orthogonal complement of the column space of $\boldsymbol{B}$, and $\boldsymbol{P}_{\boldsymbol{B}}^{\perp} = \boldsymbol{I}_n - \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B})^+ \boldsymbol{B}^\top$ is the orthogonal projection matrix onto $\mathrm{Col}(\boldsymbol{B})^\perp$.

Figure 1 provides an illustration of Proposition 1. It shows that $\boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{\xi}^{(t)} / \|\boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{\xi}^{(t)}\|$ is uniformly distributed around $\boldsymbol{B}$ while maintaining a constant angle. Since $\boldsymbol{B}$ itself reflects the intrinsic importance of each feature, this observation supports the consistency of interpreting $\boldsymbol{\xi}^{(t)}$ as feature importance.

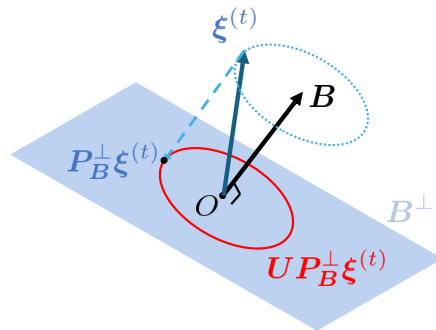

Figure 1: A schematic illustration of $\boldsymbol{\xi}^{(t)}$ for $q^* = 1$ and $n = 3$. $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ is any orthogonal matrix such that $\boldsymbol{U}\boldsymbol{B} = \boldsymbol{B}$ (i.e., rotation around $\boldsymbol{B}$).

The proof is completed by replacing the orthogonal invariance to be established for $\boldsymbol{\xi}^{(t)}$ with respect to $\boldsymbol{B}$ by an equivalent invariance of the first-layer weights $\boldsymbol{W}_1^{(t)}$ via the chain rule, and then showing recursively in t that this invariance is preserved by the update.

From spherical uniformity it follows that, letting $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$,

$$\frac{\boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{\xi}^{(t)}}{\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{\xi}^{(t)}\|} \overset{\mathrm{d}}{=} \frac{\boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{Z}}{\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{Z}\|}.$$

Together with the fact that for each null feature index $j \in S^{\mathrm{c}}$ we have $\boldsymbol{b}_j = \boldsymbol{0}_{q^*}$, this yields the following asymptotic normality:

Theorem 1. *Under Assumptions 1–4, for any null feature index $j \in S^{\mathrm{c}}$, we have*

$$\frac{\sqrt{n} \xi_j^{(t)}}{\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{\xi}^{(t)}\|} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1), \tag{4}$$

as $n \to \infty$ while $q^* = o(n)$. Furthermore, the convergence holds uniformly in $j \in S^{\mathrm{c}}$ in the sense that

$$\sup_{j \in S^{\mathrm{c}}} \sup_{u \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}\, \xi_j^{(t)}}{\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{\xi}^{(t)}\|} \leq u \right) - \Phi(u) \right| \to 0.$$

From this theorem, the asymptotic null distribution of $\xi_j^{(t)} / \|\boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{\xi}^{(t)}\|$ is identified. However, estimation of $\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp} \boldsymbol{\xi}^{(t)}\|$ is challenging since it depends on the unknown structure $\boldsymbol{B}$. In the next section, we demonstrate that by employing the data-splitting technique, the multiplicative factor that appears uniformly across all $j \in [n]$ can be ignored (*the scale-free property*), which enables valid FDR control.

Proofs of the assertions in this section are deferred to Appendix A. Figure 2 exhibits that numerical results confirm the asymptotic normality of Theorem 1.

The technical contribution underlying Proposition 1 and Theorem 1 is to extend the orthogonal-invariance-based theory of marginal asymptotic normality—originally developed by Zhao et al. (2022) and later shown to be broadly applicable by Sawaya et al. (2024)—from loss minimizers (i.e., M-estimators) to the individual iterates of loss-minimization algorithms. As a consequence, our asymptotic normality results do *not* rely on

(i) the existence or uniqueness of a loss minimizer,

(ii) convexity of the loss function,

(iii) or convergence of the optimization path.

The availability of early stopping is also appealing, as it may help reduce Type-II error. Points (ii) and (iii) are particularly important in the training of deep neural networks, where convergence to a global minimum is rarely guaranteed.

Example 3.1 (Logistic regression)**.** *The multi-index model in Definition 1 encompasses the logistic regression model* $y_i = g(\boldsymbol{\beta}^\top \boldsymbol{x}_i, \varepsilon_i)$*, where* $g(u,v) = \mathbb{1}\{(1+e^{-u})^{-1} > v\}$ *and* $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{Unif}(0,1)$ *independent of* $\boldsymbol{x}_i$*. By taking* $f_{\mathcal{W}}(\boldsymbol{x}) = \boldsymbol{w}_1^\top \boldsymbol{x}$ *and the cross-entropy loss* $\mathcal{L}(u,v) = \log(1+\exp(v)) - uv$*, our assumptions also cover the gradient descent iterates used to compute the maximum likelihood estimator. As shown by Candès and Sur (2020), the maximum likelihood estimator in logistic regression may fail to exist. Notably, because our theory applies to each iterate of SGD, it remains valid even in such settings where the optimization trajectory diverges, and convergence is not guaranteed.*

4 METHODOLOGY

In this section, we construct the actual feature selection procedure. It consists of two stages: (I) computing an importance statistic M_j for each feature that possesses desirable distributional properties, and (II) selecting features by applying an appropriate thresholding rule that controls the FDR. The properties required for the statistic to be scale-free in (I) are as follows:

(a) If $j \in S^{\mathsf{c}}$, then M_j follows (asymptotically) a distribution symmetric around zero.

(b) If $j \in S$, then M_j takes large positive values.

According to Theorem 1, $\xi_j^{(t)}$ satisfies (a) but not (b), whereas $|\xi_j^{(t)}|$ satisfies (b) but not (a). Therefore, we adopt a data-splitting approach (Dai et al., 2023a,b). Specifically, we randomly divide the data into two equal parts and, using the quantities $\xi_{j1}^{(t)}$ and $\xi_{j2}^{(t)}$ computed from each split, construct the importance statistics, for $j \in [n]$,

$$M_j = \mathsf{sign}\left(\xi_{j1}^{(t)}\xi_{j2}^{(t)}\right)\psi\left(|\xi_{j1}^{(t)}|, |\xi_{j2}^{(t)}|\right),$$

where $\psi : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a user-specified function assumed to be non-negative, symmetric, positive homogeneous, and monotone in each input, e.g., $\psi(u,v) = uv$, $\min(u,v)$, and $u+v$.

The crucial point here is that, unlike p-value based methods for FDR control, which require knowledge of the entire null distribution, this approach only relies on the symmetry under the null. As a result, no information about the asymptotic variance or convergence rate of the limiting null distribution of M_j is required.

The intuition behind why symmetry alone suffices is as follows. We can determine the cutoff for the nominal level $\alpha \in (0,1)$ as

$$\tau_\alpha = \min\left\{u > 0 : \widehat{\mathsf{FDP}}(u) \equiv \frac{\#\{j : M_j < -u\}}{\#\{j : M_j > u\} \vee 1} \leq \alpha\right\}.$$

This is expected to control the FDR because, if M_j is symmetric around zero under the null, we have

$$\begin{aligned}\mathsf{FDP}(u) &= \frac{\#\{j \in S^{\mathsf{c}} : M_j > u\}}{\#\{j : M_j > u\} \vee 1} \\ &\overset{\mathrm{d}}{=} \frac{\#\{j \in S^{\mathsf{c}} : M_j < -u\}}{\#\{j : M_j > u\} \vee 1} \\ &\leq \frac{\#\{j : M_j < -u\}}{\#\{j : M_j > u\} \vee 1} = \widehat{\mathsf{FDP}}(u).\end{aligned}$$

Overall procedure is summarized in Algorithm 1.

Algorithm 1 pseudocode for the selection procedure

Require: Nominal level $\alpha \in (0,1)$, the observation $(\boldsymbol{y}, \boldsymbol{X}) \in \mathbb{R}^m \times \mathbb{R}^{m \times n}$, the stopping time $T \in \mathbb{N}$, and $\psi : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$.

1: Split the data into two equal-sized halves $(\boldsymbol{X}^{(1)}, \boldsymbol{y}^{(1)})$ and $(\boldsymbol{X}^{(2)}, \boldsymbol{y}^{(2)})$.

2: For each part of the data, calculate $\boldsymbol{\xi}_{\cdot 1}^{(T)}$ and $\boldsymbol{\xi}_{\cdot 2}^{(T)}$ as in (2) after T update of the SGD.

3: Obtain the importance statistics $M_j = \mathsf{sign}\left(\xi_{j1}^{(T)}\xi_{j2}^{(T)}\right)\psi\left(|\xi_{j1}^{(T)}|, |\xi_{j2}^{(T)}|\right)$ for each $j \in [n]$.

4: Select features above the cutoff $\tau_\alpha = \min\{u > 0 : \widehat{\mathsf{FDP}}(u) \leq \alpha\}$.

This selection procedure asymptotically controls the FDR at a predetermined level under additional assumptions.

Assumption 5. *$\psi(\cdot,\cdot)$ is non-negative, symmetric about two inputs, and monotone in each input. Additionally, there exists $r > 0$ such that for all $a \geq 0$ and $(s,t) \in [0,\infty)^2$,*

$$\psi(as, at) \;=\; a^r\, \psi(s,t).$$

This is the formal requirement imposed on the user-specified function ψ.

Assumption 6. *In Algorithm 1, suppose that $(\boldsymbol{X}^{(1)}, \boldsymbol{y}^{(1)}) \overset{\mathrm{d}}{=} (\boldsymbol{X}^{(2)}, \boldsymbol{y}^{(2)})$. Additionally, assume that the construction of $\boldsymbol{\xi}_{\cdot 1}^{(T)}$ and $\boldsymbol{\xi}_{\cdot 2}^{(T)}$ is the same; for example, the randomness of initializations, learning rate, and loss function are common.*

This assumption is necessary for the validity of data splitting.

Assumption 7. *Let $S^+(u) = \#\{j \in S : M_j > u\}$, $S^-(u) = \#\{j \in S : M_j < -u\}$, and $S^{\pm}(u) = S^+(u) + S^-(u)$. There exist $c, \theta, \rho \in (0,1)$ such that, for $K_n = \lfloor cn \rfloor$, as $n \to \infty$,*

$$\mathbb{P}\left(S^{\pm}(u_{K_n}) \geq \theta K_n, \quad \inf_{0 \leq u \leq u_{K_n}} \frac{S^+(u)}{S^{\pm}(u)} \geq \rho \right) \to 1,$$

where u_{K_n} is the K_n-th largest magnitude among M_j's. Moreover, the following holds with the given nominal level α:

$$(\alpha\rho - (1-\rho))\,\theta \;>\; \frac{1-\alpha}{2}\,(1-\theta). \tag{5}$$

Such assumptions frequently appear in the related literature. Compared with Assumption 3.2 in Dai et al. (2023b), which requires that a fixed proportion of the true signals diverge, our assumption can be regarded as considerably weaker.

Assumption 7 requires that, within the top K_n statistics M_j's ranked by magnitude, at least a fixed fraction corresponds to non-null and, moreover, lies on the positive side. That is, among the non-null variables, at least a certain fraction is required to possess genuinely positive importance scores, and this requirement is expected to hold increasingly as the iteration t advances. Denote $n_0 = |S^{\mathsf{c}}|$.

Theorem 2. *Suppose Assumptions 1–7 hold and $n_0/n \to \pi_0 \in (0,1]$. Then, Algorithm 1 satisfies, for any given nominal level $\alpha \in (0,1)$,*

$$\mathsf{FDP} \leq \alpha + o_{\mathsf{p}}(1) \quad \text{and} \quad \limsup_{n\to\infty} \mathsf{FDR} \leq \alpha.$$

A drawback of the data-splitting method is that it effectively halves the sample size, which may reduce power. One remedy, known in the literature and also applicable here, is to aggregate the selection results obtained from multiple random splits of the data (Dai et al., 2023a), as is commonly done in stability selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013). The results obtained by applying this stabilization technique are shown in Figure 14. Related stabilization techniques include Du et al. (2023) and Ren and Barber (2024). Later, Takahashi (2025) theoretically shows that such stabilization techniques increase power in the proportional asymptotics.

5 NUMERICAL EXPERIMENTS

In this section, we empirically validate the theoretical guarantees developed in the preceding sections. We then compare the proposed method against relevant baselines. All code and scripts for reproducing our results will be made publicly available at `https://github.com/sawaya-ka/deep-feature-selection`. Further experiments are provided in Appendix D.

5.1 Marginal asymptotic normality

We numerically verify Theorem 1. The data are generated according to

$$y = g(\boldsymbol{b}_1^\top \boldsymbol{x}) + \sum_{k=2}^{q^*} \left\{ h(\boldsymbol{b}_k^\top \boldsymbol{x}) \cdot (\boldsymbol{b}_{k-1}^\top \boldsymbol{x}) \right\} + \varepsilon,$$

with $\varepsilon \sim \mathcal{N}(0,1)$, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $q^* = 8$, $g(x) = (x-2)^2$ and $h(x) = \max(x, 0)$. The vector $\boldsymbol{b}_1$ has its first half of entries equal to $2/\sqrt{n}$ and the remaining entries are zero, and $\boldsymbol{b}_k = \boldsymbol{e}_k \in \mathbb{R}^n$ for $k = 2, \ldots, q^*$.

We fixed the batch size of SGD to 128 and the learning rate to 3×10^{-3} except for Transformer. After ten update steps, we constructed a histogram of the latter half of the components of $\sqrt{n}\boldsymbol{\xi}^{(10)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(10)}\|$ and compared it with the density of a normal distribution whose mean and variance match the sample mean and sample variance of these components, as well as with the standard normal density $\mathcal{N}(0,1)$. The resulting plots are presented in Figure 2. Also, the corresponding QQ-plots are provided in Figure 6 of Appendix D.

The settings labeled $n \ll m$, $n \asymp m$, and $n \gg m$ correspond respectively to $(m,n) = (100000, 1000)$, $(2000, 1000)$, and $(10, 1000)$. The network architectures used for training in these experiments are described below. We employ dropout with a rate 0.1.

5-layer MLP. We use a 5-layer MLP consisting of

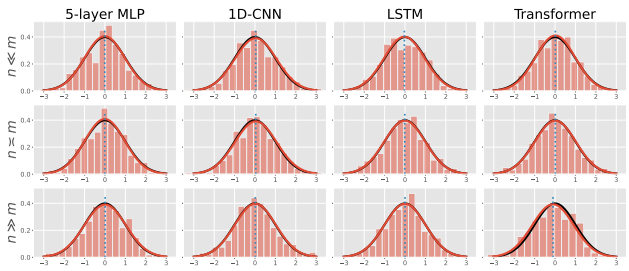

Figure 2: Histograms of the empirical distribution of $\sqrt{n}\xi_j^{(10)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(10)}\|$ for $j \in S^{\mathsf{c}}$. The solid black curve shows the $\mathcal{N}(0,1)$ density. The solid red curve represents a normal density fitted to the histograms, and the dotted blue line indicates the empirical mean.

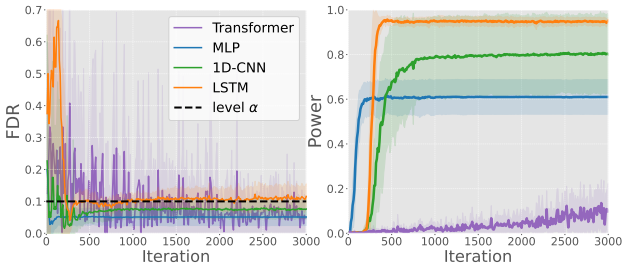

Figure 3: Results for the false discovery rate (left) and power (right) when performing feature selection at each iteration using the artificial data and models defined in Section 5.1. The solid curves represent averages over 20 independent runs, and the shaded areas indicate one standard deviation around the mean.

four hidden layers of widths $(1024, 1024, 512, 256)$ with ReLU activations and a final linear output. All weights are initialized with He-normal initialization.

1D-CNN. The 1D-CNN baseline first applies the lifting layer $\boldsymbol{W}_1^\top \boldsymbol{x}$ with $q = n$ followed by a stack of three convolutional layers with output channels $(64, 128, 128)$, kernel sizes $(11, 9, 7)$, and stride 1. Dilated convolutions with dilation rates $(1, 2, 4)$ are used to enlarge the receptive field. ReLU activation is applied after each convolution. The convolutional output is flattened and passed through a fully connected head with hidden layers $(128, 64)$ and a final linear output.

LSTM. For sequential modeling we adopt a two-layer bidirectional LSTM with hidden size 128. Each time step input is obtained by projecting the lifted representation $\boldsymbol{W}_1^\top \boldsymbol{x}$ into dimension 4. The last hidden states of the forward and backward directions are concatenated and fed into a fully connected head with hidden layer (64) and linear output. We use Xavier-normal initialization for the recurrent weights.

Transformer. The input is first lifted by a trainable dense map $\boldsymbol{W}_1^\top \in \mathbb{R}^{q\times n}$ (Xavier initialization), producing q tokens. Each token is embedded by a linear ID-specific map and normalized (LayerNorm), without positional encoding. We employ a Transformer encoder with two layers, model dimension 256, four heads, feed-forward dimension 256, and GELU activation. For sequence aggregation we use gated pooling with hidden dimension 32 and temperature $\tau = 1.0$, followed by a fully connected head with hidden size 32 and linear output. Training uses AdamW with $(\beta_1, \beta_2, \varepsilon) = (0.9, 0.95, 10^{-8})$, weight decay 0.01, batch size 128, base learning rate 3×10^{-4} with 10% warm-up, and gradient clipping at 1.0.

These experiments were conducted on a Google Cloud Platform VM equipped with a single NVIDIA A100 (40GB) GPU using PyTorch 2.6.0 with CUDA 12.4.

5.2 FDR control

We next demonstrate that Algorithm 1 is able to approximately control the false discovery rate (FDR) at or below the nominal level $\alpha = 0.1$ under the stated assumptions. We consider the setting $(m, n) = (1600, 400)$; the data-generating process and learning architectures are otherwise the same as in the previous section. We use $\psi(u, v) = u + v$.

We define the power in the feature selection problem:

$$\mathsf{Power} = \mathbb{E}\left[\frac{\#\{j \in S : j \in \hat{S}\}}{\#\{j : j \in S\}}\right].$$

We can see that the power is one minus the Type-II error.

Figure 3 reports the results. As anticipated in Assumption 7, once training progresses and the power reaches a reasonable level, the FDR is also brought under control. The trajectory of the training loss corresponding to these experiments is shown in Figure 7 of Appendix D.

Experiments were conducted on a Google Cloud Platform VM equipped with four NVIDIA Tesla T4 (16 GB each) GPUs, using PyTorch 2.8.0 (built with CUDA 12.8) and CUDA runtime 12.4.

Remark 5.1. *In a single-index model $y = g(\langle \boldsymbol{w}_\star, \boldsymbol{x}\rangle) + \varepsilon$, if the trained network well approximates the regression function then $\nabla_{\boldsymbol{x}} f_{\mathcal{W}^{(t)}}(\boldsymbol{x}) \approx g'(\langle \boldsymbol{w}_\star, \boldsymbol{x}\rangle)\, \boldsymbol{w}_\star$, so the averaged gradient $\boldsymbol{\xi}^{(t)}$ converges to $\mathbb{E}[g'(Z)]\, \boldsymbol{w}_\star$ with $Z \sim \mathcal{N}(0, \|\boldsymbol{w}_\star\|^2)$. While the factor $\mathbb{E}[g'(Z)]$ being close to zero does not affect the validity of FDR control, it cancels out the signal direction and thereby reduces the power of Algorithm 1.*

5.3 Comparison with other methods

Competing methods. We benchmark our approach against flexible variable-selection baselines that can (or aim to) control FDR. *Neural baselines* advertised as enabling FDR control via deep representations include the Neural Gaussian Mirror (NGM) (Xing et al., 2020) and DeepLINK (Zhu et al., 2021). While empirically competitive, these methods do not offer formal guarantees. All methods are run at the same nominal FDR level $\alpha = 0.1$.

We ran DeepLINK (Zhu et al., 2021) with authors' official code[1] with hyperparameters as follows: L1 penalty 10^{-3}, learning rate 10^{-3}, ELU activation, and mean squared error loss. Column-wise centering and scaling were applied by default.

Among the methods compared, NGM was thus far the most computationally demanding. To make large-scale experiments feasible, we employed several approximations: a preliminary screening step retaining at most $n/2$ variables, a subsample-based approximation of the kernel matrices (subsample size 800), and a coarse grid of 6 candidate values for the scale parameter c_j. The neural network used in NGM was a two-hidden-layer MLP with hidden widths proportional to $\log n$, trained for 60 epochs with batch size 256 and learning rate 10^{-3}.

Data generating process. We fix the ambient dimension at $n = 500$ and vary the sample size $m \in \{2000, 1000, 500\}$ to probe the effect of sample scarcity. Unless otherwise noted, the data-generating process follows Section 5.1 exactly, *except* that (because the design $\boldsymbol{X}$ is scaled by $1/\sqrt{n}$) we set the nonzero entries of the signal matrix $\boldsymbol{B}$ to 2 (rather than $2/\sqrt{n}$).

We consider four scenarios for the design matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ satisfying Assumption 1 (ii).

N(0,1). Entries are i.i.d. Gaussian: $X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1/n)$, $(i, j) \in [m] \times [n]$.

t(3). Entries are i.i.d. standardized t: $X_{ij} \overset{\text{iid}}{\sim} \frac{1}{\sqrt{n}} t(3)$, $(i, j) \in [m] \times [n]$, where $t(3)$ denotes the t-distribution with 3 degrees of freedom.

Spiked. Let $r = 2$ be the spike rank. We write $\boldsymbol{X} = \boldsymbol{V}_1 \boldsymbol{V}_2^\top + \boldsymbol{E}$, where $\boldsymbol{V}_1 \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V}_2 \in \mathbb{R}^{n\times r}$ have orthonormal columns (each drawn as r-columns from independent Haar orthogonal matrices of sizes m and n, respectively), and $\boldsymbol{E}$ has i.i.d. entries $E_{ij} \sim \mathcal{N}(0, 1/n)$.

Our inferential procedures employ sample splitting: we partition the m rows into two equal halves, $\boldsymbol{X}^{(A)} \in \mathbb{R}^{(m/2)\times n}$ and $\boldsymbol{X}^{(B)} \in \mathbb{R}^{(m/2)\times n}$. For N(0,1) and t(3), closure under row-subsampling is immediate. For Spiked, we assume a modeling convention such that, when the m rows are partitioned into two equal halves, each half independently follows the same distributional form as the original model (with common latent parameters), so that both $\boldsymbol{X}^{(A)}$ and $\boldsymbol{X}^{(B)}$ are valid Spiked samples.

[1] https://github.com/zifanzhu/DeepLINK

Table 1: Performance comparison for $m = 2000$.

	Proposed		DeepLINK		NGM	
Design	FDR	Power	FDR	Power	FDR	Power
N(0,1)	0.101	1.000	0.089	0.999	0.086	0.851
t(3)	0.072	0.843	0.081	0.870	0.063	0.599
Spiked	0.104	1.000	0.087	1.000	0.080	0.847

Table 2: Performance comparison for $m = 1000$.

	Proposed		DeepLINK		NGM	
Design	FDR	Power	FDR	Power	FDR	Power
N(0,1)	0.072	0.637	0.065	0.467	0.131	0.658
t(3)	0.061	0.168	0.064	0.085	0.144	0.401
Spiked	0.072	0.682	0.063	0.492	0.151	0.674

Tables 1–3 report the comparison of FDR and Power with $n = 500$, averaged over 20 independent runs. The entry labeled Proposed corresponds to the LSTM defined in the previous section. It can be observed that NGM fails to control the FDR except in the $m = 2000$ setting, while our method achieves relatively high power under FDR control. More detailed results are provided in the Appendix D.

6 DISCUSSION

Summary of Contributions. To the best of our knowledge, this paper provides the first theoretical guarantee for false discovery rate (FDR) control in feature selection with multilayer neural networks. This property is crucial for ensuring the reproducibility of scientific discoveries, and our results demonstrate that one can achieve both model flexibility and rigorous statistical reliability. In addition, the proposed implementation via simple data splitting is straightforward and easy to apply.

Limitations and Future Directions. Our analysis assumes that the first layer of the neural network is a dense fully connected transformation. While technically convenient, this limits the ability to exploit the spatial locality of image data or the sequential and positional structures inherent in natural language. Extending the methodology to capture richer latent

Table 3: Performance comparison for $m = 500$.

	Proposed		DeepLINK		NGM	
Design	FDR	Power	FDR	Power	FDR	Power
N(0,1)	0.047	0.040	0.019	0.003	0.242	0.380
t(3)	0.058	0.016	0.017	0.004	0.325	0.428
Spiked	0.029	0.037	0.026	0.020	0.264	0.435

structures in diverse data modalities remains an important challenge.

Moreover, we have defined the input sensitivity $\boldsymbol{\xi}^{(t)}$ as an average across instances. It would be interesting to investigate how this sensitivity can be characterized at the single-instance level, especially in domains such as computer vision where the set of pixels critical for classification may vary substantially across samples.

In this work, we considered the raw input gradients of a trained neural network as a feature importance. A promising direction for future work is to extend our analysis to more sophisticated feature attribution methods, including Integrated Gradients (Sundararajan et al., 2017), DeepLIFT (Shrikumar et al., 2017), and SmoothGrad (Smilkov et al., 2017). Adapting our FDR-control framework to such attribution methods would potentially yield more stable and interpretable feature selection.

From a theoretical standpoint, our framework relied on the $\boldsymbol{B}$-right-orthogonal invariance of the design matrix $\boldsymbol{X}$. Exploring the behavior of our algorithm under more severe correlation structures is an appealing direction for future work, as is extending the theory to regularized training regimes.

Finally, while our results establish FDR control, an important practical question is which network architectures—in terms of depth, width, use of attention, dropout, or residual connections—yield higher power as functions of the sample size and ambient dimension. Answering this question is a promising research direction, though it will likely require substantially more theoretical effort.

Acknowledgements

The author is grateful to anonymous reviewers for their constructive feedback. This work was supported by the Japan Science and Technology Agency through ACT-X (JPMJAX24CC) and the Japan Society for the Promotion of Science through Research Fellow DC1 (24KJ0841).

References

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Bates, S., Candès, E., Janson, L., and Wang, W. (2021). Metropolized knockoff sampling. *Journal of the American Statistical Association*, 116(535):1413–1427.

Benjamini, Y. and Hochberg, Y. (1995). 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.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). 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.

Candès, E. J. and Sur, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42.

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2023a). False discovery rate control via data splitting. *Journal of the American Statistical Association*, 118(544):2503–2520.

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2023b). A scale-free approach for false discovery rate control in generalized linear models. *Journal of the American Statistical Association*, 118(543):1551–1565.

Du, L., Guo, X., Sun, W., and Zou, C. (2023). False discovery rate control under general dependence by symmetrized data aggregation. *Journal of the American Statistical Association*, 118(541):607–621.

Jordon, J., Yoon, J., and van der Schaar, M. (2019). Knockoffgan: Generating knockoffs for feature selection using generative adversarial networks. In *International conference on learning representations*.

Li, Y. and Sur, P. (2023). Spectrum-aware debiasing: A modern inference framework with applications to principal components regression. *arXiv preprint arXiv:2309.07810*.

Lu, Y., Fan, Y., Lv, J., and Stafford Noble, W. (2018). Deeppink: reproducible feature selection in deep neural networks. *Advances in neural information processing systems*, 31.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473.

Rangan, S., Schniter, P., and Fletcher, A. K. (2019). Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684.

Ren, Z. and Barber, R. F. (2024). Derandomised knockoffs: leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):122–154.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Romano, Y., Sesia, M., and Candès, E. (2020). Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872.

Sawaya, K., Uematsu, Y., and Imaizumi, M. (2024). High-dimensional single-index models: Link estimation and marginal inference. *arXiv preprint arXiv:2404.17812*.

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(1):55–80.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Takahashi, T. (2025). Replica analysis for ensemble techniques in variable selection. *Journal of the Physical Society of Japan*, 94(3):031011.

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

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Xing, X., Gui, Y., Dai, C., and Liu, J. S. (2020). Ngm: Neural gaussian mirror for controlled feature selection in neural networks. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 148–152.

Xing, X., Zhao, Z., and Liu, J. S. (2023). Controlling false discovery rate using gaussian mirrors. *Journal of the American Statistical Association*, 118(541):222–241.

Zhao, Q., Sur, P., and Candès, E. J. (2022). The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance. *Bernoulli*, 28(3):1835–1861.

Zhao, Z. and Xing, X. (2022). On the testing of multiple hypothesis in sliced inverse regression. *arXiv preprint arXiv:2210.05873*.

Zhu, Z., Fan, Y., Kong, Y., Lv, J., and Sun, F. (2021). Deeplink: Deep learning inference using knockoffs with applications to genomics. *Proceedings of the National Academy of Sciences*, 118(36):e2104683118.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [No]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Proofs

A.1 Proof of Proposition 1

Proof of Proposition 1. The basic idea of the proof is inspired by the proof of Proposition 2.1 in Zhao et al. (2022). Since $\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}\|$ has the unit norm and lies in $\mathrm{Col}(\boldsymbol{B})^{\perp}$, it is sufficient to show that, for any orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ obeying $\boldsymbol{U}\boldsymbol{B} = \boldsymbol{B}$,

$$\boldsymbol{U}\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)} \overset{\mathrm{d}}{=} \boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}. \tag{6}$$

We proceed to show its sufficient condition $\boldsymbol{U}\boldsymbol{W}_1^{(t)} = \boldsymbol{W}_1^{(t)}$ since Assumption 3 implies that the input sensitivity is given by, with $\mathcal{W}_{\setminus 1}^{(t)} \equiv \mathcal{W}^{(t)} \setminus \boldsymbol{W}_1^{(t)}$,

$$\boldsymbol{\xi}^{(t)} = \sum_{i=1}^{m} \boldsymbol{W}_1^{(t)} h\left((\boldsymbol{W}_1^{(t)})^{\top}\boldsymbol{x}_i; \mathcal{W}_{\setminus 1}^{(t)}\right),$$

where $h(\cdot)$ depends on $\boldsymbol{x}_i$ only through $(\boldsymbol{W}_1^{(t)})^{\top}\boldsymbol{x}_i$, and given by

$$h\left((\boldsymbol{W}_1^{(t)})^{\top}\boldsymbol{x}_i; \mathcal{W}_{\setminus 1}^{(t)}\right) = \frac{\partial f_{\mathcal{W}^{(t)}}(\boldsymbol{x}_i)}{\partial((\boldsymbol{W}_1^{(t)})^{\top}\boldsymbol{x}_i)}.$$

For each $t \in \mathbb{N}$, denote $\boldsymbol{W}_1^{(t)} = \boldsymbol{W}_1^{(t)}(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{W}_1^{(t-1)}, \mathcal{W}_{\setminus 1}^{(t-1)})$ to clarify the dependence of SGD iterates on the sample $(\boldsymbol{y}, \boldsymbol{X})$ and the previous iterate $\mathcal{W}^{(t-1)} = (\boldsymbol{W}_1^{(t-1)}, \mathcal{W}_{\setminus 1}^{(t-1)})$. Then, Assumptions 2 and 3 yield, for each $t \in \mathbb{N}$,

$$\begin{aligned}
&\boldsymbol{W}_1^{(t)}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{U}, \boldsymbol{W}_1^{(t-1)}, \mathcal{W}_{\setminus 1}^{(t-1)}) \\
&= \boldsymbol{W}_1^{(t-1)} - \eta_t \sum_{i\in I_t} \partial_2 \mathcal{L}\left(y_i, f_{\mathcal{W}^{(t-1)}}\left(\boldsymbol{U}^{\top}\boldsymbol{x}_i\right)\right) \cdot \boldsymbol{U}^{\top}\boldsymbol{x}_i h\left((\boldsymbol{W}_1^{(t-1)})^{\top}\boldsymbol{U}^{\top}\boldsymbol{x}_i; \mathcal{W}_{\setminus 1}^{(t-1)}\right)^{\top} \\
&= \boldsymbol{U}^{\top}\boldsymbol{W}_1^{(t)}(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{U}\boldsymbol{W}_1^{(t-1)}, \mathcal{W}_{\setminus 1}^{(t-1)}),
\end{aligned} \tag{7}$$

where we use $f_{\mathcal{W}^{(t-1)}}\left(\boldsymbol{U}^{\top}\boldsymbol{x}_i\right) = f_{(\boldsymbol{U}\boldsymbol{W}_1^{(t-1)}, \mathcal{W}_{\setminus 1}^{(t-1)})}(\boldsymbol{x}_i)$ in the last equation.

By Assumption 1 (i) and $\boldsymbol{U}\boldsymbol{B} = \boldsymbol{B}$, it follows that, for each $i \in [m]$,

$$y_i = g(\boldsymbol{B}^{\top}\boldsymbol{x}_i, \varepsilon_i) = g(\boldsymbol{B}^{\top}\boldsymbol{U}^{\top}\boldsymbol{x}_i, \varepsilon_i).$$

Hence, $(\boldsymbol{y}, \boldsymbol{X}) \overset{\mathrm{d}}{=} (\boldsymbol{y}, \boldsymbol{X}\boldsymbol{U})$ by Assumption 1 (ii). Together with Assumption 4, we obtain

$$(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{U}, \boldsymbol{W}_1^{(0)}, \mathcal{W}_{\setminus 1}^{(0)}) \overset{\mathrm{d}}{=} (\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{U}\boldsymbol{W}_1^{(0)}, \mathcal{W}_{\setminus 1}^{(0)}). \tag{8}$$

This implies

$$\boldsymbol{W}_1^{(1)}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{U}, \boldsymbol{W}_1^{(0)}, \mathcal{W}_{\setminus 1}^{(0)}) \overset{\mathrm{d}}{=} \boldsymbol{W}_1^{(1)}(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{U}\boldsymbol{W}_1^{(0)}, \mathcal{W}_{\setminus 1}^{(0)}), \tag{9}$$

which follows from the fact that the identical measurable function of the random elements following the same law has again the same law. As a result, (7) for $t = 1$, (9), and the left-orthogonal invariance of $\boldsymbol{W}_1^{(0)}$ in Assumption 4 give

$$\boldsymbol{U}\boldsymbol{W}_1^{(1)}(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{W}_1^{(0)}, \mathcal{W}_{\setminus 1}^{(0)}) \overset{\mathrm{d}}{=} \boldsymbol{W}_1^{(1)}(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{W}_1^{(0)}, \mathcal{W}_{\setminus 1}^{(0)}). \tag{10}$$

Also for $t = 2$, applying (7) yields

$$\begin{aligned}
&\boldsymbol{U}\boldsymbol{W}_1^{(2)}\left(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{U}, \boldsymbol{W}_1^{(1)}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{U}, \boldsymbol{W}_1^{(0)}, \mathcal{W}_{\setminus 1}^{(0)}), \mathcal{W}_{\setminus 1}^{(1)}\right) \\
&= \boldsymbol{W}_1^{(2)}\left(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{U}\boldsymbol{W}_1^{(1)}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{U}, \boldsymbol{W}_1^{(0)}, \mathcal{W}_{\setminus 1}^{(0)}), \mathcal{W}_{\setminus 1}^{(1)}\right) \\
&= \boldsymbol{W}_1^{(2)}\left(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{W}_1^{(1)}(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{U}\boldsymbol{W}_1^{(0)}, \mathcal{W}_{\setminus 1}^{(0)}), \mathcal{W}_{\setminus 1}^{(1)}\right).
\end{aligned}$$

Therefore, we have, by (8),

$$\boldsymbol{U}\boldsymbol{W}_1^{(2)}(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{W}_1^{(1)}, \mathcal{W}_{\setminus 1}^{(1)}) \overset{\mathrm{d}}{=} \boldsymbol{W}_1^{(2)}(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{W}_1^{(1)}, \mathcal{W}_{\setminus 1}^{(1)}),$$

in the same manner as (10). We can immediately generalize this argument to any $t \in \mathbb{N}$ by the recursive argument. This implies the desired property (6). □

A.2 Proof of Theorem 1

For the proof of Theorem 1, we prepare a lemma.

Lemma A.1. *Assume the Assumption 1. Let $\boldsymbol{z} \in \mathbb{R}^n$ be a standard Gaussian random vector. For any $t > 0$, there exists a universal constant $c > 0$ such that*

$$\mathbb{P}\left(\left|\sqrt{\frac{\boldsymbol{z}^\top \boldsymbol{P}_{\boldsymbol{B}}^\perp \boldsymbol{z}}{n}} - \sqrt{\frac{n - q^*}{n}}\right| > t\right) \leq 2\exp\left(-c\left(nt^2 \wedge \sqrt{n(n-q^*)}t\right)\right).$$

Proof of Lemma A.1. Since $\mathsf{rank}(\boldsymbol{B}) = q^*$ by Assumption 1 (i), we have $\|\boldsymbol{P}_{\boldsymbol{B}}^\perp\|_F^2 = n - q^*$ and $\|\boldsymbol{P}_{\boldsymbol{B}}^\perp\|_{\mathrm{op}} = 1$. Thus, Hanson–Wright inequality (Vershynin, 2018) implies that, for any $t > 0$,

$$\mathbb{P}\left(\left|\boldsymbol{z}^\top \boldsymbol{P}_{\boldsymbol{B}} \boldsymbol{z} - (n - q^*)\right| > t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{n - q^*} \wedge t\right)\right),$$

with some constant $c > 0$. Hence, it follows that, for any $t > 0$,

$$\mathbb{P}\left(\left|\frac{\boldsymbol{z}^\top \boldsymbol{P}_{\boldsymbol{B}} \boldsymbol{z}}{n - q^*} - 1\right| > t\right) \leq 2\exp\left(-c\,(n - q^*)(t^2 \wedge t)\right),$$

with some constant $c > 0$. For $a > 0$, we have $|a^2 - 1| = |a + 1| \cdot |a - 1| \geq |a - 1|$ since $a + 1 > 1$. Using this, we obtain, for any $t > 0$,

$$\mathbb{P}\left(\left|\sqrt{\frac{\boldsymbol{z}^\top \boldsymbol{P}_{\boldsymbol{B}} \boldsymbol{z}}{n - q^*}} - 1\right| > t\right) \leq \mathbb{P}\left(\left|\frac{\boldsymbol{z}^\top \boldsymbol{P}_{\boldsymbol{B}} \boldsymbol{z}}{n - q^*} - 1\right| > t\right) \leq 2\exp\left(-c\,(n - q^*)(t^2 \wedge t)\right),$$

with some constant $c > 0$. Change-of-variable from t to $t\sqrt{n/(n - q^*)}$ completes the proof. □

Proof of Theorem 1. From Proposition 1, we obtain

$$\frac{\boldsymbol{P}_{\boldsymbol{B}}^\perp \boldsymbol{\xi}^{(t)}}{\|\boldsymbol{P}_{\boldsymbol{B}}^\perp \boldsymbol{\xi}^{(t)}\|} \overset{\mathrm{d}}{=} \frac{\boldsymbol{P}_{\boldsymbol{B}}^\perp \boldsymbol{z}}{\|\boldsymbol{P}_{\boldsymbol{B}}^\perp \boldsymbol{z}\|},$$

where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Here, j-th element of $\boldsymbol{P}_{\boldsymbol{B}}^\perp \boldsymbol{\xi}^{(t)}$ and $\boldsymbol{P}_{\boldsymbol{B}}^\perp \boldsymbol{z}$ are given by

$$\xi_j^{(t)} - \boldsymbol{b}_j^\top (\boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top \boldsymbol{\xi}^{(t)} \quad \text{and} \quad z_j - \boldsymbol{b}_j^\top (\boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top \boldsymbol{z},$$

respectively. Note that $\boldsymbol{b}_j = \mathbf{0}$ under the null by Theorem 3. Thus, for $j \in S^{\mathsf{c}}$, we have

$$\sqrt{n}\xi_j^{(t)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}\| \overset{\mathrm{d}}{=} \sqrt{n}z_j/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{z}\|. \tag{11}$$

Here, since $\mathrm{rank}(\boldsymbol{B}) = q^*$ by Assumption 1 (i), Lemma A.1 implies that $\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{z}\|/\sqrt{n} \overset{\mathrm{p}}{\to} 1$ as $n \to \infty$ while $q^* = o(n)$. This completes the proof of (4).

Next, we show the uniform convergence. Denote $\sigma_n = \sqrt{n}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{z}\|$ for convenience. Fix an arbitrary $\epsilon > 0$ and define $E_n = \{|\sigma_n - 1| < \epsilon\}$. From (11), we have

$$\Delta_n \equiv \sup_{j\in S^{\mathsf{c}}} \sup_{u\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}\,\xi_j^{(t)}}{\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}\|} \le u \right) - \Phi(u) \right| = \sup_{j\in S^{\mathsf{c}}} \sup_{u\in\mathbb{R}} \left| \mathbb{P}\left(z_j \le u\sigma_n\right) - \Phi(u) \right|.$$

Here, we have

$$\left|\mathbb{P}\left(z_j \le u\sigma_n\right) - \Phi(u)\right| \le \mathbb{P}(E_n^{\mathsf{c}}) + \left|\mathbb{P}(\sigma_n z_j < u \mid E_n) - \Phi(t)\right|. \tag{12}$$

Also, since $\sigma_n \in [1-\epsilon, 1+\epsilon]$ under E_n, it follows that

$$\mathbb{P}\left(z_j < \frac{u}{1+\epsilon}\right) \le \mathbb{P}\left(\sigma_n z_j \mid E_n\right) \le \mathbb{P}\left(z_j < \frac{u}{1-\epsilon}\right).$$

From this, we have

$$\left|\mathbb{P}(\sigma_n z_j < u \mid E_n) - \Phi(u)\right| \le \max\left\{ \Phi\left(\frac{u}{1-\epsilon}\right) - \Phi(u),\ \Phi(u) - \Phi\left(\frac{u}{1+\epsilon}\right) \right\}.$$

Taking the supremum, the mean-value theorem gives

$$\sup_{u\in\mathbb{R}} \left|\mathbb{P}(\sigma_n z_j < u \mid E_n) - \Phi(u)\right| \le \frac{\epsilon}{1-\epsilon} \cdot \frac{1}{\sqrt{2\pi e}}, \tag{13}$$

where we use the fact $\sup_{u\in\mathbb{R}} |u|\phi(u) = 1/\sqrt{2\pi e}$. Since this upper bound does not depend on $j \in [n]$, (12) and (13) yield

$$\Delta_n \le \mathbb{P}\left(|\sigma_n - 1| > \epsilon\right) + \frac{\epsilon}{1-\epsilon} \cdot \frac{1}{\sqrt{2\pi e}}.$$

The first term on the right-hand side converges to zero as $n \to \infty$ while $q^* = o(n)$ by Lemma A.1, and the second term goes to zero as $\epsilon \downarrow 0$. □

A.3 Proof of Thorem 2

Our argument is inspired by the proof of Proposition 3.2 of Dai et al. (2023b), but proceeds under weaker conditions on signal strength and dimensionality. While auxiliary lemmas overlap with Dai et al. (2023b), we provide full proofs to ensure a self-contained presentation. We prepare some notation for the proof.

- $I_u(v) \equiv \inf\{w \ge 0 : \psi(v, w) > u\}$ for any $u > 0$ and $v \ge 0$ with the convention $\inf \emptyset = +\infty$.
- $\tilde{M}_j = \mathsf{sign}(z_{j1}z_{j2})\psi(|z_{j1}|, |z_{j2}|)$ where $(z_{11}, \ldots, z_{n1}, z_{12}, \ldots, z_{n2})^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{2n})$.
- $\varsigma_{j1}^{(t)} = \sqrt{n}\xi_{j1}^{(t)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}_1^{(t)}\|$ and $\varsigma_{j2}^{(t)} = \sqrt{n}\xi_{j2}^{(t)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}_1^{(t)}\|$ for any $t \in \mathbb{N}$ and $j \in [n]$.
- $\breve{M}_j = \mathsf{sign}(\varsigma_{j1}^{(t)}\varsigma_{j2}^{(t)})\psi(|\varsigma_{j1}^{(t)}|, |\varsigma_{j2}^{(t)}|)$.
- $V^+(u) = \#\{j \in S^{\mathsf{c}} : \breve{M}_j > u\}$ and $V^-(u) = \#\{j \in S^{\mathsf{c}} : \breve{M}_j < -u\}$.
- $\tilde{F}(u) = \mathbb{P}(\tilde{M}_1 > u)$.
- $\tau_\alpha^{\sigma_n}$ is the cutoff when we apply Algorithm 1 with $\breve{M}_j$ instead of M_j.

Recall that we defined the original importance statistics as $M_j = \mathsf{sign}(\xi_{j1}^{(t)}\xi_{j2}^{(t)})\psi(|\xi_{j1}^{(t)}|, |\xi_{j2}^{(t)}|)$.

Lemma A.2 (Zero-symmetry of $\tilde{M}_j$)**.** *For each $j \in [n]$ and any Borel measurable function $\psi : [0,\infty)^2 \to [0,\infty)$, we have*

$$\tilde{M}_j \overset{\mathrm{d}}{=} -\tilde{M}_j.$$

Proof. Write $f : \mathbb{R}^2 \to \mathbb{R}$ for the measurable map

$$f(z_1, z_2) \;=\; \mathrm{sign}(z_1 z_2)\, \psi\big(|z_1|, |z_2|\big).$$

We observe the oddness under a single-coordinate reflection: for all $z_1, z_2 \in \mathbb{R}$,

$$f(-z_1, z_2) \;=\; -\, f(z_1, z_2),$$

since $\mathsf{sign}((-z_1)z_2) = -\mathsf{sign}(z_1 z_2)$ while the arguments of ψ are unchanged by taking absolute values.

Since $(z_{j1}, z_{j2})^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_2)$, its law is invariant under the orthogonal reflection

$$R := \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},$$

i.e., $(z_{j1}, z_{j2})^\top \overset{\mathrm{d}}{=} R(z_{j1}, z_{j2})^\top = (-z_{j1}, z_{j2})^\top$. Therefore, for any Borel set $A \subset \mathbb{R}$,

$$\begin{aligned}\mathbb{P}\big(\tilde{M}_j \in A\big) &= \mathbb{P}\big(f(z_{j1}, z_{j2}) \in A\big) = \mathbb{P}\big(f(-z_{j1}, z_{j2}) \in A\big) = \mathbb{P}\big(-f(z_{j1}, z_{j2}) \in A\big) \\ &= \mathbb{P}\big(f(z_{j1}, z_{j2}) \in -A\big) = \mathbb{P}\big(\tilde{M}_j \in -A\big).\end{aligned}$$

Since this holds for every Borel A, the laws of $\tilde{M}_j$ and $-\tilde{M}_j$ coincide; that is, $\tilde{M}_j \overset{\mathrm{d}}{=} -\tilde{M}_j$. □

Lemma A.3 (Invariance under common scaling)**.** *Suppose Assumption 5 holds. For any $C \in \mathbb{R} \setminus \{0\}$, define scaled importance scores for each $j \in [n]$,*

$$\tilde{\xi}_{j\cdot}^{(t)} = C\, \xi_{j\cdot}^{(t)}, \quad M_j^C = \mathsf{sign}(\tilde{\xi}_{j1}^{(t)}\tilde{\xi}_{j2}^{(t)})\, \psi\big(|\tilde{\xi}_{j1}^{(t)}|, |\tilde{\xi}_{j2}^{(t)}|\big).$$

Let the corresponding cutoff and the set of selected indices be $\tau_\alpha^C, \hat{S}_\alpha^C$. Then

$$\tau_\alpha^C = |C|^r\, \tau_\alpha, \qquad \hat{S}_\alpha^C = \hat{S}_\alpha.$$

Proof. At first, the sign factor is invariant since, for each $j \in [n]$,

$$\mathsf{sign}\big(\tilde{\xi}_{j1}^{(t)}\tilde{\xi}_{j2}^{(t)}\big) = \mathsf{sign}\big((C\xi_{j1}^{(t)})(C\xi_{j2}^{(t)})\big) = \mathsf{sign}(C^2)\, \mathsf{sign}\big(\xi_{j1}^{(t)}\xi_{j2}^{(t)}\big) = \mathsf{sign}\big(\xi_{j1}^{(t)}\xi_{j2}^{(t)}\big).$$

Also, since $|\tilde{\xi}_{jk}^{(t)}| = |C|\, |\xi_{jk}^{(t)}|$, Assumption 5 of homogeneity implies that there exists $r > 0$ such that

$$\psi(|\tilde{\xi}_{j1}^{(t)}|, |\tilde{\xi}_{j2}^{(t)}|) = \psi(|C|\, |\xi_{j1}^{(t)}|,\, |C|\, |\xi_{j2}^{(t)}|) = |C|^r\, \psi(|\xi_{j1}^{(t)}|, |\xi_{j2}^{(t)}|).$$

Hence, for all $j \in [n]$,

$$M_j^C = |C|^r\, M_j. \tag{14}$$

Then, we have, for any $u > 0$,

$$\{j : M_j^C > u\} = \{j : |C|^r M_j > u\} = \{j : M_j > u/|C|^r\},$$

$$\{j : M_j^C < -u\} = \{j : |C|^r M_j < -u\} = \{j : M_j < -u/|C|^r\}.$$

The counterpart $\widehat{\mathsf{FDP}}^C$ of $\widehat{\mathsf{FDP}}$ satisfies

$$\widehat{\mathsf{FDP}}^C(u) = \frac{\#\{j : M_j^C < -u\}}{\#\{j : M_j^C > u\} \vee 1} = \frac{\#\{j : M_j < -u/|C|^r\}}{\#\{j : M_j > u/|C|^r\} \vee 1} = \widehat{\mathsf{FDP}}\left(\frac{u}{|C|^r}\right). \tag{15}$$

From (15), it follows that

$$\{u > 0 : \widehat{\mathsf{FDP}}^C(u) \le \alpha\} = \{u > 0 : \widehat{\mathsf{FDP}}(u/|C|^r) \le \alpha\}$$
$$= \{|C|^r v : v > 0,\ \widehat{\mathsf{FDP}}(v) \le \alpha\}.$$

Therefore, we obtain

$$\tau_\alpha^C = |C|^r\, \tau_\alpha. \tag{16}$$

By (14) and (16),

$$\hat{S}_\alpha^C = \{j : M_j^C > \tau_\alpha^C\} = \{j : |C|^r M_j > |C|^r \tau_\alpha\} = \{j : M_j > \tau_\alpha\} = \hat{S}_\alpha.$$

□

Corollary A.1. *Under Assumption 5, $(M_j)_{j\in[n]}$ and $(\breve{M}_j)_{j\in[n]}$ yield the same selection result after applying Algorithm 1.*

Proof of Corollary A.1. Applying Lemma A.3 with $C = \sqrt{n}/\|\boldsymbol{P}_{\boldsymbol{B}}^\perp \boldsymbol{\xi}_1^{(t)}\|$ proves the claim if $n \ge 1$, $\boldsymbol{B} \ne \boldsymbol{0}_{n\times q^*}$, and $\boldsymbol{\xi}_1^{(t)} \ne \boldsymbol{0}_n$. Note that the convergence of $\boldsymbol{\xi}_1^{(t)}$ to zero is allowed. □

Lemma A.4. *Under the assumptions of Theorem 2, as $n \to \infty$ while $q^* = o(n)$, we have*

$$\sup_{u\in\mathbb{R},\, j\in S^{\mathsf{c}}} \left|\mathbb{P}(\breve{M}_j > u) - \mathbb{P}(\tilde{M}_j > u)\right| \to 0.$$

Proof of Lemma A.4. Define

$$\Delta_j = \sup_{u\in\mathbb{R}} |\mathbb{P}(\varsigma_{j1}^{(t)} > u) - \mathbb{P}(z_{j1} > u)| \ \vee\ \sup_{u\in\mathbb{R}} |\mathbb{P}(\varsigma_{j2}^{(t)} > u) - \mathbb{P}(z_{j1} > u)|.$$

Without loss of generality, we assume $u > 0$. Thus, by the non-negativeness of $\psi(\cdot,\cdot)$, we have

$$\{\breve{M}_j > u\} \iff \left(\left\{\psi(|\varsigma_{j1}^{(t)}|, |\varsigma_{j2}^{(t)}|) > u\right\} \cap \{\varsigma_{j1}^{(t)} > 0\}\right) \cup \left(\left\{-\psi(|\varsigma_{j1}^{(t)}|, |\varsigma_{j2}^{(t)}|) > u\right\} \cap \{\varsigma_{j1}^{(t)} \le 0\}\right).$$

Using this and the monotonicity of $\psi(\cdot,\cdot)$, for any $t \in \mathbb{N}$, we have

$$\begin{aligned}
\mathbb{P}(\breve{M}_j > u) &= \mathbb{P}\left(\varsigma_{j2}^{(t)} > I_u(\varsigma_{j1}^{(t)}),\ \varsigma_{j1}^{(t)} > 0\right) + \mathbb{P}\left(\varsigma_{j2}^{(t)} < -I_u(\varsigma_{j1}^{(t)}),\ \varsigma_{j1}^{(t)} < 0\right) \\
&\le \mathbb{P}\left(z_{j2} > I_u(\varsigma_{j1}^{(t)}),\ \varsigma_{j1}^{(t)} > 0\right) + \mathbb{P}\left(z_{j2} < -I_u(\varsigma_{j1}^{(t)}),\ \varsigma_{j1}^{(t)} < 0\right) + 2\Delta_j \\
&= \mathbb{P}\left(\mathsf{sign}(\varsigma_{j1}^{(t)} z_{j2})\psi(|\varsigma_{j1}^{(t)}|, |z_{j2}|) > u\right) + 2\Delta_j \\
&= \mathbb{P}\left(\varsigma_{j1}^{(t)} > I_u(z_{j2}),\ z_{j2} > 0\right) + \mathbb{P}\left(\varsigma_{j1}^{(t)} < -I_u(z_{j2}),\ z_{j2} < 0\right) + 2\Delta_j \\
&\le \mathbb{P}\left(z_{j1} > I_u(z_{j2}),\ z_{j2} > 0\right) + \mathbb{P}\left(z_{j1} < -I_u(z_{j2}),\ z_{j2} < 0\right) + 4\Delta_j \\
&= \mathbb{P}(\tilde{M}_j > u) + 4\Delta_j,
\end{aligned}$$

where the first inequality follows from $\varsigma_{j1}^{(t)} \overset{\mathrm{d}}{=} \varsigma_{j2}^{(t)}$ by Assumption 6, and the third equality follows from the symmetry of $\psi(\cdot,\cdot)$. Hence, Theorem 1 implies that

$$\sup_{u\in\mathbb{R},\, j\in S^{\mathsf{c}}} \left|\mathbb{P}(\breve{M}_j > u) - \mathbb{P}(\tilde{M}_j > u)\right| \le 4 \sup_{u\in\mathbb{R},\, j\in S^{\mathsf{c}}} |\Delta_j| \to 0,$$

as $n \to \infty$ with $q^* = o(n)$. □

Lemma A.5. *Under the assumptions of Theorem 1, as $n \to \infty$ while $q^* = o(n)$, we have,*

$$\sup_{i,j\in S^{\mathsf{c}},\, t_1,t_2\in\mathbb{R}} \left|\mathbb{P}\left(\frac{\sqrt{n}\xi_i^{(t)}}{\|\boldsymbol{P}_{\boldsymbol{B}}^\perp \boldsymbol{\xi}^{(t)}\|} < t_1,\ \frac{\sqrt{n}\xi_j^{(t)}}{\|\boldsymbol{P}_{\boldsymbol{B}}^\perp \boldsymbol{\xi}^{(t)}\|} < t_2\right) - \Phi(t_1)\Phi(t_2)\right| \to 0.$$

Proof of Lemma A.5. We assume, without loss of generality, that $t_1 > 0$ and $t_2 > 0$. Using the fact that $\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}\| \overset{\mathrm{d}}{=} \boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{z}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{z}\|$ where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ by Theorem 1, we have,

$$\left(\frac{\sqrt{n}\xi_i^{(t)}}{\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}\|}, \frac{\sqrt{n}\xi_j^{(t)}}{\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}\|}\right) \overset{\mathrm{d}}{=} \left(\frac{\sqrt{n}z_i}{\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{z}\|}, \frac{\sqrt{n}z_j}{\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{z}\|}\right), \tag{17}$$

for $i, j \in S^{\mathsf{c}}$. This follows from the fact that the j-th element of $\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}\|$ is $\xi_j^{(t)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}\|$ under the null since $\boldsymbol{b}_j = \boldsymbol{0}$ for $j \in S^{\mathsf{c}}$. Denote $\sigma_n = \sqrt{n}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{z}\|$. From (17), it follows that

$$\begin{aligned}\Delta_n &\equiv \sup_{i,j\in S^{\mathsf{c}},\, t_1,t_2\in\mathbb{R}} \left|\mathbb{P}\left(\frac{\sqrt{n}\xi_i^{(t)}}{\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}\|} < t_1, \frac{\sqrt{n}\xi_j^{(t)}}{\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(t)}\|} < t_2\right) - \Phi(t_1)\Phi(t_2)\right| \\ &= \sup_{i,j\in S^{\mathsf{c}},\, t_1,t_2\in\mathbb{R}} \left|\mathbb{P}\left(z_i < \sigma_n t_1, z_j < \sigma_n t_2\right) - \Phi(t_1)\Phi(t_2)\right|.\end{aligned}$$

Fix an arbitrary $\epsilon > 0$ and denote $E_n = \{|\sigma_n - 1| < \epsilon\}$. Then we have

$$\begin{aligned}&\left|\mathbb{P}\left(z_i < \sigma_n t_1, z_j < \sigma_n t_2\right) - \Phi(t_1)\Phi(t_2)\right| \\ &\le \mathbb{P}(E_n^{\mathsf{c}}) + \left|\mathbb{P}\left(z_i < \sigma_n t_1, z_j < \sigma_n t_2 \mid E_n\right) - \Phi(t_1)\Phi(t_2)\right|.\end{aligned} \tag{18}$$

Also, since $\sigma_n \in [1-\epsilon, 1+\epsilon]$ under E_n, it follows that

$$\mathbb{P}\left(z_i < \frac{t_1}{1+\epsilon}, z_j < \frac{t_2}{1+\epsilon}\right) \le \mathbb{P}\left(z_i < \sigma_n t_1, z_j < \sigma_n t_2 \mid E_n\right) \le \mathbb{P}\left(z_i < \frac{t_1}{1-\epsilon}, z_j < \frac{t_2}{1-\epsilon}\right).$$

From this, we have

$$\begin{aligned}&\left|\mathbb{P}\left(z_i < \sigma_n t_1, z_j < \sigma_n t_2 \mid E_n\right) - \Phi(t_1)\Phi(t_2)\right| \\ &\le \max\left\{\Phi\left(\frac{t_1}{1-\epsilon}\right)\Phi\left(\frac{t_2}{1-\epsilon}\right) - \Phi(t_1)\Phi(t_2),\ \Phi(t_1)\Phi(t_2) - \Phi\left(\frac{t_1}{1+\epsilon}\right)\Phi\left(\frac{t_2}{1+\epsilon}\right)\right\}.\end{aligned}$$

Taking the supremum, the mean-value theorem gives

$$\sup_{t_1,t_2\in\mathbb{R}} \left|\mathbb{P}\left(z_i < \sigma_n t_1, z_j < \sigma_n t_2 \mid E_n\right) - \Phi(t_1)\Phi(t_2)\right| \le \frac{\epsilon}{1-\epsilon}\cdot\sqrt{\frac{2}{\pi e}}, \tag{19}$$

where we use the fact $\sup_{u\in\mathbb{R}} |u|\phi(u) = 1/\sqrt{2\pi e}$. Since this upper bound does not depend on $j \in [n]$, (18) and (19) yield

$$\Delta_n \le \mathbb{P}(|\sigma_n - 1| > \epsilon) + \frac{\epsilon}{1-\epsilon}\cdot\sqrt{\frac{2}{\pi e}}.$$

The first term on the right-hand side converges to zero as $n \to \infty$ while $q^* = o(n)$ by Lemma A.1, and the second term goes to zero as $\epsilon \downarrow 0$. □

Lemma A.6. *Let n_0 be the number of null features. Under the assumptions of Theorem 2, as $n \to \infty$ while $q^* = o(n)$, we have*

$$\sup_{u\in\mathbb{R}} \mathrm{Var}\left(\frac{1}{n_0}\sum_{j\in S^{\mathsf{c}}} \mathbb{1}(\breve{M}_j > u)\right) \le \frac{1}{4n_0} + o(1).$$

Proof of Lemma A.6. We assume $u > 0$ without loss of generality. It follows that

$$\begin{aligned}&\sup_{u\in\mathbb{R}} \mathrm{Var}\left(\frac{1}{n_0}\sum_{j\in S^{\mathsf{c}}} \mathbb{1}(\breve{M}_j > u)\right) \\ &\le \frac{1}{n_0^2}\sum_{j\in S^{\mathsf{c}}} \sup_{u\in\mathbb{R}} \mathrm{Var}\left(\mathbb{1}(\breve{M}_j > u)\right) + \frac{1}{n_0^2}\sum_{j\neq j'\in S^{\mathsf{c}}} \sup_{u\in\mathbb{R}} \mathrm{Cov}\left(\mathbb{1}(\breve{M}_j > u), \mathbb{1}(\breve{M}_{j'} > u)\right),\end{aligned}$$

where $\mathbb{1}(\breve{M}_j > u)$ is a Bernoulli variable, and its variance is bounded above by $1/4$. Thus, the first term on the right-hand side is upper bounded by $1/(4n_0)$. For the first term, we have

$$\begin{aligned}
&\mathrm{Cov}\left(\mathbb{1}(\breve{M}_j > u), \mathbb{1}(\breve{M}_{j'} > u)\right) \\
&= \mathbb{P}(\breve{M}_j > u, \breve{M}_{j'} > u) - \mathbb{P}(\breve{M}_j > u)\mathbb{P}(\breve{M}_{j'} > u) \\
&\leq \left|\mathbb{P}(\breve{M}_j > u, \breve{M}_{j'} > u) - \mathbb{P}(\tilde{M}_j > u)^2\right| + \left|\mathbb{P}(\breve{M}_j > u)\mathbb{P}(\breve{M}_{j'} > u) - \mathbb{P}(\tilde{M}_j > u)^2\right|,
\end{aligned} \tag{20}$$

where the second term on the right-hand side converges to zero uniformly on $j \in S^{\mathsf{c}}$ and $u \in \mathbb{R}$ by Lemma A.4.

Repeating the argument in the proof of Lemma A.4, it follows that

$$\begin{aligned}
\mathbb{P}(\breve{M}_j > u, \breve{M}_{j'} > u) &= \mathbb{P}\left(\varsigma_{j'2}^{(t)} > I_u(\varsigma_{j'1}^{(t)}),\, \varsigma_{j2}^{(t)} > I_u(\varsigma_{j1}^{(t)}),\, \varsigma_{j'1}^{(t)} > 0,\, \varsigma_{j1}^{(t)} > 0\right) \\
&+ \mathbb{P}\left(\varsigma_{j'2}^{(t)} > I_u(\varsigma_{j'1}^{(t)}),\, \varsigma_{j2}^{(t)} < -I_u(\varsigma_{j1}^{(t)}),\, \varsigma_{j'1}^{(t)} > 0,\, \varsigma_{j1}^{(t)} < 0\right) \\
&+ \mathbb{P}\left(\varsigma_{j'2}^{(t)} < -I_u(\varsigma_{j'1}^{(t)}),\, \varsigma_{j2}^{(t)} > I_u(\varsigma_{j1}^{(t)}),\, \varsigma_{j'1}^{(t)} < 0,\, \varsigma_{j1}^{(t)} > 0\right) \\
&+ \mathbb{P}\left(\varsigma_{j'2}^{(t)} < -I_u(\varsigma_{j'1}^{(t)}),\, \varsigma_{j2}^{(t)} < -I_u(\varsigma_{j1}^{(t)}),\, \varsigma_{j'1}^{(t)} < 0,\, \varsigma_{j1}^{(t)} < 0\right) \\
&\equiv I_1 + I_2 + I_3 + I_4.
\end{aligned} \tag{21}$$

Define

$$\Delta = \sup_{j \in S^{\mathsf{c}},\, u \in \mathbb{R}} |\mathbb{P}(\varsigma_{j1}^{(t)} > u) - \mathbb{P}(z_{j1} > u)| \;\vee\; \sup_{j \in S^{\mathsf{c}},\, u \in \mathbb{R}} |\mathbb{P}(\varsigma_{j2}^{(t)} > u) - \mathbb{P}(z_{j1} > u)|.$$

Let $Q(u) = 1 - \Phi(u)$. For I_1, we have the following upper bound,

$$\begin{aligned}
I_1 &= \mathbb{E}\left[\mathbb{P}(\varsigma_{j'2}^{(t)} > I_u(x),\, \varsigma_{j2}^{(t)} > I_u(y)) \mid \varsigma_{j'1}^{(t)} = x > 0,\, \varsigma_{j1}^{(t)} = y > 0\right] \\
&\leq \mathbb{E}\left[\mathbb{P}(z_{j'2} > I_u(x),\, z_{j2} > I_u(y)) \mid \varsigma_{j'1}^{(t)} = x > 0,\, \varsigma_{j1}^{(t)} = y > 0\right] + 2\Delta \\
&= \mathbb{E}\left[Q(I_u(x))Q(I_u(y)) \mid \varsigma_{j'1}^{(t)} = x > 0,\, \varsigma_{j1}^{(t)} = y > 0\right] + 2\Delta,
\end{aligned} \tag{22}$$

where the inequality follows from Lemma A.5. Similarly, we can upper bound I_2, I_3, and I_4. Combining the four upper bounds together, we obtain an upper bound on $\mathbb{P}(\breve{M}_j > u, \breve{M}_{j'} > u)$ as

$$\mathbb{P}\left(\mathsf{sign}(z_{j2}\varsigma_{j1}^{(t)})\psi(|z_{j2}|, |\varsigma_{j1}^{(t)}|) > u,\; \mathsf{sign}(z_{j'2}\varsigma_{j'1}^{(t)})\psi(|z_{j'2}|, |\varsigma_{j'1}^{(t)}|) > u\right) + 8\Delta.$$

We can further decompose this into four terms as (21) by conditioning the signs of z_{j2} and $z_{j'2}$, and repeat the upper bound (22). This leads to

$$\mathbb{P}(\breve{M}_j > u, \breve{M}_{j'} > u) \leq \mathbb{P}(\tilde{M}_j > u)^2 + 16\Delta.$$

Similarly, we can show the corresponding lower bound. Since $\Delta \to 0$ by Lemma A.4, the covariance in (20) converges to zero. Note that we neither show nor use the joint asymptotic equivalence of $(\varsigma_{11}^{(t)}, \ldots, \varsigma_{n1}^{(t)}, \varsigma_{12}^{(t)}, \ldots, \varsigma_{n2}^{(t)})$ and $(z_{11}^{(t)}, \ldots, z_{n1}^{(t)}, z_{12}^{(t)}, \ldots, z_{n2}^{(t)})$ in the proof. □

Lemma A.7. *Under the assumptions of Theorem 2, we have, as $n \to \infty$ while $q^* = o(n)$,*

$$\sup_{u \in \mathbb{R}} \left| \frac{1}{n_0} \sum_{j \in S^{\mathsf{c}}} \mathbb{1}(\breve{M}_j > u) - \mathbb{P}(\tilde{M}_1 > u) \right| \xrightarrow{\mathrm{p}} 0.$$

Proof of Lemma A.7. We have

$$\begin{aligned}
&\sup_{u \in \mathbb{R}} \left| \frac{1}{n_0} \sum_{j \in S^{\mathsf{c}}} \mathbb{1}(\breve{M}_j > u) - \mathbb{P}(\tilde{M}_1 > u) \right| \\
&\leq \sup_{u \in \mathbb{R}} \left| \frac{1}{n_0} \sum_{j \in S^{\mathsf{c}}} \left\{\mathbb{1}(\breve{M}_j > u) - \mathbb{P}(\breve{M}_j > u)\right\} \right| + \sup_{u \in \mathbb{R}} \left| \frac{1}{n_0} \sum_{j \in S^{\mathsf{c}}} \mathbb{P}(\breve{M}_j > u) - \mathbb{P}(\tilde{M}_1 > u) \right|,
\end{aligned}$$

where the second term on the right-hand side converges to zero by Lemma A.4. Also, Chebyshev's inequality yields, for any $v \in \mathbb{R}$,

$$\sup_{u\in\mathbb{R}} \mathbb{P}\left(\left|\frac{1}{n_0}\sum_{j\in S^c}\Big\{\mathbb{1}(\breve{M}_j > u) - \mathbb{P}(\breve{M}_j > u)\Big\}\right| > v\right) \le \frac{1}{v^2}\sup_{u\in\mathbb{R}} \mathrm{Var}\left(\frac{1}{n_0}\sum_{j\in S^c}\mathbb{1}(\breve{M}_j > u)\right),$$

where the supremum of variance converges to zero by Lemma A.6. This completes the proof. □

Lemma A.8. *Under the assumptions of Theorem 2, there exists a constant $\delta_0 = \delta_0(\alpha, c, \theta, \rho, \pi_0) > 0$ such that*

$$\mathbb{P}\big(\,\tilde{F}(\tau_\alpha^{\sigma_n}) \ge \delta_0\,\big) \;\longrightarrow\; 1.$$

Moreover, one can take explicitly

$$\delta_* := \frac{(1-\theta)c}{2\pi_0}, \qquad U := \frac{(\alpha\rho - (1-\rho))\,\theta c}{\pi_0(1-\alpha)}, \qquad \delta_0 := \frac{\delta_* + U}{2} \in (0,1).$$

Proof. Define the generalized inverse $\tilde{F}^{\leftarrow}(\delta) := \inf\{u \ge 0 :\ \tilde{F}(u) < \delta\}$ (so $\tilde{F}(\tilde{F}^{\leftarrow}(\delta)) \ge \delta$). Note that $\tau_\alpha^{\sigma_n}$, V^+, and V^- are defined at the top of this section.

Step 1: An upper bound for $\tilde{F}(u_{K_n})$. By Assumption 7, $S^{\pm}(u_{K_n}) \ge \theta K_n$ with probability $1 - o(1)$; hence the number of nulls among the top K_n magnitudes satisfies

$$V^+(u_{K_n}) + V^-(u_{K_n}) \;\le\; (1-\theta)K_n.$$

By Lemma A.7, $V^+(u) + V^-(u) = 2n_0\tilde{F}(u) + o_p(n_0)$ uniformly in u. Since n_0 and n are of the same order,

$$2n_0\tilde{F}(u_{K_n}) \;\le\; (1-\theta)K_n + o_p(n),$$

and dividing both sides by n yields

$$\tilde{F}(u_{K_n}) \;\le\; \delta_* + o_p(1). \tag{23}$$

Step 2: Construct a subthreshold u_0. Let $\delta_0 := (\delta_* + U)/2$, which satisfies $\delta_* < \delta_0 < U$ by (5). By (23) and the monotonicity of $\tilde{F}$, we have $u_0 := \tilde{F}^{\leftarrow}(\delta_0) < u_{K_n}$ with probability $1 - o(1)$. Hence, by monotonicity again,

$$S^{\pm}(u_0) \;\ge\; S^{\pm}(u_{K_n}) \;\ge\; \theta K_n, \tag{24}$$

with probability approaching one. Moreover, Assumption 7 (applied for all $u \le u_{K_n}$) gives

$$S^+(u_0) \;\ge\; \rho\, S^{\pm}(u_0), \qquad S^-(u_0) \;\le\; (1-\rho)\, S^{\pm}(u_0),$$

with probability approaching one. Lemma A.7 implies

$$V^+(u_0) \;=\; n_0\,\delta_0 + o_p(n_0). \tag{25}$$

Step 3: Upper bound for $\widehat{\mathsf{FDP}}^{\sigma_n}(u_0)$. Let $\widehat{\mathsf{FDP}}^{\sigma_n}(u)$ denote the version of $\widehat{\mathsf{FDP}}(u)$ with $\breve{M}_j$ substituted for M_j in the definition. Using (24)–(25), Lemma A.2, and $K_n = cn + o(n)$,

$$\widehat{\mathsf{FDP}}^{\sigma_n}(u_0) = \frac{1 + V^-(u_0) + S^-(u_0)}{V^+(u_0) + S^+(u_0)} \;\le\; \frac{1 + n_0\delta_0 + (1-\rho)S^{\pm}(u_0) + o_p(n)}{n_0\delta_0 + \rho S^{\pm}(u_0) + o_p(n)}.$$

Divide numerator and denominator by n and pass to lim sup using $n_0/n \to \pi_0$ and (24):

$$\limsup_{n\to\infty}\ \widehat{\mathsf{FDP}}^{\sigma_n}(u_0) \;\le\; \frac{\pi_0\delta_0 + (1-\rho)\theta c}{\pi_0\delta_0 + \rho\theta c} \;=:\; \Psi(\delta_0).$$

By the definition of U and the equivalence

$$\Psi(\delta) \le \alpha \iff \pi_0(1-\alpha)\,\delta \le (\alpha\rho - (1-\rho))\,\theta c,$$

we have $\Psi(\delta_0) < \alpha$ because $\delta_0 < U$. Hence there exists $\varepsilon > 0$ such that

$$\mathbb{P}\big(\widehat{\mathsf{FDP}}^{\sigma_n}(u_0) \le \alpha - \varepsilon\big) \;\to\; 1. \tag{26}$$

Step 4: Compare $\tau_\alpha^{\sigma_n}$ to u_0. By definition $\tau_\alpha^{\sigma_n} := \inf\{u > 0 : \widehat{\mathsf{FDP}}^{\sigma_n}(u) \le \alpha\}$. From (26) we obtain $\tau_\alpha^{\sigma_n} \le u_0$ with probability $1 - o(1)$. Since $\tilde{F}$ is nonincreasing and $\tilde{F}(\tilde{F}^{\leftarrow}(\delta_0)) \ge \delta_0$, we conclude

$$\tilde{F}(\tau_\alpha^{\sigma_n}) \;\ge\; \tilde{F}(u_0) \;\ge\; \delta_0,$$

with probability $1 - o(1)$, which proves the claim. □

Proof of Theorem 2. To begin with, we have, by Corollary A.1,

$$\begin{aligned}
\mathsf{FDR} &= \mathbb{E}\left[\frac{\#\{j \in S^{\mathsf{c}} : \check{M}_j > \tau_\alpha^{\sigma_n}\}}{\#\{j : \check{M}_j > \tau_\alpha^{\sigma_n}\} \vee 1}\right] = \mathbb{E}\left[\frac{V^-(\tau_\alpha^{\sigma_n})}{V^+(\tau_\alpha^{\sigma_n}) + S^+(\tau_\alpha^{\sigma_n})}\right] \\
&\le \mathbb{E}\left[\frac{1 + V^-(\tau_\alpha^{\sigma_n}) + S^-(\tau_\alpha^{\sigma_n})}{V^+(\tau_\alpha^{\sigma_n}) + S^+(\tau_\alpha^{\sigma_n})} + \frac{|V^+(\tau_\alpha^{\sigma_n}) - V^-(\tau_\alpha^{\sigma_n})|}{V^+(\tau_\alpha^{\sigma_n}) + S^+(\tau_\alpha^{\sigma_n})}\right] \\
&\le \alpha + \mathbb{E}\left[\frac{|V^+(\tau_\alpha^{\sigma_n}) - V^-(\tau_\alpha^{\sigma_n})|}{V^+(\tau_\alpha^{\sigma_n}) + S^+(\tau_\alpha^{\sigma_n})}\right],
\end{aligned} \tag{27}$$

where the last inequality follows from $\widehat{\mathsf{FDP}}^{\sigma_n}(\tau_\alpha^{\sigma_n}) \le \alpha$ by construction. Since $\mathsf{FDR} = 0$ when $V^+(\tau_\alpha^{\sigma_n}) + S^+(\tau_\alpha^{\sigma_n}) = 0$, we consider the case $V^+(\tau_\alpha^{\sigma_n}) + S^+(\tau_\alpha^{\sigma_n}) \ge 1$. By Lemma A.7 and Lemma A.2, we have, as $n \to \infty$,

$$\frac{1}{n_0}\left|V^+(\tau_\alpha^{\sigma_n}) - V^-(\tau_\alpha^{\sigma_n})\right| \le \frac{1}{n_0}\left|V^+(\tau_\alpha^{\sigma_n}) - n_0\tilde{F}(\tau_\alpha^{\sigma_n})\right| + \frac{1}{n_0}\left|V^-(\tau_\alpha^{\sigma_n}) - n_0\tilde{F}(\tau_\alpha^{\sigma_n})\right| = o_p(1). \tag{28}$$

Also, since $n_0^{-1}V^+(\tau_\alpha^{\sigma_n}) = \tilde{F}(\tau_\alpha^{\sigma_n}) + o_p(1)$ by Lemma A.7, Lemma A.8 implies that, as $n \to \infty$,

$$\frac{1}{n_0}V^+(\tau_\alpha^{\sigma_n}) \ge \delta_0 + o_p(1). \tag{29}$$

Therefore, from (28) and (29), we have

$$\frac{|V^+(\tau_\alpha^{\sigma_n}) - V^-(\tau_\alpha^{\sigma_n})|}{V^+(\tau_\alpha^{\sigma_n}) + S^+(\tau_\alpha^{\sigma_n})} = o_p(1). \tag{30}$$

Since the left-hand side is bounded by one, (27), (30), and the bounded convergence theorem yield $\mathsf{FDR} \le \alpha + o(1)$. □

A.4 A necessary and sufficient condition for the conditional null

For probability measures μ and ν on $\mathbb{R}$, let $W_1(\mu,\nu)$ denote the 1-Wasserstein distance and $d_{\mathrm{TV}}(\mu,\nu)$ the total variation distance. Write $\mathsf{BL}_1 := \{\varphi : \mathbb{R} \to \mathbb{R} : \ \|\varphi\|_\infty \le 1,\ \mathrm{Lip}(\varphi) \le 1\}$ for bounded 1-Lipschitz functions.

We consider the multi-index model

$$y \;=\; g(\boldsymbol{B}^\top \boldsymbol{x},\ \varepsilon), \qquad \boldsymbol{x} \in \mathbb{R}^n,\ y \in \mathbb{R},\ \boldsymbol{B} \in \mathbb{R}^{n \times q^*},$$

with $\varepsilon \perp\!\!\!\perp \boldsymbol{x}$, $1 \le q^* < n$, and $\mathrm{rank}(\boldsymbol{B}) = q^*$. Let the j-th row be $\boldsymbol{b}_j^\top \in \mathbb{R}^{q^*}$, and set $\boldsymbol{u} := \boldsymbol{B}^\top \boldsymbol{x} \in \mathbb{R}^{q^*}$.

Assumption 8 (Minimal thickness of the projected regressor)**.** *There exists a nonempty open set $O \subset \mathbb{R}^{q^*}$ such that the law of $\boldsymbol{u}$ admits a Lebesgue density strictly positive on O.*

Assumption 9 (Local kernel Lipschitzness in W_1)**.** *Let $K(u) := \mathcal{L}(y \mid \boldsymbol{u} = u)$ be the conditional law (a stochastic kernel). There exists $L > 0$ such that*

$$W_1\big(K(u), K(u')\big) \;\le\; L\,\|u - u'\| \qquad \textit{for all } u, u' \in O.$$

Assumption 10 (Bounded-Lipschitz nondegeneracy)**.** *For every nonzero $\boldsymbol{v} \in \mathbb{R}^{q^*}$ there exist $\varphi \in \mathsf{BL}_1$ and a measurable set $A_{\boldsymbol{v}} \subset O$ with positive Lebesgue measure such that the directional derivative $D_{\boldsymbol{v}} m_\varphi(u)$ exists for Lebesgue-a.e. $u \in A_{\boldsymbol{v}}$ and is not a.e. zero on $A_{\boldsymbol{v}}$, where $m_\varphi(u) \equiv \mathbb{E}[\varphi(y) \mid \boldsymbol{u} = u]$.*

Assumption 11 (Conditional thickness of x_j given $\boldsymbol{x}_{-j}$)**.** *For the fixed index $j \in \{1, \dots, n\}$ under consideration, write $\boldsymbol{x} = (x_j, \boldsymbol{x}_{-j})$. For almost every $\boldsymbol{a}_{-j} \in \mathbb{R}^{n-1}$, the conditional law $\mathcal{L}(x_j \mid \boldsymbol{x}_{-j} = \boldsymbol{a}_{-j})$ has a Lebesgue density that is strictly positive on some nonempty open set $K(\boldsymbol{a}_{-j}) \subset \mathbb{R}$.*

Definition 3 (Conditional null for variable j)**.** *We say that the conditional null holds for the index $j \in [n]$ if*

$$y \perp\!\!\!\perp x_j \mid \boldsymbol{x}_{-j}.$$

Lemma A.9. *Under Assumption 9, for every $\varphi \in \mathsf{BL}_1$, the map $m_\varphi(u) = \mathbb{E}[\varphi(y) \mid \boldsymbol{u} = u]$ is L-Lipschitz on O. In particular, m_φ is differentiable almost everywhere on O.*

Proof. By the Kantorovich–Rubinstein duality for W_1 on Polish spaces,

$$|m_\varphi(u) - m_\varphi(u')| = \Big| \int \varphi \, dK(u) - \int \varphi \, dK(u') \Big| \le W_1(K(u), K(u')) \le L\|u - u'\|,$$

since $\varphi \in \mathsf{BL}_1$ is 1-Lipschitz and bounded. Rademacher's theorem ensures almost-everywhere differentiability of Lipschitz maps $m_\varphi : O \to \mathbb{R}$. □

Lemma A.10. *Fix $j \in \{1, \dots, n\}$. Under Assumptions 8, 11, and the conditional null for j in Definition 3, fix $\varphi \in \mathsf{BL}_1$ and write*

$$\boldsymbol{w} := \sum_{k \neq j} \boldsymbol{b}_k x_k, \qquad \boldsymbol{v} := \boldsymbol{b}_j x_j,$$

so that $\boldsymbol{u} = \boldsymbol{w} + \boldsymbol{v}$. Then for almost every $\boldsymbol{a}_{-j} \in \mathbb{R}^{n-1}$, there exists a nonempty open set

$$G(\boldsymbol{a}_{-j}) \ \subset \ \boldsymbol{w}(\boldsymbol{a}_{-j}) + \operatorname{span}\{\boldsymbol{b}_j\}$$

such that $m_\varphi(u)$ is almost everywhere constant on $G(\boldsymbol{a}_{-j})$.

Proof. By $\boldsymbol{x} \perp\!\!\!\perp \varepsilon$ and the definition of K,

$$\mathbb{E}[\varphi(y) \mid x_j, \boldsymbol{x}_{-j}] \ = \ \mathbb{E}[\varphi(y) \mid \boldsymbol{u}] \ = \ m_\varphi(\boldsymbol{u}).$$

The conditional null $y \perp\!\!\!\perp x_j \mid \boldsymbol{x}_{-j}$ implies

$$\mathbb{E}[\varphi(y) \mid x_j, \boldsymbol{x}_{-j}] = \mathbb{E}[\varphi(y) \mid \boldsymbol{x}_{-j}],$$

hence

$$m_\varphi(\boldsymbol{w} + \boldsymbol{v}) = h_\varphi(\boldsymbol{x}_{-j}) \quad \text{a.s.}$$

for some measurable $h_\varphi(\cdot)$.

By Assumption 11, for almost every $\boldsymbol{a}_{-j}$ the conditional support of $x_j \mid \boldsymbol{x}_{-j} = \boldsymbol{a}_{-j}$ contains a nonempty open set $K(\boldsymbol{a}_{-j}) \subset \mathbb{R}$. Consider the linear map

$$T : \ \mathbb{R} \to \mathbb{R}^{q^*}, \qquad t \mapsto \boldsymbol{b}_j t.$$

Its image is the subspace $\operatorname{span}\{\boldsymbol{b}_j\}$, and T is open onto its image in finite dimensions. Therefore, for almost every $\boldsymbol{a}_{-j}$, the image

$$H(\boldsymbol{a}_{-j}) := T\big(K(\boldsymbol{a}_{-j})\big)$$

is a nonempty open set inside $\operatorname{span}\{\boldsymbol{b}_j\}$ (with respect to the subspace topology). Consequently,

$$G(\boldsymbol{a}_{-j}) := \boldsymbol{w}(\boldsymbol{a}_{-j}) + H(\boldsymbol{a}_{-j})$$

is a nonempty open set within the affine subspace $\boldsymbol{w}(\boldsymbol{a}_{-j}) + \operatorname{span}\{\boldsymbol{b}_j\}$ on which m_φ is almost everywhere constant (equal to $h_\varphi(\boldsymbol{a}_{-j})$). □

Theorem 3. *Fix $j \in \{1, \dots, n\}$. Under Assumptions 8, 9, 10, and 11, the following are equivalent:*

$$y \perp\!\!\!\perp x_j \mid \boldsymbol{x}_{-j} \qquad \Longleftrightarrow \qquad \boldsymbol{b}_j = \boldsymbol{0}\,.$$

Proof. ($\Leftarrow$) If $\boldsymbol{b}_j = \boldsymbol{0}$, then

$$\boldsymbol{u} = \sum_{k \neq j} \boldsymbol{b}_k x_k$$

is $\sigma(\boldsymbol{x}_{-j})$-measurable. By $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{\varepsilon}$, the conditional law

$$\mathcal{L}(y \mid x_j, \boldsymbol{x}_{-j}) = \mathcal{L}\big(g(\boldsymbol{u}, \varepsilon) \mid x_j, \boldsymbol{x}_{-j}\big) = K(\boldsymbol{u})$$

does not depend on x_j, hence $y \perp\!\!\!\perp x_j \mid \boldsymbol{x}_{-j}$.

($\Rightarrow$) Suppose $y \perp\!\!\!\perp x_j \mid \boldsymbol{x}_{-j}$. Fix $\varphi \in \mathsf{BL}_1$. By Lemma A.10, for almost every $\boldsymbol{a}_{-j}$, m_φ is almost everywhere constant on a nonempty open set inside the affine subspace $\boldsymbol{w}(\boldsymbol{a}_{-j}) + \operatorname{span}\{\boldsymbol{b}_j\}$. By Assumption 8, these affine pieces intersect O on sets of positive Lebesgue measure in $\mathbb{R}^{q^*}$; by Lemma A.9, m_φ is locally Lipschitz on O, hence (Rademacher) directionally differentiable almost everywhere on O. Consequently,

$$D_{\boldsymbol{v}} m_\varphi(u) = 0 \quad \text{for Lebesgue-a.e. } u \in O \;\text{ and all } \boldsymbol{v} \in \operatorname{span}\{\boldsymbol{b}_j\}.$$

If $\operatorname{span}\{\boldsymbol{b}_j\} \neq \{\boldsymbol{0}\}$, then there exists a nonzero $\boldsymbol{v} \in \operatorname{span}\{\boldsymbol{b}_j\}$. Assumption 10 then yields some $\tilde{\varphi} \in \mathsf{BL}_1$ and a measurable $A_{\boldsymbol{v}} \subset O$ of positive Lebesgue measure such that $D_{\boldsymbol{v}} m_{\tilde{\varphi}}(u)$ exists for Lebesgue-a.e. $u \in A_{\boldsymbol{v}}$ and is not almost everywhere zero on $A_{\boldsymbol{v}}$, which contradicts the conclusion above (applied to $\tilde{\varphi}$). Thus necessarily $\operatorname{span}\{\boldsymbol{b}_j\} = \{\boldsymbol{0}\}$, i.e., $\boldsymbol{b}_j = \boldsymbol{0}$. □

Remark A.1 (TV-variant for classification)**.** *Assumption 9 can be replaced by the TV version $d_{\mathrm{TV}}(K(u), K(u')) \leq L\|u - u'\|$ on O; then $|m_\varphi(u) - m_\varphi(u')| \leq d_{\mathrm{TV}}(K(u), K(u'))$ for $\varphi \in \mathsf{BL}_1$, and Lemma A.9 and Theorem 3 remain valid with the same proof. In particular, for the ordinal binary classification model $y = \mathbb{1}\{h(u) + \varepsilon > 0\}$ (e.g., logistic regression and Probit model) with $q^* = 1$, $h(\cdot)$ locally Lipschitz and ε independent with bounded density, the kernel is TV-Lipschitz and Assumption 10 holds with $\varphi(y) = y$ whenever the class-probability $p(u) = \mathbb{P}(\varepsilon > -h(u))$ is not a.e. flat in any nonzero direction.*

B On the right-orthogonal invariance

In this section, we delineate what kinds of random designs fall into the class of *right-orthogonally invariant* (ROI) matrices, and what kinds do not. ROI is sometimes assumed in the literature of approximate message passing algorithms in the proportional asymptotics where n and m diverge with $m/n \to (0, \infty)$ (Rangan et al., 2019; Li and Sur, 2023). Subsequently, we discuss in what sense the $\boldsymbol{B}$-ROI in Assumption 1 (ii) is relaxed.

B.1 Definition and basic consequences

Definition 4 (Right-orthogonal invariance)**.** *A random matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ is* ROI *if*

$$\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{X}\boldsymbol{U} \qquad (\forall \boldsymbol{U} \in O(n)).$$

If $\mathbb{E}\|\boldsymbol{X}\|_F^2 < \infty$, then ROI implies the isotropy of the column Gram:

$$\mathbb{E}[\boldsymbol{X}^\top \boldsymbol{X}] = c\,\boldsymbol{I}_n, \qquad c = \frac{1}{n}\,\mathbb{E}\|\boldsymbol{X}\|_F^2, \tag{31}$$

which is a necessary (but not sufficient) condition for ROI. There are several closure properties.

Lemma B.1 (Left-multiplicative closure)**.** *Let $\boldsymbol{A}$ and $\boldsymbol{Z}$ be independent random matrices. If $\boldsymbol{Z}$ is ROI, then $\boldsymbol{X} := \boldsymbol{A}\boldsymbol{Z}$ is ROI.*

Lemma B.2 (Right Haar mixer)**.** *For any (possibly deterministic) $\boldsymbol{Y}$ and $\boldsymbol{Q} \sim \mathrm{Haar}(O(n))$ independent of $\boldsymbol{Y}$, $\boldsymbol{X} := \boldsymbol{Y}\boldsymbol{Q}$ is ROI.*

Lemma B.3 (Orthogonally conjugate mixture)**.** *Let $\boldsymbol{\Lambda}$ be a symmetric positive-definite random matrix, and suppose $(\boldsymbol{X}\boldsymbol{U} \mid \boldsymbol{\Lambda}) \overset{\mathrm{d}}{=} (\boldsymbol{X} \mid \boldsymbol{U}^\top \boldsymbol{\Lambda} \boldsymbol{U})$ and $\boldsymbol{\Lambda} \overset{\mathrm{d}}{=} \boldsymbol{U}^\top \boldsymbol{\Lambda} \boldsymbol{U}$ for all $\boldsymbol{U}$. Then the marginal $\boldsymbol{X}$ is ROI.*

B.2 Canonical ROI examples

Denote the Stiefel manifold $V_{n,r} \equiv \{\boldsymbol{W} : \boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_r\}$.

(E1) *Matrix-normal with isotropic columns.* If $\boldsymbol{X} \sim \mathcal{MN}(0, \boldsymbol{\Sigma}_{\text{row}}, \boldsymbol{I}_n)$, then $\boldsymbol{X}$ is ROI. Conversely, $\mathcal{MN}(0, \boldsymbol{\Sigma}_{\text{row}}, \boldsymbol{\Sigma}_{\text{col}})$ with $\boldsymbol{\Sigma}_{\text{col}} \not\propto \boldsymbol{I}_n$ is not ROI.

(E2) *Elliptical rows (after whitening).* If each row is elliptical $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$ with $\boldsymbol{z}_i$ spherically symmetric, then $\boldsymbol{Z} := \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ is ROI.

(E3) *Spiked with Haar loadings plus isotropic noise.* Let $\boldsymbol{X} = \alpha \boldsymbol{V}\boldsymbol{W}^\top + \boldsymbol{E}$, where $\boldsymbol{W} \in V_{n,r}$ is Haar and $\boldsymbol{E}$ is ROI (e.g., i.i.d. Gaussian). Then $\boldsymbol{X}$ is ROI (by left Haar-invariance of $\boldsymbol{W}$ and Lemma B.1).

(E4) *Linear multi-layer with an ROI rightmost factor.* If $\boldsymbol{X} = \boldsymbol{X}_1\boldsymbol{X}_2 \cdots \boldsymbol{X}_L$ with $\boldsymbol{X}_L$ i.i.d. standard Gaussian, then $\boldsymbol{X}$ is ROI (Lemma B.1).

(E5) *VAR with orthogonally invariant covariance mixing.* With $\boldsymbol{X}_{i,\cdot} = \sum_{k=1}^{\nu} \alpha_k \boldsymbol{X}_{i-k,\cdot} + \boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_i \mid \boldsymbol{\Sigma} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} \sim \text{InvWishart}(\boldsymbol{I}_n)$, one has $(\boldsymbol{X}\boldsymbol{U} \mid \boldsymbol{\Sigma}) \overset{\text{d}}{=} (\boldsymbol{X} \mid \boldsymbol{U}^\top \boldsymbol{\Sigma} \boldsymbol{U})$ and $\boldsymbol{\Sigma}$ is orthogonally invariant, hence $\boldsymbol{X}$ is ROI by Lemma B.3.

(E6) *Stiefel-uniform columns and random right projection.* If $\boldsymbol{Q} \in V_{n,p}$ is uniform and $\boldsymbol{X} = s\,\boldsymbol{Q}$, then $\boldsymbol{X}\boldsymbol{U} \overset{\text{d}}{=} \boldsymbol{X}$. More generally, for any $\boldsymbol{Y}$ independent of $\boldsymbol{Q} \sim$Haar, $\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{Q}$ is ROI (Lemma B.2).

B.3 Non-ROI archetypes (counterexamples)

(N1) *Anisotropic Gaussian across columns.* $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\text{col}})$ with $\boldsymbol{\Sigma}_{\text{col}} \not\propto \boldsymbol{I}_n$ violates (31).

(N2) *Columnwise scaling heterogeneity.* $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{D}$ with i.i.d. isotropic $\boldsymbol{Z}$ and non-scalar diagonal $\boldsymbol{D}$ has $\mathbb{E}[\boldsymbol{X}^\top \boldsymbol{X}] = \mathbb{E}[\boldsymbol{Z}^\top \boldsymbol{Z}]\boldsymbol{D}^2 \not\propto \boldsymbol{I}_n$.

(N3) *Toeplitz/AR(1) column covariance.* $\boldsymbol{\Sigma}_{\text{col}} = (\rho^{|i-j|})$ breaks ROI already at the second moment.

(N4) *Rademacher i.i.d. entries.* Invariance holds only for the finite hyperoctahedral group (sign flips and permutations), not for all $\boldsymbol{U} \in O(n)$.

(N5) *Blockwise variance mixtures across columns.* Different column blocks having different scales violate (31).

B.4 Relaxing ROI to the stabilizer

We relax the ROI assumption to invariance under the stabilizer

$$\mathcal{G}_{\boldsymbol{B}} := \{\, \boldsymbol{U} \in O(n) : \boldsymbol{U}\boldsymbol{B} = \boldsymbol{B} \,\}.$$

Definition 5 ($\boldsymbol{B}$-ROI)**.** *A random matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ is $\boldsymbol{B}$-ROI if $\boldsymbol{X} \overset{\text{d}}{=} \boldsymbol{X}\boldsymbol{U}$ for all $\boldsymbol{U} \in \mathcal{G}_{\boldsymbol{B}}$.*

Let $r = \mathsf{rank}(\boldsymbol{B})$ and take $\boldsymbol{Q} = [\boldsymbol{Q}_B \;\; \boldsymbol{Q}_\perp]$ with $\mathsf{Col}(\boldsymbol{Q}_B) = \mathsf{Col}(\boldsymbol{B})$. Then $\mathcal{G}_{\boldsymbol{B}} = \{\, \boldsymbol{Q}\,\text{diag}(\boldsymbol{I}_r, \boldsymbol{R})\,\boldsymbol{Q}^\top : \boldsymbol{R} \in O(n-r) \,\}$ and $\boldsymbol{U}\boldsymbol{P}_{\boldsymbol{B}}^\perp = \boldsymbol{P}_{\boldsymbol{B}}^\perp \boldsymbol{U}$ for all $\boldsymbol{U} \in \mathcal{G}_{\boldsymbol{B}}$.

If $\mathbb{E}\|\boldsymbol{X}\|_F^2 < \infty$ and we write $\boldsymbol{Q}^\top\, \mathbb{E}[\boldsymbol{X}^\top \boldsymbol{X}]\, \boldsymbol{Q} = \left(\begin{smallmatrix} \boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{C}^\top & \boldsymbol{D} \end{smallmatrix}\right)$, then $\boldsymbol{B}$-ROI forces

$$\boldsymbol{C} = \boldsymbol{0}, \qquad \boldsymbol{D} = c\,\boldsymbol{I}_{n-r},$$

while $\boldsymbol{A} \in \mathbb{R}^{r \times r}$ is arbitrary. In contrast, ROI requires $\mathbb{E}[\boldsymbol{X}^\top \boldsymbol{X}] = c\,\boldsymbol{I}_n$. Thus $\boldsymbol{B}$-ROI is strictly weaker unless $r = 0$ (then it coincides with ROI). At the distributional level, $\boldsymbol{B}$-ROI is equivalent to: for all $\boldsymbol{U} \in O(n-r)$,

$$\big(\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{B}},\ \boldsymbol{X}\boldsymbol{P}_{\boldsymbol{B}}^\perp\big) \overset{\text{d}}{=} \big(\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{B}},\ \boldsymbol{X}\boldsymbol{P}_{\boldsymbol{B}}^\perp \boldsymbol{U}\big),$$

i.e. the conditional law of $\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{B}}^\perp$ given $\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{B}}$ is ROI.

New important examples under $\boldsymbol{B}$-ROI.

- **Fixed-loading spike + isotropic noise:** $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{\Lambda}^\top + \boldsymbol{E}$ with $\mathsf{Col}(\boldsymbol{\Lambda}) = \mathsf{Col}(\boldsymbol{B})$ and $\boldsymbol{E}$ isotropic on $\mathsf{Col}(\boldsymbol{B})^\perp$. Here $\boldsymbol{\Lambda}$ may be deterministic (no Haar randomness needed).

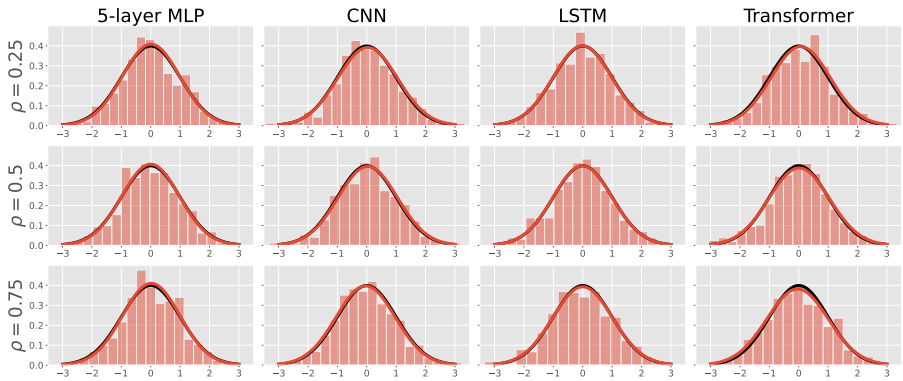

Figure 4: Histograms of the empirical distribution of $\sqrt{n}\xi_j^{(10)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(10)}\|$ for $j \in S^{\mathsf{c}}$. The solid black curve shows the $\mathcal{N}(0,1)$ density. The solid red curve represents a normal density fitted to the histograms.

- **Anisotropy/discreteness only along $\mathsf{Col}(\boldsymbol{B})$:** $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$ with $\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q} = \mathrm{diag}(\boldsymbol{\Sigma}_B,\ \sigma^2\boldsymbol{I}_{n-r})$, where $\boldsymbol{\Sigma}_B$ is arbitrary SPD; or $\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{B}}$ is discrete/binary while $\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{B}}^{\perp}$ is continuous isotropic (Gaussian/t/elliptical).
- **Row dependence with conjugate mixing on the complement:** VAR-type rows with innovations covariance $\boldsymbol{\Sigma}$ satisfying $\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q} = \mathrm{diag}(\boldsymbol{\Sigma}_B,\ \sigma^2\boldsymbol{I}_{n-r})$.
- **Partial random right projection:** $\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{Q}_\perp$ with $\boldsymbol{Q}_\perp \in V_{n,n-r}$ uniform and $\mathsf{Col}(\boldsymbol{Q}_\perp) = \mathsf{Col}(\boldsymbol{B})^\perp$.

When $r = n$, $\mathcal{G}_{\boldsymbol{B}} = \{\boldsymbol{I}_n\}$ and the assumption is vacuous; $\mathsf{Col}(\boldsymbol{B})^\perp = \{0\}$ so our directional statements are degenerate. Conversely, $r = 0$ reduces to ROI.

C Elliptical designs

This section presents the asymptotic normality and feature selection results for designs that violate the $\boldsymbol{B}$-right orthogonal invariance assumption, and provides numerical evidence that our theoretical results remain valid in more general settings.

As an instance of elliptical distributions, we examine a design where each row of $\boldsymbol{X}$ is independently drawn from a multivariate normal distribution with an AR(1) covariance structure. That is, $\boldsymbol{x}_i \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $(\boldsymbol{\Sigma})_{ij} = \rho^{|i-j|}$, $\rho > 0$ for any $i \in [n]$. We conducted experiments with $(m, n) = (2000, 1000)$. To address potential instability induced by strong correlations, we set the learning rate and weight decay to 10^{-3} and 10^{-4}, respectively, for the MLP and 1D-CNN, while keeping all other configurations identical to those in Section 5.1. Figure 4 presents the results, showing that the asymptotic normality of Theorem 1 for null variables is numerically preserved, irrespective of the correlation strength among features.

Next, we examine the iteration-wise evolution of the FDR and Power observed during the numerical experiments. We fix the correlation parameter at $\rho = 0.5$, set the learning rate and weight decay as described above, and keep all other settings identical to those in Section 5.2. The results are shown in Figure 5, indicating that FDR control is successfully achieved despite the presence of feature correlations. On the other hand, the detection power of LSTM begins to decay after a certain number of iterations, suggesting that early stopping could be beneficial.

D Additional numerical experiments

This section provides additional and more detailed results complementing the experiments presented in Section 5.

QQ-plots. As further evidence supporting the asymptotic normality demonstrated in Section 5.1, Figure 6 shows the QQ-plots of $\sqrt{n}\xi_j^{(10)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(10)}\|$ for $j \in S^{\mathsf{c}}$.

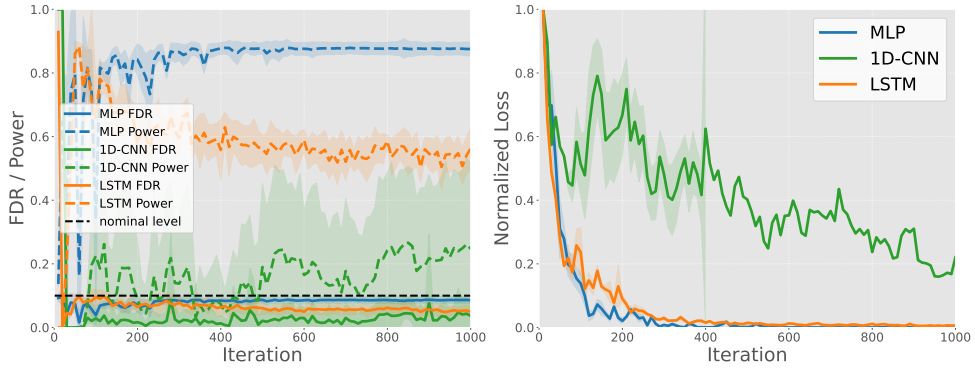

Figure 5: Results for the FDR/power (left) and the training loss (right) when performing feature selection at each iteration. The solid curves represent averages over 20 independent runs, and the shaded areas indicate one standard deviation around the mean.

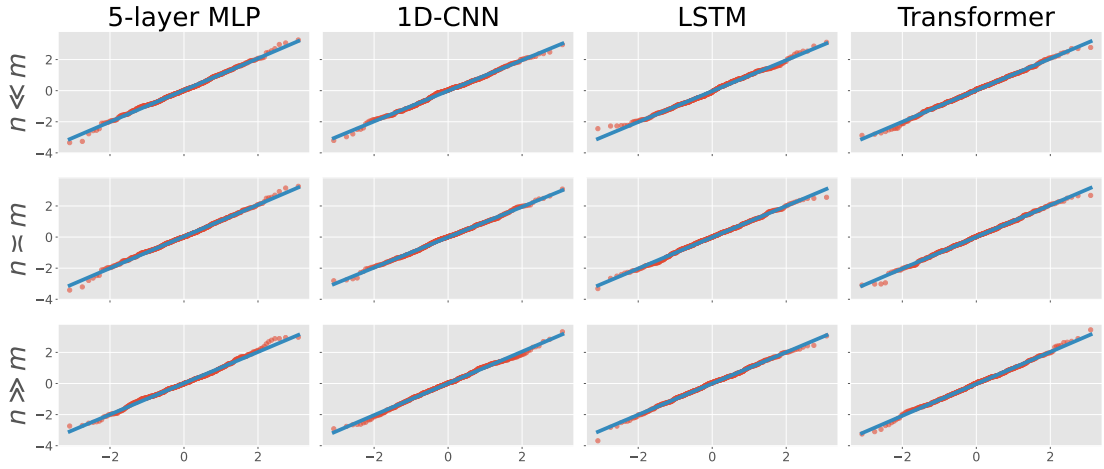

Figure 6: QQ-plots of $\sqrt{n}\xi_j^{(10)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(10)}\|$ under the settings of Figure 2.

Loss trajectories. Figure 7 presents the evolution of the training loss corresponding to the FDR and Power trajectories shown in Figure 3.

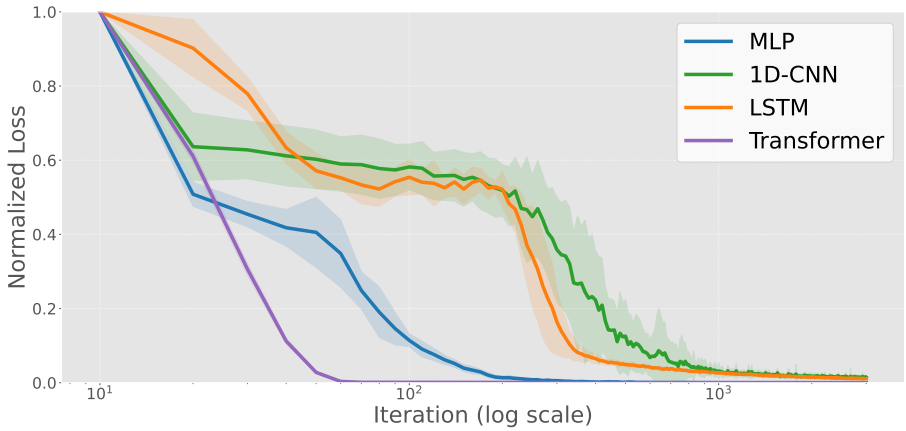

Figure 7: Training loss trajectories of each method across iterations (corresponding to the methods plotted in Figure 3).

Detailed tables. Next, we report the detailed results corresponding to the comparisons in Tables 1–3. In addition to the results of LSTM (which corresponds to Proposed in the main text), we include those of the MLP and 1D-CNN variants of our proposed framework. Furthermore, the values in parentheses below each entry denote the standard deviations computed over 20 independent random seeds. Tables 4–6 show the results. These results suggest that the occasional exceedance of the nominal FDR level $\alpha = 0.1$ by our proposed method is likely due to random variation arising from averaging over a relatively small number of 20 trials. In contrast, the fact that NGM consistently yields FDR values above 0.1 with small standard deviations indicates a genuine failure in FDR control.

Table 4: Performance comparison for $m = 2000$.

	MLP		1D-CNN		LSTM		DeepLINK		NGM	
Design	FDR	Power	FDR	Power	FDR	Power	FDR	Power	FDR	Power
N(0,1)	0.075 (0.019)	0.830 (0.049)	0.094 (0.028)	0.998 (0.002)	0.101 (0.025)	1.000 (0.000)	0.089 (0.021)	0.999 (0.000)	0.086 (0.015)	0.851 (0.013)
t(3)	0.043 (0.042)	0.363 (0.305)	0.047 (0.044)	0.512 (0.382)	0.072 (0.040)	0.843 (0.343)	0.081 (0.030)	0.870 (0.218)	0.063 (0.031)	0.599 (0.151)
Spiked	0.076 (0.034)	0.808 (0.070)	0.099 (0.028)	0.998 (0.003)	0.104 (0.026)	1.000 (0.000)	0.087 (0.021)	1.000 (0.000)	0.080 (0.016)	0.847 (0.017)

Table 5: Performance comparison for $m = 1000$.

	MLP		1D-CNN		LSTM		DeepLINK		NGM	
Design	FDR	Power	FDR	Power	FDR	Power	FDR	Power	FDR	Power
N(0,1)	0.060 (0.030)	0.186 (0.069)	0.075 (0.031)	0.546 (0.093)	0.072 (0.024)	0.637 (0.089)	0.065 (0.021)	0.467 (0.122)	0.131 (0.031)	0.658 (0.041)
t(3)	0.053 (0.057)	0.080 (0.068)	0.040 (0.039)	0.122 (0.111)	0.061 (0.043)	0.168 (0.130)	0.064 (0.076)	0.085 (0.107)	0.144 (0.612)	0.401 (0.127)
Spiked	0.047 (0.036)	0.187 (0.066)	0.070 (0.023)	0.571 (0.094)	0.072 (0.021)	0.682 (0.089)	0.063 (0.028)	0.492 (0.126)	0.151 (0.031)	0.674 (0.046)

Classification problem. Next, as an application to a different data-generating process, we consider a multi-class classification problem. Each entry of $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ is drawn from an i.i.d. standard Gaussian distribution. We

Table 6: Performance comparison for $m = 500$.

	MLP		1D-CNN		LSTM		DeepLINK		NGM	
Design	FDR	Power	FDR	Power	FDR	Power	FDR	Power	FDR	Power
N(0,1)	0.087 (0.092)	0.031 (0.029)	0.087 (0.079)	0.057 (0.033)	0.047 (0.068)	0.040 (0.034)	0.019 (0.58)	0.003 (0.010)	0.242 (0.049)	0.380 (0.093)
t(3)	0.074 (0.133)	0.012 (0.015)	0.041 (0.098)	0.009 (0.010)	0.058 (0.098)	0.016 (0.012)	0.017 (0.061)	0.004 (0.013)	0.325 (0.04)	0.428 (0.076)
Spiked	0.116 (0.223)	0.028 (0.033)	0.138 (0.214)	0.059 (0.046)	0.029 (0.059)	0.037 (0.036)	0.026 (0.065)	0.020 (0.049)	0.264 (0.039)	0.435 (0.056)

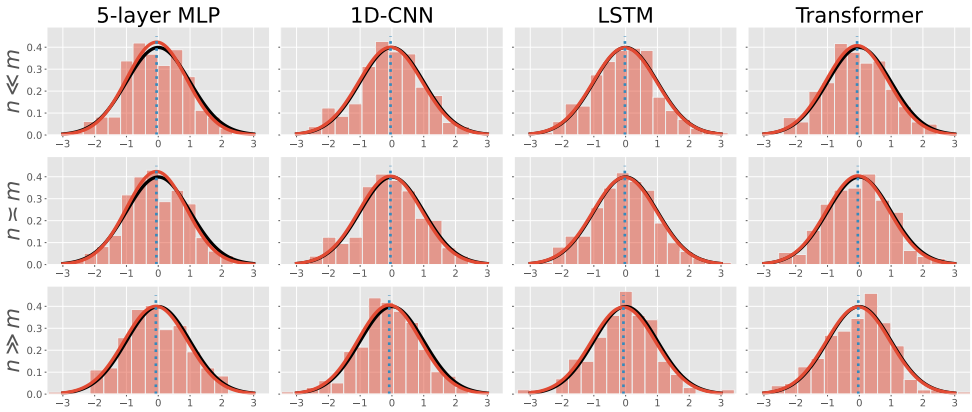

Figure 8: Histograms of the empirical distribution of $\sqrt{n}\xi_j^{(10)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(10)}\|$ for $j \in S^{\mathsf{c}}$ under the multi-class classification model. The solid black curve shows the $\mathcal{N}(0,1)$ density. The solid red curve represents a normal density fitted to the histograms, and the dotted blue line indicates the empirical mean.

construct a weight matrix $\boldsymbol{B} \in \mathbb{R}^{n\times 3}$ by drawing a Gaussian matrix in $\mathbb{R}^{(n/2)\times 3}$, orthonormalizing its columns via QR, and embedding it into the top $n/2$ coordinates while filling zeros in S^{c}. For $K = 3$ classes, let the class-k score be

$$h_k(\boldsymbol{x}) = \alpha_k \sin\left(\omega_k\left(\boldsymbol{B}_{\cdot 1}^{\top}\boldsymbol{x}\right)\right) + \beta_k \cos\left(\nu_k\left(\boldsymbol{B}_{\cdot 2}^{\top}\boldsymbol{x}\right)\right) + \gamma_k(\boldsymbol{B}_{\cdot 3}^{\top}\boldsymbol{x}) + b_k,$$

where $(\alpha_k, \beta_k, \gamma_k)$ control the relative contributions, (ω_k, ν_k) set the frequencies and b_k balances class prior. Class probabilities follow a softmax with temperature $\tau > 0$,

$$\Pr(y_i = k \mid \boldsymbol{x}_i) = \frac{\exp\left(h_k(\boldsymbol{x}_i)/\tau\right)}{\sum_{\ell=1}^{K}\exp\left(h_\ell(\boldsymbol{x}_i)/\tau\right)},$$

and labels are sampled accordingly. We keep τ and the amplitudes fixed across trials unless stated otherwise and vary the random seed to average over data realizations.

The configurations of the MLP, 1D-CNN, and LSTM models are the same as those used in Section 5. The choices of m and n follow the same setting as well. Figures 8 and 9 present histograms and QQ-plots that confirm the asymptotic normality of the proposed statistics. Figure 10 presents the results of feature selection when $m = 4000$ and $n = 400$. In this setting, the power remains nearly zero for all methods, and in some cases the training loss does not decrease. This behavior is likely due to the non-null distribution of $\xi_j^{(t)}$ not being sufficiently separated from its null counterpart, suggesting that further investigation is needed to determine whether this issue can be mitigated through adjustments to the network architecture or optimization strategy.

Sensitivity to the choice of $\psi(u, v)$. We consider the same setting as in Figure 3 and vary the aggregation function $\psi(u, v)$ among $\min(u, v)$, $\max(u, v)$, uv, and $u + v$. The resulting performance for each architecture is shown in Figures 11–13, labeled as min, max, prod, and mean, respectively. While Dai et al. (2023a) proved in Proposition 1 that $\psi(u, v) = u + v$ is optimal, our results are broadly consistent with this finding.

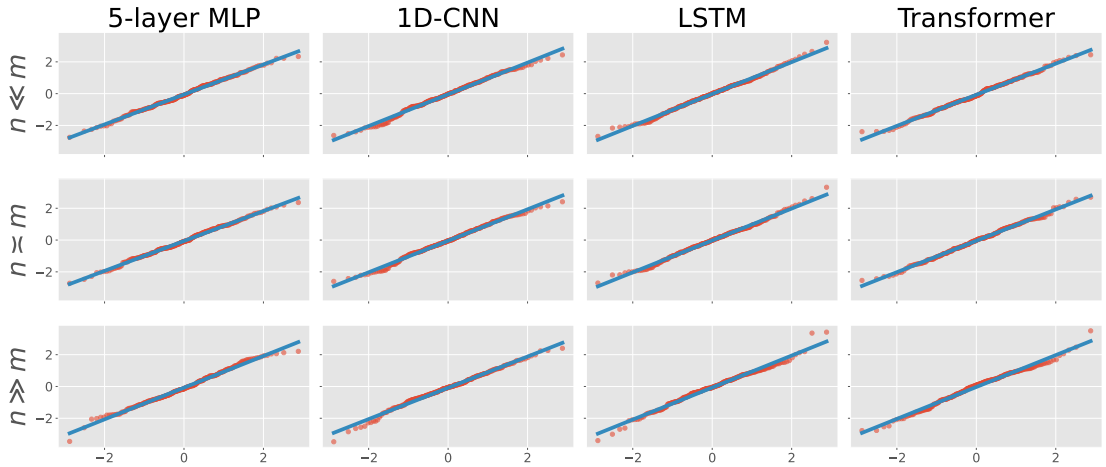

Figure 9: QQ-plots of $\sqrt{n}\xi_j^{(10)}/\|\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{\xi}^{(10)}\|$ under the multi-class classification model.

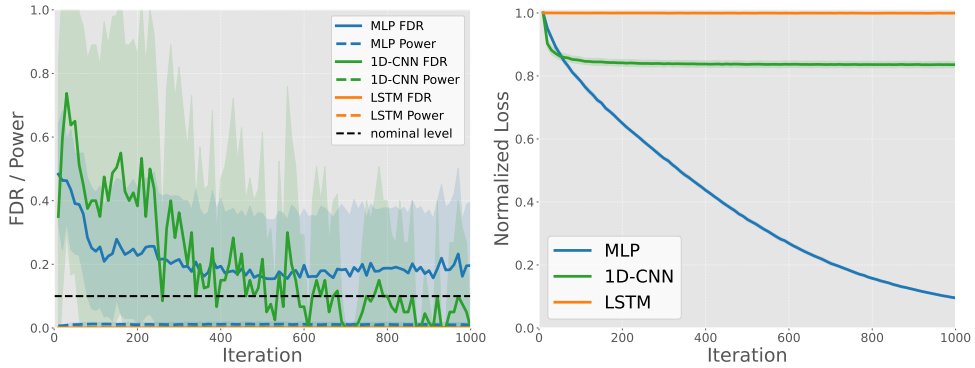

Figure 10: Results for the FDR/power (left) and the training loss (right) when performing feature selection at each iteration under the multi-class classification model. The solid curves represent averages over 20 independent runs, and the shaded areas indicate one standard deviation around the mean.

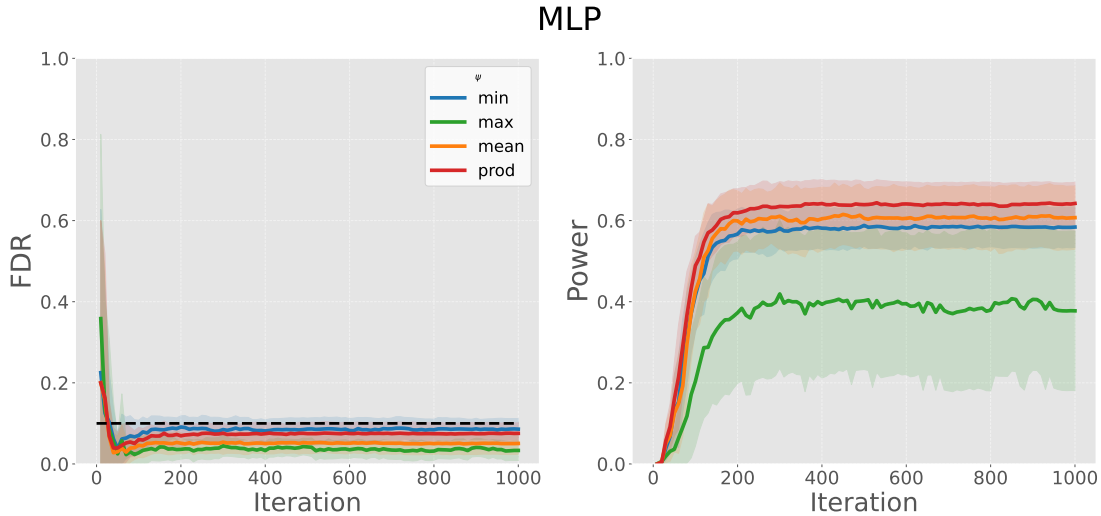

Figure 11: Under the same setting as in Figure 3, we present the false discovery proportion and power of the selected features at each SGD step for the MLP architecture, for each choice of $\psi(u, v) \in \{\min(u, v), \max(u, v), u+v, uv\}$.

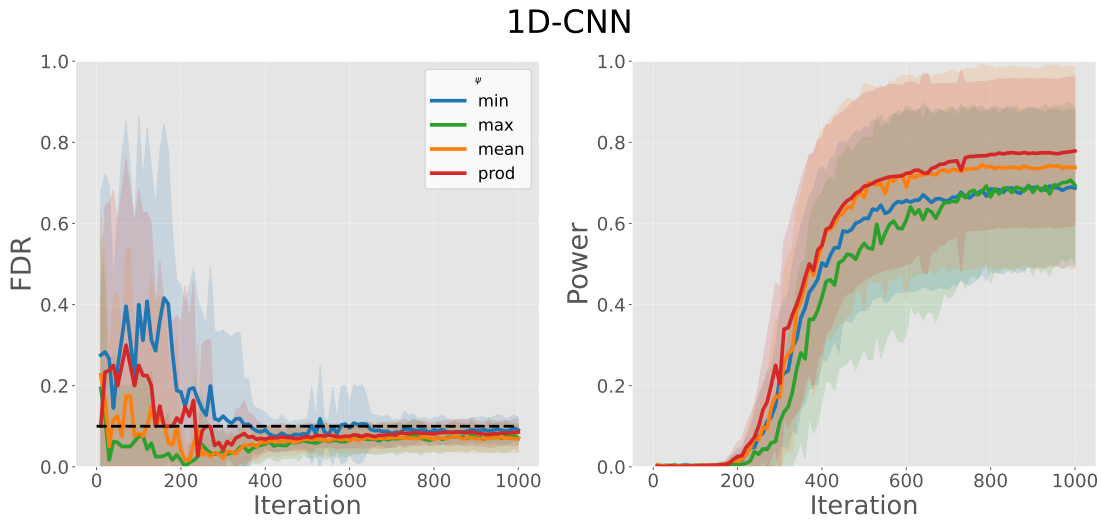

Figure 12: Under the same setting as in Figure 3, we present the false discovery proportion and power of the selected features at each SGD step for the 1D-CNN architecture, for each choice of $\psi(u, v) \in \{\min(u, v), \max(u, v), u + v, uv\}$.

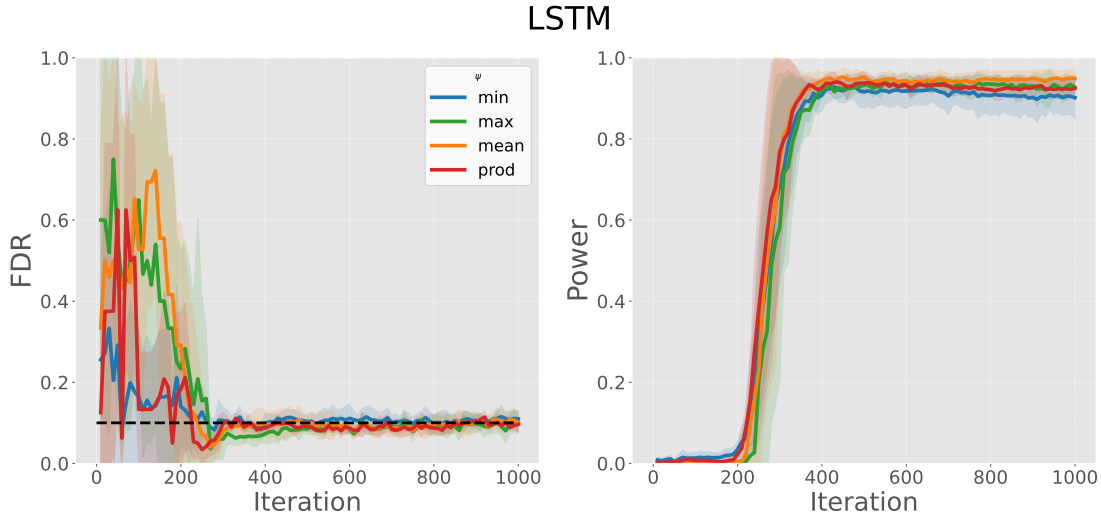

Figure 13: Under the same setting as in Figure 3, we present the false discovery proportion and power of the selected features at each SGD step for the LSTM architecture, for each choice of $\psi(u, v) \in \{\min(u, v), \max(u, v), u + v, uv\}$.

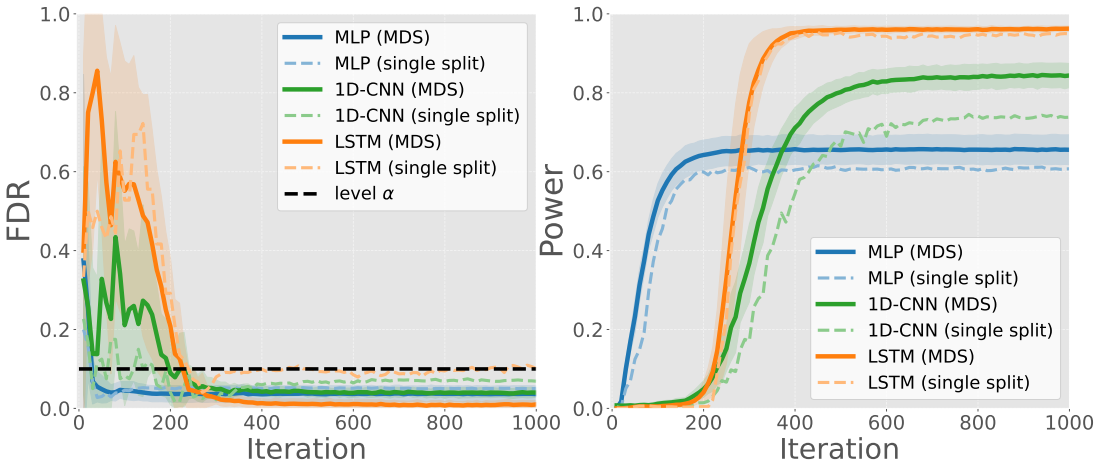

Figure 14: Under the same setting as in Figure 3 (plotted as dotted lines), we present the false discovery proportion and power of the selected features at each SGD step for the multiple data-splitting method (Dai et al., 2023a) over 20 random splits for each architecture.

Stabilizing technique. Finally, we present the results of multiple data splitting (Dai et al., 2023a), which enhances stability and power by aggregating feature selection results across multiple random splits. Specifically, the results in Figure 3 are depicted by dashed lines, and the corresponding aggregated results over 20 random splits under the same setting are overlaid and shown in Figure 14. We can see numerically that this approach simultaneously reduces the false discovery rate while improving power. The computation required approximately 9 hours on a Google Cloud Platform VM equipped with four NVIDIA Tesla T4 (16,GB each) GPUs, using PyTorch 2.8.0 (built with CUDA 12.8) and CUDA runtime 12.4.