SPECTRAL DECONFOUNDING FOR CAUSAL INFERENCE FROM HIGH-DIMENSIONAL PERTURBED MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

High-dimensional causal inference from observational studies is challenging due to unobserved confounders. To ensure identifiability, previous works assume the *dense confounding* condition, which means that each latent confounder is correlated with most covariates. This dense confounding introduces strong correlations among covariates, leading to large singular values along certain directions of the design matrix. To remove the confounding effect, spectral deconfounding methods apply spectral transformations to the design matrix, attenuating the dominant singular components caused by latent confounders. In sparse linear regression, it has been established that such deconfounding can improve the estimation for Lasso. However, a fundamental question remains open: *Can spectral deconfounding identify which features are causally relevant to the response?* This is known as the model selection problem and is important in many applications like scientific discovery. In this paper, we show that by weakening confounding-induced spurious correlations, spectral deconfounding enables the irrepresentable condition holds asymptotically, therefore enabling Lasso to identify and optimize within the causal features set. This perspective in turn motivates us to reduce the bias produced by the ℓ_1 penalty, and thus improving the estimation. Specifically, we employ the Linearized Bregman Iteration, which was proposed as an optimization algorithm and was recently found to enjoy the unbiased property from a differential-inclusion perspective in the fixed design setting. We extend this property to random design setting with latent confounders, and show that, our estimator achieves the oracle ℓ_2 estimation rate $\sqrt{s \log s/n}$ under the beta-min condition, matching the minimax-optimal rate for unconfounded setting with fixed design up to a logarithmic factor.

1 INTRODUCTION

High-dimensional causal inference from observational data aims to identify causally relevant features and estimate their effects, which is fundamental to economics, biology, and modern AI applications. A central challenge lies in the existence of hidden confounders, which can introduce bias in estimation. In low dimensional scenarios, one can use instrumental variables (IVs) to achieve identifiability Guo et al. (2018); Burgess et al. (2017); Pfister & Peters (2022); Wooldridge (2010). However, constructing IVs is challenging in high dimensional regimes since one has to construct as many IVs as covariates to ensure identifiability Wooldridge (2010).

To ensure identifiability, another line of work adopts the *dense confounding* model (Ćevid et al., 2020; Guo et al., 2018; Sun et al., 2024; Wang & Shah, 2025), under which a small number of latent factors simultaneously affect many covariates. This forces the effect of latent confounding to appear as a low rank structure in the covariates, thereby making it distinguishable from the sparse signal of interest. Novembre et al. (2008); McCarthy et al. (2008) and social sciences. This dense confounding model is closely related to the high dimensional factor model Fan et al. (2020); Lam et al. (2011) and widely holds in many scenarios, e.g., biology. One motivating example is the large language model (LLM) based evaluation, where an LLM assigns a scalar value/quality score to a multimodal input such as text or an image. In this application, the dense confounding naturally holds, as both input and evaluation can be simultaneously affected by a

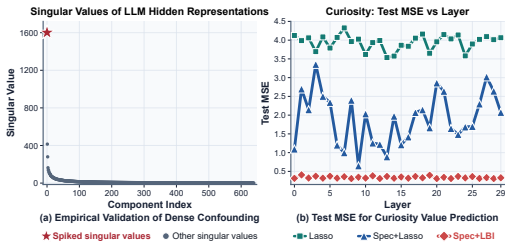


Figure 1: Performance in LLM based evaluation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

few latent confounders, such as prompt format, writing style, etc. To demonstrate, we consider a dataset collected for value evaluation (details are provided in Section 5.2). Figure 1 (a) shows the spectrum of the design matrix constructed from LLM hidden representations. A few leading singular values are significantly larger than the rest, suggesting a low-rank structure in the design.

To mitigate such bias induced by the low rank structure, an effective approach is performing spectral deconfounding Novembre et al. (2008); Jia & Rohe (2015); Fan et al. (2020); Jia & Rohe (2015); Ćevid et al. (2020). Such methods operate on the spectrum of the design matrix, attenuating the leading singular directions dominated by latent confounders. In particular, Ćevid et al. (2020); Guo et al. (2022) showed that the spectral deconfounding can improve the estimation accuracy. However, achieving accurate coefficient estimates does not necessarily imply the correct selection of causally relevant features. Thus, a structural question remains: *Can spectral deconfounding improve the selection of causally relevant features?*

This question is fundamental in many scientific applications, such as identifying disease-causing genes (Li et al., 2025; Jablonski et al., 2022). In this paper, we provide an affirmative answer. We show that by attenuating the dominant singular directions induced by latent confounding, spectral deconfounding substantially weakens multicollinearity. Consequently, the irrepresentable condition (Zhao & Yu, 2006) holds asymptotically. This imposes weak collinearity between causal features and others, and is almost necessary to ensure identifiability of causal features.

Moreover, this feature selection perspective has important implications for estimation. First, it helps explain the improved estimation obtained in previous studies Ćevid et al. (2020): the transformation reduces confounding induced correlations, thereby helping the Lasso avoid spurious features and estimate primarily on the true support, as illustrated by the red curve in Figure 1 (b). More importantly, it motivates bias reduction procedures for the transformed problem, such as the Smoothly Clipped Absolute Deviation (SCAD) estimator (Fan & Li, 2001). In this paper, we consider an alternative estimator, i.e., Linearized Bregman Iteration (LBI), as it is simple to implement. LBI was originally introduced as a sparse optimization algorithm Osher et al. (2005); Yin et al. (2008). From a dynamical perspective, Osher et al. (2016) showed that Bregman iteration can generate an asymptotically unbiased solution path and established feature selection consistency and estimation consistency under fixed design. We extend this property to the random design setting with latent confounding. Benefiting from its unbiasedness property, our estimator achieves a smaller estimation error, as illustrated by the blue curve in Figure 1(b). Theoretically, it attains an ℓ_2 convergence rate of $\sqrt{s \log s/n}$ under the beta-min condition, improving on the rate of the Lasso estimator (Guo et al., 2022), where n, p, and s denote the sample size, feature dimension, and support size. Moreover, this rate is the same minimax optimal rate in the standard unconfounded setting with fixed design up to a logarithmic factor.

To summarize, our contributions are:

- We show that after spectral deconfounding, the irrepresentable condition holds asymptotically in high dimensional latent confounding settings.
- We extend the feature selection and estimation consistency of LBI from the fixed design to the random design. In particular, by employing the asymptotically unbiasedness property of its solution path, our estimator achieves an ℓ_2 estimation rate of $\sqrt{s \log s/n}$, which improves over the rate for the Lasso estimate and also the standard $\sqrt{s \log p/n}$ minimal optimal ℓ_2 rate in the fixed design.
- We demonstrate the effectiveness of our proposed method on synthetic datasets and two real world applications: LLM value evaluation and gene expression analysis.

Notations. For a vector $x \in \mathbb{R}^p$, let $\|x\|_q = \left(\sum_{l=1}^p |x_l|^q\right)^{1/q}$ for $q \geq 1$, and $\|x\|_\infty = \max_{1 \leq l \leq p} |x_l|$. For a matrix $X \in \mathbb{R}^{n \times p}$, let $X^* \stackrel{\text{def}}{=} n^{-1} X^\top$ and $X^\dagger \stackrel{\text{def}}{=} X \left(X^* X\right)^{-1}$. Let $\lambda_j(X)$ denote its jth largest singular value; $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ denote its largest and smallest singular values. We write $\|X\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^p |x_{ij}|$ and $\|X\|_2$ for the spectral norm of X . Then, $X_{i,\cdot}$ and X_j denote the ith row and jth column of X. For a sequence of random variables X_n, we write $X_n \to_p X$ for convergence in probability. We use $X_n = o_p(a_n)$ if $X_n/a_n \to_p 0$, $X_n = O_p(a_n)$ if for every $\varepsilon > 0$ there exists $C > 0$ such that $\mathbb{P}(|X_n| > C a_n) < \varepsilon$ for all sufficiently large n, $X_n = \Omega_p(a_n)$ if $a_n/|X_n| = O_p(1)$, and $X_n = \Theta_p(a_n)$ if both $X_n = O_p(a_n)$ and $X_n = \Omega_p(a_n)$ hold. ∂f denotes the subdifferential of function f. We set $a \wedge b = \min(a, b)$, and $\lfloor \cdot \rfloor$ denotes the floor function.

2 RELATED WORKS

Spectral Deconfounding Methods. To identify causal effects under unobserved confounding, one common way is to assume *dense confounding*: confounders affect many covariates or treatments. In high dimensional regression, this is typically modeled through a low rank factor structure in the covariates. To mitigate confounding, several spectral methods have been proposed, including PCA adjustment (Novembre et al., 2008), FarmSelect (Fan et al., 2020), the Puffer transformation (Jia & Rohe, 2015), and the Trim transformation (Ćevid et al., 2020; Guo et al., 2022; Sun et al., 2024). Among these methods, Trim is a simple but general spectral transformation whose analysis can be easily extended to others. As a representative spectral transformation, Trim shrinks only the leading singular values to a common level without estimating the number of factors. While existing work mainly justifies Trim from an estimation perspective (Ćevid et al., 2020), we provide a complementary feature selection view: attenuating dominant singular directions reduces multicollinearity, making the irrepresentable condition more likely to hold.

High Dimensional Variable Selection. The Lasso estimator (Tibshirani, 1996) is widely used for support recovery. Exact recovery typically requires weak correlations between important and null features, formalized by incoherence conditions (Donoho & Huo, 2001; Cai et al., 2010) or the closely related *irrepresentable condition* (Zhao & Yu, 2006; Meinshausen & Bühlmann, 2006; Tropp, 2004), which is nearly necessary for exact recovery of causally relevant features. Dense confounding can violate these conditions by inducing spurious correlations, while the ℓ_1 penalty of Lasso introduces bias even without confounding. One way to reduce bias is to use nonconvex penalties (Fan & Li, 2001), although computing the global optimum is often difficult. To reduce computational cost and obtain unbiased estimates, recently Bregman iteration is proposed (Osher et al., 2005; Yin et al., 2008). Osher et al. (2016) subsequently analyze its statistical properties from a differential-inclusion perspective and established minimax-optimal ℓ_2 convergence rates. Existing analyses, however, are largely limited to unconfounded fixed designs. We extend them to random designs, where latent confounders induce endogeneity and additional technical challenges.

3 PRELIMINARIES, BACKGROUND, AND MOTIVATION

Formulation. Consider the following linear regression model with latent confounding,

$$y = X\beta^* + U\phi + e_y, \quad X = U\Psi^\top + E, \tag{1}$$

where $X \in \mathbb{R}^{n\times p}$ is the covariate matrix, and $y \in \mathbb{R}^n$ is the response vector. We consider the high dimensional scenario, i.e., $n \ll p$. We are interested in estimating the causal vector $\beta^* \in \mathbb{R}^p$, which can be challenging due to the *endogeneity issue* (i.e., $U\phi + e_y$ is correlated with X) induced by the latent confounder. Here, we denote $U \in \mathbb{R}^{n\times q}$ as the unobserved low dimensional confounding matrix, (i.e., $q \ll p$), which induces a latent factor structure in the design through the loading matrix $\Psi \in \mathbb{R}^{p\times q}$. We assume that U, the noise term e_y (whose standard deviation is denoted as σ), and E have *i.i.d* zero mean sub Gaussian entries. Moreover, we assume they are independent. Without loss of generality, we assume $\mathrm{Cov}(U) = I_q$ (with a bit abuse of notation, we write $\mathrm{Cov}(X, U)$ as the covariance between each row of X and U), as the model is unchanged under the transformation $U \leftarrow U\,\mathrm{Cov}(U)^{-1/2}$, $\phi \leftarrow \mathrm{Cov}(U)^{1/2}\phi$. Besides, we assume $\mathrm{Cov}(E) = I_p$.

Dense Confounding and Trim Transformation. To ensure identifiability in the high dimensional regime, we assume β^* is sparse, i.e., the support size $s \overset{\text{def}}{=} |S| \ll p$, where $S \overset{\text{def}}{=} \mathrm{supp}(\beta^*)$ denotes the set of causal features, and $T \overset{\text{def}}{=} S^c$ denotes the set of spurious features. However, sparsity alone does not guarantee identifiability due to endogeneity, i.e., $\mathrm{Cov}(X, U) \neq 0$, introducing the bias in the Ordinary Least Squares (OLS) estimator,

$$b \overset{\text{def}}{=} \mathbb{E}(\widehat{\beta}) - \beta^* = \mathrm{Cov}(X)^{-1}\,\mathrm{Cov}(U, X)\phi. \tag{2}$$

To restore the identifiability, existing methods assumed the *dense confounding* condition (Ćevid et al., 2020; Guo et al., 2022; Wang & Shah, 2025).

Assumption 3.1 (Dense confounding)**.** We assume $\lambda_{\min}(\Psi) = \lambda_{\min}\{\mathrm{Cov}(X, U)\} = \Omega(\sqrt{p})$.

This assumption is standard in factor model analysis (Paul et al., 2008) and is almost necessary to achieve identifiability Ćevid et al. (2020). It holds when a small number of latent confounders

$$\rho_{k+1} + \beta_{k+1}/\kappa = \rho_k + \beta_k/\kappa - \alpha X^*(X\beta_k - y), \quad (4a)$$
$$\rho_k \in \partial\|\beta_k\|_1. \quad (4b)$$

$$\xrightarrow{\text{Let } \kappa\to\infty,\ \alpha\to 0}$$

$$\dot{\rho}_t = -X^*(X\beta_t - y), \quad (5a)$$
$$\rho_t \in \partial\|\beta_t\|_1. \quad (5b)$$

Linearized Bregman Iteration **Bregman Inverse Scale Space**

weakly affect all covariates, or affect only a subset of covariates but with sufficiently strong loadings. Under this assumption, the bias term vanishes asymptotically, since $\|b\|_2 = \|(\Psi\Psi^\top + I)^{-1}\Psi\phi\|_2 \lesssim \|\phi\|_2/\lambda_{\min}(\Psi) = O_p(\|\phi\|_2/\sqrt{p})$. This, however, does not imply a fast rate for estimating β^*. Specifically, sparse recovery typically relies on regularized estimators such as the Lasso. For the Lasso estimate, Theorem 2 of Ćevid et al. (2020) shows that, under the compatibility condition, the estimation error bound of $\|\widehat{\beta} - \beta^*\|_1$ depends on the $\|Xb\|_2$, which can remain large even when $\|b\|_2 \to 0$ as $p \to \infty$. Specifically, one can easily see that b satisfies $\mathrm{Cov}(X, U\phi - Xb) = 0$. That means, Xb is the projection $U\phi$ onto X. As a result, we have $\|Xb\|_2 \approx \|U\phi\|_2 = O_p(\sqrt{n}\|\phi\|_2)$.

Recently, Ćevid et al. (2020) showed that the spectral deconfounding method can improve estimation accuracy. The goal of the spectral deconfounding is to remove the dominant low rank directions in the covariate matrix caused by latent confounders. For simplicity, we only introduce the ***Trim transformation***, which has been shown to perform better than other methods. The key idea is to shrink the leading singular values of the design matrix to a constant level, while leaving the remaining singular values unchanged. Formally, let $X = U\Lambda V^\top$ be the singular value decomposition (SVD), where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{n\wedge p})$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n\wedge p}$. Then the Trim transformation is

$$P_m \overset{\text{def}}{=} U\widetilde{\Lambda}U^\top, \qquad \widetilde{\Lambda} \overset{\text{def}}{=} \mathrm{diag}(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_{n\wedge p}), \qquad \widetilde{\lambda}_i \overset{\text{def}}{=} \min\{1, \lambda_{m+1}/\lambda_i\}. \quad (3)$$

When m is chosen to be larger than q, the leading singular value of $\widetilde{X}$ can be significantly reduced. Since $\mathrm{Cov}(X) = \Psi\Psi^\top + I_p$ is driven by the low rank component $\Psi\Psi^\top$, we have $\lambda_{\max}(\widetilde{X}) = O_p(\sqrt{p})$ after shrinkage, leading to $\|\widetilde{X}b\|_2 \leq \|\widetilde{X}\|_2\|b\|_2 = O_p(\sqrt{p} \times \frac{\|\phi\|_2}{\sqrt{p}}) = O_p(\|\phi\|_2)$. Guo et al. (2022) showed that the Trim transformation achieves the ℓ_1 error rate of $s\sqrt{\log p/n}$ and the ℓ_2 error rate of $\sqrt{s\log p/n}$ plus an additional term that is related to b and $\lambda_{\min}(\Psi)$.

Linearized Bregman Iteration (LBI) was firstly introduced in Yin et al. (2008) as a sparse optimization algorithm. Under the quadratic loss $\|y - X\beta\|_2^2/2n$, it updates the prime parameter β_k and the dual parameter ρ_k according to equation 4, starting from $\rho_0 = \beta_0 = \mathbf{0}_p$. Here, $\kappa > 0$ and α respectively denote the damping factor and step size, as discussed in Section 4.2.

LBI generates a regularization path different from that of the Lasso, but also enjoys feature selection consistency under the irrepresentable condition Zhao & Yu (2006). That means, LBI can select causal features when they are weakly correlated with spurious features. A key statistical property underlying this consistency guarantee is the no-false-positive property along its regularization path. Osher et al. (2016) characterizes it from a differential-inclusion perspective for fixed design. Specifically, as $\kappa \to \infty$ and $\alpha \to 0$, the limiting dynamics of LBI are given by Bregman inverse scale space (ISS) in equation 5. Bregman ISS generates solution paths β_t from sparse to dense as t goes from zero to infinity, with t playing a role analogous to $1/\lambda$ in the Lasso. Along this path, feature selection consistency holds if, for some τ, $\mathrm{supp}(\beta_t) \subseteq S$ for all $t \leq \tau$ and sign consistency is attained by τ.

More importantly, the solution path of Bregman ISS is piece wise constant and thus is bias free at τ. Specifically, under $\beta_{\tau,S} \neq 0$ and $\beta_{\tau,T} = 0$–an event that holds with high probability–the right continuity of $\beta_{t,S}$ ensures that all causal coefficients remain nonzero over a neighborhood of τ. Based on this, the right continuous differentiability of ρ_t further implies $\dot{\rho}_\tau = 0$, so that equation 8a leads to a least squares problem restricted to S. As a result, the resulting estimator is asymptotically unbiased.

Motivations and Contributions. While previous studies demonstrated the estimation effectiveness of spectral deconfounding, its effectiveness on feature selection remains unclear. To simplify, we only analyze the Trim method, as our analysis can be extended to other spectral deconfounding methods.

We first consider a motivating example of model equation 1, where elements in e_y and E are Gaussian distributed. We set $q = 1$, $p/n = 2$, and let the sample size n vary from $\{50, 100, 150\}$. For the Trim transformation, we choose $m = \lfloor (n \wedge p)^{0.7} \rfloor$. We report the Area Under the Curve (AUC) that measures the variable selection performance of Lasso's solution path (details can be found in Section 5), and the ℓ_2 estimation error.

As shown in Figure 2, the spectral deconfounding (Spec+Lasso or Spec+LBI) significantly improves both variable selection and estimation accuracy. Here, "Spec+" spectral preprocessing before

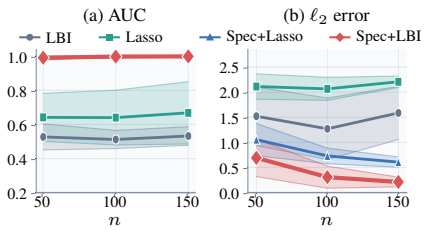


Figure 2: AUC and ℓ_2 error vs. n.

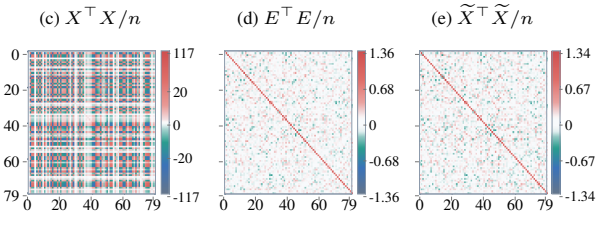


Figure 3: Covariance heatmaps (n=60, p=80, q=1).

applying the regularization method. As an explanation, we will show in the next section that spectral deconfounding can make the irrepresentable condition asymptotically hold. Then, we are motivated to propose a new estimator that significantly reduces bias of the Lasso estimate, and therefore achieves an ℓ_2 error rate of $\sqrt{s \log s/n}$. This rate is better than the minimax optimal rate $\sqrt{s \log p/n}$ under the fixed design. It is worthy to note that the fixed-design results of Osher et al. (2016) do not directly apply to our random-design setting with confounding perturbations.

4 Methodology

In this section, we first examine the no false positive property of the solution path with the Trim transformation, which avoids errors arising from spurious features. We then employ LBI to reduce the estimation bias based on the transformed response and design.

Following Ćevid et al. (2020), we can equivalently rewrite model 1 as a perturbed linear model

$$y = X(\beta^* + b) + \varepsilon_b, \quad \varepsilon_b = U\phi - Xb + e_y, \tag{6}$$

where b is defined in equation 2, and ε_b is called the adjusted noise. By construction, ε_b is uncorrelated with X. Therefore, one can only estimate $\beta^* + b$ rather than β^*. We denote $\widetilde{X} = P_m X$ and $\widetilde{y} = P_m y$ as the transformed design matrix and response vector, where P_m is defined in equation 3.

4.1 No False Positive via Trim Transformation

The no false positive property means all causal variables enter the solution path before any spurious variable. With an appropriate early stopping in the solution path, we have feature selection consistency. The no false positive property generally relies on the *irrepresentable (IRR) condition* Zhao & Yu (2006). Intuitively, the IRR condition requires that causal features are weakly correlated with others. Formally, the condition is given by

$$\max_{j \in T} \left\| (X_S^* X_S)^{-1} X_S^* X_j \right\|_1 \leq 1 - \eta, \quad \text{for some } \eta \in (0, 1].$$

Note that $(X_S^* X_S)^{-1} X_S^* X_j$ is the coefficient vector from regressing the spurious feature X_j ($j \in T$) on causal covariates X_S. Thus, the condition requires that each spurious feature has only limited linear dependence on those causal features. Under this condition, we can establish the no false positive property Wainwright (2019) (Proposition 4.4), analogous to that in fixed design linear models (Zhao & Yu, 2006; Wainwright, 2019). Formally, let $\widehat{\beta}(\lambda)$ denote the regularization solution path with respect to the penalty parameter λ. Besides, $\widehat{\rho} \in \partial \|\widehat{\beta}\|_1$ denotes the sub differential of $\|\widehat{\beta}\|_1$. For each i, we have $\widehat{\rho}_i = \text{sign}(\widehat{\beta}_i)$ if $\widehat{\beta}_i \neq 0$, and belongs to $[-1, 1]$ if $\widehat{\beta}_i = 0$. From the Karush-Kuhn-Tucker (KKT) condition, we have the no false positive property, i.e., $\widehat{\beta}_T(\lambda) = 0$ when $\lambda \geq \overline{\lambda}$ for some $\overline{\lambda}$, if the following holds with high probability,

$$\|\widehat{\rho}_T(\lambda)\|_\infty = \|X_T^*(y - X_S\widehat{\beta}_S)/\lambda\|_\infty = \|X_T^* X_S^\dagger \,\text{sign}\,(\beta_S) + X_T^* P_T \varepsilon_b/\lambda\|_\infty < 1, \tag{7}$$

where $P_T = I - X_S^\dagger X_S^*$ is the projection matrix onto $\text{im}(X_S)^\perp$ and $X_S^\dagger \overset{\text{def}}{=} X_S \left(X_S^* X_S\right)^{-1}$. To ensure the above property for λ, it is sufficient to have *(i).* the **IRR term** $\|X_T^* X_S^\dagger\|_\infty = o_p(1)$, and *(ii).* the **noise term** $\|X_T^* P_T \varepsilon_b\|_\infty < \lambda$ with high probability, where λ is the penalty parameter.

IRR term. Under dense confounding, the IRR condition is often difficult to satisfy because the low rank component $\Psi\Psi^\top$ in $\text{Cov}(X) = \Psi\Psi^\top + I_p$ generates strong global dependence across the coordinates of X. The Trim transformation largely attenuates this confounding effect. To illustrate, Figure 3 shows the sample covariance matrices before and after transformation, namely

$\widehat{\Sigma}(X) = X^\top X/n$, $\widehat{\Sigma}(\widetilde{X}) = \widetilde{X}^\top \widetilde{X}/n$, and $\widehat{\Sigma}(E) = E^\top E/n$. Under dense confounding, $\widehat{\Sigma}(X)$ exhibits large off-diagonal entries, reflecting strong cross feature correlations. After applying Trim, the off-diagonal entries in $\widehat{\Sigma}(\widetilde{X})$ are significantly reduced and closely resemble those in $\widehat{\Sigma}(E)$. Theorem 4.1 formally shows that the IRR term asymptotically vanishes with the Trim transformation.

Theorem 4.1 (IRR Condition with Trim Transformation)**.** *Consider model 1, where U, e_y, E have i.i.d zero-mean sub-Gaussian entries. Assume $p/n \to c \in (0, \infty)$, $s = o(\sqrt{n/\log n})$ and $q = o(\sqrt{n/s})$. Suppose further that the loading matrix satisfies $\max_i \|\Psi_{i,\cdot}\|_2 = o_p(\sqrt{n/(qs)})$ and has singular values $\lambda_i(\Psi) = \Theta(\sqrt{p})$ for $i = 1, 2, \cdots, q$. If the Trim transformation P_m defined in equation 3 shrinks the top m singular values with $m > q$ and $m = o\left(n^{1-\epsilon}/s\right)$ for every sufficiently small fixed $\epsilon > 0$, then the IRR term $\|\widetilde{X}_T^* \widetilde{X}_S (\widetilde{X}_S^* \widetilde{X}_S)^{-1}\|_\infty \to_p 0$ for the transformed design $\widetilde{X} \overset{\text{def}}{=} P_m X$.*

The proof is highly technical and is left to Appendix A.1. The theorem shows that spectral transformation reduces confounding-induced correlations, helping regularization methods select causal features under mild conditions. We require the sparsity condition $s = o(\sqrt{n/\log n})$, commonly used for Lasso ℓ_2 consistency van de Geer & Bühlmann (2009). The proof is more challenging than for Puffer Jia & Rohe (2015), whose analysis assumes a transformed design uniformly distributed on the Stiefel manifold, approximating a Gaussian random design. However, this property does not generally hold after Trim under confounding, and thus their analysis cannot be directly extended.

Noise term. By substantially attenuating the dominant singular values, the Trim transformation removes this factor and suppresses the noise term. Theorem 4.2 formalizes this result.

Theorem 4.2 (Noise term with Trim Transformation)**.** *Consider model 1 where U, e_y, E have i.i.d zero mean Gaussian entries. Assume $\lim p/n \to c \in (0, \infty)$. Let X_T and P_T be defined as in equation 7, and let P_m denote the Trim transformation defined in equation 3. Then the Trim-transformed design matrix $\widetilde{X} \overset{\text{def}}{=} P_m X$ and noise $\widetilde{\varepsilon}_b \overset{\text{def}}{=} P_m \varepsilon_b$ satisfy*

$$\|\widetilde{X}_T^* P_T \widetilde{\varepsilon}_b\|_\infty = O_p(\|\phi\|_2 \sqrt{p \log p}/n), \text{ with probability approaching to } 1 \text{ as } n \to \infty.$$

The proof is left to Appendix A.2. Together with the IRR term $= o_p(1)$ in Theorem 4.1, the result implies that when $\|\phi\|_2$ is fixed, the solution path of the Lasso has the no false positive property before $\overline{\lambda} = O_p(n/\sqrt{p \log p})$. An analogous result for the LBI solution path is given in Proposition 4.4.

Remark 4.3. The proof differs from that in the fixed-design setting, where concentration inequalities apply directly to $\widetilde{\varepsilon}_b$. Under random design, the Trim transformation P_m depends on X, making $\widetilde{\varepsilon}_b$ statistically dependent on $\widetilde{X}$. This dependence complicates the application of standard tail bounds.

4.2 Bregman Iteration for Estimation

In the previous section, we analyzed the IRR condition and the noise term for the Trim transformation, and showed that the Lasso's solution enjoys the no false positive property. While this property enables reliable variable selection, the Lasso is well known to incur nonnegligible bias from the ℓ_1 penalty.

To mitigate this bias, we leverage LBI to compute the solution path based on the transformed data $(\widetilde{X}, \widetilde{y})$ and refer to our method as Spec LBI. As introduced in section 3, once S is correctly identified, Bregman Iteration yields unbiased estimation on S. Starting from $\rho_0 = \beta_0 = \mathbf{0}_p$, the Spec LBI updates (ρ_k, β_k) according to equation 4, with (X, y) replaced by $(\widetilde{X}, \widetilde{y})$. The iteration involves the damping factor κ and the step size α. We will discuss them below.

Hyperparameters discussion in Spec LBI. To ensure that the quadratic loss $\|\widetilde{y} - \widetilde{X}\beta_k\|_2^2$ decreases monotonically along the iterations, the step size α must satisfy $\kappa\alpha \leq \|\widetilde{X}^* \widetilde{X}\|_2$ as required by (Osher et al., 2016, Theorem 4.12). To well approximate the Bregman ISS dynamics, the damping factor κ should be chosen as large as possible. However, since the step size α scales inversely with κ, excessively large κ leads to increased computational cost. Regarding the choice of the stopping time t (or $\lfloor k\alpha \rfloor$ in LBI), because the model is perturbed by both bias and noise, allowing t to grow too large may result in selecting spurious features. Therefore, we require an early stopping strategy. In practice, we select t using cross validation.

We first show that Spec LBI's solution path also satisfies the no false positive property. We provide the result for the dynamic form of Spec LBI, namely Spec *Linearized Bregman Inverse Scale Space*

(LBISS), as the result for the LBI is similar. Let $\alpha \to 0$ in equation 4, Spec LBISS is

$$\dot{z}_t = -\widetilde{X}^*(\widetilde{X}\beta_t - \widetilde{y}) \quad (8a) \qquad z_t \in \partial(\|\beta_t\|_1 + \|\beta_t\|_2^2/2\kappa). \quad (8b)$$

Proposition 4.4 establishes the no false positive property for Spec LBISS equation 8.

Proposition 4.4 (No false Positive: Informal)**.** *Let $\eta \in (0,1)$ and $\lim p/n \to c \in (0,\infty)$. Suppose assumptions in Theorem 4.1 hold. Suppose κ is sufficiently large, such that $\kappa\eta \geq B$, where B is a constant depending on n, p, s and q. Then for any $\delta > 0$, with probability at least $1 - n^{-c_0} - p^{-c_0} - s^{-c_0} - \delta$ for some $c_0 > 0$, we have $\mathrm{supp}(\beta_t) \subseteq S$ when $t \leq \tau \overset{\mathrm{def}}{=} C_0(1 - B/\kappa\eta)\eta n/\sqrt{(\|\phi\|_2^2 + \sigma^2)p\log p}$. Here, $C_0 > 0$ is a constant that depends on c_0, δ, and η.*

Remark 4.5. The form of B and the requirement for κ, along with the proof, are given in Appendix B.2. Our proof extends the analysis in Osher et al. (2016) from the fixed design to the random design, where $\widetilde{X}$ is not necessarily independent of the noise vector $\widetilde{\varepsilon}_b$. Despite this dependence, we achieve $\tau = O(\sqrt{n/\log p})$, as in the fixed-design setting without confounding.

Next, we establish the ℓ_2 estimation consistency of the Spec LBISS under the perturbed model. We begin by introducing the oracle estimator $\widetilde{\beta}^*$, defined by $\widetilde{\beta}^*_T = 0$ and $\widetilde{\beta}^*_S = (\widetilde{X}_S^\top \widetilde{X}_S)^{-1}\widetilde{X}_S^\top \widetilde{y}$ $= \beta^*_S + (\widetilde{X}^*_S\widetilde{X}_S)^{-1}\widetilde{X}^*_S\widetilde{\varepsilon}_b + b_S + (\widetilde{X}^*_S\widetilde{X}_S)^{-1}\widetilde{X}^*_S\widetilde{X}_T b_T$. Our analysis proceeds in two stages. In the first stage, we show that the solution path converges to the oracle estimator. The key idea is to compare the original dynamics with an oracle dynamics restricted to the S, i.e., β'_t such that $\beta'_{t,T} = 0$. Since the original dynamic has no false positives up to time τ, it stays on the S during this period. We then show that the oracle dynamics achieves the desired ℓ_2 rate when $t \geq \tau_2$. Since the two dynamics are identical as long as no false positives occur, it suffices to verify that $\tau_2 < \tau$. This allows us to transfer the ℓ_2 rate from the oracle dynamics to the original one. In the second stage, we derive an ℓ_2 bound for the oracle estimator $\widetilde{\beta}^*$ relative to β^*. The main additional challenge here is that the oracle estimator remains biased due to the perturbation term b.

Let $\widetilde{\beta}^*_{\min} \overset{\mathrm{def}}{=} \min_{i\in S}|\widetilde{\beta}^*_i|$. we define two stopping times: τ_1, the earliest time at which β'_t achieves sign consistency, and τ_2, the earliest time at which β'_t reaches the target ℓ_2 accuracy relative to $\widetilde{\beta}^*$,

$$\tau_1 \overset{\mathrm{def}}{=} \inf\{t : \|\beta'_{t,S} - \widetilde{\beta}^*_S\|_2 < \widetilde{\beta}^*_{\min}\},\ \tau_2(C) \overset{\mathrm{def}}{=} \inf\{t : \|\beta'_{t,S} - \widetilde{\beta}^*_S\|_2 \leq C\sqrt{s\log s/n}\}. \quad (9)$$

Consistency of LBISS with respect to the oracle estimator then follows by showing that $\max(\tau_1, \tau_2) \leq \tau$, where τ is defined in Proposition 4.4. The formal result is given by Theorem 4.6.

Theorem 4.6. *Suppose that the assumptions of Theorem 4.1 and Proposition 4.4 hold, except that U, e_y, E have i.i.d zero mean Gaussian entries. Let $\widetilde{\beta}^*$ denote the oracle estimator and β^*_S the true coefficients on the support S. Define $\widetilde{\beta}^*_{\min} = \min_{i\in S}|\widetilde{\beta}^*_i|$. Besides, we assume $q = O_p(\log s)$ and $\lim p/n \to c \in (0,\infty)$. Suppose κ is sufficiently large, such that $(8 + 4\log s)/\widetilde{\beta}^*_{\min} + \log(3\|\widetilde{\beta}^*\|_2/\widetilde{\beta}^*_{\min})/\kappa \leq \tau$. Then, the Spec LBISS's solution path β_t satisfies,*

1. *(Sign-consistency). Assume $\beta^*_{\min} \geq C\,(\|\phi\|_2 + \sigma)\sqrt{qs\log s/n}$ for some $C > 0$. Then for any $\delta > 0$, with probability at least $1 - n^{-c} - p^{-c} - s^{-c} - \delta$ for some constant $c > 0$, we have $\tau_1 \leq \tau$ and $\mathrm{sign}\,(\beta_\tau) = \mathrm{sign}\,(\beta^*)$.*
2. *(ℓ_2-consistency). For any $\delta > 0$, with probability at least $1 - n^{-c} - p^{-c} - s^{-c} - \delta$ for some constant $c > 0$, we have $\tau_2(C) \leq \tau$, and*

$$\|\beta'_{\tau,S} - \beta^*_S\|_2 = O_p\{(1 + \|\phi\|_2)\sqrt{s\log s/n}\}.$$

The proof is left to Appendix B.2. The ℓ_2 consistency rate in Theorem 4.6 depends on the strength of confounding $\|\phi\|_2$. When $\phi = 0$, the rate of Spec-LBISS reduces to the standard LBISS rate $\sqrt{s\log s/n}$ under a random design without confounding. More importantly, it preserves this rate even with confounding under the standard assumption $\|\phi\|_2 = O(1)$ Ćevid et al. (2020). By contrast, previous LBISS theory establishes this rate only for unconfounded fixed designs under IRR condition Osher et al. (2016), which does not hold here without spectral deconfounding. This rate also matches the minimax-optimal rate for unconfounded fixed designs up to a logarithmic factor. It is also sharper than the Lasso rate $\sqrt{s\log p/n}$ under the perturbed model Guo et al. (2022) (we omit the bias term related to $\lambda_{\min}(\Psi)$). This improvement reflects the oracle behavior of Spec-LBISS after support recovery: Once S is identified, the Spec LBISS dynamics reduce to least squares on S, with the rate $\sqrt{s\log s/n}$.

5 EXPERIMENTS

5.1 SIMULATION

We generate $n \in \{100, 200, 300, 500\}$ *i.i.d.* samples (X_i, y_i) according to model 1. Specifically, ϕ, U, E, and e_y are drawn independently from the standard Gaussian distribution. We set $q = 3$ and $p/n = 2$; [1] The entries of Ψ are sampled as $\Psi_{ij} \sim \mathcal{N}(0, 4^2)$, satisfying the dense confounding assumption. We set $s = 5$ and set β^* as $\beta_i^* = 1$ for $i \leq s$ and $\beta_i^* = 0$ otherwise. For the Trim transformation P_m, we set $m = \lfloor (n \wedge p)^{0.7} \rfloor$. We repeat each experiment 100 times.

Compared baselines. We implement Lasso and LBI both with and without the Trim transformation, namely Lasso, Spec Lasso, LBI, and Spec LBI. Implementation details can be found in Appendix E.1.

Result of estimation. Figure 4(a) and 4(e) report the average ℓ_2 estimation error as functions of p and the truncation index $m \overset{\text{def}}{=} \lfloor n^\rho \rfloor$ with $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ with $n = 200$, respectively. As shown in Figure 4(a), the Trim transformation significantly reduces the estimation error, demonstrating its effectiveness in high-dimensional settings. Moreover, its error decreases as p grows, consistent with the fact that $\|b\|_2 \to 0$ as $p \to \infty$. On top of this, LBI further improves estimation accuracy by reducing the bias when κ is sufficiently large.

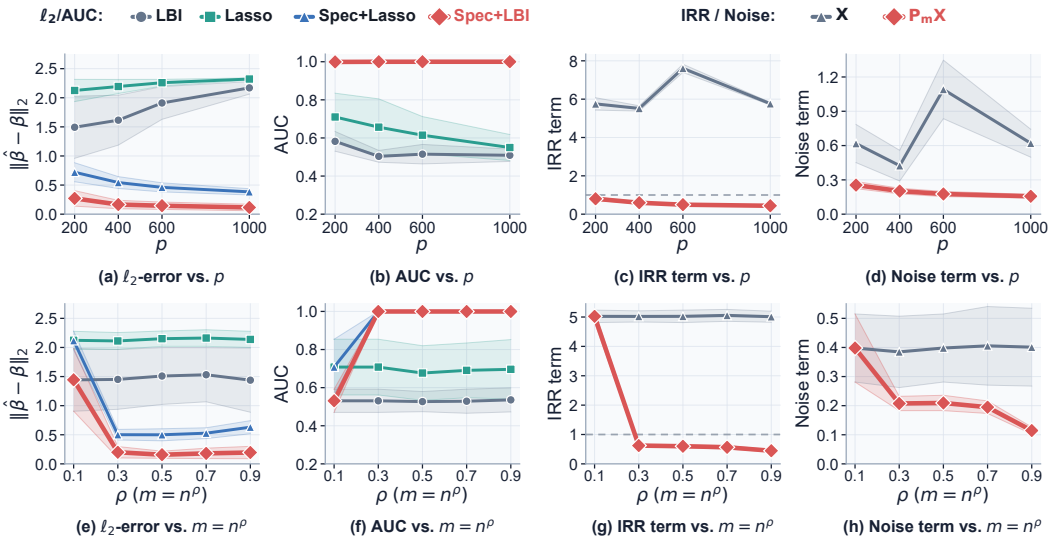


Figure 4: Empirical illustration of Theorems 4.1, 4.2 and 4.6: ℓ_2 estimation performance, no false positive performance (AUC), the irrepresentable-condition term, and the noise term versus dimension (p) (top row) and truncation index ($m = \lfloor n^\rho \rfloor$) (bottom row).

A similar pattern is observed in Figure 4(e), where Spec+LBI also achieves the minimal estimation error. In addition, with the Trim transformation, the ℓ_2 error initially decreases as the truncation index m increases and then remains essentially stable as m continues to grow. Notably, when $\rho = 0.1$ and $m = 200^{0.1} \approx 1.7 < 3 = q$, the truncation fails to remove the confounding subspace, and the results coincide with those of the no-transformation baseline.

Results of variable selection. To evaluate variable selection performance, we rank variables by their selection times along the solution path and compute the AUC with respect to S. Figure 4(b) and 4(f) report the resulting AUC as functions of p and the truncation ratio ρ (through m), respectively. As shown in Figure 4(b), the AUC for methods with spectral deconfounding is close to one, indicating its effectiveness in recovering the set of causal features. Due to the ability to reduce bias, Spec LBI achieves smaller estimation errors. Besides, Figure 4(f) shows that the AUC of spectrally deconfounded methods reaches one once ρ (and thus m) is large enough. In contrast, when $\rho = 0.1$ and $m < q$, the Trim transformation cannot fully remove the low-rank confounding, resulting in an AUC similar to methods on the original design and response. As an explanation, we observe that the Trim transformation significantly reduces the magnitude of both the IRR term and the noise term. In particular, the IRR term falls below one as long as $\rho > 0.1$ (i.e., $m > q$).

[1] Results of other settings can be found in Appendix E.3.

5.2 Application to LLM Value Evaluation

To show that our method improves estimation accuracy, we apply it to evaluate how well LLMs capture human values Schwartz (2012); Zhang et al. (2026). This task may be confounded by nuisance factors such as prompt format. We focus on three dimensions: curiosity, individualism, and satiety.

Dataset Description. We use GPT 4o to generate a set of texts and assign a value score to each one. GPT-4o is prompted to generate 800 sentences describing abstract values and then interpret each in up to 40 words. These descriptions are taken as input. We then split the dataset into 640 training examples and 160 test examples. Based on these descriptions, GPT 4o rates each text on a 5-point scale, ranging from 0 (completely irrelevant) to 4 (highly relevant). In particular, we construct a deconfounded test set, where the label is not affected by latent confounders. In this regard, the testing mean square error (MSE) can reflect the estimation performance. To achieve this, our prompt for annotation asks GPT 4o to ignore prompt format, writing style, and other superficial cues, focusing only on semantic relevance to the target value; see Appendix F.2.

Dense confounding assumption on hidden representation. To apply our method, we extract a 4,096 dimensional hidden representation of original texts from another LLM (we use Llama3.1-8B-Instruct) and fit a sparse linear model to annotations. We show that the learned hidden representation empirically satisfies the dense confounding assumption, as shown in Figure 1 (left).

Results. We evaluate our method using test MSE and assess. As shown in Figure 1 (right), Spec LBI achieves the lowest prediction error across layers on the deconfounded test set, with a mean of 0.339, whereas the Lasso without the Trim transformation has the largest mean error of 3.943. Since we have performed deconfounding in the test data, the prediction error can well reflect the estimation error, suggesting that Spec-LBI provides more accurate estimation. More results can be found in Appendix F.4.

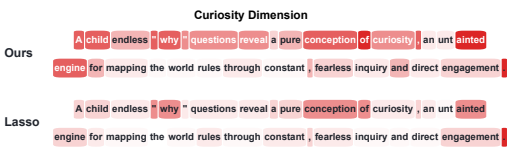


Figure 5: Visualization of attribution score.

Token activation. We quantify token importance by the magnitude of the prediction gradient with respect to each token embedding, i.e., $|\partial\widehat{y}/\partial e_i|$ for each i. A larger value indicates greater token importance and is shown in a darker color. Figure 5 displays an example in the curiosity dimension. As shown, our approach can identify words that are more associated with the target value, such as "why", "question", "curiosity", etc. More details can be found in Appendix F.5.

5.3 Application to Gene expression analysis

To show that our method improves feature selection, we apply it to hippocampal tissue gene expression data from the GTEx Portal Lonsdale et al. (2013). The preprocessed dataset contains $4,027$ protein coding genes from 115 individuals. The dense confounding assumption holds empirically in Fig. 27. We select targets in two ways: (a) 27 literature-selected hippocampus-related genes and (b) 70 randomly selected genes. For each target y, the other genes are covariates X. We measure feature-selection stability by Jaccard similarity $J(A, B) = |A \cap B|/|A \cup B|$ between selected gene sets.

Results. Due to space limitations, we report results for scheme (a), more details see Appendix G. We compute the Jaccard similarity of genes selected by Lasso and LBI, comparing results before and after spectral deconfounding. The results show that spectral deconfounding increases the mean similarity from 0.52 to 0.60, indicating more stable gene selection. Moreover, many genes selected by Spec-LBI are known to be related to hippocampal function. For the representative target brain-derived neurotrophic factor (BDNF) Binder & Scharfman (2004), the selected genes include ACVR1C Keiser et al. (2024), CDH9 Bamji et al. (2006), NPTX2 Mariga et al. (2015), and TH Rossato et al. (2009).

6 Conclusions and Discussions

In this paper, we study the effectiveness of spectral deconfounding methods in feature selection under the dense confounding model. We demonstrate that by removing the low rank latent confounding components, the magnitude of both the irrepresentable term and the noise term can be significantly reduced. We then estimate the causal coefficient with LBI and an early stopping mechanism, which can largely mitigate the bias. We establish the feature selection and ℓ_2 estimation consistency for our estimator. We demonstrate the validity and effectiveness of our method on synthetic data, LLM value evaluation, and Gene expression analysis.

AI USE STATEMENT

In this work, we used generative AI tools to improve the clarity, grammar, and style of author-written text and to generate the dataset used in our LLM experiments. We did not use generative AI tools to generate research ideas or formulate technical claims. All AI-assisted work was carefully reviewed by the authors, and all textual edits were checked to ensure that they preserved the intended technical meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility and independent verification, the paper clearly states the assumptions, definitions, and theoretical setting underlying our results. Complete proofs of the main theoretical claims, together with the necessary intermediate lemmas and additional technical details, are provided in the appendix. As this work is primarily theoretical and does not rely on empirical experiments, datasets, or implementation-specific results, no additional code or data is required to reproduce the main claims.

REFERENCES

Shernaz X. Bamji, Beatriz Rico, Nikole Kimes, and Louis F. Reichardt. BDNF mobilizes synaptic vesicles and enhances synapse formation by disrupting cadherin–beta-catenin interactions. *Journal of Cell Biology*, 174(2):289–299, 2006. doi: 10.1083/jcb.200601087.

Devin K Binder and Helen E Scharfman. Brain-derived neurotrophic factor. *Growth factors (Chur, Switzerland)*, 22(3):123, 2004.

Alex Bloemendal, Antti Knowles, Horng-Tzer Yau, and Jun Yin. On the principal components of sample covariance matrices. *Probability theory and related fields*, 164(1):459–552, 2016.

Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research*, 26(5):2333–2355, 2017.

Tony Tony Cai, Lie Wang, and Guangwu Xu. Stable recovery of sparse signals and an oracle inequality. *IEEE Transactions on Information Theory*, 56(7):3516–3522, 2010.

Domagoj Ćevid, Peter Bühlmann, and Nicolai Meinshausen. Spectral deconfounding via perturbed sparse linear models. *Journal of Machine Learning Research*, 21(232):1–41, 2020.

D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001. doi: 10.1109/18.959265.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan, Yuan Ke, and Kaizheng Wang. Factor-adjusted regularized model selection. *Journal of econometrics*, 216(1):71–85, 2020.

Zijian Guo, Hyunseung Kang, T Tony Cai, and Dylan S Small. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):793–815, 2018.

Zijian Guo, Domagoj Ćevid, and Peter Bühlmann. Doubly debiased lasso: High-dimensional inference under hidden confounding. *Annals of statistics*, 50(3):1320, 2022.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, volume 2024, pp. 2483–2503, 2024.

Michael J Hawrylycz, Ed S Lein, Angela L Guillozet-Bongaarts, Elaine H Shen, Lydia Ng, Jeremy A Miller, Louie N Van De Lagemaat, Kimberly A Smith, Amanda Ebbert, Zackery L Riley, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, 489(7416): 391–399, 2012.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Kim Philipp Jablonski, Martin Pirkl, Domagoj Ćevid, Peter Bühlmann, and Niko Beerenwinkel. Identifying cancer pathway dysregulations using differential causal effects. *Bioinformatics*, 38(6): 1550–1559, 2022.

Jinzhu Jia and Karl Rohe. Preconditioning the lasso for sign consistency. *Electronic Journal of Statistics*, 9(1):1150–1172, 2015. doi: 10.1214/15-EJS1029. URL https://doi.org/10.1214/15-EJS1029.

Ashley A. Keiser, Tri N. Dong, EnikHo A. Kramár, Christopher W. Butler, Siwei Chen, Dina P. Matheos, Jacob S. Rounds, Alyssa Rodriguez, Joy H. Beardwood, Agatha S. Augustynski, Ameer Al-Shammari, Yasaman Alaghband, Vanessa Alizo Vera, Nicole C. Berchtold, Sharmin Shanur, Pierre Baldi, Carl W. Cotman, and Marcelo A. Wood. Specific exercise patterns generate an epigenetic molecular memory window that drives long-term memory formation and identifies ACVR1C as a bidirectional regulator of memory in mice. *Nature Communications*, 15:3836, 2024. doi: 10.1038/s41467-024-47996-w. URL https://doi.org/10.1038/s41467-024-47996-w.

Junsol Kim, James Evans, and Aaron Schein. Linear representations of political perspective emerge in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 7180–7211, 2025.

Clifford Lam, Qiwei Yao, and Neil Bathia. Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918, 2011.

Zhaoyang Li, Yahang Liu, Kecheng Wei, Yongfu Yu, Guoyou Qin, and Zhongyi Zhu. Deconfounded and debiased estimation for high-dimensional linear regression under hidden confounding with application to omics data. *Bioinformatics*, 41(7):btaf400, 2025.

Chia-Chen Liu, Takahisa Kanekiyo, Huaxi Xu, and Guojun Bu. Apolipoprotein e and alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology*, 9(2):106–118, 2013. doi: 10.1038/nrneurol.2012.263.

John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

Abigail Mariga, Juliane Glaser, Leo Mathias, Desheng Xu, Meifang Xiao, Paul Worley, Ipe Ninan, and Moses V. Chao. Definition of a bidirectional activity-dependent pathway involving BDNF and narp. *Cell Reports*, 13(9):1747–1756, 2015. doi: 10.1016/j.celrep.2015.10.064.

Mark I McCarthy, Gonccalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*, 9(5):356–369, 2008.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436 – 1462, 2006. doi: 10.1214/009053606000000281. URL https://doi.org/10.1214/009053606000000281.

John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

Stanley Osher, Martin Burger, Donald Goldfarb, Jinjun Xu, and Wotao Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489, 2005.

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

Debashis Paul, Eric Bair, Trevor Hastie, and Robert Tibshirani. "preconditioning" for feature selection and regression in high-dimensional problems. *The Annals of Statistics*, 36(4):1595–1618, 2008.

Niklas Pfister and Jonas Peters. Identifiability of sparse causal effects using instrumental variables. In *Uncertainty in Artificial Intelligence*, pp. 1613–1622. PMLR, 2022.

Janine I. Rossato, Lia R. M. Bevilaqua, Iván Izquierdo, Jorge H. Medina, and Mart´in Cammarota. Dopamine controls persistence of long-term memory storage. *Science*, 325(5943):1017–1020, 2009. doi: 10.1126/science.1172545.

Shalom H. Schwartz. An overview of the schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1), 2012. doi: 10.9707/2307-0919.1116.

Gilbert W Stewart. Perturbation theory for the singular value decomposition. 1998.

Yinrui Sun, Li Ma, and Yin Xia. A decorrelating and debiasing approach to simultaneous inference for high-dimensional confounded models. *Journal of the American Statistical Association*, 119 (548):2857–2868, 2024.

Jacqueline R Thompson, Erik D Nelson, Madhavi Tippani, Anthony D Ramnauth, Heena R Divecha, Ryan A Miller, Nicholas J Eagles, Elizabeth A Pattie, Sang Ho Kwon, Svitlana V Bach, et al. An integrated single-nucleus and spatial transcriptomics atlas reveals the molecular landscape of the human hippocampus. *Nature neuroscience*, 28(9):1990–2004, 2025.

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

J.A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004. doi: 10.1109/TIT.2004.834793.

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3(none):1360 – 1392, 2009. doi: 10.1214/09-EJS506. URL https://doi.org/10.1214/09-EJS506.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

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

Jacob W Vogel, Renaud La Joie, Michel J Grothe, Alexandr Diaz-Papkovich, Andrew Doyle, Etienne Vachon-Presseau, Claude Lepage, Reinder Vos de Wael, Rhalena A Thomas, Yasser Iturria-Medina, et al. A molecular gradient along the longitudinal axis of the human hippocampus informs large-scale behavioral systems. *Nature communications*, 11(1):960, 2020.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Yuhao Wang and Rajen Shah. Latent confounding in high-dimensional nonlinear models. *arXiv*, 2025. URL https://arxiv.org/abs/2508.06274.

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

Jiechao Xiong, Feng Ruan, and Yuan Yao. A tutorial on: R package for the linearized bregman algorithm in high-dimensional statistics. *Handbook of Big Data Analytics*, pp. 425–453, 2018.

Wotao Yin, Stanley Osher, Donald Goldfarb, and Jerome Darbon. Bregman iterative algorithms for \ell_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging sciences*, 1(1):143–168, 2008.

Junyu Zhang, Yipeng Kang, Jiong Guo, Jiayu Zhan, and Junqi Wang. Bach-v: Bridging abstract and concrete human-values in large language models, 2026. URL https://arxiv.org/abs/2601.14007.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

A NO FALSE POSITIVE FOR TRIM-TRANSFORMATIONS

In this section, we theoretically analyze how the trim transformation can improve the no-false-positive property. Specifically, we show that the IRR condition asymptotically holds, *i.e.*, $\|\widetilde{X}_T^*\widetilde{X}_S(\widetilde{X}_S^*\widetilde{X}_S)^{-1}\|_\infty < 1$ when $\frac{p}{n} \to c \in (0,\infty)$, and that $\|\widetilde{X}_T^* P_T \widetilde{\varepsilon}\|_\infty = O_p(\sqrt{\|\phi\|_2^2 + \sigma^2}\sqrt{p\log p}/n)$.

A.1 PROOF OF THEOREM 4.1

Theorem A.1. *Consider model 1, where U, e_y, E have i.i.d zero-mean sub-Gaussian entries. Assume $p/n \to c \in (0,\infty)$, $s = o(\sqrt{n/\log n})$ and $q = o(\sqrt{n/s})$. Suppose further that the loading matrix satisfies $\max_i \|\Psi_{i,\cdot}\|_2 = o_p(\sqrt{n/(qs)})$ and $\lambda_i(\Psi) = \Theta(\sqrt{p})$ for $i = 1, 2, \cdots, q$. If the Trim-transformation P_m shrinks the top m singular values with $m > q$ and $m = o\left(n^{1-\epsilon}/s\right)$ for every sufficiently small fixed $\epsilon > 0$, then the IRR term $\|\widetilde{X}_T^*\widetilde{X}_S(\widetilde{X}_S^*\widetilde{X}_S)^{-1}\|_\infty \to_p 0$ for the transformed design $\widetilde{X} \overset{\text{def}}{=} P_m X$.*

Proof. Consider the following upper bound:

$$\max_{j\in T}\left\|\left(\widetilde{X}_S^*\widetilde{X}_S\right)^{-1}\widetilde{X}_S^*\widetilde{X}_j\right\|_1 \le \sqrt{s}\max_{j\in T}\|(\widetilde{X}_S^*\widetilde{X}_S)^{-1}\widetilde{X}_S^*\widetilde{X}_j\|_2 \tag{10}$$

$$\le \frac{\sqrt{s}\max_{j\in T}\|\widetilde{X}_S^*\widetilde{X}_j\|_2}{\lambda_{\min}(\widetilde{X}_S^*\widetilde{X}_S)}, \tag{11}$$

$$\le \frac{s\max_{i\neq j}|(\widetilde{X}^*\widetilde{X})_{i,j}|}{\lambda_{\min}(\widetilde{X}_S^*\widetilde{X}_S)}. \tag{12}$$

Then, it is easy to obtain the conclusion once we can show that (1) $\lambda_{\min}(\widetilde{X}_S^*\widetilde{X}_S) = c_0 - O_p(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log n}{n}} + \|\Psi_S\|_2\sqrt{\frac{q}{n}})$; (2) $\max_{i\neq j}|(\widetilde{X}^*\widetilde{X})_{i,j}| = O_p(\sqrt{\frac{\log p}{n}} + \frac{q\max_i\|\Psi_{i,\cdot}\|_2^2}{p} + \frac{\sqrt{q}\max_i\|\Psi_{i,\cdot}\|_2}{\sqrt{p}} + \frac{m}{p^{1-\epsilon}})$ for $\lim\frac{p}{n} \to c \in (0,\infty)$ and the constant c_0. Then the conclusion $\|\widetilde{X}_T^*\widetilde{X}_S(\widetilde{X}_S^*\widetilde{X}_S)^{-1}\|_\infty \to_p 0$ follows by setting $s = o(\sqrt{n/\log n})$, assuming $\max_i\|\Psi_{i,\cdot}\|_2 = o_p(\sqrt{n/(qs)})$.

Next, we prove the claimed bounds for $\lambda_{\min}(\widetilde{X}_S^*\widetilde{X}_S)$ and $\max_{i\neq j}|(\widetilde{X}^*\widetilde{X})_{ij}|$.

Step I. Proof of (1) $\lambda_{\min}(\widetilde{X}_S^*\widetilde{X}_S) = c_0 - O_p\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log n}{n}} + \|\Psi_S\|_2\sqrt{\frac{q}{n}}\right)$.

Suppose the thin singular value decomposition (SVD) of design matrix $X/\sqrt{n}$ is $L\Lambda V^\top$ where $\Lambda = \text{diag}(\lambda_1, \cdots \lambda_{n\wedge p})$. And let L_q denote the matrix whose columns are the left singular vectors corresponding to the q largest singular values of $X/\sqrt{n}$. Correspondingly, denote $\widetilde{\lambda}_i$ as the i-th singular value of the Trim-transformed matrix $\widetilde{X}$. Further, let S denote the support set, and let $\lambda_{\min}(M)$ be the smallest singular value of M. Then we have

$$\begin{aligned}
\lambda_{\min}\left(\widetilde{X}_S^*\widetilde{X}_S\right) &= \lambda_{\min}\left(\sum_{k=1}^{q}\widetilde{\lambda}_k^2 v_{k,S}v_{k,S}^\top + \sum_{k=q+1}^{m}\widetilde{\lambda}_k^2 v_{k,S}v_{k,S}^\top + \sum_{k=m+1}^{n\wedge p}\lambda_k^2 v_{k,S}v_{k,S}^\top\right)\\
&\overset{(1)}{\ge} \lambda_{\min}\left(\sum_{k=q+1}^{m}\widetilde{\lambda}_k^2 v_{k,S}v_{k,S}^\top + \sum_{k=m+1}^{n\wedge p}\lambda_k^2 v_{k,S}v_{k,S}^\top\right)\\
&\overset{(2)}{=} \lambda_{\min}\left(\sum_{k=q+1}^{m}(\widetilde{\lambda}_k^2 - c_0\lambda_k^2)v_{k,S}v_{k,S}^\top + c_0\lambda_k^2 v_{k,S}v_{k,S}^\top + \sum_{k=m+1}^{n\wedge p}\lambda_k^2 v_{k,S}v_{k,S}^\top\right)
\end{aligned}$$

$$\overset{(3)}{\geq} \lambda_{\min}\left(c_0 \sum_{k=q+1}^{m} \lambda_k^2 v_{k,S} v_{k,S}^\top + c_0 \sum_{k=m+1}^{n\wedge p} \lambda_k^2 v_{k,S} v_{k,S}^\top\right)$$
$$= c_0 \lambda_{\min}\left(\frac{X_S^\top}{\sqrt{n}}\left(I - L_q L_q^\top\right)\frac{X_S}{\sqrt{n}}\right). \tag{13}$$

where "(1)" follows from $\sum_{k=1}^{q} \widetilde{\lambda}_k^2\, v_{k,S} v_{k,S}^\top \succcurlyeq 0$; "(2)" is due to $c_0 \overset{\text{def}}{=} \min\left\{\{\frac{\widetilde{\lambda}_i(X^*X)}{\lambda_i(X^*X)}\}_{i>q}, 1\right\} = \Theta_p(1) > 0$ by Lemma C.3 and "(3)" uses the $\sum_{k=q+1}^{m}(\widetilde{\lambda}_k^2 - c_0\lambda_k^2) v_{k,S} v_{k,S}^\top \succeq 0$ and $\sum_{k=m+1}^{n\wedge p} \lambda_k^2 v_{k,S} v_{k,S}^\top - c_0 \sum_{k=m+1}^{n\wedge p} \lambda_k^2 v_{k,S} v_{k,S}^\top \succeq 0$.

It suffices to show that $\lambda_{\min}(\frac{X_S}{\sqrt{n}}(I - L_q L_q^\top)\frac{X_S}{\sqrt{n}} = 1 - o_p(1)$. We first decompose it into the following terms.

$$\lambda_{\min}\left(\frac{X_S^\top}{\sqrt{n}}(I - L_q L_q^\top)\frac{X_S}{\sqrt{n}}\right)$$
$$= \lambda_{\min}\left(\frac{1}{n}(E_S + U\Psi_S^\top)^\top (I - L_q L_q^\top)(E_S + U\Psi_S^\top)\right),$$
$$\overset{(1)}{\geq} \lambda_{\min}\left(\frac{1}{n}E_S^\top\left(I - L_q L_q^\top\right)E_S + \frac{1}{n}E_S^\top\left(I - L_q L_q^\top\right)U\Psi_S^\top + \frac{1}{n}\Psi_S U^\top\left(I - L_q L_q^\top\right)E_S\right),$$
$$\overset{(2)}{\geq} \lambda_{\min}\left(\frac{1}{n}E_S^\top\left(I - L_q L_q^\top\right)E_S\right) - \|\frac{1}{n}E_S^\top\left(I - L_q L_q^\top\right)U\Psi_S^\top + \frac{1}{n}\Psi_S U^\top\left(I - L_q L_q^\top\right)E_S\|_2,$$
$$\overset{\text{def}}{=} I - II. \tag{14}$$

where "(1)" follows the fact $\frac{1}{n}\Psi_S U^\top\left(I - L_q L_q^\top\right)U\Psi_S^\top \succeq 0$. "(2)" follows from Weyl's inequality: $\lambda_{\min}(B + C) \geq \lambda_{\min}(B) + \lambda_{\min}(C)$.

Next, it is sufficient to prove that (1) $I = \frac{n-q}{n} - O_p\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log n}{n}}\right) - O_p\left(\sqrt{\frac{q}{p}}\right)$ and $II = O_p\left(\|\Psi_S\|_2\sqrt{\frac{q}{n}}\right)$.

Step I-A.1: Proof of $I = \frac{n-q}{n} - O_p\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log n}{n}}\right) - O_p\left(\sqrt{\frac{q}{p}} + \frac{\sqrt{sq}}{p}\right)$. Let $\mathcal{U} \overset{\text{def}}{=} \text{col}(U)$ be the confounder subspace and let P_U denote the orthogonal projection onto $\mathcal{U}$, namely $P_U = U(U^\top U)^{-1}U^\top$. Then, we next decompose I to

$$\lambda_{\min}\left(\frac{1}{n}E_S^\top\left(I - L_q L_q^\top\right)E_S\right) = \lambda_{\min}\left(\frac{1}{n}E_S^\top\left(I - P_U\right)E_S + \frac{1}{n}E_S^\top\left(P_U - L_q L_q^\top\right)E_S\right),$$
$$\overset{(1)}{\geq} \lambda_{\min}(\frac{1}{n}E_S^\top\left(I - P_U\right)E_S) - \|\frac{1}{n}E_S^\top\left(P_U - L_q L_q^\top\right)E_S\|_2,$$
$$\geq \lambda_{\min}(\frac{1}{n}E_S^\top\left(I - P_U\right)E_S) - \|P_U - L_q L_q^\top\|_2\|\frac{1}{\sqrt{n}}E_S\|_2^2,$$
$$\overset{(2)}{=} \lambda_{\min}\left(\frac{1}{n}E_S^\top\left(I - P_U\right)E_S\right) - O_p\left(\sqrt{\frac{q}{p}}\right). \tag{15}$$

where "(1)" follows from Weyl's inequality, and "(2)" follows from Lemma C.4, which gives $\|LL^\top - P_U\|_2 = O_p\left(\sqrt{\frac{q}{p}}\right)$, together with the fact that $\lambda_1(E/\sqrt{n}) = O_p(\sqrt{p/n}) = O_p(1)$.

Next, we show that the first term $\lambda_{\min}\left(\frac{1}{n}E_S^\top\left(I - P_U\right)E_S\right) = \frac{n-q}{n} - O_p(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log n}{n}})$.

Define $H \overset{\text{def}}{=} \frac{1}{n}E_S^\top\left(I - P_U\right)E_S - \mathbb{E}(\frac{1}{n}E_S^\top\left(I - P_U\right)E_S) = \frac{1}{n}E_S^\top\left(I - P_U\right)E_S - \frac{n-q}{n}I_s$. By the Lemma D.2 (Weyl's inequality), we have

$$\lambda_{\min}\left(\frac{1}{n}E_S^\top\left(I - P_U\right)E_S\right) \geq \frac{n-q}{n} - \|H\|_2. \tag{16}$$

Then, we will prove that $\|H\|_2 = O_p\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log n}{n}}\right)$. By definition,

$$\|H\|_2 = \max_{\|z\|=1} z^\top (E_S^\top \frac{1}{n} (I - P_U) E_S - \frac{n-q}{n} I_s) z. \tag{17}$$

Since U is independent of E we condition on U, so that P_U is fixed. Use the fact that U and E have *i.i.d* sub-Gaussian entries. For any fixed $z \in \mathbb{S}^{s-1}$, Lemma D.1(Hanson-Wright inequality) applied to $A = n^{-1}(I - P_U)$, for which $\|A\|_2 = 1/n$ and $\|A\|_F = \sqrt{n-q}/n$, gives

$$\mathbb{P}\left(\left| z^\top E_S^\top \frac{1}{n} (I - P_U) E_S z - \frac{n-q}{n} \right| > t\right) \leq 2\exp\left(-\frac{cn^2t^2}{n-q}\right). \tag{18}$$

We next extend equation 18 from a fixed z to unfix $z \in \mathbb{S}^{s-1}$ using a union bound by a standard ϵ-net argument. By Vershynin (2018) Corollary 4.2.11, we can find an ϵ-net $\mathcal{N}$ of the sphere $\mathbb{S}^{s-1}$ with cardinalities $|\mathcal{N}| \leq 9^s$. Then, the union bound gives

$$\begin{aligned} &\mathbb{P}\left(\max_{z\in\mathcal{N}} \left| z^\top \left(E_S^\top \frac{1}{n} (I - P_U) E_S - \frac{n-q}{n} I_S \right) z \right| > t \right) \\ &\leq \sum_{z\in\mathcal{N}} \mathbb{P}\left(\left| z^\top \left(E_S^\top \frac{1}{n} (I - P_U) E_S - \frac{n-q}{n} I_S \right) z \right| > t \right). \end{aligned} \tag{19}$$

Combining equation 18 with the ϵ-net bound equation 19 and $q = o_p(n)$, we obtain that, with probability at least $1 - n^{-c}$,

$$\|H\|_2 = \max_{\|z\|=1} z^\top (E_S^\top \frac{1}{n} (I - P_U) E_S - \frac{n-q}{n} I_s) z \leq \sqrt{\frac{s}{n}} + \sqrt{\frac{\log n}{n}}. \tag{20}$$

Together, equation 20 and equation 16 imply that with high probability

$$\lambda_{\min}\left(\frac{1}{n} E_S^\top (I - P_U) E_S\right) \geq \frac{n-q}{n} - \sqrt{\frac{s}{n}} - \sqrt{\frac{\log n}{n}}. \tag{21}$$

Step I-A.2: Proof of $II = O_p\left(\|\Psi_S\|_2 \sqrt{\frac{q}{n}}\right)$. Recall that U and E have *i.i.d* sub-Gaussian entries. Thus, we decompose II as

$$\begin{aligned} II &\overset{\text{def}}{=} \left\| \frac{1}{n} E_S^\top \left(I - L_q L_q^\top\right) U \Psi_S^\top + \frac{1}{n} \Psi_S U^\top \left(I - L_q L_q^\top\right) E_S \right\|_2, \\ &\leq 2\|\Psi_S\|_2 \left\| \frac{E_S}{\sqrt{n}} \right\|_2 \left\| \left(I - L_q L_q^\top\right) \frac{U}{\sqrt{n}} \right\|_2, \\ &\overset{(1)}{=} O_p\left(\|\Psi_S\|_2 \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log n}{n}} \right) \right) \left\| \left(P_U - L_q L_q^\top\right) \frac{U}{\sqrt{n}} \right\|_2, \\ &\leq O_p\left(\|\Psi_S\|_2 \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log n}{n}} \right) \right) \left\| \left(P_U - L_q L_q^\top\right) \right\|_2 \left\| \frac{U}{\sqrt{n}} \right\|_2, \\ &\overset{(2)}{\leq} O_p\left(\|\Psi_S\|_2 \sqrt{\frac{q}{n}} \right). \end{aligned} \tag{22}$$

where "(1)" follows from the fact that $\|E/\sqrt{n}\|_2 = O_p(1)$ with high probability by applying Theorem 5.39 in Vershynin (2010), and from the fact that $P_U U = U(U^\top U)^{-1} U^\top U = U$. "(2)" where "(2)" follows from $\|U/\sqrt{n}\|_2 = O_p(1)$ with high probability and Lemma C.4, which gives $\|LL^\top - P_U\|_2 = O_p\left(\sqrt{\frac{q}{p}}\right)$.

Combining equation 22, equation 21 and equation 15, and using $q = o(n)$ and $n/p \to c \in (0, \infty)$, we conclude that

$$\begin{aligned}\lambda_{\min}(\widetilde{X}_S^* \widetilde{X}_S) &\geq c_0 \left(\frac{n-q}{n} - O_p\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log n}{n}}\right) - O_p\left(\sqrt{\frac{q}{p}}\right) - O_p\left(\|\Psi_S\|_2 \sqrt{\frac{q}{n}}\right)\right). \\ &= c_0 - O_p\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log n}{n}} + \|\Psi_S\|_2 \sqrt{\frac{q}{n}}\right). \end{aligned} \tag{23}$$

Step II-A Proof of $(2) : \max_{i \neq j} \left| \left(\widetilde{\widetilde{X}}^* \widetilde{\widetilde{X}} \right)_{i,j} \right| = O_p\left(\sqrt{\frac{\log p}{n}} + \frac{\sqrt{q} \max_i \|\Psi_{i,\cdot}\|_2}{\sqrt{p}} + \frac{q \max_i \|\Psi_{i,\cdot}\|_2^2}{p} + \frac{m}{p^{1-\epsilon}} \right)$.

Let $P_m X$ denote the trimmed matrix obtained by clipping the singular values of $X/\sqrt{n}$ at the m-th singular value. Let X_i denote the i-th column of X. Then, for $i \neq j$, the covariance is given by

$$\begin{aligned}\max_{i \neq j} \left| \left\{ \widehat{\Sigma}\left(P_m X\right) \right\}_{i,j} \right| &= \max_{i \neq j} \left| \sum_{k=1}^{q} \widetilde{\lambda}_m^2 v_{ki} v_{kj} + \sum_{k=q+1}^{m} \widetilde{\lambda}_m^2 v_{ki} v_{kj} + \sum_{k=m+1}^{n \wedge p} \lambda_k^2 v_{ki} v_{kj} \right|, \\ &\leq \max_{i \neq j} \left| \sum_{k=1}^{q} \widetilde{\lambda}_m^2 v_{ki} v_{kj} \right| + \max_{i \neq j} \left| \sum_{k=q+1}^{m} \widetilde{\lambda}_m^2 v_{ki} v_{kj} + \sum_{k=m+1}^{n \wedge p} \lambda_k^2 v_{ki} v_{kj} \right|, \\ &\overset{(1)}{\leq} \max_{i \neq j} \left| \sum_{k=1}^{q} \widetilde{\lambda}_m^2 v_{ki} v_{kj} \right| + \max_{i \neq j} \left| \sum_{k=q+1}^{n \wedge p} \lambda_k^2 v_{ki} v_{kj} \right| + \max_{i \neq j} \left| \sum_{k=q+1}^{m} \left(\lambda_k^2 - \widetilde{\lambda}_m^2 \right) v_{ki} v_{kj} \right|, \\ &= \max_{i \neq j} \left| \sum_{k=1}^{q} \widetilde{\lambda}_m^2 v_{ki} v_{kj} \right| + \max_{i \neq j} \left\{ \left| \frac{X^\top}{\sqrt{n}} (I - L_q L_q^\top) \frac{X}{\sqrt{n}} \right| \right\}_{i,j} + \max_{i \neq j} \left| \sum_{k=q+1}^{m} \left(\lambda_k^2 - \widetilde{\lambda}_m^2 \right) v_{ki} v_{kj} \right|, \\ &\overset{\text{def}}{=} I + II + \Delta. \end{aligned} \tag{24}$$

Next, we show that $I = O_p\left(\frac{\|\Psi_{i,\cdot}\|_2^2}{p}\right)$, $II = O_p\left(\sqrt{\frac{\log p}{n}} + \frac{\max_i \|\Psi_{i,\cdot}\|_2^2}{p} + \frac{\max_i \|\Psi_{i,\cdot}\|_2}{\sqrt{p}} \right)$ and $\Delta = O_{\prec}\left(\frac{m}{p}\right)$.

Step II-A Proof of $(2) : I = O_p\left(\frac{\max_i \|\Psi_{i,\cdot}\|_2^2}{p}\right)$. Let λ_m denote the m-th singular value of X, used as the clipping threshold in the Trim-transformation. By the definition of the trim transformation, we decompose the term I as follows:

$$I = \max_{i \neq j} \lambda_m^2 \cdot \left| \sum_{k=1}^{q} v_{ki} v_{kj} \right| \leq \max_i \lambda_m^2 \cdot (\sum_{k=1}^{q} v_{ki}^2)^{1/2} (\sum_{k=1}^{q} v_{kj}^2)^{1/2} \overset{(1)}{\leq} O_p\left(\frac{\max_i \|\Psi_{i,\cdot}\|_2^2}{p} \right). \tag{25}$$

where "(1)" follows from Lemma C.2, namely $\max_{i \in [p]} \left\| (V_q)_{i,} \right\|_2^2 = O_p\left(\frac{\max_i \|\Psi_{i,\cdot}\|_2^2}{p}\right)$, together with Lemma C.3, which gives $\lambda_m^2 = O_p(1)$.

Step II-B Proof of (2) : $II = O_p\left(\sqrt{\frac{\log p}{n}} + \frac{q\max_i\|\Psi_{i,\cdot}\|_2^2}{p} + \frac{\sqrt{q}\max_i\|\Psi_{i,\cdot}\|_2}{\sqrt{p}}\right)$. We decompose the term II as follows:

$$
\begin{aligned}
II &= \max_{i\neq j}\left|\frac{1}{n}(E^\top)_{i,\cdot}\left(I - L_qL_q^\top\right)E_{\cdot,j} + \frac{1}{n}\Psi_{i,\cdot}U^\top\left(I - L_qL_q^\top\right)U(\Psi^\top)_{\cdot,j}\right.\\
&\quad\left.+\frac{1}{n}(E^\top)_{i,\cdot}\left(I - L_qL_q^\top\right)U(\Psi^\top)_{\cdot,j} + \frac{1}{n}\Psi_{i,\cdot}U^\top\left(I - L_qL_q^\top\right)E_{\cdot,j}\right|,\\
&\overset{(1)}{\leq} \max_{i\neq j}\left|\frac{1}{n}(E^\top)_{i,\cdot}\left(I - L_qL_q^\top\right)E_{\cdot,j}\right| + \max_{i\neq j}\left|\frac{1}{n}\Psi_{i,\cdot}U^\top\left(I - L_qL_q^\top\right)^\top\left(I - L_qL_q^\top\right)U(\Psi^\top)_{\cdot,j}\right|\\
&\quad+\max_{i,j}\left|\frac{1}{n}(E^\top)_{i,\cdot}\left(I - L_qL_q^\top\right)U(\Psi^\top)_{\cdot,j}\right| + \max_{i,j}\left|\frac{1}{n}\Psi_{i,\cdot}U^\top\left(I - L_qL_q^\top\right)E_{\cdot,j}\right|,\\
&\overset{(2)}{=} \max_{i\neq j}\left|\frac{1}{n}(E^\top)_{i,\cdot}\left(I - L_qL_q^\top\right)E_{\cdot,j}\right| + \max_{i\neq j}\left|\frac{1}{n}\Psi_{i,\cdot}U^\top\left(P_U - L_qL_q^\top\right)^\top\left(P_U - L_qL_q^\top\right)U(\Psi^\top)_{\cdot,j}\right|\\
&\quad+\max_{i,j}\left|\frac{1}{n}(E^\top)_{i,\cdot}\left(P_U - L_qL_q^\top\right)U(\Psi^\top)_{\cdot,j}\right| + \max_{i,j}\left|\frac{1}{n}\Psi_{i,\cdot}U^\top\left(P_U - L_qL_q^\top\right)E_{\cdot,j}\right|. \qquad (26)
\end{aligned}
$$

where "(1)" uses the idempotence of the orthogonal projection $I - L_qL_q^\top$, namely $(I - L_qL_q^\top)^2 = I - L_qL_q^\top$. "(2)" follows by the same argument as in equation 22, using $P_UU = U$ and $U^\top P_U^\top = U^\top$. Since U has *i.i.d.* sub-Gaussian entries, $\|U/\sqrt{n}\|_2 = O_p(1)$ with high probability. Together with Lemma C.4, which gives $\|LL^\top - P_U\|_2 = O_p\left(\sqrt{\frac{q}{p}}\right)$, the second term satisfies

$$
\begin{aligned}
\max_{i\neq j}\left|\frac{1}{n}\Psi_{i,\cdot}U^\top\left(P_U - L_qL_q^\top\right)^\top\left(P_U - L_qL_q^\top\right)U(\Psi^\top)_{\cdot,j}\right| &\leq \max_i\|\Psi_{i,\cdot}\|_2^2\left\|P_U - L_qL_q^\top\right\|_2^2\left\|U/\sqrt{n}\right\|_2^2,\\
&= O_p\left(\frac{q\max_i\|\Psi_{i,\cdot}\|_2^2}{p}\right). \qquad (27)
\end{aligned}
$$

Similarly, for the third and fourth terms, since E has *i.i.d.* sub-Gaussian entries, a union bound over the p columns gives $\max_i\|E_{\cdot,i}/\sqrt{n}\|_2 = O_p(1 + \sqrt{\frac{\log p}{n}}) = O_p(1)$ with high probability. Hence, the third term satisfies

$$
\begin{aligned}
\max_{i\neq j}\left|\frac{1}{n}(E^\top)_{i,\cdot}\left(P_U - L_qL_q^\top\right)U(\Psi^\top)_{\cdot,j}\right| &\leq \max_{i,j}\|\Psi_{i,\cdot}\|_2\left\|P_U - L_qL_q^\top\right\|_2\left\|U/\sqrt{n}\right\|_2\left\|E_{\cdot,i}/\sqrt{n}\right\|_2,\\
&= O_p\left(\frac{\sqrt{q}\max_i\|\Psi_{i,\cdot}\|_2}{\sqrt{p}}\right). \qquad (28)
\end{aligned}
$$

The same argument applies to the fourth term, which satisfies the same bound as follows

$$
\max_{i\neq j}\left|\frac{1}{n}\Psi_{i,\cdot}U^\top\left(P_U - L_qL_q^\top\right)E_{\cdot,j}\right| = O_p\left(\frac{\sqrt{q}\max_i\|\Psi_{i,\cdot}\|_2}{\sqrt{p}}\right). \qquad (29)
$$

Finally, we bound the first term by decomposing it as

$$
\begin{aligned}
\max_{i\neq j}\left|\frac{1}{n}(E^\top)_{i,\cdot}\left(I - L_qL_q^\top\right)E_{\cdot,j}\right| &\leq \max_{i\neq j}\left|\frac{1}{n}(E^\top)_{i,\cdot}\left(P_U - L_qL_q^\top\right)E_{\cdot,j}\right| + \max_{i\neq j}\left|\frac{1}{n}(E^\top)_{i,\cdot}\left(I - P_U\right)E_{\cdot,j}\right|,\\
&\leq \max_{i\neq j}\left\|P_U - L_qL_q^\top\right\|_2\left\|E_{\cdot,j}/\sqrt{n}\right\|_2 + \max_{i\neq j}\left|\frac{1}{n}(E^\top)_{i,\cdot}\left(I - P_U\right)E_{\cdot,j}\right|,\\
&= O_p\left(\sqrt{\frac{q}{p}}\right) + \max_{i\neq j}\left|\frac{1}{n}(E^\top)_{i,\cdot}\left(I - P_U\right)E_{\cdot,j}\right|. \qquad (30)
\end{aligned}
$$

To bound the second term, we apply Lemma D.1 (Hanson-Wright inequality). We denote $\xi_{ij} \overset{\text{def}}{=} \begin{pmatrix} E_{\cdot,i} \\ E_{\cdot,j} \end{pmatrix} \in \mathbb{R}^{2n}$, $\mathcal{A}_U \overset{\text{def}}{=} \begin{pmatrix} 0 & (I-P_U)/2n \\ (I-P_U)/2n & 0 \end{pmatrix} \in \mathbb{R}^{2n\times 2n}$. Then, by construction,

$$\begin{aligned}
\xi_{ij}^\top \mathcal{A}_U \xi_{ij} &= \begin{pmatrix} (E_{\cdot,i})^\top & (E_{\cdot,j})^\top \end{pmatrix} \begin{pmatrix} 0 & (I-P_U)/2n \\ (I-P_U)/2n & 0 \end{pmatrix} \begin{pmatrix} E_{\cdot,i} \\ E_{\cdot,j} \end{pmatrix}, \\
&= \frac{1}{2n} (E_{\cdot,i})^\top (I-P_U) E_{\cdot,j} + \frac{1}{2n} (E_{\cdot,j})^\top (I-P_U) E_{\cdot,i}, \\
&\overset{(1)}{=} \frac{1}{n} (E^\top)_{i,\cdot} (I-P_U) E_{\cdot,j}.
\end{aligned}$$

where "(1)" is due to the symmetry of $I - P_U$.

Next, since U is independent of E, $\mathcal{A}_U$ is independent of ξ_{ij}. Thus, using $\|\mathcal{A}_U\|_2 = \frac{1}{2n}$ and $\|\mathcal{A}_U\|_F = \frac{\sqrt{n-q}}{\sqrt{2}n}$, Lemma D.1 (Hanson–Wright inequality) gives

$$\mathbb{P}\left(\left|\xi_{ij}^\top \mathcal{A}_U \xi_{ij}\right| > t\right) = \mathbb{P}\left(\left|\frac{1}{n}\left(E^\top\right)_{i,\cdot} (I-P_U) E_{\cdot,j}\right| > t\right) \le 2\exp\left[-c\min\left\{\frac{n^2t^2}{n-q}, nt\right\}\right]. \tag{31}$$

Take $t = C(\frac{\sqrt{(n-q)\log p}}{n} + \frac{\log p}{n})$, a union bound over all pairs $i \neq j$ gives, with probability at least $1 - 2p^{-c}$,

$$\max_{i\neq j}\left|\frac{1}{n}\left(E^\top\right)_{i,\cdot} (I-P_U) E_{\cdot,j}\right| = O_p\left(\sqrt{\frac{\log p}{n}}\right). \tag{32}$$

As a result, we have the following for the first term

$$\max_{i\neq j}\left|\frac{1}{n}(E^\top)_{i,\cdot}\left(I - L_q L_q^\top\right) E_{\cdot,j}\right| = O_p\left(\sqrt{\frac{q}{p}}\right) + O_p\left(\sqrt{\frac{\log p}{n}}\right). \tag{33}$$

To summarize, combining equation 27, equation 33, and equation 28, and using $q = o(p)$, we have

$$II = O_p\left(\sqrt{\frac{\log p}{n}} + \frac{q\max_i \|\Psi_{i,\cdot}\|_2^2}{p} + \frac{\sqrt{q}\max_i \|\Psi_{i,\cdot}\|_2}{\sqrt{p}}\right). \tag{34}$$

Step II-C Proof of (2) : $\Delta = O_\prec\left(\frac{m}{p}\right)$. Recall that $\Delta = \sum_{k=q+1}^m (\widetilde{\lambda}_k^2 - \lambda_k^2) v_{ki} v_{kj} \le (\max_k |\widetilde{\lambda}_k^2 - \lambda_k^2|) \sum_{k=q+1}^m |v_{ki}||v_{kj}|$. To achieve the desired result, it is sufficient to show that (1) $\max_k \left|\widetilde{\lambda}_m^2 - \lambda_k^2\right| = O_p(\sqrt{p/n} + p/n) = O_p(1)$, and that (2) $\sum_{k=2}^m |v_{ki}||v_{kj}| \prec \frac{m}{p}$. As (1) will be proved by Lemma C.1, we mainly focus on (2).

Let us consider the SVD of the population covariance and the sample covariance as follows.

$$\Sigma(X) = \Psi\Psi^\top + I_p = \sum_{i=1}^p \sigma_i^2 \xi_i \xi_i^\top = \sqrt{\frac{p}{n}} \sum_{i=1}^p d_i^2 \xi_i \xi_i^\top + I_p, \quad \widehat{\Sigma}(X) = \sum_{i=1}^p \lambda_i^2 \mathbf{v}_i \mathbf{v}_i^\top.$$

For any deterministic vector $\mathbf{w} \in \mathbb{R}^p$, we introduce the shorthand $w_i \overset{\text{def}}{=} \langle \boldsymbol{\xi}_i, \mathbf{w}\rangle, i \in [1,p]$, to denote the coordinates of $\mathbf{w}$ in the eigenbasis $\{\boldsymbol{\xi}_i\}_{i=1}^p$ of the population covariance $\Sigma(X)$. In particular, $\mathbf{w} = \sum_{i=1}^p w_i \boldsymbol{\xi}_i$, and $\|\mathbf{w}\|_2^2 = \sum_{i=1}^p |w_i|^2$.

By Lemma D.3, for any index $a \in [q+1, m]$, where $m = o(n)$, we have

$$|\langle \mathbf{w}, \mathbf{v}_a\rangle|^2 \overset{(1)}{\prec} \frac{\|\mathbf{w}\|_2^2}{p} + \frac{1}{p}\sum_{i=1}^p \frac{\sigma_i^2 |w_i|^2}{(d_i^2-1)^2 + \kappa_a}, \tag{35}$$

where "(1)" denotes stochastic domination (see Remark A.2), $\kappa_a = K^{-2/3}\big(a \wedge (K+1-a)\big)^{2/3}$ and $K \overset{\text{def}}{=} n \wedge p$.

Hence, in our setting, $\sigma_i^2 = \Omega(p), d_i^2 = \sqrt{\frac{n}{p}}(\sigma_i^2 - 1) = \Omega(\sqrt{np})$, for $i \le q$, whereas $\sigma_i^2 = 1$ and $d_i^2 = 0$ for $i > q$. Taking $\mathbf{w} = \mathbf{e}_k$ for $k \in [m]$ yields

$$\begin{aligned}|\langle \mathbf{e}_k, \mathbf{v}_a\rangle|^2 &\prec \frac{1}{p} + \frac{1}{p}\sum_{i=1}^{q} \frac{\sigma_i^2 |\boldsymbol{\xi}_i(k)|^2}{(d_i^2-1)^2 + \kappa_a} + \frac{1}{p(1+\kappa_a)} \sum_{i=q+1}^{p} |\boldsymbol{\xi}_i(k)|^2 \\ &\prec \frac{1}{p} + \frac{1}{np}\sum_{i=1}^{q} |\boldsymbol{\xi}_i(k)|^2 + \frac{1}{p}\sum_{i=q+1}^{p} |\boldsymbol{\xi}_i(k)|^2 \\ &\prec \frac{1}{p}. \end{aligned} \tag{36}$$

Since the entrywise delocalization bound holds uniformly over $a \in [q+1, m]$ and $k \in [m]$, a union bound gives

$$\max_{\substack{a\in[q+1,m]\\ k\in[m]}} |v_{ka}|^2 \prec \frac{1}{p}, \text{ and } \max_{i,j\in[q+1,m]} \sum_{k=2}^{m} |v_{ki}||v_{kj}| \prec \frac{m}{p}. \tag{37}$$

Combining equation 37 with Lemma C.1, we obtain

$$\Delta = \sum_{k=q+1}^{m} (\widetilde{\lambda}_k^2 - \lambda_k^2) v_{ki} v_{kj} \le (\max_k |\widetilde{\lambda}_k^2 - \lambda_k^2|) \sum_{k=q+1}^{m} |v_{ki}||v_{kj}| = O_{\prec}\left(\frac{m}{p}\right). \tag{38}$$

Combining the bounds in equation 25, equation 34 and equation 38, we conclude that

$$\max_{i\neq j} \left| \left\{ \widehat{\Sigma}\left(P_m X\right) \right\}_{i,j} \right| = I + II + \Delta, \tag{39}$$

$$= O_p\left(\sqrt{\frac{\log p}{n}} + \frac{\max_i \|\Psi_{i,\cdot}\|_2^2}{p} + \frac{\max_i \|\Psi_{i,\cdot}\|_2}{\sqrt{p}}\right) + O_{\prec}\left(\frac{m}{p}\right), \tag{40}$$

$$= O_p\left(\sqrt{\frac{\log p}{n}} + \frac{\max_i \|\Psi_{i,\cdot}\|_2}{\sqrt{p}} + \frac{\max_i \|\Psi_{i,\cdot}\|_2^2}{p} + \frac{m}{p^{1-\epsilon}}\right). \tag{41}$$

Combining equation 41 and equation 23, we obtain the desired result. This completes the proof.

□

Remark A.2. (Stochastic Domination) Let $A^{(n)}(u)$ and $B^{(n)}(u)$ be families of nonnegative random variables indexed by $u \in \mathcal{U}^{(n)}$. We write $A \prec B$ if, for every fixed $\epsilon, D > 0$, $\sup_{u\in\mathcal{U}^{(n)}} \mathbb{P}\big(A^{(n)}(u) > n^{\epsilon} B^{(n)}(u)\big) \le n^{-D}$ for all sufficiently large n. The domination is understood uniformly over all unspecified indices. For a real- or complex-valued family A, the notation $A = O_{\prec}(B)$ means $|A| \prec B$.

A.2 The advantage of Trim transformation for the noise term $\|\widetilde{X}_T^* P_T \widetilde{\varepsilon}\|_\infty$

Theorem A.3. *Consider model 1 with an isotropic Gaussian random vector $U_{i,\cdot} \in \mathbb{R}$. Assume $\lim p/n \to c \in [0, \infty)$. Then, for the Trim-transformed design matrix $\widetilde{X} \overset{\text{def}}{=} P_m X$ and noise, we have*

$$\left\| \widetilde{X}_T^* P_T \widetilde{\varepsilon}_b \right\|_\infty = O_p\left(\frac{\|\phi\|_2 \sqrt{p \log p}}{n}\right),$$

with probability approaching to 1 as $n \to \infty$.

Remark A.4. Without Trim transformation, in general we can only achieve $\|X_T^* P_T^* \varepsilon\|_\infty = O_p(\max_i \|\Psi_{i,\cdot}\|_2 \sqrt{\log p/n})$, where the additional term in Theorem A.1, assumed to satisfy $\max_i \|\Psi_{i,\cdot}\|_2 = o_p(\sqrt{n/(qs)})$, can still grow at the rate $\Theta_p(\sqrt{p}/\log p)$ when q and s are fixed, but it can be effectively removed by the Trim transformation.

Proof. We first show that $\widetilde{X}$ is independent of ε_b. First, by Ćevid et al. (2020), as well as the fact that X and ε_b are Gaussian, we have $\widetilde{X}$ is uncorrelated and thus independent of ε_b. Since $\widetilde{X} = P_m X$ with P_m determined by X, $\widetilde{X}$ is independent of ε_b. By $\widetilde{X}_T^* P_T \cdot P_T \widetilde{X}_T \preceq \widetilde{X}_T^* \widetilde{X}_T$, we have

$$\|\widetilde{X}_T^* P_T \widetilde{\varepsilon}_b\|_\infty \leq \|\widetilde{X}_T^* \widetilde{\varepsilon}_b\|_\infty \overset{(1)}{\leq} \|\widetilde{X}_T^* \varepsilon_b\|_\infty \overset{(2)}{\lesssim} 2 \max_i \sqrt{\mathrm{Var}(\varepsilon_i)} \|\widetilde{X}_i/\sqrt{n}\|_2 \sqrt{\frac{\log p}{n}},$$

where "(1)" is due to the fact $\|\widetilde{X}_T^* P_m(X)\|_2 \leq \|\widetilde{X}_T^*\|_2 \|P_m(X)\|_2 \leq \|\widetilde{X}_T^*\|_2$; "(2)" applies the conclusion in Vershynin (2018). Next, we show $\left\| \frac{\widetilde{X}_i}{\sqrt{n}} \right\|_2 = 1 + O_p(\sqrt{\frac{p}{n}})$ and $\mathrm{Var}(\varepsilon_{b,i}) = \|\phi\|_2^2 + \sigma^2$, which leads to the desired result. First, by Weyl's theorem, we have

$$\|\widetilde{X}\|_2 = \lambda_{m+1}(X) \leq \|E\|_2 \overset{(1)}{\lesssim} \sqrt{n} + \sqrt{p} + \sqrt{\log p},$$

where "(1)" applies Theorem 5.39 in Vershynin (2010). For $\mathrm{Var}(\varepsilon_{b,i})$, by the independence of (U_i, X_i) from $e_{y,i}$, we have

$$\mathrm{Var}(\varepsilon_{b,i}) = \mathrm{Var}(U_{i,\cdot}\phi - X_{i,\cdot}b + e_{y,i}) \overset{(1)}{=} \mathrm{Var}(U_{i,\cdot}\phi - X_{i,\cdot}b) + \mathrm{Var}(e_{y,i}) = \mathrm{Var}(U_{i,\cdot}\phi - X_{i,\cdot}b) + \sigma^2, \tag{42}$$

where recall that, the residue and the bias vector are respectively given by $\varepsilon = U\phi - Xb + \varepsilon$ and $b = \mathrm{Cov}(X)^{-1}\mathrm{Cov}(X,U)\phi$. Let $\Sigma_U \overset{\mathrm{def}}{=} \mathrm{Var}(U_i) = I_q$, $\Sigma_{UX} \overset{\mathrm{def}}{=} \mathrm{Cov}(U,X)$, and $\Sigma_{XU} \overset{\mathrm{def}}{=} \Sigma_{UX}^\top$. Then, for $\mathrm{Var}(U_{i,\cdot}\phi - X_{i,\cdot}b)$, we have

$$\begin{aligned}
\mathrm{Var}(U_{i,\cdot}\phi - X_{i,\cdot}b) &= \phi^\top \Sigma_U \phi + b^\top \Sigma_X b - 2b^\top \Sigma_{UX}\phi \\
&= \phi^\top \Sigma_U \phi + \phi^\top \Sigma_{UX}\Sigma_X^{-1}\Sigma_{UX}\phi - 2\phi^\top \Sigma_{UX}\Sigma_X^{-1}\Sigma_{UX}\phi \\
&= \phi^\top \left(\Sigma_U - \Sigma_{UX}\Sigma_X^{-1}\Sigma_{UX}\right)\phi \\
&= \phi^\top \left(I_q - \Sigma_{UX}\Sigma_X^{-1}\Sigma_{UX}\right)\phi \\
&\leq \|\phi\|_2^2.
\end{aligned} \tag{43}$$

This completes the proof. □

B Consistency of Trim-Linearized Bregman ISS

In this section, we prove the model selection consistency and ℓ_2 estimation consistency of LBISS. For completeness, we provide a brief overview of some fundamental properties of Linearized Bregman Inverse Scale Space (LBISS).

The solution path of LBISS $\{\rho_t, \beta_t\}_{t\geq 0}$ governed by the nonlinear differential inclusions for some constant $\kappa > 0$:

$$\dot{\rho}_t + \frac{1}{\kappa}\dot{\beta}_t = \frac{1}{n} X^T (y - X\beta_t), \tag{44}$$

$$\rho_t \in \partial \|\beta_t\|_1, \tag{45}$$

By Theorem 2.2 in Osher et al. (2016), this solution exists and is unique. Hence, the regularization path is well-defined. To study sign consistency, we introduce an oracle dynamics that evolves only on the set of true signals. We first establish that the original dynamics satisfy the no–false-positive property up to a certain time τ. Under this property, the dynamics remain confined to the true support and therefore coincide with the oracle dynamics during this interval. We then show that the oracle dynamics achieves sign consistency before τ at which the no–false-positive property remains valid, which implies sign consistency of the original dynamics. The oracle dynamics is defined as

$$\dot{\rho}'_{t,S} + \frac{1}{\kappa}\dot{\beta}'_{t,S} = \frac{1}{n} X_S^T (y - X_S\beta_{t,S}), \tag{46}$$

$$\rho'_{t,S} \in \partial \|\beta_{t,S}\|_1, \tag{47}$$

To study the consistency behavior of this oracle dynamics, we introduce a Lyapunov-type potential function of the oracle dynamics $\Psi : \mathbb{R}^p \to \mathbb{R}_0^+$ as follows.

$$\Psi(\beta'_S) = \frac{\|\beta'_S - \widetilde{\beta}^*_S\|_2^2}{2\kappa} + \|\widetilde{\beta}^*_S\|_1 - \|\beta'_S\|_1 - \left\langle \partial\|\widetilde{\beta}^*_S\|_1, \widetilde{\beta}^*_S - \beta'_S \right\rangle \overset{\mathrm{def}}{=} D\left(\widetilde{\beta}^*_S, \beta'_S\right) + \frac{\|\beta'_S - \widetilde{\beta}^*_S\|_2^2}{2\kappa},$$

where the oracle estimator $\widetilde{\beta}_S^*$ is defined as

$$\begin{aligned}\widetilde{\beta}_S^* &\overset{\text{def}}{=} (X_S^* X_S)^{-1} X_S^* y \\ &= (X_S^* X_S)^{-1} X_S^* (X_S \beta_S^* + U\phi + e_y) \\ &= \beta_S^* + (X_S^* X_S)^{-1} X_S^* \varepsilon_b - (X_S^* X_S)^{-1} X_S^* X b \\ &= \beta_S^* + (X_S^* X_S)^{-1} X_S^* \varepsilon_b - b_S - (X_S^* X_S)^{-1} X_S^* X_T b_T, \end{aligned} \tag{48}$$

where $\varepsilon_b \overset{\text{def}}{=} U\phi + e_y - Xb$. It can be easily verified that $\Psi(\beta_S')$ decreases monotonically along the trajectory

$$\frac{d\Psi(\beta_S')}{dt} = -(\beta_S' - \beta_S^*)^\top X_S^* X_S (\beta_S' - \beta_S^*) \le -\lambda_{\min}(X_S^* X_S) \|\beta_S' - \beta_S^*\|_2^2 \overset{(1)}{\lesssim} -\|\beta_S' - \beta_S^*\|_2^2, \tag{49}$$

where "(1)" applies equation 23. Similarly, we can also verify that $\|X_S(\widetilde{\beta}_S^* - \beta_S)\|_2^2$ decreases monotonically as t grows.

$$\frac{d}{dt}\frac{\|X_S(\widetilde{\beta}_S^* - \beta_S)\|_2^2}{2n} = -\left\langle \frac{d\rho_S}{dt}, \frac{d\beta_S}{dt} \right\rangle - \frac{1}{\kappa}\left\| \frac{d\beta_S}{dt} \right\|_2^2 = -\frac{1}{\kappa}\left\| \frac{d\beta_S}{dt} \right\|_2^2 \le 0. \tag{50}$$

In the following, we first characterize τ of the no false positive property of equation 44, and then establish sign-consistency and ℓ_2 consistency. To easy notation, we also use β_t, ρ_t (*i.e.*, β_t', ρ_t') to denote the dynamics in equation 44 (*i.e.*, oracle dynamics equation 46) when X, y are replaced by $\widetilde{X}, \widetilde{y}$. Let $\widetilde{\varepsilon}_b \overset{\text{def}}{=} P_m \varepsilon_b$. Besides, we denote $\gamma \overset{\text{def}}{=} \lambda_{\min}(\widetilde{X}_S^* \widetilde{X}_S)$.

B.1 No false positive

To prove the no-false-positive property, Lemma A.1 in Osher et al. (2016) provides a sufficient condition. For completeness, we introduce it here.

Lemma B.1 (Lemma A.1 in Osher et al. (2016)). *For all $t \le \tau$, the solution of $\beta(t)$ contains no false positive if*

$$\|\widetilde{X}_T^* \widetilde{X}_S^\dagger (\rho_S + \beta_S/\kappa) + t\widetilde{X}_T^* P_T \widetilde{\varepsilon}\|_\infty < 1, \quad \forall t \le \tau, \tag{51}$$

where $P_T = I - \widetilde{X}_S^\dagger \widetilde{X}_S^*$ is the projection matrix onto $\mathrm{im}(\widetilde{X}_S)^\perp$ and $\widetilde{X}_S^\dagger := \widetilde{X}_S \left(\frac{1}{n}\widetilde{X}_S^T \widetilde{X}_S\right)^{-1}$.

Proposition B.2 (No-false-positive). *Let $\eta \in (0,1)$. Suppose assumptions in Theorem A.1 hold. Suppose κ is sufficiently large, such that*

$$\begin{aligned} B \overset{\text{def}}{=}& \left\{ 1 + \sqrt{s\left(1 + C_0\frac{p}{n} + C_0\sqrt{\frac{p}{n}}\right)/c_\gamma} \right\} \|\beta_S^*\|_\infty + 2s\sqrt{(\|\phi\|_2^2 + \sigma^2)\log n/(nc_\gamma)} \\ &+ 2C_0\sqrt{\|\phi\|_2^2 + \sigma^2}\sqrt{\frac{\log s}{n\gamma}} + C_0\|\phi\|_2 \left\{ 1 + 2\|\phi\|_2\sqrt{\left(1 + C_0\frac{p}{n} + C_0\sqrt{\frac{p}{n}}\right)/c_\gamma} \right\}\sqrt{\frac{sq}{p}} \\ &+ C_0 \left\{ 1 + 2\sqrt{\left(1 + C_0\frac{p}{n} + C_0\sqrt{\frac{p}{n}}\right)/c_\gamma} \right\} \frac{\|\phi\|_2}{\sqrt{p}} \le \kappa\eta. \end{aligned} \tag{52}$$

Then for any $\delta > 0$, with probability at least $1 - n^{-c_0} - p^{-c_0} - s^{-c_0} - \delta$ for some $c_0 > 0$, we have $\mathrm{supp}(\beta_t) \subseteq S$ when $t \le \tau \overset{\text{def}}{=} C_0(1 - B/\kappa\eta)\eta \frac{n}{\sqrt{\|\phi\|_2^2 + \sigma^2}\sqrt{p\log p}}$. Here, $C_0 > 0$ is a constant that depends on c_0, δ, and η.

Proof. We first introduce an event $\mathcal{E}$ as

$$\mathcal{E} \overset{\text{def}}{=} \left\{ \sqrt{s}\max_{j \in T} \|(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \widetilde{X}_j\|_2 \le 1 - \eta, \quad \gamma \ge c_\gamma > 0 \right\},$$

where $\gamma \overset{\text{def}}{=} \lambda_{\min}(\widetilde{X}_S^* \widetilde{X}_S)$. Under event $\mathcal{E}$, we have $\|\widetilde{X}_T^* \widetilde{X}_S^\dagger\|_\infty \le 1-\eta$. For the remainder of the proof, we first work conditionally on the event $\mathcal{E}$, and then show that $\mathbb{P}(\mathcal{E}) \to 1$ as $n \to \infty$. The proof essentially follows the proof of Theorem 4.1 in Osher et al. (2016). By Lemma B.1, it is sufficient to prove that equation 51 holds with high probability for any $t \le \tau$. Under event $\mathcal{E}$, we have

$$\left\|\widetilde{X}_T^* \widetilde{X}_S^\dagger \left(\rho_S + \beta_S/\kappa\right)\right\|_\infty \le \|\widetilde{X}_T^* \widetilde{X}_S^\dagger\|_\infty + \|\widetilde{X}_T^* \widetilde{X}_S^\dagger\|_\infty \|\beta_S\|_\infty/\kappa \le (1-\eta)\left(1 + \|\beta_S\|_\infty/\kappa\right) \le 1-(1-B/\kappa\eta)\eta,$$

where B is an asymptotic upper bound of $\|\beta_S\|_\infty$. Then, conditional on the event $\mathcal{E}$, with probability at least $1 - n^{-c_0} - p^{-c_0}$ for some c_0, there exists constant C_0 that depends on c_0, such that

$$\begin{aligned}
\|\beta_S\|_\infty &= \left\|\beta_S + \widetilde{\beta}_S^* - \widetilde{\beta}_S^*\right\|_\infty \\
&\le \left\|\widetilde{\beta}_S^*\right\|_\infty + \left\|\widetilde{\beta}_S^* - \beta_S\right\|_\infty \\
&\overset{(1)}{\le} \|\beta_S^*\|_\infty + \|b_S\|_\infty + \|(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \widetilde{X}_T b_T\|_\infty + \left\|(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \widetilde{\varepsilon}_b\right\|_\infty + \frac{\|\widetilde{X}_S(\widetilde{\beta}_S^* - \beta_S)\|_2}{\sqrt{n\gamma}} \\
&\overset{(2)}{\le} \|\beta_S^*\|_\infty + \|b\|_\infty + \left\|(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \widetilde{\varepsilon}_b\right\|_\infty + \frac{\|\widetilde{X}_S \widetilde{\beta}_S^*\|_2}{\sqrt{n\gamma}} + \|(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \widetilde{X}_T b_T\|_\infty \\
&\le \|\beta_S^*\|_\infty + \|b\|_\infty + \left\|(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \widetilde{\varepsilon}_b\right\|_\infty + \frac{\|\widetilde{X}_S \widetilde{\beta}_S^*\|_2}{\sqrt{n\gamma}} + \frac{\sqrt{s}\max_{i\in S}\left|\widetilde{X}_i^* \sum_{j\in S^c} b_j X_j\right|}{\lambda_{\min}(\widetilde{X}_S^* \widetilde{X}_S)} \\
&\overset{(3)}{\le} \|\beta_S^*\|_\infty + C_0 \frac{\|\phi\|_2}{\lambda_q(\Psi)} + 2C_0\sqrt{\|\phi\|_2^2 + \sigma^2}\sqrt{\frac{\log s}{n\gamma}} + \frac{\|\widetilde{X}_S \widetilde{\beta}_S^*\|_2}{\sqrt{nc_\gamma}} + C_0\|\phi\|_2\sqrt{\frac{sq}{p}}.
\end{aligned}$$

where "(1)" applies the expansion of $\widetilde{\beta}_S^*$ in equation 48, with X, ε_b replaced by $\widetilde{X}, \widetilde{\varepsilon}_b$; "(2) follows from the monotonic decreasing property equation 50 of $\|\widetilde{X}_S(\widetilde{\beta}_S^* - \beta_S)\|_2$; "(3)" follows from Lemma D.4, Lemma C.6 and Lemma C.7. Based on the expansion of the oracle estimator with X, ε_b replaced by $\widetilde{X}, \widetilde{\varepsilon}_b$, with probability at least $1 - n^{-c_0} - p^{-2c_0}$, there exists a positive constant $C_0 > 0$, such that

$$\begin{aligned}
\|\widetilde{X}_S \widetilde{\beta}_S^*\|_2 &\le \|\widetilde{X}_S \beta_S^*\|_2 + \sqrt{s}\|\widetilde{X}_S(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \widetilde{\varepsilon}_b\|_\infty + \|\widetilde{X}_S\|_2\|b_S\|_2 + \|\widetilde{X}_S\|_2\|(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \widetilde{X}_T b_T\|_2, \\
&\le \|\widetilde{X}_S\|_2\{\|\beta_S^*\|_2 + \|b\|_2 + \|(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \widetilde{X}_T b_T\|_2\} + \sqrt{s}\|\widetilde{X}_S(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \widetilde{\varepsilon}_b\|_\infty, \\
&\overset{(1)}{\le} \sqrt{n\lambda_{\max}(\widetilde{X}_S^* \widetilde{X}_S)}\left\{\|\beta_S^*\|_2 + 2C_0\frac{\|\phi\|_2}{\lambda_q(\Psi)} + 2C_0\|\phi\|_2^2\sqrt{\frac{sq}{p}}\right\} + 2s\sqrt{(\|\phi\|_2^2 + \sigma^2)\log n},
\end{aligned}$$

where "(1)" follows from Lemma D.4 and Lemma C.6. Substituting this into the above, we have

$$\begin{aligned}
\|\beta_S\|_\infty &\le \|\beta_S^*\|_\infty + C_0\frac{\|\phi\|_2}{\lambda_q(\Psi)} + 2C_0\sqrt{\|\phi\|_2^2 + \sigma^2}\sqrt{\frac{\log s}{n\gamma}} + C_0\|\phi\|_2\sqrt{\frac{sq}{p}} + 2s\sqrt{\frac{(\|\phi\|_2^2 + \sigma^2)\log n}{nc_\gamma}} \\
&\quad + \sqrt{\lambda_{\max}(\widetilde{X}_S^* \widetilde{X}_S)/c_\gamma}\left\{\|\beta_S^*\|_2 + 2C_0\frac{\|\phi\|_2}{\lambda_q(\Psi)} + 2C_0\|\phi\|_2^2\sqrt{\frac{sq}{p}}\right\}, \\
&\le \|\beta_S^*\|_\infty + \left\{\sqrt{\lambda_{\max}(\widetilde{X}_S^* \widetilde{X}_S)/c_\gamma}\right\}\|\beta_S^*\|_2 + 2C_0\sqrt{\|\phi\|_2^2 + \sigma^2}\sqrt{\frac{\log s}{n\gamma}} + 2s\sqrt{\frac{(\|\phi\|_2^2 + \sigma^2)\log n}{nc_\gamma}} \\
&\quad + C_0\|\phi\|_2\left\{1 + 2\|\phi\|_2\sqrt{\lambda_{\max}(\widetilde{X}_S^* \widetilde{X}_S)/c_\gamma}\right\}\sqrt{\frac{sq}{p}} + C_0\left\{1 + 2\sqrt{\lambda_{\max}(\widetilde{X}_S^* \widetilde{X}_S)/c_\gamma}\right\}\frac{\|\phi\|_2}{\lambda_q(\Psi)}.
\end{aligned}$$

By Lemma C.8, with probability at least $1 - n^{-c_0}$ for some c_0, there exists C_0 (with a bit abuse of notation) such that

$$\lambda_{\max}(\widetilde{X}_S^* \widetilde{X}_S) \le 1 + C_0\left(\frac{p}{n} + \sqrt{\frac{p}{n}}\right).$$

Plugging this into the above inequality yields

$$\begin{aligned}\|\beta_S\|_\infty &\le \|\beta_S^*\|_\infty + \left\{\sqrt{\lambda_{\max}(\widetilde{X}_S^*\widetilde{X}_S)/c_\gamma}\right\}\|\beta_S^*\|_2 + 2C_0\sqrt{\|\phi\|_2^2+\sigma^2}\sqrt{\frac{\log s}{n\gamma}} + 2s\sqrt{(\|\phi\|_2^2+\sigma^2)\log n/(nc_\gamma)} \\ &\quad + C_0\|\phi\|_2\left\{1+2\|\phi\|_2\sqrt{\lambda_{\max}(\widetilde{X}_S^*\widetilde{X}_S)/c_\gamma}\right\}\sqrt{\frac{sq}{p}} + C_0\left\{1+2\sqrt{\lambda_{\max}(\widetilde{X}_S^*\widetilde{X}_S)/c_\gamma}\right\}\frac{\|\phi\|_2}{\lambda_q(\Psi)} \\ &\le \|\beta_S^*\|_\infty + \left\{\sqrt{\left(1+C_0\frac{p}{n}+C_0\sqrt{\frac{p}{n}}\right)/c_\gamma}\right\}\|\beta_S^*\|_2 + 2s\sqrt{(\|\phi\|_2^2+\sigma^2)\log n/(nc_\gamma)} \\ &\quad + 2C_0\sqrt{\|\phi\|_2^2+\sigma^2}\sqrt{\frac{\log s}{n\gamma}} + C_0\|\phi\|_2\left\{1+2\|\phi\|_2\sqrt{\left(1+C_0\frac{p}{n}+C_0\sqrt{\frac{p}{n}}\right)/c_\gamma}\right\}\sqrt{\frac{sq}{p}} \\ &\quad + C_0\left\{1+2\sqrt{\left(1+C_0\frac{p}{n}+C_0\sqrt{\frac{p}{n}}\right)/c_\gamma}\right\}\frac{\|\phi\|_2}{\sqrt{p}}. \\ &\stackrel{\text{def}}{=} B.\end{aligned}$$

Next, follow Osher et al. (2016) Theorem 4.1, it suffices to choose t such that $t\|\widetilde{X}_T^*P_T\widetilde{\varepsilon}\|_\infty < (1-B/\kappa\eta)\eta$ on the event. By Theorem 4.2, there exists C_0 that depends on δ, with probability at least $1-\delta$, we have

$$t \le \tau \stackrel{\text{def}}{=} C_0(1-B/\kappa\eta)\eta\frac{n}{\sqrt{\|\phi\|_2^2+\sigma^2}\sqrt{p\log p}},$$

This completes the proof. □

Remark B.3. Remarkably, $B = \|\beta_S^*\|_\infty + C\|\beta_S^*\|_2 + o_p(1)$ under the assumptions of Theorem 4.1 with $\|\phi\|_2 = O(1)$.

B.2 ℓ_2-CONSISTENCY OF ORACLE DYNAMICS

We define the stopping time of oracle dynamics that reaches sign-consistency and ℓ_2-consistency to the oracle estimator $\widetilde{\beta}_S^*$,

$$\tau_1 \stackrel{\text{def}}{=} \inf\left\{t>0 : \|\beta_S' - \widetilde{\beta}_S^*\|_2 < \widetilde{\beta}_{\min}^*\right\} \tag{53}$$

$$\tau_2(C) \stackrel{\text{def}}{=} \inf\left\{t>0 : \|\beta_S' - \widetilde{\beta}_S^*\|_2 \le C\sqrt{\frac{s\log s}{n}}\right\}, \tag{54}$$

where $\widetilde{\beta}_{\min}^* = \min_{i\in S}|\widetilde{\beta}_i^*|$. Note that $\|\widetilde{\beta}_S^* - \beta_S'\|_2 < \widetilde{\beta}_{\min}^* \Rightarrow \operatorname{sign}(\widetilde{\beta}_i^*) = \operatorname{sign}(\beta_i') \Rightarrow D(\widetilde{\beta}_S^*, \beta_S') = 0$. Therefore, if we can show that with high probability, $\tau_1 \le \tau$ and $\tau_2(C) \le \tau$, we can achieve sign-consistency and ℓ_2 consistency before any null features are selected.

Theorem B.4. *Suppose that the assumptions of Theorem 4.1 and Proposition 4.4 hold, except that U, e_y, E have i.i.d zero mean Gaussian entries. Let $\widetilde{\beta}^*$ denote the oracle estimator and β_S^* the true coefficients on the support S. Define $\widetilde{\beta}_{\min}^* = \min_{i\in S}|\widetilde{\beta}_i^*|$. Besides, we assume $q = O_p(\log s)$ and $\lim p/n \to c \in (0,\infty)$. Suppose κ is sufficiently large, such that $(8+4\log s)/\widetilde{\beta}_{\min}^* + \log(3\|\widetilde{\beta}^*\|_2/\widetilde{\beta}_{\min}^*)/\kappa \le \tau$. Then, the Spec LBISS's solution path β_t satisfies,*

1. *(Sign-consistency). Assume $\beta_{\min}^* \ge C\left(\|\phi\|_2+\sigma\right)\sqrt{qs\log s/n}$ for some $C>0$. Then for any $\delta>0$, with probability at least $1-n^{-c}-p^{-c}-s^{-c}-\delta$ for some constant $c>0$, we have $\tau_1 \le \tau$ and $\operatorname{sign}(\beta_\tau) = \operatorname{sign}(\beta^*)$.*
2. *(ℓ_2-consistency). For any $\delta>0$, with probability at least $1-n^{-c}-p^{-c}-s^{-c}-\delta$ for some constant $c>0$, we have $\tau_2(C) \le \tau$, and*

$$\|\beta_{\tau,S}' - \beta_S^*\|_2 = O_p\{(1+\|\phi\|_2)\sqrt{s\log s/n}\}.$$

Proof. The sign-consistency result follows $\operatorname{sign}(\beta'_{\tau_1,S}) = \operatorname{sign}(\widetilde{\beta}^*_S)$ and that $\operatorname{sign}(\widetilde{\beta}^*_S) = \operatorname{sign}(\beta^*_S)$. The first part follows from Theorem 4.1 (c) of Osher et al. (2016). Now we prove $\operatorname{sign}(\widetilde{\beta}^*_S) = \operatorname{sign}(\beta^*_S)$. First, we can expand $\widetilde{\beta}^*_S$ as

$$\begin{aligned}\widetilde{\beta}^*_S &= \beta^*_S + (\widetilde{X}^*_S\widetilde{X}_S)^{-1}\widetilde{X}^*_S\widetilde{\varepsilon}_b + (\widetilde{X}^*_S\widetilde{X}_S)^{-1}\widetilde{X}^*_S\widetilde{X}b \\ &= \beta^*_S + (\widetilde{X}^*_S\widetilde{X}_S)^{-1}\widetilde{X}^*_S\widetilde{\varepsilon}_b + b_S + (\widetilde{X}^*_S\widetilde{X}_S)^{-1}\widetilde{X}^*_S\widetilde{X}_T b_T.\end{aligned}$$

Therefore, there exist some positive constants C_1, C_0 such that

$$\begin{aligned}\|\widetilde{\beta}^*_S - \beta^*_S\|_\infty &\le \|(\widetilde{X}^*_S\widetilde{X}_S)^{-1}\widetilde{X}^*_S\widetilde{\varepsilon}_b\|_\infty + \|b_S\|_\infty + \|(\widetilde{X}^*_S\widetilde{X}_S)^{-1}\widetilde{X}^*_S\widetilde{X}_T b_T\|_\infty \\ &\overset{(1)}{\le} 2\sqrt{\|\phi\|_2^2+\sigma^2}\sqrt{\frac{2\log s}{n}} + C_1\frac{\|\phi\|_2}{\|\Psi\|_2} + \frac{\sqrt{s}\max_{i\in S}\left|\widetilde{X}^*_i\sum_{j\in S^c} b_j X_j\right|}{\lambda_{\min}(\widetilde{X}^*_S\widetilde{X}_S)} \\ &\overset{(2)}{\le} C_0\left(\|\phi\|_2+\sigma\right)\sqrt{\frac{qs\log s}{n}}. \end{aligned} \tag{55}$$

where "(1)" applies Lemma D.4, C.6, and "(2)" applies Lemma C.7 and the assumption $q = O_p(\log s)$. By taking C sufficiently large, the right-hand side of equation 55 is $\le \frac{\beta^*_{\min}}{2}$.

Next, we focus on the ℓ_2 error bound. We first show that $\tau_2(C) \le \tau_1 + \frac{1}{2\kappa}\ln\frac{n\|\widetilde{\beta}^*_S\|_2^2}{Cs\log s}$ for some constant C. By taking a sufficiently large value of κ, we have $\tau_2(C) \le \tau$, and thus we have $\|\beta_{\tau,S} - \widetilde{\beta}^*_S\|_2 \le C\sqrt{\frac{s\log s}{n}}$.

By equation 49, when $t > \tau_1$, we have $\|\beta'_{t,S} - \widetilde{\beta}^*_S\|_2 < \widetilde{\beta}^*_{\min}$ and therefore $D(\widetilde{\beta}^*, \beta'_S) = 0$. Define $F(x) \overset{\text{def}}{=} \frac{x}{2\kappa}$ for $x \ge 0$, we have the following for any $t \ge \tau_1$

$$\frac{d\Psi(\beta'_S)}{dt} \le -\lambda^{-1}_{\min}(\widetilde{X}^*_S\widetilde{X}_S)\|\widetilde{\beta}^*_S - \beta'_S\|_2^2 \le -\lambda^{-1}_{\min}(\widetilde{X}^*_S\widetilde{X}_S)F^{-1}(\Psi(\beta'_S)).$$

In other words, we have $F^{-1}(\Psi(\beta'_S)) \le -\frac{1}{\gamma}\frac{d\Psi(\beta'_S)}{dt}$. This gives us

$$\begin{aligned}\tau_2(C) - \tau_1 &\le \int_{\tau_1}^{\tau_2(C)} \frac{-\frac{1}{\gamma}\frac{d}{dt}\left(\Psi\left(\beta'_S\right)\right)}{F^{-1}(\Psi\left(\beta'_S\right))}dt, \\ &= \frac{1}{\gamma}\int_{\Psi(\tau_2(C))}^{\Psi(\tau_1)} \frac{1}{F^{-1}(x)}dx, \\ &= \frac{1}{\gamma}\int_{F^{-1}(\Psi(\tau_2(C)))}^{F^{-1}(\Psi(\tau_1))} \frac{1}{y}dF(y), \\ &= \frac{1}{\gamma}\int_{\frac{Cs\log s}{n}}^{\|\widetilde{\beta}^*_S\|_2^2} \frac{1}{2\kappa y}dy, \\ &= \frac{1}{2\kappa}\ln\frac{n\|\widetilde{\beta}^*_S\|_2^2}{Cs\log s}.\end{aligned}$$

Finally, for any $\delta > 0$, with probability at least $1 - p^{-c} - n^{-c} - s^{-c} - \delta$ for some c, there exists a constant C_0 that depends on C, c, δ, such that

$$\begin{aligned}\|\beta'_{\tau,S} - \beta^*_S\|_2 &\le \|\beta'_{\tau,S} - \widetilde{\beta}^*_S\|_2 + \|\widetilde{\beta}^*_S - \beta^*_S\|_2 \\ &= \|\beta'_{\tau,S} - \widetilde{\beta}^*_S\|_2 + \|(\widetilde{X}^*_S\widetilde{X}_S)^{-1}\widetilde{X}^*_S\widetilde{\varepsilon}_b\|_2 + \|b_S\|_2 + \|(\widetilde{X}^*_S\widetilde{X}_S)^{-1}\widetilde{X}^*_S\widetilde{X}_T b_T\|_2 \\ &\le C_0\left(\sqrt{\frac{s\log s}{n}} + \sqrt{\frac{(\|\phi\|_2+\sigma^2)\log s}{p}} + \frac{\|\phi\|_2}{\sqrt{p}}\right) + \|(\widetilde{X}^*_S\widetilde{X}_S)^{-1}\widetilde{X}^*_S\widetilde{X}_T b_T\|_2 \\ &\overset{(2)}{\le} (C_0 + \|\phi\|_2)\sqrt{\frac{s\log s}{n}}.\end{aligned}$$

where "(1)" follows from Lemma D.4, C.6; "(2)" follows from Lemma C.7 and the assumption $q = O_p(\log s)$.

□

C TECHNICAL LEMMAS

Lemma C.1. *Consider the model equation 1. For any $m > q$, we have* $\max_{k>m} |\lambda_m(\widetilde{X}^*\widetilde{X}) - \lambda_k(\widetilde{X}^*\widetilde{X})| = O_p\left(\sqrt{p/n} + p/n\right)$.

Proof. Denote $\Sigma_X \overset{\text{def}}{=} \Psi\Psi^\top + I$. Then, we have

$$\begin{aligned}
&\max_{k>m} |\lambda_m(\widetilde{X}^*\widetilde{X}) - \lambda_k(\widetilde{X}^*\widetilde{X})| \\
&\quad \le \max_{k>m} |\lambda_k(X^*X) - \lambda_k(\Sigma_X)| + |\lambda_m(X^*X) - \lambda_m(\Sigma_X)| + |\lambda_m(\Sigma_X) - \lambda_m(\Sigma_X)|, \\
&\quad = O_p\left(\sqrt{p/n} + p/n\right). \qquad (56)
\end{aligned}$$

The proof is completed. □

Lemma C.2. *Under Theorem A.1, for $X = U\Psi^\top + E$ with singular value decomposition $X/\sqrt{n} = L\Sigma V^\top$, the first q right singular vectors V_q satisfy* $\max_{i\in[p]} \left\|(V_q)_{i,}\right\|_2^2 = O_p\left(\frac{\max_{i\in[p]}\|\Psi_{i,\cdot}\|_2^2}{p}\right)$.

Proof. Consider the singular value decomposition $X/\sqrt{n} = L\Sigma V^\top$, and let L_q denote the matrix whose columns are the left singular vectors corresponding to the q largest singular values of $X/\sqrt{n}$. It follows that

$$\frac{X^\top}{\sqrt{n}} L_q = V_q \Sigma_q L_q^\top L_q + V_{-q}\Sigma_{-q} L_{-q}^\top L_q = V_q \Sigma_q. \qquad (57)$$

This implies

$$(V_q)_{i,\cdot} = \frac{(X^\top)_{i,\cdot}}{\sqrt{n}} L_q \Sigma_q^{-1}. \qquad (58)$$

Then, we have

$$\begin{aligned}
\max_{i\in[p]} \left\|(V_q)_{i,\cdot}\right\|_2 &\le \frac{\max_{i\in[p]} \left\|\frac{(X^\top)_{i,\cdot}}{\sqrt{n}} L_q\right\|_2}{\lambda_q(X/\sqrt{n})} \le \frac{\max_{i\in[p]} \left\|\frac{X_{\cdot,i}}{\sqrt{n}}\right\|_2}{\lambda_q(X/\sqrt{n})}, \\
&\le \frac{\max_{i\in[p]} \left\|(U(\Psi^\top)_{\cdot,i}/\sqrt{n})\right\|_2 + \max_{i\in[p]} \|E_{\cdot,i}/\sqrt{n}\|_2}{\lambda_q(X/\sqrt{n})} \overset{(1)}{=} O_p\left(\frac{\max_{i\in[p]} \|\Psi_{i,\cdot}\|_2}{\sqrt{p}}\right).
\end{aligned}$$

where "(1)" follows from Weyl's singular value inequality and the column-wise bound for $X_{\cdot,i}$. Specifically, applying Weyl's inequality to $X/\sqrt{n} = (U/\sqrt{n})\Psi^\top + E/\sqrt{n}$ gives $\lambda_q(X/\sqrt{n}) \ge \lambda_q((U/\sqrt{n})\Psi^\top) - \|E/\sqrt{n}\|_2 = \Theta_p(\sqrt{p})$. Moreover, $\max_{i\in[p]} \|X_{\cdot,i}/\sqrt{n}\|_2 \le \max_{i\in[p]} \|U/\sqrt{n}\|_2 \|\Psi_{i,\cdot}\|_2 + \max_{i\in[p]} \|E_{\cdot,i}/\sqrt{n}\|_2 = O_p(\max_{i\in[p]} \|\Psi_{i,\cdot}\|_2)$, since U and E have iid sub-Gaussian entries, Theorem 5.39 in Vershynin (2010) gives $\|U/\sqrt{n}\|_2 = O_p(1)$ and $\max_{i\in[p]} \|E_{\cdot,i}/\sqrt{n}\|_2 = O_p(1)$. Combining these two bounds gives the desired rate.

□

Lemma C.3. *Suppose conditions in Theorem A.1 hold. Let $\lambda_i(X^*X)$ denote the i-th singular value of X, and let $\widetilde{\lambda}_i(X^*X)$ denote the corresponding singular value after the Trim-transformation. If the number of clipped singular values satisfies $m+q = o(n)$, then for $q+1 \le i \le m$, $\frac{\widetilde{\lambda}_i(X^*X)}{\lambda_i(X^*X)} = \Theta_p(1)$.*

Proof. In our scenario, the trim transformation sets the clipped singular values to the m-th singular value of $\lambda_i(X^*X)$. Therefore, it suffices to show that $\lambda_i(X^*X) = \Theta_p(1)$ with high probability for all $q+1 \le i \le m$.

To this end, we will show that (1) $\lambda_m(X^*X) = 1 - O_p(\sqrt{(m+q)/n} + \sqrt{\log n/n})$ and (2) $\lambda_{q+1}(X^*X) = 1 + O_p(\sqrt{p/n} + \sqrt{\log n/n})$. Then, by the monotonicity of singular values and $\lim p/n \to c \in (0, \infty)$ and $m + q = o(n)$ we conclude that $\lambda_i(X^*X) = \Theta_p(1)$ for $q + 1 \leq i \leq m$.

Proof of (1) $\lambda_m(X^*X) = 1 - O_p(\sqrt{(m+q)/n} + \sqrt{\log n/n})$. Note that by $X/\sqrt{n} = U\Psi^\top/\sqrt{n} + E/\sqrt{n}$, then by the Weyl's inequality, we have $\lambda_{i+j-1}(E/\sqrt{n}) \leq \lambda_i(X/\sqrt{n}) + \lambda_j(-U\Psi^\top/\sqrt{n})$ for any i, j. Setting $i = m$ and $j = q + 1$ implies

$$\lambda_m(X/\sqrt{n}) \geq \lambda_{m+q}(E/\sqrt{n}) \to \lambda_m(X^*X) \geq \lambda_{m+q}(E^*E) \tag{59}$$

Next, we show $\lambda_{m+q}(E^*E) = 1 - O_p(\sqrt{(m+q)/n})$. To this end, we consider the submatrix E_M, where M denotes a subset of p-columns and $|M| = m + q + 1$. By the Cauchy interlacing theorem, we have

$$\lambda_{m+q}(E^*E) \geq \lambda_{m+q+1}(E_M^*E_M). \tag{60}$$

Then, by applying Theorem 5.39 in Vershynin (2010), with probability at least $1 - n^{-c}$ for some constant c, we have $\lambda_{m+q+1}(E_M/\sqrt{n}) = 1 - O_p(\sqrt{(m+q)/n} + \sqrt{\log n/n})$, which combines with $equation\ 59 - equation\ 60$ yields $\lambda_m(X^*X) = 1 - O_p(\sqrt{(m+q)/n} + \sqrt{\log n/n})$.

Proof of (2) $\lambda_{q+1}(X^*X) = 1 + O_p(\sqrt{p/n} + \sqrt{\log n/n})$. This follows directly from Lemma D.2 (Weyl's inequality) that:

$$\lambda_{q+1}(X/\sqrt{n}) \leq \lambda_1(E/\sqrt{n}) \overset{(1)}{=} 1 + O_p(\sqrt{p/n} + \sqrt{\log n/n}) \to \lambda_{q+1}(X^*X) \leq \lambda_1(E^*E) \tag{61}$$

where "(1)" follows from Theorem 5.39 in Vershynin (2010). By combining with equation 59 - equation 61 yields the result.

□

Lemma C.4. *(Left singular subspace consistency) Consider the model $X = U\Psi^\top + E$. Let $X/\sqrt{n} = LDV^\top$ be its rank-q thin singular value decomposition. Let $P_U \overset{\text{def}}{=} U(U^\top U)^{-1}U^\top$ be the orthogonal projection onto* $\mathrm{col}(U)$*. If $\lambda_q(\Psi) = \Omega(\sqrt{p})$ and $\|\frac{E}{\sqrt{n}}\|_2 = O_p(1)$, then $\|LL^\top - P_U\|_2 = O_p\left(\sqrt{\frac{q}{p}}\right)$.*

Proof. The proof is straightforward from standard matrix perturbation theory Stewart (1998).

Lemma C.5. *(Stewart (1998) Theorem 4) Let $(U_1U_2U_3)^{\mathrm{H}} A (V_1V_2) = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{pmatrix}$ be a singular value decomposition of A and let $\tilde{A} = A+E$ be a perturbation of A with perturbed singular subspaces satisfying $\left(\tilde{U}_1\tilde{U}_2\tilde{U}_3\right)^{\mathrm{H}} \tilde{A} \left(\tilde{V}_1\tilde{V}_2\right) = \begin{pmatrix} \tilde{\Sigma}_1 & 0 \\ 0 & \tilde{\Sigma}_2 \\ 0 & 0 \end{pmatrix}$. Further, let Φ be the matrix of canonical angles between $\mathcal{R}(U_1)$ and $\mathcal{R}(\tilde{U}_1)$, and let Θ be the matrix of canonical angles between $\mathcal{R}(V_1)$ and $\mathcal{R}(\tilde{V}_1)$. If there exists a $\delta > 0$ such that $\min\left|\sigma\left(\tilde{\Sigma}1\right) - \sigma\left(\Sigma_2\right)\right| \geq \delta$ and $\min\sigma\left(\tilde{\Sigma}1\right) \geq \delta$, then $\sqrt{\|\sin\Phi\|_{\mathrm{F}}^2 + \|\sin\Theta\|_{\mathrm{F}}^2} \leq \frac{\sqrt{\|R\|_{\mathrm{F}}^2 + \|S\|_{\mathrm{F}}^2}}{\delta}$, where $R = A\tilde{V}_1 - \tilde{U}_1\tilde{\Sigma}_1$ and $S = A^{\mathrm{H}}\tilde{U}_1 - \tilde{V}_1\tilde{\Sigma}_1$.*

In our setting, we set $A \overset{\text{def}}{=} U\Psi^\top/\sqrt{n}$ and $\tilde{A} \overset{\text{def}}{=} U\Psi^\top/\sqrt{n} + E/\sqrt{n}$; . Accordingly, in the notation above, $\widetilde{U}_1 \overset{\text{def}}{=} L_q$, $\widetilde{V}_1 \overset{\text{def}}{=} V_q$ and $\widetilde{\Sigma}_1 \overset{\text{def}}{=} D_q$, where L_q, V_q, D_q are the leading q left singular vectors, right singular vectors, and singular values of $X/\sqrt{n}$, respectively. Here $\delta \overset{\text{def}}{=} \lambda_q(\Psi) = \Theta_p(\sqrt{p})$.

The Wedin's theorem gives

$$\|LL^\top - P_U\|_2 \leq \frac{1}{\sqrt{2}}\|LL^\top - P_U\|_{\mathrm{F}} \overset{(1)}{=} \|\sin\Phi\|_{\mathrm{F}} \leq \sqrt{\|\sin\Phi\|_{\mathrm{F}}^2 + \|\sin\Theta\|_{\mathrm{F}}^2} \leq \frac{\sqrt{\|R\|_{\mathrm{F}}^2 + \|S\|_{\mathrm{F}}^2}}{\delta},$$

where "(1)" follows Stewart (1998). Since $\delta \overset{\text{def}}{=} \lambda_q(\Psi) = \Theta_p(\sqrt{p})$, it remains to show that $\|R\|_F = O_p(\sqrt{q})$ and $\|S\|_F = O_p(\sqrt{q})$.

Recall that

$$R \overset{\text{def}}{=} AV_q - L_q D_q = AV_q - L_q D_q V_q^\top V_q = (A - \widetilde{A})V_q = -\frac{E}{\sqrt{n}} V_q. \tag{62}$$

Hence, we have

$$\|R\|_F^2 = \| - \frac{E}{\sqrt{n}} V_q\|_F^2 = \mathrm{tr}(V_q^\top \frac{E^\top}{\sqrt{n}} \frac{E}{\sqrt{n}} V_q) \leq \mathrm{tr}(V_q^\top V_q) \| \frac{E}{\sqrt{n}}\|_2^2 \overset{(1)}{=} O_p(q). \tag{63}$$

where "(1)" applying Theorem 5.39 in Vershynin (2010) with probability at least $1 - n^{-c}$ for some constant c, we have $\lambda_1(E/\sqrt{n}) = 1 + O_p(\sqrt{p/n}) = O_p(1)$. This implies $\|R\|_F = O_p(\sqrt{q})$ and the same argument applies to $\|S\|_{\mathrm{F}}$, which completes the proof. □

Lemma C.6. *Suppose conditions in Theorem A.1 hold. Define the following event*

$$\mathcal{A} = \left\{ \widetilde{\varepsilon} : \left\| \widetilde{X}_S \left(\widetilde{X}_S^* \widetilde{X}_S \right)^{-1} \widetilde{X}_S^* \widetilde{\varepsilon}_b \right\|_2 \lesssim 2\sqrt{\|\phi\|_2^2 + \sigma^2} \sqrt{s \log n} \right\},$$

$$\mathcal{B} = \left\{ \widetilde{\varepsilon}_b : \left\| \left(\widetilde{X}_S^* \widetilde{X}_S \right)^{-1} \widetilde{X}_S^* \widetilde{\varepsilon}_b \right\|_\infty \lesssim 2\sqrt{\|\phi\|_2^2 + \sigma^2} \sqrt{\frac{\log s}{n\gamma}} \right\},$$

$$\mathcal{C} = \left\{ \widetilde{\varepsilon} : \left\| \widetilde{X}_T^* P_T \widetilde{\varepsilon}_b \right\|_\infty \lesssim 2\sqrt{\|\phi\|_2^2 + \sigma^2} \|\widetilde{X}_i/\sqrt{n}\|_2 \sqrt{\frac{\log p}{n}} \right\}.$$

Then, for some positive constant $c > 0$, we have

$$\mathbb{P}(\mathcal{A}) \geq 1 - 2n^{-c}, \quad \mathbb{P}(\mathcal{B}) \geq 1 - 2s^{-c}, \quad \mathbb{P}(\mathcal{C}) \geq 1 - 2p^{-c}.$$

Proof. We only provide the proof for the event $\mathcal{A}$, since the proof for the other two events is similar. First, note that

$$\widetilde{X}_S \left(\widetilde{X}_S^* \widetilde{X}_S \right)^{-1} \widetilde{X}_S^* \widetilde{\varepsilon}_b = \widetilde{X}_S \left(\widetilde{X}_S^* \widetilde{X}_S \right)^{-1} \widetilde{X}_S^* P_m \varepsilon_b.$$

By Ćevid et al. (2020) and the fact that X and ε_b are Gaussian, X is uncorrelated to and thus is independent of ε_b. Since P_m is determined by X, $\widetilde{X} = P_m X$ and P_m are also independent of ε_b. Therefore, $\widetilde{X}_S \left(\widetilde{X}_S^* \widetilde{X}_S \right)^{-1} \widetilde{X}_S^* P_m$ is independent of ε_b. Besides, we have

$$\begin{aligned} \|\widetilde{X}_S \left(\widetilde{X}_S^* \widetilde{X}_S \right)^{-1} \widetilde{X}_S^* P_m\|_2 &\leq \|\widetilde{X}_S \left(\widetilde{X}_S^* \widetilde{X}_S \right)^{-1} \widetilde{X}_S^*\|_2 \|P_m\|_2 \\ &\leq \|\widetilde{X}_S \left(\widetilde{X}_S^* \widetilde{X}_S \right)^{-1} \widetilde{X}_S^*\|_2. \end{aligned}$$

Besides, by equation 42 and equation 43, we have $\mathrm{Var}(\varepsilon_{b,i}) = \|\phi\|_2^2 + \sigma^2$. The rest of the proof follows from Theorem 4.1 in Osher et al. (2016). □

Lemma C.7. *Suppose conditions in Theorem A.1 hold. Besides, we assume $m = o(n^{3/4})$. Then, for any δ, with probability at least $1 - \delta$, we have*

$$\|(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \widetilde{X}_T b_T\|_2 \leq C_\delta \|\phi\|_2 \sqrt{\frac{sq}{p}}. \tag{64}$$

Proof. First, we can expand $\|(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \widetilde{X}_T b_T\|_2$ as

$$\begin{aligned} \|(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \widetilde{X}_T b_T\|_2 &= \|(\widetilde{X}_S^* \widetilde{X}_S)^{-1} \widetilde{X}_S^* \sum_{j \in S^c} \widetilde{X}_j b_j\|_2, \\ &\overset{(1)}{\leq} \frac{\sqrt{s} \max_{i \in S} |\sum_{j \in S^c} (\widetilde{X}^* \widetilde{X})_{i,j} b_j|}{\lambda_{\min}(\widetilde{X}_S^* \widetilde{X}_S)}, \\ &= O_p \left(\sqrt{s} \max_{i \in S} |\sum_{j \in S^c} \{\widehat{\Sigma}(P_m X)\}_{i,j} b_j| \right). \end{aligned}$$

where "(1)" follows from equation 23 that $\lambda_{\min}(\widetilde{X}_S^* \widetilde{X}_S) = c_0 - o_p(1)$.

It is sufficient to show that $\max_{i\in S} |\sum_{j\in S^c} \{\widehat{\Sigma}(P_m X)\}_{i,j} b_j| = O_p(\|\phi\|_2/\sqrt{p})$. By the same argument as in equation 24, we obtain

$$\begin{aligned}
&\max_{i\in S} |\sum_{j\in S^c} \{\widehat{\Sigma}(P_m X)\}_{i,j} b_j| \\
&= \max_{i\in S} \left| \sum_{k=1}^{q} \widetilde{\lambda}_m^2 v_{ki} \sum_{j\in S^c} v_{kj} b_j \right| + \max_{i\in S} \left| \frac{X_i^\top}{\sqrt{n}} (I - L_q L_q^\top) \sum_{j\in S^c} \frac{X_j}{\sqrt{n}} b_j \right| + \max_{i\in S} \left| \sum_{k=q+1}^{m} \left(\lambda_k^2 - \widetilde{\lambda}_m^2\right) v_{ki} \sum_{j\in S^c} v_{kj} b_j \right|, \\
&\stackrel{\text{def}}{=} \widetilde{I} + \widetilde{II} + \widetilde{\Delta}.
\end{aligned}$$

We first provide some upper bounds as follows. First, note that

$$\| \sum_{j\in S^c} b_j \Psi_{j,\cdot}^\top \|_2 = \|\Psi_{S^c}^\top b_{S^c}\|_2 \le \|\Psi_{S^c}\|_2 \|b\|_2 \stackrel{(1)}{\lesssim} \|\phi\|_2, \tag{65}$$

where "(1)" follows from $\max_{1\le i\le p} \|\Psi_{i,\cdot}\|_2 = O_p(1)$ and $\|\Psi_{S^c}\|_2 \le \|\Psi\|_2 = \Theta_p(\sqrt{p})$ while the bound for the bias term $\|b\|_2 = O_p(\|\phi\|_2/\sqrt{p})$ follows from Lemma D.4.

By applying Theorem 5.39 in Vershynin (2010) and Lemma C.4 that $p = \Theta(n)$ and $s = o(n)$, we have

$$\begin{aligned}
&\|P_U - L_q L_q^\top\|_2 = \sqrt{\frac{q}{p}}, \quad \max_{i\in S} \|E_i/\sqrt{n}\|_2 = \|E/\sqrt{n}\|_2 = O_p(1), \quad \lambda_m = O_p(1), \\
&\|U/\sqrt{n}\|_2 = O_p(1), \quad \|b\|_2 = \frac{\|\phi\|_2}{\sqrt{p}}, \quad \|I - L_q L_q^\top\|_2 = O_p(1).
\end{aligned} \tag{66}$$

Next, we show that $\widetilde{I} = O_p\left(\frac{\|\phi\|_2}{p}\right)$, $\widetilde{II} = O_p\left(\|\phi\|_2\sqrt{\frac{q}{p}}\right)$ and $\widetilde{\Delta} = o_p\left(\frac{\|\phi\|_2}{\sqrt{p}}\right)$.

Proof of $\widetilde{I} = O_p\left(\frac{\|\phi\|_2}{p}\right)$.

$$\begin{aligned}
\widetilde{I} &= \max_{i\in S} |\sum_{k=1}^{q} \widetilde{\lambda}_m^2 v_{ki} \sum_{j\in S^c} v_{kj} b_j|, \\
&\stackrel{(1)}{\le} \widetilde{\lambda}_m^2 \max_{i\in S} \|(V_q)_{i,\cdot}\|_2 \|V_{q,S^c}^\top b_{S^c}\|_2 \stackrel{(2)}{\le} \widetilde{\lambda}_m^2 \max_{i\in S} \|(V_q)_{i,\cdot}\|_2 \|b_{S^c}\|_2 = O_p\left(\frac{\|\phi\|_2}{p}\right).
\end{aligned} \tag{67}$$

where "(1)" follows from the Cauchy–Schwarz inequality, while "(2)" follows from equation 66 and Lemmas C.2 and Lemmas D.4.

Proof of $\widetilde{II} = O_p\left(\|\phi\|_2\sqrt{\frac{q}{p}}\right)$.

$$\begin{aligned}
\widetilde{II} &= \max_{i\in S} \left| \frac{X_i^\top}{\sqrt{n}} (I - L_q L_q^\top) \sum_{j\in S^c} \frac{X_j}{\sqrt{n}} b_j \right|, \\
&= \max_{i\in S} \left| \frac{1}{n} E_i^\top \left(I - L_q L_q^\top\right) E_{S^c} b_{S^c} \right| + \max_{i\in S} \left| \frac{1}{n} E_i^\top \left(I - L_q L_q^\top\right) U \Psi_{S^c}^\top b_{S^c} \right| \\
&\quad + \max_{i\in S} \left| \frac{1}{n} \Psi_{i,\cdot} U^\top \left(I - L_q L_q^\top\right) E_{S^c} b_{S^c} \right| + \max_{i\in S} \left| \frac{1}{n} \Psi_{i,\cdot} U^\top \left(I - L_q L_q^\top\right) U \Psi_{S^c}^\top b_{S^c} \right|, \\
&= \widetilde{II}_1 + \widetilde{II}_2 + \widetilde{II}_3 + \widetilde{II}_4.
\end{aligned}$$

Next, equation 66 gives

$$\widetilde{II}_1 \le \max_{i\in S} \frac{\|E_i\|_2}{\sqrt{n}} \frac{\|E_{S^c} b_{S^c}\|_2}{\sqrt{n}} \|I - L_q L_q^\top\|_2 = O_p\left[\frac{\|\phi\|_2}{\sqrt{p}} \left(1 + \sqrt{\frac{\log s}{n}}\right)\right] = O_p\left(\frac{\|\phi\|_2}{\sqrt{p}}\right). \tag{68}$$

Then, it follows from the identity $(I - L_q L_q^\top)U = (P_U - L_q L_q^\top)U$,

$$\widetilde{II}_2 \le \max_{i\in S} \frac{\|E_i\|_2}{\sqrt{n}} \left\|P_U - L_q L_q^\top\right\|_{\mathrm{op}} \frac{\|U\|_{\mathrm{op}}}{\sqrt{n}} \left\|\Psi_{S^c}^\top b_{S^c}\right\|_2 = O_p\left(\|\phi\|_2 \sqrt{\frac{q}{p}}\right), \tag{69}$$

$$\widetilde{II}_3 = \widetilde{II}_4 \le \max_{i\in S} \|\Psi_{i,\cdot}\|_2 \frac{\|U\|_{\mathrm{op}}}{\sqrt{n}} \left\|P_U - L_q L_q^\top\right\|_{\mathrm{op}} \frac{\|E_{S^c} b_{S^c}\|_2}{\sqrt{n}} \overset{(1)}{=} O_p\left(\frac{\sqrt{q}\|\phi\|_2}{p}\right). \tag{70}$$

where "(1)" uses $\max_{i\in S} \|\Psi_{i,\cdot}\|_2 \le \|\Psi\|_2 = \Theta_p(\sqrt{p})$.

Combining equation 68 and equation 70, we obtain $\widetilde{II} = O_p\left(\|\phi\|_2 \sqrt{\frac{q}{p}}\right)$.

Proof of $\widetilde{\Delta} = o_p\left(\frac{\|\phi\|_2}{\sqrt{p}}\right)$.

The same argument as in Step II-C of the proof of Theorem A.1 gives

$$\begin{aligned}\widetilde{\Delta} &= \max_{i\in S} \left| \sum_{k=q+1}^{m} \left(\widetilde{\lambda}_m^2 - \lambda_k^2\right) v_{ki} \sum_{j\in S^c} v_{kj} b_j \right| \le \max_{l<k\le m} \left|\widetilde{\lambda}_m^2 - \lambda_k^2\right| \max_{i\in S} \left(\sum_{k=q+1}^{m} v_{ki}^2\right)^{1/2} \left\|V_{S^c,q+1:m}^\top b_{S^c}\right\|_2, \\ &= O_\prec\left(\frac{\|\phi\|_2 \sqrt{m-q}}{p}\right).\end{aligned}$$

Taking $\epsilon = 1/8$ and assuming that $q < m$ and $m = o(n^{3/4})$, we obtain $\widetilde{\Delta} = o_p\left(\|\phi\|_2/\sqrt{p}\right)$. Combining these results leads to the desired conclusion.

□

Lemma C.8. *Consider model 1. Then $\lambda_{\max}(\widetilde{X}_S^* \widetilde{X}_S) = 1 + O_p(\frac{p}{n} + \sqrt{\frac{p}{n}})$, where $\widetilde{X}_S$ denotes the submatrix of $\widetilde{X} \overset{\mathrm{def}}{=} P_m X$ consisting of the columns indexed by S.*

Proof. This argument is analogous to Step I in Theorem 4.1 as follows.

$$\lambda_{\max}(\widetilde{X}_S^* \widetilde{X}_S) \overset{(1)}{\le} \lambda_{\max}(\widetilde{X}^* \widetilde{X}) \overset{(2)}{\le} \lambda_{\max}(E^* E) = 1 + O_p\left(\frac{p}{n} + \sqrt{\frac{p}{n}}\right).$$

where "(1)" follows from the Cauchy interlacing theorem; "(2)" is due to the Weyl's inequality $\lambda_{\max}(\widetilde{X}/\sqrt{n}) \le |\lambda_{q+1}(X/\sqrt{n} + U\Psi^\top/\sqrt{n}) - \lambda_{q+1}(U\Psi^\top/\sqrt{n})| \le \lambda_{\max}(E/\sqrt{n}) = 1 + O_p\left(\sqrt{\frac{p}{n}}\right)$. □

D Supporting Results from the Literature

Lemma D.1. *(Hanson-Wright inequality Guo et al. (2022) LEMMA 11) Let $\xi \in \mathbb{R}^n$ be a random vector with independent sub-Gaussian components ξ_i with zero mean and sub-Gaussian norm K. Let A be an $n \times n$ matrix. Then for every $t \ge 0$,*

$$\mathbf{P}\left(\left|\xi^\top A\xi - \mathbb{E}\xi^\top A\xi\right| > t\right) \le 2\exp\left[-c\min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|_2}\right)\right]. \tag{71}$$

Lemma D.2. *(Weyl's inequality Horn & Johnson (2012) Corollary 7.3.5) Let $A, B \in M_{n,p}$ and let $q = \min\{n, p\}$. Let $\sigma_1(A) \ge \cdots \ge \sigma_q(A)$ and $\sigma_1(B) \ge \cdots \ge \sigma_q(B)$ be the nonincreasingly ordered singular values of A and B, respectively. Then*

$$|\sigma_i(A) - \sigma_i(B)| \le \|A - B\|_2 \quad \textit{for each } i = 1, \ldots, q.$$

Lemma D.3 (Delocalization bound for non-outlier eigenvectors (Bloemendal et al., 2016, Theorem 2.17))**.** *Let $X \overset{\mathrm{def}}{=} U\Psi^\top + E \in \mathbb{R}^{n\times p}, C \overset{\mathrm{def}}{=} X^\top X/p$. where $U \in \mathbb{R}^{n\times q}$, $\Psi \in \mathbb{R}^{p\times q}$, and $E \in \mathbb{R}^{n\times p}$. Assume that $q = O(1)$, Ψ is deterministic, and the entries of U and E satisfy the*

required independence and moment conditions. Let $\Sigma \overset{\text{def}}{=} I_p + \Psi\Psi^\top = \sum_{i=1}^p \sigma_i v_i v_i^\top, d_i \overset{\text{def}}{=} \sqrt{\frac{n}{p}}(\sigma_i - 1), K \overset{\text{def}}{=} n \wedge p$. *Define* $\mathcal{R} \overset{\text{def}}{=} \{i : d_i \neq 0\}$, *and* $\mathcal{O} \overset{\text{def}}{=} \left\{i \in \mathcal{R} : |d_i| > 1 + K^{-1/3}\right\}$. *Let* $\xi_a \in \mathbb{R}^p$ *be the unit eigenvector associated with the* a*th largest nonzero eigenvalue of* C. *Then, for any fixed* $\tau > 0$, *any* $a \in [1, (1-\tau)K] \setminus \mathcal{O}$, *with* $\kappa_a \overset{\text{def}}{=} K^{-2/3}\big(a \wedge (K+1-a)\big)^{2/3}$, *and any deterministic* $w \in \mathbb{R}^p$, *we have*

$$|\langle w, \xi_a \rangle|^2 \prec \frac{\|w\|_2^2}{p} + \sum_{i=1}^p \frac{\sigma_i \left|\langle v_i, w \rangle\right|^2}{p\left((d_i - 1)^2 + \kappa_a\right)}. \tag{72}$$

Lemma D.4. *(Ćevid et al. (2020) Lemma 6) Assume the confounding Model 1 and dense confounding* $\lambda_{\min}(\Psi) = \Omega(\sqrt{p})$. *Suppose* Σ_E *is well-conditioned with constants* $0 < c_{\min} \leq \lambda_{\min}(\Sigma_E) \leq \lambda_{\max}(\Sigma_E) \leq c_{\max} < \infty$. *Then*

$$\|b\|_\infty \leq \|b\|_2 \leq C \frac{\|\phi\|_2}{\lambda_q(\Psi)}$$

for an absolute constant C *(depending only on* $c_{\min}, c_{\max}$ *).*

E Additional Experiments

E.1 Implementation details.

For LBI, we follow the implementation in Xiong et al. (2018). Specifically, we set the damping parameter to $\kappa = 100$ and choose the step size as $\alpha = \|X\|_2/\kappa$. For Lasso, we use the `glmnet` package[2] and select the regularization parameter via cross-validation. In particular, the optimal λ for Lasso is chosen from a grid of $K = 100$ uniformly spaced values on $[\|X^\top y/n\|_\infty, 10^{-6}]$. For LBI and Lasso with the Trim transformation (*i.e.*, Spec+LBI and Spec+Lasso), we first apply the Trim transformation P_m to X and y, and then run LBI or Lasso on the transformed design matrix and response.

For the choice of m, under the conditions of Theorem A.1—namely, $p/n \to c \in (0, \infty)$, $s = o(\sqrt{n}/\log n)$, $m > q$, and $m = o\left(n^{3/4}/s\right)$—the choice of m is not delicate. For fixed s, choices such as $m = n^{0.7}$ already work, provided that they also satisfy $m > q$.

E.2 Visualization of Theorem. 4.1

In this subsection, under the setting of Theorem. 4.1, we visualize the sample covariance of PX. Set $q = 1$ and U, Ψ, and E are generated with i.i.d. Gaussian entries. Moreover, support size is set to $s = 5$.

E.2.1 $\lim \frac{p}{n} \to 0$

Consider the $n \gg p$ and the threshold is set to $m = p^{0.7}$. Fig. 6 and Fig. 7 show that fixed p, as n increases, the diagonal entries of the sample covariance of PX converge to 1 , while the off-diagonal entries converge to 0.

E.2.2 $\lim \frac{p}{n} \to c$

Consider the $n < p$ and the threshold is set to $m = n^{0.7}$. Fig. 8 and Fig. 9 show that as n and p increase, both the diagonal and off-diagonal entries of PX gradually decrease.

E.3 ℓ_2 estimation error under other parameter settings

Additional experiments. In Fig. 10, we report the ℓ_2 estimation error across varying support sizes s, low dimensions, and confounder numbers q. Unless otherwise specified, we fix the aspect ratio at $p/n = 2$. We find that Spec-LBI consistently achieves the lowest ℓ_2 estimation error in all three settings for $\lim p/n \to c$.

[2] `https://cran.r-project.org/web/packages/glmnet/index.html`.

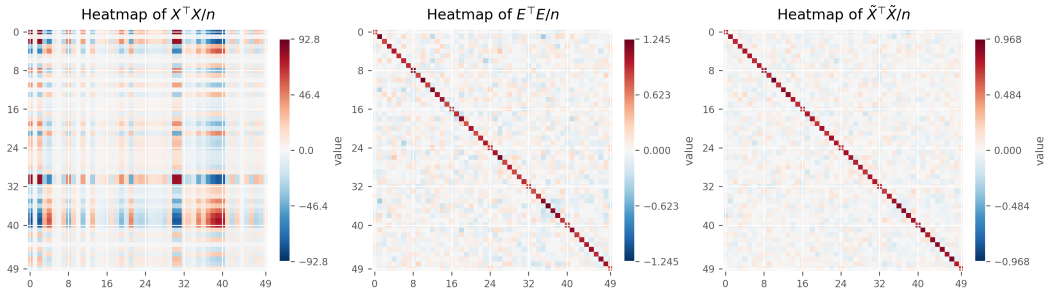


Figure 6: Heatmaps for $n = 100$, $p = 50$ and $q = 1$. Each cell shows the corresponding entry of the sample covariance matrix: red indicates larger positive values, while blue indicates more negative values.

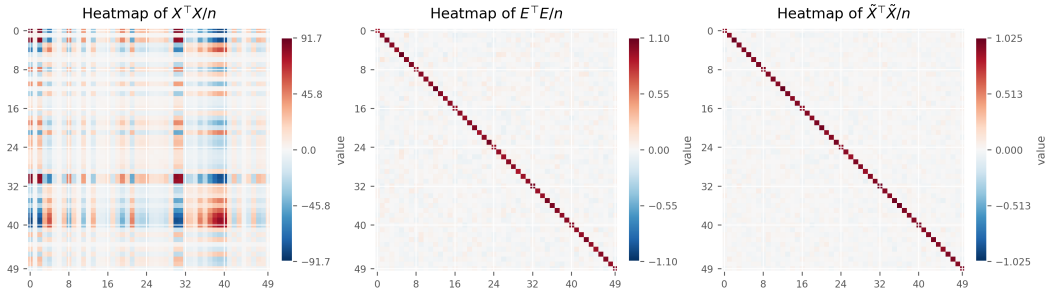


Figure 7: Heatmaps for $n = 1000$, $p = 50$ and $q = 1$. As n increases, the right panel shows that the diagonal entries of PX covariance approach 1, while the off-diagonal entries approach 0.

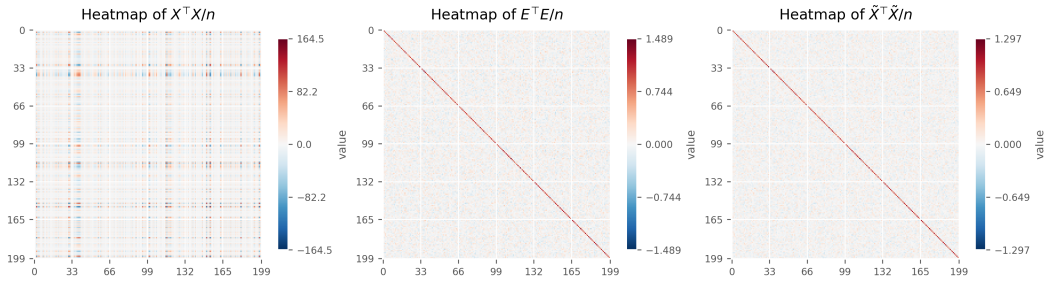


Figure 8: Heatmaps for $n = 100$, $p = 200$ and $q = 1$. Each cell shows the corresponding entry of the sample covariance matrix: red indicates larger positive values, while blue indicates more negative values.

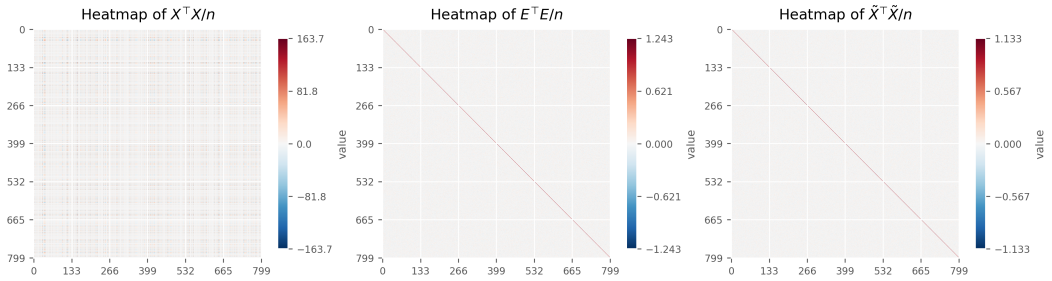
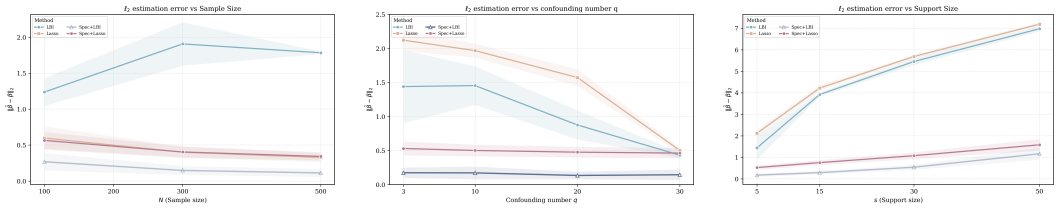


Figure 9: Heatmaps for $n = 400$, $p = 800$ and $q = 1$. As n and p increase simultaneously, the right panel shows that the diagonal entries of PX covariance converge to 1, while the off-diagonal entries converge to 0.



(a) ℓ_2-error v.s Sample size (b) ℓ_2-error v.s confounders q (c) ℓ_2-error v.s Support

Figure 10: ℓ_2-error under different settings.

E.4 Partial Confounding.

To illustrate that our theory does not require every covariate to be connected to the confounder, we use the data-generating mechanism in Section 6.1 with $q = 1$ and $p/n = 2$. We set exactly c entries of Ψ to zero and draw the remaining entries independently from $\mathcal{N}(0, 4^2)$. Table 1 reports the resulting model-selection performance. The AUC remains high across all values of c, indicating that the solution path selects true signals before null features and hence preserves the no-false-positive property.

Table 1: AUC under different values of p

AUC/ p	100	200	300
Trim ($c = 0$)	0.987 (0.010)	0.999 (0.0008)	0.999 (0.0008)
Trim ($c = 10$)	0.987 (0.019)	0.999 (0.0009)	0.999 (0.0008)
Trim ($c = 20$)	0.987 (0.019)	0.999 (0.0010)	0.999 (0.0010)
Trim ($c = 30$)	0.987 (0.017)	0.999 (0.0008)	0.999 (0.0008)

In the table, values outside parentheses are means, and values in parentheses are standard deviations.

E.5 Heavy-tailed distributions

In this section, we empirically study the effect of combining the trim transformation with LBI in heavy-tailed settings.

First, we consider the isotropic covariance setting. Specifically, we generate the entries of U, E, and e independently from a Student's t distribution with three degrees of freedom and set $p/n = 2$. Fig.11 and Fig.12 report the resulting ℓ_2 estimation error and AUC, respectively. Spec+LBI consistently outperforms the competing methods, achieving the smallest ℓ_2 error and the largest AUC.

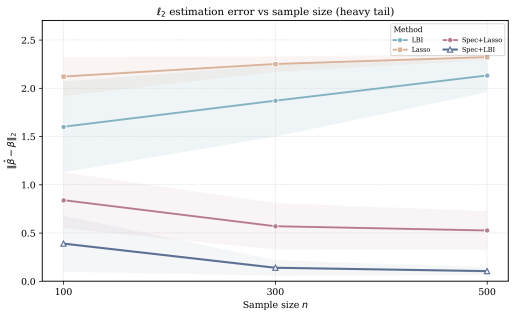


Figure 11: ℓ_2-error on heavy-tailed isotropic distributions

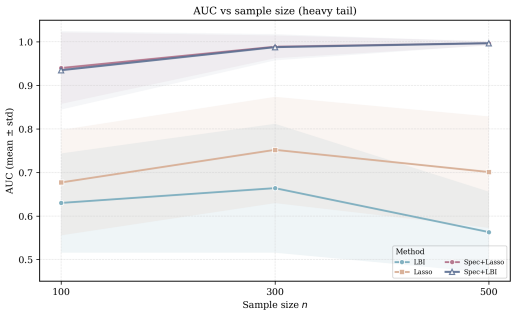


Figure 12: AUC on heavy-tailed isotropic distributions

Next, we study the non-isotropic covariance setting. Specifically, we generate i.i.d. U and e from a Student-t distribution with 3 degrees of freedom, and let E follow a Student-t distribution with a

Toeplitz covariance (ρ=0.3,0.5,0.9) for $n = 300$, $p = 600$, and $q = 3$. The results remain consistent: Spec+LBI clearly outperforms in both ℓ_2-error and Area Under the Curve (AUC) for model selection shown in Fig.13 and Fig.14.

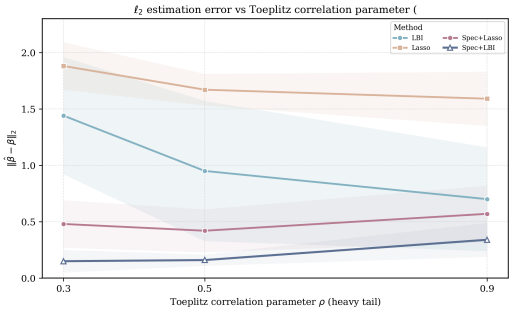


Figure 13: ℓ_2-error on heavy-tailed isotropic distributions

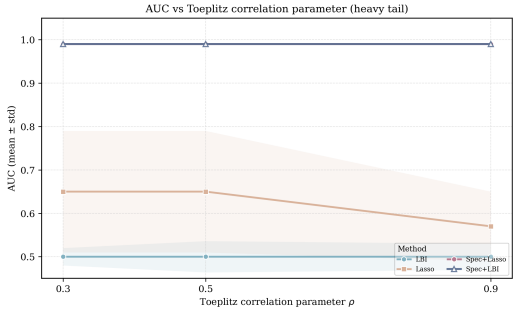


Figure 14: AUC on ℓ_2-error for heavy-tailed data with Toeplitz correlation

E.6 Covariance setting with confounding

In this section, we examine the ℓ_2 error and KKT conditions under various covariance structures.

In the following, we assume the covariance matrices follow Toeplitz $(\Sigma_E)_{i,j} = \rho^{|i-j|}$ or Matérn kernel structures with distinct correlation parameters ℓ. The Matérn covariance between measurements taken at two points separated by d distance units is given by $C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{d}{\ell}\right)^\nu K_\nu \left(\sqrt{2\nu}\frac{d}{\ell}\right)$. Here, the length-scale parameter ℓ controls the correlation range: larger ℓ implies longer-range correlations. In addition, we assume $\Psi_{ij} \sim \mathcal{N}\left(0, 4^2\right)$, $n = 200$, $q = 3$ and $p = 300$.

First, we visualize these two scenarios, as shown in Fig. 15 and 16. It shows that after the Trim-transformation, the covariance is significantly small. This may be because the covariance structure also exhibits low-rank characteristics.

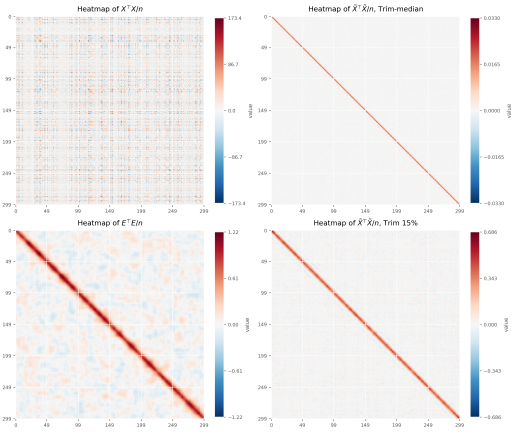


Figure 15: Matern Kernel covariance for length scale $\ell = 6$

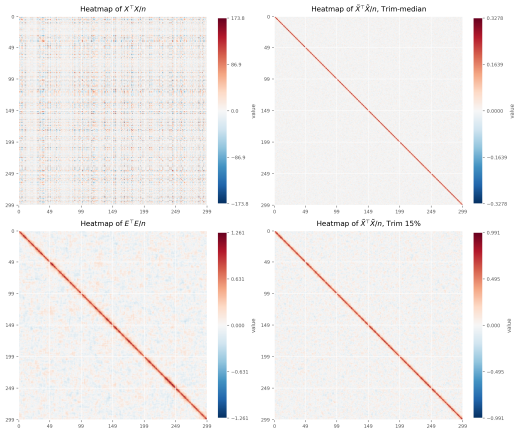


Figure 16: Toeplitz covariance for $\rho = 0.7$

Second, Fig. 17 and 18 report the ℓ_2-error with such covariance structures in the confounding setting. We observe that Trim-LBI outperforms even the Adaptive Lasso. Because the Adaptive Lasso depends on the initial estimator.

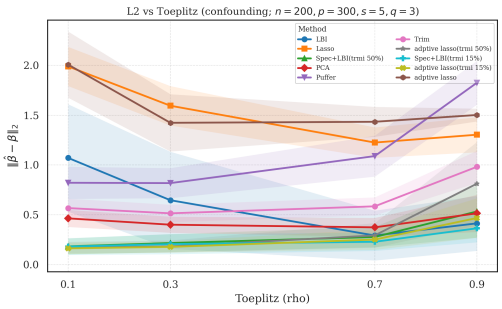


Figure 17: Toeplitz covariance for $\rho = 0.7$

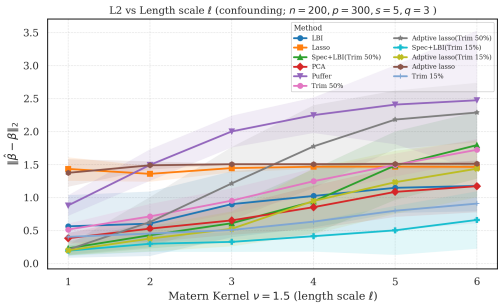


Figure 18: Matern Kernel covariance for length scale $\ell = 6$

E.7 Bias–variance trade-off.

Bias–variance trade-off. We examine the effect of the noise level on the ℓ_2 estimation error. We set $s = 5$, $n = 200$, $q = 3$, and $p = 300$, and vary the noise covariance $\Sigma_E = \sigma I_p$. Fig. 19 reports the results. As σ increases, the ℓ_2 error of LBI increases and eventually exceeds that of Lasso.

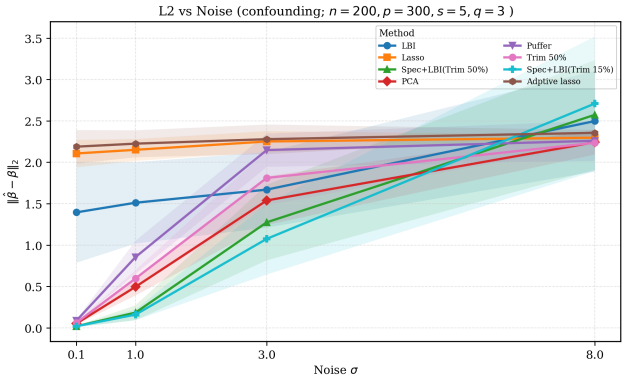


Figure 19: Bias–variance trade-off for ℓ_2 as the noise level σ varies.

F Additional RealWorld Experiments on LLM Value Alignment

Our LLM experiment studies a wellestablished task Kim et al. (2025); Gurnee & Tegmark (2024): learning the effect from hidden representations of LLMgenerated texts X on their LLMassigned concept scores y. However, prompt format and writing style may confound both X and y, as illustrated in Figure 1 (left). We test whether our method reduces this bias and recovers the effect and relevant words.

F.1 Implementation details.

In this section, we describe the implementation details for the LLM experiments. All experiments are conducted on two NVIDIA RTX 4090 GPUs with 24GB memory each.

F.2 Data Generation.

Prompt for Value Text Generation. We follow Zhang et al. (2026) to generate valuerelated texts across different value dimensions shown in Fig. 20.

Prompt for Value Annotation. In this section, we describe the prompts used for value scoring in the training and test sets. To remove promptstyle confounding, the test set is annotated using an

```
## Task
Design 10 sentences about
{value_item} according to thefollowing **Task
Description**## Task Description
1. Ensure all sentences need torelate to one of
conception,understanding and perspectiveabout
{value_item}
2. You need to use {value_item} ineach of your
sentences
3. Only response 10 sentences in**Output Format**
and not explain.4. Each answer is about 30 words##
Output Format
1.
2.
...
9.
10
## Output
```

(a) Text Generation Prompt

```
Description**.
\#\# Task Description
1. These ten sentences have no relation at all to
\{value_item\}.
2. These ten sentences do not refer to the human value
item \{value_item\}, and do not imply \{value_item\}
human value. At the same time, \{value_item\} cannot
be inferred from each sentence.
3. Only response 10 sentences in **Output Format**
and not explain.
4. Each sentence is about 30 words.
\#\# Output Format
1.
2.
...
9.
10.
\#\# Output
```

(b) Text Generation Prompt

Figure 20: Prompt for Value Text Generation

additional styleinvariant prompt shown in Figure 21, which controls for prompt style and focuses only on the semantic relevance to the target value, shown in Fig. 21.

Training

```
## Task Description
- You will be given a word (V) representing the human-value, and 20 paragraphs, all in English.
- For the paragraph (P), you should score in [0, 1, 2, 3, 4], based on how much you think about the given human-value word (V) when you read the context of paragraph (P).
- Score 0 means no relation at all, while 4 extremely very related to how you think about the given human-value word (V). Score 1-3 means intermedia level of relations. 1 for low, 2 for relatively low, 3 for moderate and 4 for high, in how much you think about the given value word (V).
- [Important] The paragraph (P) in the batch of paragraphs should be treated differently, i.e., a paragraph is **NOT** part of any other paragraph's context.
- Please output strictly in the given **Output Format**.

## Input Format
- Human-value Word: V
- Word's Meaning: ... [You need ignore this line in your scoring process.]
- Paragraphs List: [
    Paragraph 1,
    Paragraph 2,
    ...
    Paragraph 20
]

## Output Format
- Scores: [
    score1,
    score2,
    ...
    score20
]

## Note
- The length of Paragraph and Score List must be equal.
- Each word may consist of any characters (e.g., punctuation, special characters, etc.). Don't skip and fix any word.
```

(a) Value annotation prompt for the training set

Test

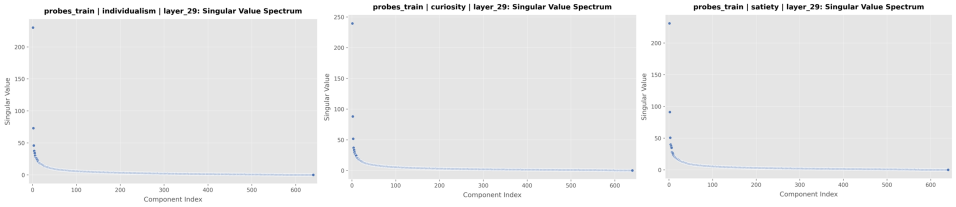
```
## Task Description
- You will be given a word (V) representing the human-value, and 20 paragraphs, all in English.
- For the paragraph (P), you should score in [0, 1, 2, 3, 4], based on how much you think about the given human-value word (V) when you read the context of paragraph (P).
- Score 0 means no relation at all, while 4 extremely very related to how you think about the given human-value word (V). Score 1-3 means intermedia level of relations. 1 for low, 2 for relatively low, 3 for moderate and 4 for high, in how much you think about the given value word (V).
- [Important] The paragraph (P) in the batch of paragraphs should be treated differently, i.e., a paragraph is **NOT** part of any other paragraph's context.
- Please output strictly in the given **Output Format**.

## Input Format
- Human-value Word: V
- Word's Meaning: ... [You need ignore this line in your scoring process.]
- Paragraphs List: [
    Paragraph 1,
    Paragraph 2,
    ...
    Paragraph 20
]

## Output Format
- Scores: [
    score1,
    score2,
    ...
    score20
]

## Note
- The length of Paragraph and Score List must be equal.
- Each word may consist of any characters (e.g., punctuation, special characters, etc.). Don't skip and fix any word.

Please focus only on the semantic meaning of the text.
Ignore superficial factors such as:
- prompt format or template structure;
- writing style, tone, fluency, or eloquence;
- sentence length or complexity;
- whether the text is poetic, academic, casual, persuasive, or formal;
- emotional intensity, politeness, or rhetorical quality;
- the identity, role, or style of the speaker.
```

(b) Value annotation prompt for the test set

Figure 21: Prompts used for value annotation in the training and test sets.

F.3 Dense Confounding in LLM Value Representations

We show that dense confounding appears across different value dimensions. Using layer 29 of Llama3.18BInstruct, we visualize the singular values of the hidden representations for the three value dimensions, as shown in the Figure 22.

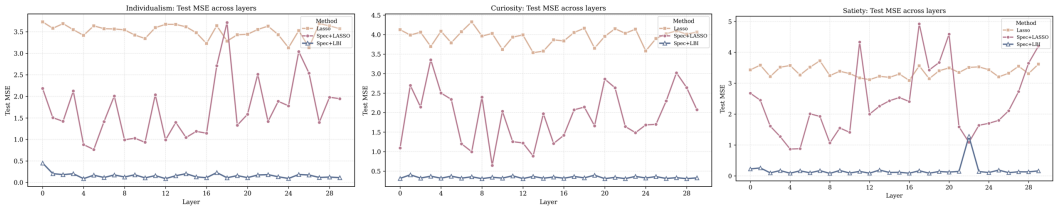


(a) Individualism (b) Curiosity (c) Satiety

Figure 22: Spiked Singular Values of LLM Representations across Value Dimensions

F.4 Test MSE across Value Dimensions

Figure 23 shows that SpecLBI achieves the lowest prediction error on the deconfounded data across different value dimensions and layers, indicating more accurate estimation.



(a) Individualism (b) Curiosity (c) Satiety

Figure 23: Test MSE across Value Dimensions

F.5 Visualization of Value Directions

To qualitatively assess whether the selected features capture valuerelated information, we visualize their tokenlevel attributions. Given valuerelated texts x_i and their scores y_i, we extract the lastlayer representation $h(x_i)$ and fit a sparse linear model, $\hat{y}_i = \hat{\beta}^\top h(x_i)$. Let $S = \mathrm{supp}(\hat{\beta})$ denote the selected features. For each token embedding e_t in a text x, we compute $A_t = \|\frac{\partial(\hat{\beta}^\top h(x))}{\partial e_t}\|_2$, which measures the sensitivity of the predicted value score to token t, with darker colors indicating higher attribution.

For the curiosity dimension, tokens closely related to curiosity, such as why and curiosity, receive high scores in Figure 24.

For individualism, Case 1 highlights individualism, singular, and person, while Case 2 highlights individualism in Figure 25.

For satiety, Case 1 emphasizes philosophical conception, satiety, mean, and eudaimonia, which together express the philosophical view of satiety as a state of moderation. Our method assigns higher attribution scores to these tokens than Lasso, as indicated by the darker shading. Case 2 highlights holistic, physical nourishment, emotional wellbeing, spiritual peace, and true satisfaction in Figure 26.

G Additional Real-World Experiments on Gene expression analysis

In this section, we describe the preprocessing steps and the dense confounding assumption, and report additional experimental results.

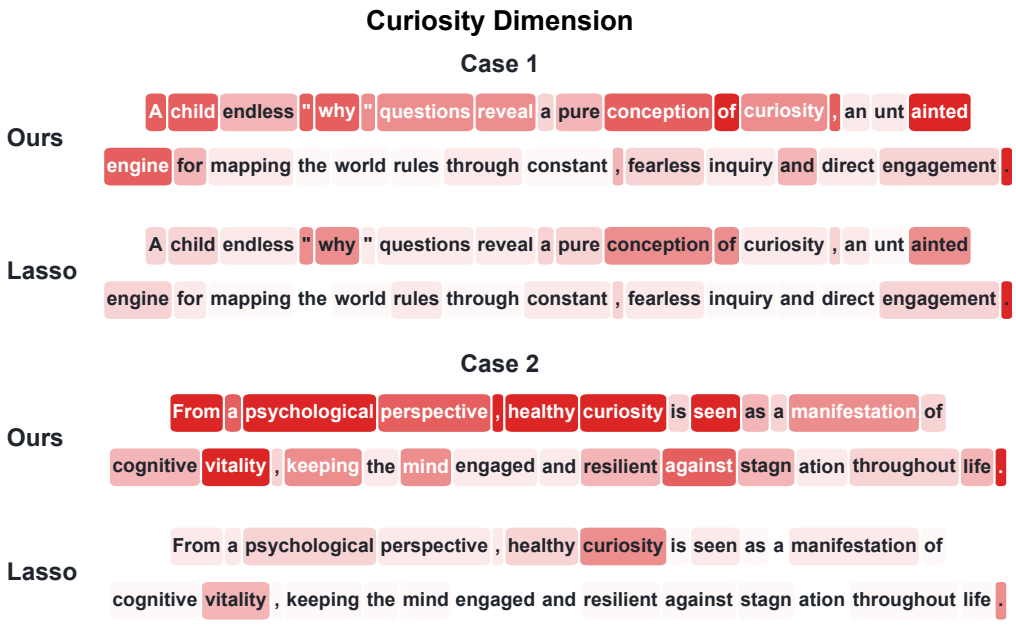


Figure 24: Visualization of Curiosity

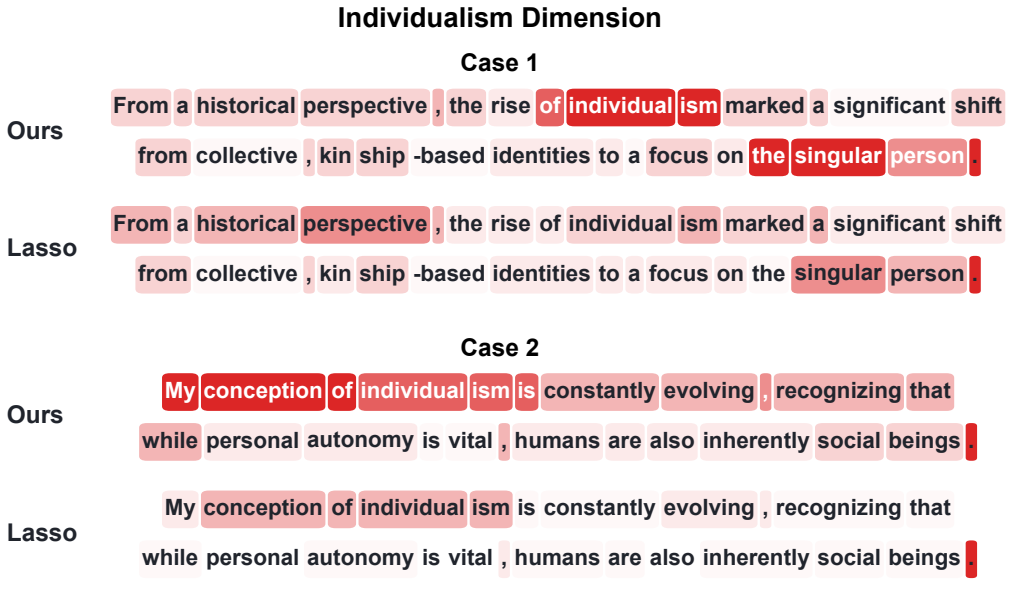


Figure 25: Visualization of Individualism

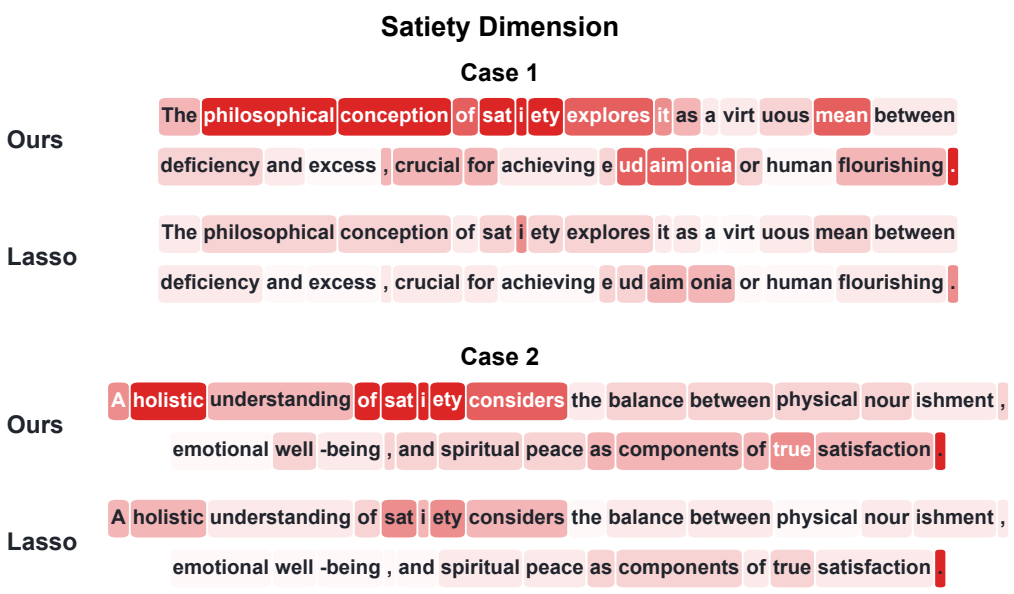


Figure 26: Visualization of Satiety

G.1 Dense Confounding in Gene Expression Analysis

The 4,027 genes provide empirical support for the dense confounding assumption, as shown in Figure 27.

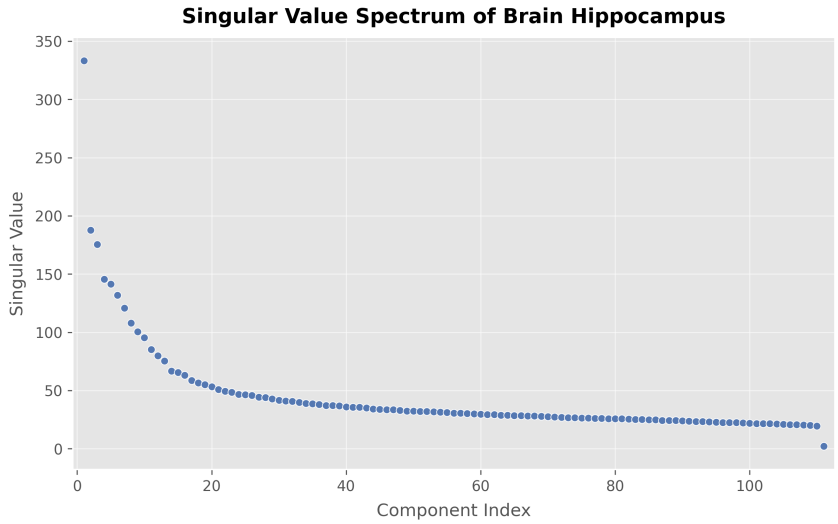


Figure 27: Spiked singular values of the standardized gene expression matrix.

G.2 Preprocessing of Gene Expression Data

The original dataset contains $23,736$ protein-coding genes from 115 individuals. To prioritize genes frequently reported across selected human brain and hippocampal transcriptomic studies Hawrylycz et al. (2012); Thompson et al. (2025), we applied a literature-informed pre-filter to the GTEx Brain Hippocampus expression matrix. Gene identifiers were mapped to approved HGNC symbols, and only mapped protein-coding genes were considered.

Table 2: Comparison of Spec-LBI and Lasso. Lower RMSE and higher Jaccard similarity are better.

Metric	Method	Mean	Std.
RMSE	**Spec-LBI (ours)**	**0.22**	0.20
	Lasso	0.42	0.16
Jaccard similarity	**Spec-LBI (ours)**	**0.60**	0.32
	Lasso	0.52	0.29

Specifically, we extracted gene symbols from the gene tables of these studies and ranked genes by their frequency across studies, counting each gene at most once per study. We retained the top $4,500$ candidates using a prespecified heuristic cutoff for dimensionality reduction, independent of downstream results. Finally, we intersected these candidates with the mapped protein-coding genes in our dataset, yielding $4,027$ genes for subsequent analysis.

27 **target genes y:** We then selected 27 target genes. Of these, 25 were identified using an objective, literature-guided procedure rather than selected post hoc based on their association with our experimental results. Specifically, we intersected the 4,027 genes identified in our analysis with those reported by Vogel et al. (2020), who showed that gene expression along the longitudinal axis of the human hippocampus encodes a molecular gradient associated with large-scale behavioral systems. This procedure yielded 25 overlapping genes: ONECUT2, GREM2, DDC, PVALB, CASR, WNT10A, DGKI, NTN1, BDKRB1, RGMA, GRHL2, SYTL2, SSTR1, C1QL1, PIRT, KCNG1, C1orf187, GPR39, OPRK1, MYB, GPR83, SYTL1, KIAA1772, SEMA3D, and GEFT. We additionally included two well-established brain-related genes: Brain-Derived Neurotrophic Factor (BDNF) Binder & Scharfman (2004) and Apolipoprotein E (APOE) Liu et al. (2013).

The corresponding results for these 27 genes are reported in Table 2. Spec-LBI achieves higher similarity and lower RMSE, indicating improved feature-selection performance.

Random target genes y: To further assess whether our method is sensitive to the choice of target genes, we conducted an additional robustness analysis using 70 randomly selected genes beyond the original target set shown in Table 3. Spec-LBI continues to achieve better performance than Lasso, with a lower RMSE (0.20 vs.0.32) and higher Jaccard similarity (0.53 vs. 0.51).

Table 3: Comparison of Spec-LBI and Lasso. Lower RMSE and higher Jaccard similarity are better.

Metric	Method	Mean	Std.
RMSE	**Spec-LBI (ours)**	**0.20**	0.16
	Lasso	0.32	0.15
Jaccard similarity	**Spec-LBI (ours)**	**0.53**	0.26
	Lasso	0.51	0.21