Supplementary Material for submission "A Conditional Randomization Test for Sparse Logistic Regression in High-Dimension"

A Proofs of theoretical results in Section 3

We first present some technical lemmas that are useful for the proof of the main theorem. From now on, let $\precsim$ and $\succsim$ denote inequalities with a hidden constant factor, *i.e.* $x \precsim y$ means that with high probability, there exists an absolute constant $C > 0$ such that $x \leq Cy$, and vice versa. As mentioned in the main text, in what follows, without writing it explicitly, we consider $p = p(n)$.

Lemma A.1 (Lemma E.1, [21])**.** *Assume Assumption 3.1, under the logistic model, we have*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \precsim s^* \sqrt{\frac{\log p}{n}} \quad \textit{and} \quad \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 \precsim \sqrt{\frac{s^* \log p}{n}},$$

where $s^* = \|\boldsymbol{\beta}^0\|_0$. In addition, we also have

$$\frac{1}{n}\sum_{i=1}^{n} g''(\mathbf{X}_{i,*}\boldsymbol{\beta}^0)[\mathbf{X}_{i,-j}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)]^2 \precsim \frac{s^* \log(p)}{n} \ ,$$

where $g(x) = 1/(1 + \exp(x))$ is the sigmoid function.

Lemma A.2 (Lemma E.2 [21], concentration of the gradient and Hessian of the logistic loss function)**.** *Assume Assumption 3.1 holds, under the logistic model, we have, with $\mathbf{v}^* \overset{\textit{def.}}{=} (1, -\mathbf{w}^{0,j}) \in \mathbb{R}^p$,*

$$\|\nabla \ell(\boldsymbol{\beta}^0)\|_\infty \precsim \sqrt{n^{-1}\log p}, \quad \textit{and}$$
$$\|\mathbf{v}^{*\top} \nabla^2 \ell(\boldsymbol{\beta}^0) - \mathbb{E}_{\boldsymbol{\beta}^0}[\mathbf{v}^{*\top} \nabla^2 \ell(\boldsymbol{\beta}^0)]\|_\infty \precsim \sqrt{n^{-1}\log p}.$$

Lemma A.3 (Lemma E.3, [21])**.** *Assume Assumption 3.1 holds, under logistic model, we have*

$$\|\hat{\boldsymbol{\beta}}^{d\mathbf{x}_{*,j}} - \mathbf{w}^{0,j}\|_1 \precsim (s' \vee s^*)\sqrt{\frac{\log p}{n}},$$

where $s^* = \|\boldsymbol{\beta}^0\|_0$ and $s' = \|\mathbf{w}^{0,j}\|_0$. In addition, we also have

$$\frac{1}{n}\sum_{i=1}^{n} g''(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})[\mathbf{X}_{i,-j}(\hat{\boldsymbol{\beta}}^{d\mathbf{x}_{*,j}} - \mathbf{w}^{0,j})]^2 \precsim \frac{(s' \vee s^*)\log(p)}{n}.$$

Lemma A.4 (Lemma E.4, [21], local smoothness conditions on the loss function)**.** *Let $\hat{\boldsymbol{\beta}}^{\textit{null}} = (0, \hat{\boldsymbol{\beta}}_{-j}) \in \mathbb{R}^p$, where $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}^0$. It holds that*

$$|\mathbf{v}^{*\top}[\nabla \ell(\boldsymbol{\beta}) - \nabla \ell(\boldsymbol{\beta}^0) - \nabla^2 \ell(\boldsymbol{\beta}^0)(\boldsymbol{\beta} - \boldsymbol{\beta}^0)]| \precsim \frac{(s^* \vee s')\log p}{n} \ ,$$
$$|(\hat{\mathbf{v}} - \mathbf{v}^*)^\top[\nabla \ell(\boldsymbol{\beta}) - \nabla \ell(\boldsymbol{\beta}^0)] \ | \precsim \frac{(s^* \vee s')\log p}{n} \ .$$

for both $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{\textit{null}}$ and $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, where $\hat{\mathbf{v}} \overset{\textit{def.}}{=} (1, (\hat{\boldsymbol{\beta}}^{d\mathbf{x}_{*,j}})^\top)$.

Remark A.1. *We make a slight abuse of notation in the definition of $\hat{v}$ and v^*, which formally corresponds to the $j = 1$ case. We use the fact that permuting the corresponding variable indices position of $\hat{v}$, v^*, and $\nabla \ell(\beta)$ simultaneously does not change the value of $v^T \nabla \ell(\beta)$, and hence will not change the proofs.*

Proof of Theorem 3.1. The following proof is an adaptation from [21]. Notice that our version of the proof is shorter, with specific consideration on sparse logistic regression, and with elaboration on the convergence rate of the decorreleated test score, which is missing from [21].

Denote $\hat{\mathbf{v}} \overset{\text{def.}}{=} (1, -(\hat{\boldsymbol{\beta}}^{d\mathbf{x}_{*,j}})^\top)$, then the decorrelated test score can be written in a more general from as

$$T_j^{\text{decorr}} = n^{1/2}\hat{\mathbf{I}}_{j|-j}^{-1/2}\left(\nabla_j \ell(\hat{\boldsymbol{\beta}}) - (\hat{\boldsymbol{\beta}}^{d\mathbf{x}_{*,j}})^\top \nabla_{\boldsymbol{\beta}_{-j}} \ell(\hat{\boldsymbol{\beta}})\right) = n^{1/2}\hat{\mathbf{I}}_{j|-j}^{-1/2}\hat{\mathbf{v}}^\top \nabla \ell(\hat{\boldsymbol{\beta}}) \ . \tag{14}$$

Moreover, denote $\hat{\boldsymbol{\beta}}^{\text{null}} \overset{\text{def.}}{=} (0, \hat{\boldsymbol{\beta}}_{-j})$ and $\mathbf{v}^* \overset{\text{def.}}{=} (1, -\mathbf{w}^{0,j})$, then we have, under the null hypothesis,

$$n^{1/2}|\hat{\mathbf{v}}^\top \nabla(\hat{\boldsymbol{\beta}}^{\text{null}}) - \mathbf{v}^{*\top}\nabla\ell(\boldsymbol{\beta}^0)| \leq \underbrace{n^{1/2}|\mathbf{v}^{*\top}\{\nabla\ell(\boldsymbol{\beta}^0) - \nabla\ell(\hat{\boldsymbol{\beta}}^{\text{null}})\}|}_{A_1} + \underbrace{n^{1/2}|(\hat{\mathbf{v}} - \mathbf{v}^*)^\top \nabla\ell(\hat{\boldsymbol{\beta}}^{\text{null}})|}_{A_2}$$

where we use the triangle inequality in the last step. From Lemma A.4, we have

$$\begin{aligned} A_1 &\leq n^{1/2}\left(|\mathbf{v}^{*\top}\nabla^2\ell(\boldsymbol{\beta}^0)(\hat{\boldsymbol{\beta}}^{\text{null}} - \boldsymbol{\beta}^0)| + \mathcal{O}_{\mathbb{P}}\left(\frac{(s^* \vee s')\log p}{n}\right)\right) \\ &\leq n^{1/2}\left(\|\hat{\boldsymbol{\beta}}^{\text{null}} - \boldsymbol{\beta}^0\|_1 \|\mathbf{v}^{*\top}\nabla^2\ell(\boldsymbol{\beta}^0)\|_\infty + \mathcal{O}_{\mathbb{P}}\left(\frac{(s^* \vee s')\log p}{n}\right)\right) \\ &\precsim \frac{(s^* \vee s')\log p}{\sqrt{n}} \end{aligned}$$

where the second inequality is by Holdër inequality, and the last inequality is due to Lemma A.1 and A.2. Similarly, we can bound A_2, by using Lemma A.3 and Lemma A.4

$$\begin{aligned} A_2 &\leq n^{1/2}\left(|(\hat{\mathbf{v}} - \mathbf{v}^*)^\top \nabla\ell(\boldsymbol{\beta}^0)| + \mathcal{O}_{\mathbb{P}}\left(\frac{(s^* \vee s')\log p}{n}\right)\right) \\ &\leq n^{1/2}\left(\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 \|\nabla\ell(\boldsymbol{\beta}^0)\|_\infty + \mathcal{O}_{\mathbb{P}}\left(\frac{(s^* \vee s')\log p}{n}\right)\right) \precsim \frac{(s^* \vee s')\log p}{\sqrt{n}} \end{aligned}$$

This implies that,

$$n^{1/2}|\hat{\mathbf{v}}^\top \nabla(\hat{\boldsymbol{\beta}}^{\text{null}}) - \mathbf{v}^{*\top}\nabla\ell(\boldsymbol{\beta}^0)| \precsim n^{-1/2}(s^* \vee s')\log(p)\,. \tag{15}$$

The remaining part of the proof is to bound $\hat{\mathbf{I}}_{j|-j} - \mathbf{I}_{j|-j}$, where, by definition

$$\mathbf{I}_{j|-j} = \mathbb{E}\left\{g''(\mathbf{X}_{i,*}\boldsymbol{\beta}^0)\left[\mathbf{X}_{i,j} - \mathbf{X}_{i,-j}\mathbf{w}^{0,j}\right]\,\mathbf{X}_{i,j}\right\}$$

Evaluating the difference between $\hat{\mathbf{I}}_{j|-j}$ and $\mathbf{I}_{j|-j}$ gives

$$\begin{aligned} &\hat{\mathbf{I}}_{j|-j} \text{ - } \mathbf{I}_{j|-j} \\ &= \frac{1}{n}\sum_{i=1}^n g''(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})\left[\mathbf{X}_{i,j} - \mathbf{X}_{i,-j}\hat{\boldsymbol{\beta}}^{d\mathbf{x}_{*,j}}\right]\,\mathbf{X}_{i,j} - \mathbb{E}\left\{g''(\mathbf{X}_{i,*}\boldsymbol{\beta}^0)\left[\mathbf{X}_{i,j} - \mathbf{X}_{i,-j}\mathbf{w}^{0,j}\right]\,\mathbf{X}_{i,j}\right\} \\ &= \left(\frac{1}{n}\sum_{i=1}^n g''(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})\mathbf{X}_{i,j}^2 - \mathbb{E}\left\{g''(\mathbf{X}_{i,*}\boldsymbol{\beta}^0)\mathbf{X}_{i,j}^2\right\}\right) \\ &\quad + \left(\frac{1}{n}\sum_{i=1}^n g''(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})\mathbf{X}_{i,-j}\hat{\boldsymbol{\beta}}^{d\mathbf{x}_{*,j}}\,\mathbf{X}_{i,j} - \mathbb{E}\left\{g''(\mathbf{X}_{i,*}\boldsymbol{\beta}^0)\mathbf{X}_{i,-j}\mathbf{w}^{0,j}\,\mathbf{X}_{i,j}\right\}\right) \\ &\leq \underbrace{\left(\frac{1}{n}\sum_{i=1}^n g''(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})\mathbf{X}_{i,j}^2 - \mathbb{E}\left\{g''(\mathbf{X}_{i,*}\boldsymbol{\beta}^0)\mathbf{X}_{i,j}^2\right\}\right)}_{C} + \underbrace{\left|\frac{1}{n}\sum_{i=1}^n g''(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})\mathbf{X}_{i,-j}(\hat{\boldsymbol{\beta}}^{d\mathbf{x}_{*,j}} - \mathbf{w}^{0,j})\,\mathbf{X}_{i,j}\right|}_{B_1} \\ &\quad + \underbrace{\left|\frac{1}{n}\sum_{i=1}^n [g''(\mathbf{X}_{i,*}\boldsymbol{\beta}^0) - g''(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})]\mathbf{X}_{i,-j}\mathbf{w}^{0,j}\,\mathbf{X}_{i,j}\right|}_{B_2} + \\ &\quad \underbrace{\left|\frac{1}{n}\sum_{i=1}^n g''(\mathbf{X}_{i,*}\boldsymbol{\beta}^0)\mathbf{X}_{i,-j}\mathbf{w}^{0,j}\,\mathbf{X}_{i,j} - \mathbb{E}\left\{g''(\mathbf{X}_{i,*}\boldsymbol{\beta}^0)\mathbf{X}_{i,-j}\mathbf{w}^{0,j}\,\mathbf{X}_{i,j}\right\}\right|}_{B_3}, \end{aligned}$$

where the last step follows from triangle inequality.

We have, by Cauchy-Schwartz inequality, by Lemma A.3, and by the fact that $g''(x) \in (0,1)$ for every $x \in \mathbb{R}$; and $\mathbf{X}_{i,-j}$, $\mathbf{X}_{i,j}$ is sub-exponential by Assumption 3.1:

$$B_1 \leq \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} g''(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})((\hat{\boldsymbol{\beta}}^{d_{\mathbf{X}_{*,j}}} - \mathbf{w}^{0,j})^\top \mathbf{X}_{i,-j})^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} g''(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})\mathbf{X}_{i,j}^2\right)}$$
$$\precsim \sqrt{\frac{(s^* \vee s')\log(p)}{n}}.$$

Similarly, to bound B_2, we have, again by Cauchy-Schwartz inequality,

$$B_2 \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}[g''(\mathbf{X}_{i,*}\boldsymbol{\beta}^0) - g''(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})]^2 \left(\mathbf{X}_{i,-j}\mathbf{w}^{0,j}\ \mathbf{X}_{i,j}\right)^2}$$
$$\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}[g''(\mathbf{X}_{i,*}\boldsymbol{\beta}^0)\mathbf{X}_{i,*}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)]^2 \left(\mathbf{X}_{i,-j}\mathbf{w}^{0,j}\ \mathbf{X}_{i,j}\right)^2},$$

where the second inequality comes from using the self-concordance property of the sigmoid function (discussed at length in [2] and extended further in [22]), that is, $|g''(t_1) - g''(t)| \leq |t_1 - t|g''(t)$ for a fixed constant t, and for every $t_1 \in \mathbb{R}$ such that t_1 converges to t, with $t_1 = \hat{\boldsymbol{\beta}}$, and $t = \boldsymbol{\beta}^0$. By Assumption 3.1-A3 that $\mathbf{X}_{i,j}$ is sub-exponential, applying Bernstein inequality leads to

$$B_2 \precsim \sqrt{\frac{s^* \log p}{n}}.$$

To bound B_3, by direct application of Hoeffding inequality, we have $B_3 \precsim \sqrt{\frac{(s^* \vee s')\log p}{n}}$. This implies

$$|\hat{\mathbf{I}}_{j|-j} - \mathbf{I}_{j|-j}| \precsim \sqrt{\frac{(s^* \vee s')\log p}{n}}\,. \tag{16}$$

Putting Equation (15) and (16) together, we have, under the null hypothesis,

$$T_j^{\text{decorr}} \xrightarrow{\mathcal{D}} n^{1/2}\,\mathbf{I}_{j|-j}^{-1/2}\mathbf{v}^{*\top}\nabla\ell(\boldsymbol{\beta}^0) \overset{\text{def.}}{=} T_j^*\,,$$

with convergence rate $\mathcal{O}(n^{-1/2})$. Finally, by noting that we can decompose $\nabla\ell(\boldsymbol{\beta}^0) = \frac{1}{n}\sum_{i=1}^{n}\nabla\ell_i(\boldsymbol{\beta}^0)$, and each $\nabla\ell_i(\boldsymbol{\beta}^0)$ has bounded first, second, and third moment, a direct application of Berry-Esseen theorem give convergence in distribution of T_j^* to a standard normal law, with rate $\mathcal{O}(n^{-1/2})$.

We also arrive at the second conclusion of Theorem 3.1 by noting that it is a straightforward by-product of the result on normality of the distribution of decorrelated test score under null hypothesis, based on the formula for the p-values of CRT-logit algorithm.

□

Proof of Corollary 3.1. The proof of this result is a straightforward adaptation from [6]. For shorter notation, we denote $\widehat{\mathcal{S}} \overset{\text{def.}}{=} \widehat{\mathcal{S}}_{\text{BY-CRT}}$ and $\widehat{k} \overset{\text{def.}}{=} \widehat{k}_{BY}$. If we denote $\bar{\alpha} \overset{\text{def.}}{=} \dfrac{\alpha}{p\sum_{i=1}^{p} 1/i} \in (0,1)$, then step 1 in the procedure defined in Definition B.2 is equivalent to finding $\widehat{k}$ such that

$$\widehat{k} = \max\left\{k \in [p] \mid \hat{p}_{(k)} \leq k\bar{\alpha}\right\}. \tag{17}$$

For every $i, j, k \in [p]$, let us define

$$p_{i,j,k} = \begin{cases} \mathbb{P}\left(\widehat{p}_i \in ((j-1)\bar{\alpha}, j\bar{\alpha}]\ , i \in \widehat{S} \text{ and } |\widehat{S}| = k\right) & \text{if } j \geq 2 \\ \mathbb{P}\big(\widehat{p}_i \in [0,\bar{\alpha}]\,,\ i \in \widehat{S} \text{ and } |\widehat{S}| = k\big) & \text{if } j = 1. \end{cases} \tag{18}$$

Then, since $i \in \widehat{\mathcal{S}}$ and $|\widehat{\mathcal{S}}| = k$ implies that $\widehat{p}_i \leq \widehat{p}_{\widehat{k}} \leq \widehat{k}\bar{\alpha} = k\bar{\alpha}$, we have

$$\frac{|\widehat{S} \cap \mathcal{S}^c|}{|\widehat{S}| \vee 1} = \sum_{k=1}^{p} \mathbb{1}_{|\widehat{S}|=k} \frac{\sum_{i \in \mathcal{S}^c} \mathbb{1}_{i \in \widehat{S}}}{k} = \sum_{i \in \mathcal{S}^c} \sum_{k=1}^{p} \frac{1}{k} \mathbb{1}_{|\widehat{S}|=k \text{ and } i \in \widehat{S}}$$
$$= \sum_{i \in \mathcal{S}^c} \sum_{k=1}^{p} \frac{1}{k} \mathbb{1}_{|\widehat{S}|=k \text{ and } i \in \widehat{S} \text{ and } 0 \leq \widehat{p}_i \leq k\bar{\alpha}} .$$

Taking the expectation and writing that

$$\mathbb{1}_{0 \leq \widehat{p}_i \leq k\bar{\alpha}} = \mathbb{1}_{\widehat{p}_i \in [0, \bar{\alpha}]} + \sum_{j=2}^{k} \mathbb{1}_{\widehat{p}_i \in ((j-1)\bar{\alpha}, j\bar{\alpha}]} ,$$

we get

$$\mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^c|}{|\widehat{\mathcal{S}}| \vee 1}\right] = \sum_{i \in \mathcal{S}^c} \sum_{k=1}^{p} \frac{1}{k} \sum_{j=1}^{k} p_{i,j,k} = \sum_{i \in \mathcal{S}^c} \sum_{j=1}^{p} \sum_{k=j}^{p} \frac{1}{k} p_{i,j,k}$$
$$\leq \sum_{i \in \mathcal{S}^c} \sum_{j=1}^{p} \sum_{k=j}^{p} \frac{1}{j} p_{i,j,k} = \underbrace{\sum_{j=1}^{p} \frac{1}{j} \sum_{i \in \mathcal{S}^c} \sum_{k=j}^{p} p_{i,j,k}}_{A} .$$

Denote $F(j) \overset{\text{def.}}{=} \sum_{i \in \mathcal{S}^c} \sum_{j'=1}^{j} \sum_{k=1}^{p} p_{i,j',k}$ for all $j \in \{1, \ldots, p\}$, and remark that $p_{i,j',k} = 0$ if $j' > k$, by definition of $\widehat{\mathcal{S}}_{\text{BY-CRT}}$. We then have

$$A = F(1) + \sum_{j=2}^{p} \frac{1}{j} \big[F(j) - F(j-1) \big] = \sum_{j=1}^{p-1} \left(\frac{1}{j} - \frac{1}{j+1} \right) F(j) + \frac{F(p)}{p} .$$

This leads to

$$\mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^c|}{|\widehat{\mathcal{S}}| \vee 1}\right] \leq \sum_{j=1}^{p-1} \left(\frac{1}{j} - \frac{1}{j+1} \right) F(j) + \frac{F(p)}{p} \tag{19}$$

By the definition of $p_{i,j,k}$ in Eq. (18), we have

$$F(j) = \sum_{i \in \mathcal{S}^c} \mathbb{P}(\widehat{p}_i \leq j\bar{\alpha} \text{ and } i \in \widehat{\mathcal{S}}) \leq \sum_{i \in \mathcal{S}^c} \mathbb{P}(\widehat{p}_i \leq j\bar{\alpha}).$$

Therefore

$$\mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^c|}{|\widehat{\mathcal{S}}| \vee 1}\right] \leq \sum_{i \in \mathcal{S}^c} \sum_{j=1}^{p-1} \frac{\mathbb{P}(\widehat{p}_i \leq j\bar{\alpha})}{j(j+1)} + \sum_{i \in \mathcal{S}^c} \frac{\mathbb{P}(\widehat{p}_i \leq p\bar{\alpha})}{p}$$

Taking the limit where $n \to \infty$ and p fixed, we have, using the result in Theorem 3.1,

$$\limsup_{n \to \infty} \mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^c|}{|\widehat{\mathcal{S}}| \vee 1}\right] \leq \sum_{i \in \mathcal{S}^c} \left(\sum_{j=1}^{p-1} \frac{1}{j+1} + 1 \right) \bar{\alpha}$$
$$= \left(\sum_{j=1}^{p} \frac{1}{j} \right) |\mathcal{S}^c| \bar{\alpha} .$$

We conclude the proof by noting that $\bar{\alpha} \overset{\text{def.}}{=} \dfrac{\alpha}{p \sum_{j=1}^{p} 1/j}$. □

B Controlling False Discovery Rate Procedures

Definition B.1 (Benjamini-Hochberg procedure [5]). *Let $\alpha \in (0, 1)$ be the predefined FDR control level. Let $\hat{p}_1, \ldots, \hat{p}_m$ be output p-values from an inference algorithm,* e.g. *Algorithm 1. We reorder them ascendingly, denoted by $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \cdots \leq \hat{p}_{(p)}$ and $\mathcal{H}_0^{(1)}, \ldots, \mathcal{H}_0^{(p)}$, then*

1. *Find* $\widehat{k}_{BH}$ *such that*

$$\widehat{k}_{BH} = \max\left\{k \in [p] \mid \hat{p}_{(k)} \leq \frac{k\alpha}{p}\right\}.$$

2. *If* $\widehat{k}_{BH}$ *exists, take* $\widehat{S} = \{j \in [p] : \hat{p}_{(j)} \leq \hat{p}_{\widehat{k}_{BH}}\}$. *Otherwise* $\widehat{S} = \emptyset$.

Definition B.2 (Benjamini-Yekutieli procedure [6])**.** *Let* $\alpha \in (0, 1)$ *be the predefined FDR control level. Let* $\hat{p}_1, \ldots, \hat{p}_m$ *be output p-values from Algorithm 1. We reorder them ascendingly, denoted by* $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \cdots \leq \hat{p}_{(p)}$ *and* $\mathcal{H}_0^{(1)}, \ldots, \mathcal{H}_0^{(p)}$*, then*

1. *Find* $\widehat{k}_{BY}$ *such that*

$$\widehat{k}_{BY} = \max\left\{k \in [p] \mid \hat{p}_{(k)} \leq \frac{k\alpha}{p\sum_{i=1}^{p} 1/i}\right\}.$$

2. *If* $\widehat{k}_{BY}$ *exists, take* $\widehat{S} = \{j \in [p] : \hat{p}_{(j)} \leq \hat{p}_{\widehat{k}_{BY}}\}$. *Otherwise* $\widehat{S} = \emptyset$.

C Setting the ℓ_1-Regularization Parameter of the $\mathbf{X}_{*,j}$-distillation

A core issue is the dependency of the statistical power and FDR of CRT-logit on the ℓ_1- regularization parameter λ_{dx} when doing Lasso distillation on x_j in Eq. (10). One might choose the reference value $\lambda_{\text{univ}} = \sqrt{n^{-1}\log p}$ with theoretical validity, as suggested in [21, 28]. However, experimental results in Fig. 5 show that at $\lambda_{dx} = \lambda_{\text{univ}}$ (or $\log_{10}\lambda/\lambda_{\text{univ}} = 0.0$ with the labeling of the figure), we do not have the best possible FDR/Power with CRT-logit inference. For this experiment, we average the inference results of 100 simulations (with similar setting in Section 4.1) for different values of n and λ_{dx}, with p fixed. There is a clear phase transition in both FDR and average power when the regularization parameter λ_{dx} increases. In other words, we have found empirically that both FDR and power of the method are sensitive to the ℓ_1-regularization parameter. Preferably, one wants to return a high statistical power while controlling FDR under predefined level. Hence, it is necessary to choose λ_{dx} wisely. In practice, we advise using cross-validation for $\mathbf{X}_{*,j}$-distillation operator, as defined by Eq. (10). This means we would have to find p different values of λ_{dx} with cross-validation, and we reemphasize the importance of the screening step to reduce the number of computations.

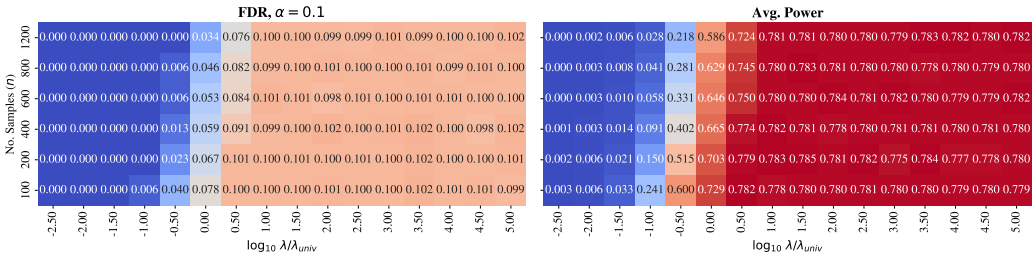


Figure 5: **FDR/Average Power of 100 runs of simulations while varying the number of samples and ℓ_1 regularization parameter and fixing the number of variables**. Note: λ_{dx} is scaled with the factor $\lambda_{\text{univ}} = \sqrt{\log(p)/n}$, *e.g.* the first value for regularization grid is $\lambda_{dx} = 10^{-2}\lambda_{\text{univ}}$. Default parameter (similar settings in Section 4.1): $p = 400$, SNR=3.0 (signal-to-noise ration), $\rho = 0.5$ (feature correlation), $\kappa = 0.05$ (sparsity). FDR is controlled at level $\alpha = 0.1$.

D Pseudocode for CRT-logit and Related Algorithms

Algorithm 2: Conditional Randomization Test [10]

INPUT dataset $(\mathbf{X}, \mathbf{y})$, with $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$, number of sampling runs B, test statistic T_j, conditional distribution $P_{j|-j}$ for each $j = 1, \ldots, p$;
OUTPUT vector of p-values $\{\hat{p}_j\}_{j=1}^p$;
for $j = 1, 2, \ldots, p$ **do**
    Compute test statistics T_j for original variable;
    for $b = 1, 2, \ldots, B$ **do**
        1. Generate $\tilde{\mathbf{X}}_{*,j}^{(b)}$, a knockoff sample from $P_{j|-j}$;
        2. Compute $\tilde{T}_j^{(b)}$ for knockoff variables;
    end
    Compute the empirical p-value

$$\widehat{p}_j = \frac{1 + \sum_{b=1}^{B} \mathbf{1}_{\tilde{T}_j^{(b)} \geq T_j}}{1 + B}$$

end

Algorithm 3: Lasso-Distillation Conditional Randomization Test [19]

INPUT dataset $(\mathbf{X}, \mathbf{y})$, $X \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$, test statistic T_j for each $j = 1, \ldots, p$;
OUTPUT vector of p-values $\{p_j\}_{j=1}^p$;
$\hat{\mathcal{S}}^{\text{SCREENING}} = \{j : j \in [p], \hat{\beta}_j^{\text{MLE}} \neq 0\}$ `// Using Eq.` (2)
for $j \in \hat{\mathcal{S}}^{SCREENING}$ **do**
    1. Distill information of $\mathbf{X}_{-j}$ to $\mathbf{X}_{*,j}$ and to $\mathbf{y}$ by finding:
        - $\hat{\boldsymbol{\beta}}^{d_y,j}(\lambda) \leftarrow$ `solve_sparse_logistic_cv`$(\mathbf{X}_{-j}, \mathbf{y})$ `// Using Eq.` (2)
        - $\hat{\boldsymbol{\beta}}^{d\mathbf{x}_{*,j}}(\lambda) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \frac{1}{2} \|\mathbf{X}_{*,j} - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2 + \lambda_{dx} \|\boldsymbol{\beta}\|_1$ `// with` λ_{dx} `set using cross-validation`
    2. Obtain test statistic:

$$T_j = \sqrt{n} \frac{\langle \mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y,j}, \mathbf{X}_{*,j} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d\mathbf{x}_{*,j}} \rangle}{\left\|\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y,j}\right\|_2 \left\|\mathbf{X}_{*,j} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d\mathbf{x}_{*,j}}\right\|_2}$$

    3. Compute (two-sided) p-value

$$\widehat{p}_j = 2\left[1 - \Phi\left(T_j\right)\right]$$

end

Algorithm 4: Holdout Randomization Test [26]

INPUT dataset $(\mathbf{X}, \mathbf{y})$, with $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$, number of sampling runs B, test statistic T_j, conditional distribution $P_{j|-j}$ for each $j = 1, \ldots, p$, empirical risk $L(\cdot)$;
OUTPUT vector of p-values $\{\hat{p}_j\}_{j=1}^p$;
$(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}), (\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}}) \leftarrow$ `data_splitting`$(\mathbf{X}, \mathbf{y})$;
$\hat{f}_\theta \leftarrow$ `model_fitting`$(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$;
for $j = 1, 2, \ldots, p$ **do**
    $T_j \leftarrow L(\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}}, \hat{f}_\theta(\mathbf{X}_{\text{test}}))$;
    for $b = 1, 2, \ldots, B$ **do**
        1. Generate $\tilde{\mathbf{X}}_{*,j}^{(b)} \sim P_{j|-j}$;
        2. $\tilde{T}_j^{(b)} \leftarrow L(\tilde{\mathbf{X}}_{*,j}^{(b)}, \mathbf{y}_{\text{test}}, \hat{f}_\theta(\tilde{\mathbf{X}}_{*,j}^{(b)}))$;
    end
    Compute the empirical p-value

$$\widehat{p}_j = \frac{1 + \sum_{b=1}^{B} \mathbf{1}_{\tilde{T}_j^{(b)} \geq T_j}}{1 + B}$$

end

E Time complexity of Related Methods

We present the time complexity of benchmarked methods in Table 2.

Table 2: **Time complexities of related methods with CRT-logit**, where p is the dimension size (number of variables), B is the number of sampling runs, and $\hat{k} \ll p$ the cardinality of the screening set (see Section D for more details).

Methods	Time (Iteration) Complexity	References
Debiased Lasso	$\mathcal{O}(p^4)$	[33, 28, 16]
Knockoff Filter	$\mathcal{O}(p^3)$	[4, 10]
CRT	$\mathcal{O}(Bp^4)$	[10]
HRT	$\mathcal{O}(p^3 + Bp^2)$	[26]
dCRT (with screening)	$\mathcal{O}(\hat{k}p^3)$	[19]
CRT-logit (with screening)	$\mathcal{O}(\hat{k}p^3)$	**(this work)**

F Additional Details on Experiments in Section 4

F.1 Preprocessing of the brain-imaging dataset

The Human Connectome Project dataset (HCP) is a collection of brain imaging data on healthy young adult subjects with age ranging from 22 to 35. The participants performed different tasks while being scanned by a magnetic resonance imaging (MRI) device to record blood oxygenation level dependent (BOLD) signals of the brain. The aim of this analysis is to investigate which areas of the brain can predict cognitive activity across participants, while taking into account the information from other brain regions. The brain imaging modalities include, among others, resting-state fMRI (R-fMRI) and task-evoked fMRI (T-fMRI). In this work, we only deal with decoding the task-evoked fMRI dataset. The four classification problems we are working with are as follows.

- Relational: predict whether the participant matches figures or identified feature similarities.
- Gambling: predict whether the participant gains or loses gambles.
- Emotion: predict whether the participant watches an angry face or a geometric shape.
- Social: predict whether the participant watches a movie with social behavior or not.

To perform dimension reduction, we apply a clustering scheme that preserves the spatial structure of the data. This is achieved with data-driven parcellation along with a spatially constrained clustering algorithm, following the conclusions by [29] and [27]. The hierarchical clustering scheme that we use recursively merges pair of clusters of features based on a criterion that minimized the within-cluster variance. This algorithm is implemented in `scikit-learn` [23], a popular package for applied machine learning.

G Extra experiment: application on genome-wide association study with Human Brain Cancer Dataset

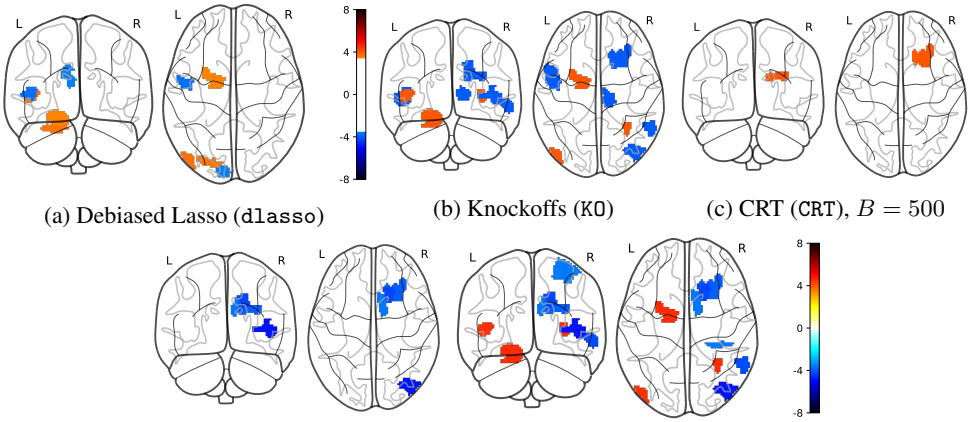
Description The last in our benchmark is a Genome-wide Association Study (GWAS) on the The Cancer Genome Atlas (TCGA) dataset [30, 31]. We choose to analyze the Glioma cohort, which consists of $n = 1026$ patients across a wide age range, diagnosed with this type of brain tumor, with a total of $p = 24776$ genes in the data matrix, recorded as copy number variations (CNVs) at the gene level in log ratio format. As with the brain-imaging inference in Section 4.3, we use clustering to reduce the dimension to $C = 1000$ clusters. However, we use different criterion to merge variables (genes) to clusters of variables, which is the pairwise Linkage Disequilibrium, following [1, Section 4] (with available R library). For the response, a long-term survivor (LTS) is defined as a patient who survived more than five years after diagnosis and would be labeled $y = 0$, and otherwise would be a short-term survivor (STS), labeled $y = 1$. The objective is to identify significant genes that contribute to classification of the LTS/STS status. Similar to the Human Connectome Project dataset, there is no real ground-truth for the TCGA Glioma. However, we have the list of mutations and the frequency of those detected in the diagnosed patients. We therefore select the 1000 most frequent gene mutations that appeared in this list, *i.e.* the ground truth list consists of 1000 genes (variables).

Table 3: **List of detected genes associated with Glioma Cancer from the TCGA dataset.** $n = 1026$, $p = 24776$ (clustered to $C = 1000$). Empty line (—) signifies no detection. Methods listed in the table are the clustering version. Commonly detected genes between methods are put in bold text. Most detected genes are listed in the mutant list database that can be found in the recorded patients [30].

Methods	Detected Genes
dLasso	—
KO	**ABCC10**, **ANK3**, CDH23, PTEN, **SPEN**, **SVIL**, ZMIZ1
dCRT	**ANK3**, **ANKRD30A**, CDH23, PTEN, RET, **SPEN**,ZMIZ1
CRT-logit	**ABCC10**, **ANKRD30A**, BCOR, EPHA3, PPL, SPAG17, **SPEN**, **SVIL**, USP9X
Original CRT	**ABCC10**, BCOR, EPHA3, **SPEN**, **SVIL**
HRT	**ABCC10**, **SPEN**

Result The result from Table 3 shows that CRT-logit finds the largest number of genes. Moreover, most of selected genes in this table are detected in the list of mutated genes found on recorded patients. Some genes are detected by all the benchmarked methods, most prominently `SPEN`, which is found on over 10 % of patients in the cohort. Furthermore, this gene is known to be associated not only with brain cancer, but also with other types of cancer in The Human Protein Atlas project [17]. Note that, in the absence of a ground-truth, this does not guarantee all genes found are associated with glioma, but this experiment demonstrates the capability of CRT-logit in GWAS studies.

H Example of decoding maps in semi-realistic brain-analysis experiment of Section 4.3



(a) Debiased Lasso (`dlasso`) (b) Knockoffs (`KO`) (c) CRT (`CRT`), $B = 500$

(d) Distilled-CRT (`dCRT`) (e) CRT-logit (our method)

Figure 6: **Decoding maps of Relational task in semi-realistic HCP900 experiment, using 400 subjects and dimension reduction to 1000 clusters (*i.e.* one random seed for generating labels $\mathbf{y}$)**. We omit Holdout Randomization Test (`HRT`) as the method does not select any brain region. For `dlasso`, `dCRT` and `CRT-logit`, we plot the test-statistics; for `KO` the sign of selected coefficients, and for `CRT` the $-log_{10}$ of the empirical p-values.

I Ineffectiveness of CRT in extremely high-dimensional problems

When the number of observations n is too small compared to the number of variables p, *e.g.* when $n/p < 0.2$ as shown in Figure 7, the inference problem becomes too ill-posed. Indeed, the statistical power of both the original dCRT and our proposed solution CRT-logit decrease dramatically from a large value in the easy setting ($p < n$) to zero when $p > 1600$. The failure to detect any significant variable when the dimension of the problem becomes too high hints on future direction of performing statistical inference on clusters of variables. For instance, the works of [9, 12] have provided detailed discussions on this matter.

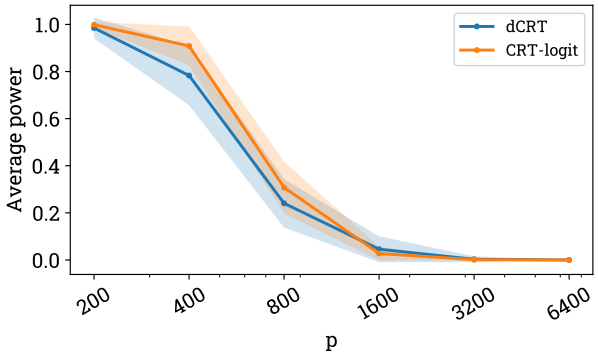


Figure 7: **FDR/Average Power of 100 runs of simulations while varying the number of variables p and fixing the number of observations** $n = 400$. Default parameter: SNR $= 2.0, \rho = 0.5, \kappa = 0.04$. FDR is controlled at level $\alpha = 0.1$. The experimental setup is similar to Section 4.1. Both methods (`dCRT`: original dCRT and `CRT-logit`: our version of CRT) perform well in easy settings where $n \geq p$, but cannot detect any variables when p becomes large compared to n.