Statistica Sinica Preprint No: SS-2025-0220	
Title	High-dimensional Inference Via Hybrid Orthogonalization
Manuscript ID	SS-2025-0220
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0220
Complete List of Authors	Yang Li, Zemin Zheng, Jia Zhou and Ziwei Zhu
Corresponding Authors	Jia Zhou
E-mails	tszhjia@mail.ustc.edu.cn
Notice: Accepted author version.	

High-Dimensional Inference Via Hybrid Orthogonalization

Yang Li[1], Jia Zhou*[2], Zemin Zheng[1] and Ziwei Zhu[3]

[1] *University of Science and Technology of China*

[2] *Hefei University of Technology,* [3] *Radix Trading*

The past decade has witnessed a surge of endeavors in statistical inference for high-dimensional sparse regression, particularly via de-biasing or relaxed orthogonalization. Nevertheless, these techniques typically require a more stringent sparsity condition than is needed for estimation consistency, which seriously limits their practical applicability. To alleviate such a constraint, we propose to exploit the strong predictors to residualize the design matrix before performing de-biasing-based inference on the parameters of interest. This leads to a hybrid orthogonalization technique (HOT) that performs strict orthogonalization against the strong predictors but relaxed orthogonalization against the others. Under an approximately sparse model with a mixture of strong and weak signals, we establish the asymptotic normality of the HOT estimator while accommodating as many strong signals as consistent estimation allows. The efficacy of the proposed method is also demonstrated through numerical studies.

Key words and phrases: Large-scale inference, Relaxed orthogonalization, Hybrid orthogonalization, Feature screening, Partially penalized regression.

1. Introduction

Feature selection is crucial to a myriad of modern machine learning problems in high-dimensional settings, including bi-clustering analysis with high-throughput genomic data, collaborative filtering in sophisticated recommendation systems, and compression of deep neural networks, among others. However, characterizing the uncertainty of feature se-

*Corresponding author

†Yang Li and Jia Zhou contributed equally to this work.

lection is challenging: the curse of high dimensionality of covariates can undermine the validity of classical inference procedures (Fan et al., 2019; Sur et al., 2020), which hold only under low-dimensional setups.

When the response depends on the explanatory variables through a structured model, such as a linear model, a series of statistical inference methods have been proposed to mitigate the curse of dimensionality and derive valid p-values for these explanatory variables. Among them, a line of research performs statistical inference after model selection with significance statements conditional on the selected model (Tibshirani et al., 2016; Tian and Taylor, 2018). Another complementary class of methods, which is more closely related to this work, proposes to de-bias the penalized estimators by inverting the Karush-Kuhn-Tucker condition of the corresponding optimization problem or, equivalently, by using low-dimensional projections (Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014), thereby establishing unconditional significance statements. The major advantage of such inference procedures is that they do not require the true signals to be uniformly strong, thereby accommodating general magnitudes of true regression coefficients.

There is no free lunch though: the asymptotic normality of the de-biased estimators comes at the price of a strict sparsity constraint: $s = o(\sqrt{n}/\log p)$, where n is the sample size, p is the dimension of the parameter of interest, and s denotes the number of nonzero coefficients. This is more restrictive than $s = o(n/\log p)$, which suffices for consistent estimation under L_2-loss. When the true regression coefficients are strictly sparse (L_0-sparse), the minimax results in Cai and Guo (2017) show that $s = o(\sqrt{n}/\log p)$ is necessary for the asymptotic normality of the de-biased estimator. Interestingly, if the row-wise maximum sparsity s_Φ of the precision matrix of the predictors satisfies $s_\Phi = o(\sqrt{n}/\log p)$, Javanmard and Montanari (2018) developed a novel leave-one-out

proof technique to prove that the constraint on s can be relaxed to a nearly optimal one: $s = o\{n/(\log p)^2\}$. Beyond these theoretical advances, Bellec and Zhang (2022) suggested de-biasing the lasso (Tibshirani, 1996) with degrees-of-freedom adjustment and showed that when $s_\Phi = o(\sqrt{n}/\log p)$, the adjusted de-biasing scheme allows $s = o(n/\log p)$. More recently, Shinkyu and Sueishi (2025) proposed a small tuning parameter selection procedure for the de-biased lasso, and proved asymptotic normality provided $s = o(\sqrt{n/\log p})$. For approximately sparse (capped L_1-sparse, see (2.2)) structures, it remains unknown if the sparsity constraint can be relaxed similarly.

Our work is mainly motivated by a key observation that despite embracing signals of varying strengths, the existing de-biasing inference methods (Zhang and Zhang, 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014) treat all the covariates equally, regardless of their signal strength. Then the estimation errors on all the signal coordinates can accumulate and jeopardize the asymptotic validity of the resulting estimators. Note that when the magnitude of a regression coefficient is of order $\Omega\{\sqrt{(\log p)/n}\}$, it can be identified by feature screening techniques (Fan and Lv, 2008; Li et al., 2012; Wang and Leng, 2016) or variable selection methods (Wainwright, 2009; Zhang, 2010; Zhang and Zhang, 2014) with high probability. Then a natural question arises: if those strong signals can be identified in the first place, is the number of them still limited to $o(\sqrt{n}/\log p)$ for valid inference?

In this paper, we intend to address the above question by developing and analyzing a new inference methodology that takes full advantage of the strong signals. We propose to first identify the strong signals and eliminate their impact on the subsequent inference. Specifically, we propose a new hybrid orthogonalization technique (HOT) that residualizes the features of interest such that they are strictly orthogonal to the strong features and approximately orthogonal to the remaining weak ones. The resulting HOT estimator

thus avoids the bias of estimating large coefficients in the original de-biasing step and substantially relaxes the requirement on the sparsity of the regression coefficient vector. To the best of our knowledge, such a hybrid inference procedure is new.

1.1 Contributions and organization

We summarize our main contributions as follows:

1. We construct the HOT estimator via exact orthogonalization against strong signals followed by relaxed orthogonalization against the remaining signals. Moreover, we establish the asymptotic normality of the HOT estimator, which allows the number of strong signals s_1 to be of order $o(n/s_\Phi)$. One crucial ingredient of our analysis is verifying that the sign-restricted cone invertibility factor of the residualized design matrix that is yielded by the exact orthogonalization step is well above zero.

2. We propose another equivalent formulation of the HOT estimator through a partially penalized least squares problem (see Proposition 1). Analyzing this form also leads to the asymptotic normality of the HOT estimator, but with a different requirement on s_1: $s_1 = o(n/\log p)$. This matches the sparsity requirement for consistent estimation in terms of the Euclidean norm. Combining this sparsity constraint with the first one gives a further relaxed constraint: $s_1 = o\{\max(n/\log p, n/s_\Phi)\}$.

3. We allow the maximum row-sparsity of the precision matrix to be as large as $o(n/\log p)$, which is more relaxed than $o(\sqrt{n}/\log p)$ imposed by Javanmard and Montanari (2018) and Bellec and Zhang (2022). Table 1 below summarizes the sparsity conditions required for the asymptotic normality by different methods.

4. We show that the HOT method allows the number of weak signals to be of order $o(\sqrt{n}/\log p)$, which is the same order as the sparsity constraint required by existing

de-biasing methods.

5. Last but not least, we show that the asymptotic efficiency of the HOT estimator is the same as that of the existing de-biasing methods, indicating no efficiency loss due to the hybrid orthogonalization.

Table 1: Sparsity conditions on the regression coefficient vector $\boldsymbol{\beta}$ and the maximum row-sparsity of the precision matrix $\boldsymbol{\Phi}$ required by different methods.

	Sparsity of $\boldsymbol{\beta}$	Sparsity of $\boldsymbol{\Phi}$
Cai and Guo (2017)	$s = o(\sqrt{n}/\log p)$	–
Javanmard and Montanari (2018)	$s = o\{n/(\log p)^2\}$	$s_\Phi = o(\sqrt{n}/\log p)$
Bellec and Zhang (2022)	$s = o(n/\log p)$	$s_\Phi = o(\sqrt{n}/\log p)$
Shinkyu and Sueishi (2025)	$s = o(\sqrt{n/\log p})$	–
HOT	$s_1 = o(\max\{n/\log p, n/s_\Phi\})$	$s_\Phi = o(n/\log p)$

The rest of this paper is organized as follows. Section 2 presents the motivation and the proposed HOT methodology. We establish comprehensive theoretical properties to ensure the validity of the HOT method in Section 3. Numerical studies are provided in Sections 4 and 5.

1.2 Notation

For $\mathbf{x} = (x_1, \ldots, x_p)^\top$, denote by $\|\mathbf{x}\|_q = (\sum_{j=1}^p |x_j|^q)^{1/q}$ the L_q-norm for $q \in (0, \infty)$. Given two sets S_1 and S_2, we use $S_1 \backslash S_2$ to denote $S_1 \cap S_2^c$. For any matrix $\mathbf{X}$, denote by $\mathbf{X}_{S_1}$ the submatrix of $\mathbf{X}$ with columns in S_1, and $\mathbf{X}_{S_1 S_2}$ the submatrix of $\mathbf{X}$ with rows in S_1 and columns in S_2. We write $\mathbf{X}_{-j}$ for the submatrix of $\mathbf{X}$ with columns indexed by the set $\{1, 2, \ldots, p\} \backslash \{j\}$. For a symmetric matrix $\mathbf{A}$, write $\rho(\mathbf{A}) = \max_i \mathbf{A}_{ii}$ for the maximum diagonal element of $\mathbf{A}$. For two series $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$

if there exists a universal constant C such that $a_n \leq Cb_n$ when n is sufficiently large. We write $a_n = \Omega(b_n)$ if there exists a universal constant c such that $a_n \geq cb_n$ when n is sufficiently large. We say $a_n \asymp b_n$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$, and $a_n \gg b_n$ if $\limsup_{n\to\infty} \frac{b_n}{a_n} = 0$.

2. Efficient inference via hybrid orthogonalization

2.1 Setup and motivation

Consider the following high-dimensional linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ is an n-dimensional vector of responses, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is an $n \times p$ design matrix with p covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional unknown regression coefficient vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is an n-dimensional random error vector. Our main goal is to conduct pointwise statistical inference for the components of the parameter vector $\boldsymbol{\beta}$. For example, we might want to construct confidence intervals for individual coefficients or test an individual null hypothesis $H_{0,j} : \beta_j = 0$ versus the alternative $H_{A,j} : \beta_j \neq 0$.

Under a typical high-dimensional sparse setup, the number of features p is allowed to grow exponentially fast with the sample size n, while the true regression coefficient vector $\boldsymbol{\beta}$ remains parsimonious. Given that the strict sparsity and uniform signal strength assumptions are often violated in practice, we instead adopt the weaker notion of the capped L_1 sparsity due to Zhang and Zhang (2014):

$$s_{L_1} := \sum_{j=1}^{p} \min\left\{|\beta_j|/(\sigma\lambda_0), 1\right\}, \tag{2.2}$$

where σ is the noise level, and $\lambda_0 = \sqrt{2\log(p)/n}$ matches the order of the maximum spurious correlation $\|n^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}\|_\infty$. Specifically, a signal gets fully counted towards the

sparsity allowance only when its order of magnitude exceeds $\sigma\lambda_0$, which allows it to stand out from spurious predictors and be identified through feature screening or selection (Fan and Lv, 2008; Wainwright, 2009; Zhang, 2010). The weaker signals instead make discounted contributions to s_{L_1}; the magnitude of each contribution depends on the corresponding signal strength. Therefore, depending on whether the signal strength exceeds $\sigma\lambda_0$, the sparsity parameter s_{L_1} can naturally be decomposed into two parts s_1 and s_2 that correspond to the strong and weak signals respectively.

To motivate our approach, we first introduce the low-dimensional projection estimator (LDPE) $\widehat{\boldsymbol{\beta}}^{\mathrm{L}} = (\widehat{\beta}_1^{\mathrm{L}}, \cdots, \widehat{\beta}_p^{\mathrm{L}})^\top$ proposed in Zhang and Zhang (2014). Consider an initial estimator $\widehat{\boldsymbol{\beta}}^{(\mathrm{init})} = (\widehat{\beta}_1^{(\mathrm{init})}, \ldots, \widehat{\beta}_p^{(\mathrm{init})})^\top$ obtained through the scaled lasso regression (Sun and Zhang, 2012) with tuning parameter λ_0:

$$(\widehat{\boldsymbol{\beta}}^{(\mathrm{init})}, \widehat{\sigma}) \in \underset{\boldsymbol{\beta},\sigma}{\arg\min} \left\{ (2\sigma n)^{-1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + 2^{-1}\sigma + \lambda_0 \|\boldsymbol{\beta}\|_1 \right\}. \tag{2.3}$$

Then for $1 \le j \le p$, define the LDPE as

$$\widehat{\beta}_j^{\mathrm{L}} := \widehat{\beta}_j^{(\mathrm{init})} + \mathbf{z}_j^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(\mathrm{init})}) / \mathbf{z}_j^\top \mathbf{x}_j,$$

where $\mathbf{z}_j$ is a relaxed orthogonalization of $\mathbf{x}_j$ against $\mathbf{X}_{-j}$ and can be generated as the residual of the corresponding lasso regression.

To see the asymptotic distribution of LDPE, note that

$$\widehat{\beta}_j^{\mathrm{L}} - \beta_j = \frac{\mathbf{z}_j^\top \boldsymbol{\varepsilon}}{\mathbf{z}_j^\top \mathbf{x}_j} + \sum_{k \ne j} \frac{\mathbf{z}_j^\top \mathbf{x}_k (\beta_k - \widehat{\beta}_k^{(\mathrm{init})})}{\mathbf{z}_j^\top \mathbf{x}_j}. \tag{2.4}$$

The first term on the right-hand side (RHS) dictates the asymptotic distribution of the LDPE and is of order $O_{\mathbb{P}}(n^{-1/2})$. The second term on the RHS is responsible for the bias of the LDPE. Define the bias factor $\eta_j := \max_{k \ne j} |\mathbf{z}_j^\top \mathbf{x}_k| / \|\mathbf{z}_j\|_2$ and the noise factor $\tau_j := \|\mathbf{z}_j\|_2 / |\mathbf{z}_j^\top \mathbf{x}_j|$. Then the bias term of the LDPE is bounded as

$$\sum_{k \ne j} \frac{|\mathbf{z}_j^\top \mathbf{x}_k (\beta_k - \widehat{\beta}_k^{(\mathrm{init})})|}{|\mathbf{z}_j^\top \mathbf{x}_j|} \le \left(\max_{k \ne j} \frac{|\mathbf{z}_j^\top \mathbf{x}_k|}{\|\mathbf{z}_j\|_2} \right) \frac{\|\mathbf{z}_j\|_2}{|\mathbf{z}_j^\top \mathbf{x}_j|} \left\| \widehat{\boldsymbol{\beta}}^{(\mathrm{init})} - \boldsymbol{\beta} \right\|_1 = \eta_j \tau_j \|\widehat{\boldsymbol{\beta}}^{(\mathrm{init})} - \boldsymbol{\beta}\|_1. \tag{2.5}$$

When the design is Gaussian, $\eta_j \asymp \sqrt{\log p}$ and $\tau_j \asymp n^{-1/2}$ under typical realizations of $\mathbf{X}$.

Therefore, in order to achieve asymptotic normality of $\widehat{\beta}_j^{\mathrm{L}}$, we require the bias term in (2.5) to be $o_{\mathbb{P}}(n^{-1/2})$, which entails that

$$\eta_j \|\widehat{\boldsymbol{\beta}}^{(\mathrm{init})} - \boldsymbol{\beta}\|_1 = o_{\mathbb{P}}(1).$$

Under mild regularity conditions, Zhang and Zhang (2014) showed that $\|\widehat{\boldsymbol{\beta}}^{(\mathrm{init})} - \boldsymbol{\beta}\|_1 = O_{\mathbb{P}}(s_{L_1}\sqrt{\log(p)/n})$. Combining this with the aforementioned order of η_j suggests that one requires

$$s_{L_1} = o(\sqrt{n}/\log p) \tag{2.6}$$

to achieve the asymptotic normality of $\widehat{\beta}_j^{\mathrm{L}}$. A similar sparsity constraint is required by other de-biasing methods such as Javanmard and Montanari (2014) and van de Geer et al. (2014), implying that valid inference via the de-biased estimators typically requires the true model to be sparser than is required for L_2-consistent estimation.

The advantage of this inference procedure is that it applies to general magnitudes of the true regression coefficients β_j under sparse or approximately sparse structures as long as the aforementioned accuracy of the initial estimate holds. The downside, however, is that it requires the stringent sparsity assumption (2.6). Looking back at (2.4) and the definition of $\mathbf{z}_j$, we find that the construction of $\mathbf{z}_j$ in LDPE does not discriminate between strong and weak signals. Then even with an initial estimate to offset the impacts of strong signals in the bias term, the estimation errors of the s_1 large coefficients can aggregate to be of order $s_1\sqrt{\log(p)/n}$, which gives rise to a harsh constraint on the number of strong predictors.

To address this issue, we propose to construct the vector $\mathbf{z}_j$ by a hybrid orthogonalization technique after identifying large coefficients by feature screening or selection procedures. Intuitively, an ideal hybrid orthogonalization vector $\mathbf{z}_j$ should be orthogonal to the strong predictors so that their impact can be completely removed in the bias term

even without resorting to any initial estimate. At the same time, it should be a relaxed orthogonalization against the remaining covariate vectors because some weak signals may have magnitudes above $n^{-1/2}$ and invalidate the asymptotic normality.

2.2 Unveiling the mystery of the hybrid orthogonalization

The first step of hybrid orthogonalization is residualizing the target predictor and the weak predictors using the strong predictors through exact orthogonalization. After obtaining those residualized covariate vectors, we then construct the relaxed orthogonalization vector via penalized regression. Specifically, given a pre-screened set $\widehat{S}$ of strong signals and any target index $1 \leq j \leq p$, the HOT vector $\mathbf{z}_j$ can be constructed in two steps:

Step 1. Exact orthogonalization. For $k \in \{j\} \cup \widehat{S}^c$, we take $\boldsymbol{\psi}_k^{(j)}$ as the exact orthogonalization of $\mathbf{x}_k$ against $\mathbf{X}_{\widehat{S}\setminus\{j\}}$. That is,

$$\boldsymbol{\psi}_k^{(j)} = (\mathbf{I}_n - \mathbf{P}_{\widehat{S}\setminus\{j\}})\mathbf{x}_k, \tag{2.7}$$

where $\mathbf{P}_A = \mathbf{X}_A(\mathbf{X}_A^\top \mathbf{X}_A)^{-1}\mathbf{X}_A^\top$ is the projection matrix of the column space of $\mathbf{X}_A$. Therefore, for any $k \in \{j\} \cup \widehat{S}^c$, the projected vector $\boldsymbol{\psi}_k^{(j)}$ is orthogonal to $\mathbf{X}_{\widehat{S}\setminus\{j\}}$. The submatrix of the strong covariates $\mathbf{X}_{\widehat{S}\setminus\{j\}}$ will then be discarded in the construction of $\mathbf{z}_j$ after this step. It is worth noticing that when $j \in \widehat{S}^c$, for any $k \in \widehat{S}^c$, $\boldsymbol{\psi}_k^{(j)} = (\mathbf{I}_n - \mathbf{P}_{\widehat{S}})\mathbf{x}_k$, which is the residual vector of projecting $\mathbf{x}_k$ onto the space spanned by all covariate vectors in $\widehat{S}$. However, $\boldsymbol{\psi}_k^{(j)}$ varies across $j \in \widehat{S}$, since the corresponding projection space only includes the other covariate vectors in $\widehat{S}$.

Note that the calculation of $\boldsymbol{\psi}_k^{(j)}$ entails the requirement $|\widehat{S}| < n$. In practice, feature screening procedures such as sure independence screening (SIS) (Fan and Lv, 2008) and high-dimensional ordinary least-squares projection (HOLP) (Wang and Leng, 2016) naturally yield such a pre-screened set by choosing a submodel with size less than n. More

theoretical details about $|\widehat{S}|$ will be discussed later in Definition 1.

Step 2. Relaxed orthogonalization. The hybrid orthogonalization vector $\mathbf{z}_j$ is constructed as the residual of the following lasso regression based on $\{\psi_k^{(j)}\}_{k\in\{j\}\cup\widehat{S}^c}$:

$$\begin{aligned}\mathbf{z}_j &= \boldsymbol{\psi}_j^{(j)} - \boldsymbol{\psi}_{\widehat{S}^c\setminus\{j\}}^{(j)}\widehat{\boldsymbol{\omega}}_j; \\ \widehat{\boldsymbol{\omega}}_j &= \underset{\mathbf{b}_{\widehat{S}^c\setminus\{j\}}}{\arg\min}\Big\{(2n)^{-1}\big\|\boldsymbol{\psi}_j^{(j)} - \boldsymbol{\psi}_{\widehat{S}^c\setminus\{j\}}^{(j)}\mathbf{b}_{\widehat{S}^c\setminus\{j\}}\big\|_2^2 + \lambda_j \sum_{k\in\widehat{S}^c\setminus\{j\}} v_{jk}|b_k|\Big\},\end{aligned} \tag{2.8}$$

where $\boldsymbol{\psi}_{\widehat{S}^c\setminus\{j\}}^{(j)}$ is the matrix consisting of columns $\boldsymbol{\psi}_k^{(j)}$ for $k\in\widehat{S}^c\setminus\{j\}$, $v_{jk} = \|\boldsymbol{\psi}_k^{(j)}\|_2/\sqrt{n}$, b_k denotes the kth component of $\mathbf{b}\in\mathbb{R}^p$ and λ_j the regularization parameter. It is clear that $\mathbf{z}_j$ is also orthogonal to the matrix $\mathbf{X}_{\widehat{S}\setminus\{j\}}$ of strong predictors since it is a linear combination of the projected vectors $\boldsymbol{\psi}_k^{(j)}$, $k\in\{j\}\cup\widehat{S}^c$. Moreover, together with (2.7), we have for any $k\in\{j\}\cup\widehat{S}^c$ that

$$\mathbf{z}_j^\top\boldsymbol{\psi}_k^{(j)} = \mathbf{z}_j^\top(\mathbf{I}_n - \mathbf{P}_{\widehat{S}\setminus\{j\}})\mathbf{x}_k = \mathbf{z}_j^\top\mathbf{x}_k. \tag{2.9}$$

Therefore, the resulting bias factor η_j^{H} is the same as η_j in Zhang and Zhang (2014), given that

$$\eta_j^{\mathrm{H}} = \max_{k\in\widehat{S}^c\setminus\{j\}}|\mathbf{z}_j^\top\boldsymbol{\psi}_k^{(j)}|/\|\mathbf{z}_j\|_2 = \max_{k\in\widehat{S}^c\setminus\{j\}}|\mathbf{z}_j^\top\mathbf{x}_k|/\|\mathbf{z}_j\|_2 = \max_{k\neq j}|\mathbf{z}_j^\top\mathbf{x}_k|/\|\mathbf{z}_j\|_2 = \eta_j.$$

In the sequel, we drop the superscript $^{\mathrm{H}}$ from η_j^{H} to simplify notation. Through this two-step hybrid orthogonalization, $\mathbf{z}_j$ is a strict orthogonalization against the strong predictors but a relaxed orthogonalization against the others.

With $(\mathbf{z}_j)_{j\in[p]}$, the HOT estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1,\ldots,\widehat{\beta}_p)^\top$ is defined coordinate-wise as

$$\widehat{\beta}_j = \frac{\mathbf{z}_j^\top\mathbf{y}}{\mathbf{z}_j^\top\mathbf{x}_j}. \tag{2.10}$$

Plugging model (2.1) into the definition above, we obtain that

$$\widehat{\beta}_j - \beta_j = \frac{\mathbf{z}_j^\top\boldsymbol{\varepsilon}}{\mathbf{z}_j^\top\mathbf{x}_j} + \sum_{k\neq j}\frac{\mathbf{z}_j^\top\mathbf{x}_k\beta_k}{\mathbf{z}_j^\top\mathbf{x}_j} = \frac{\mathbf{z}_j^\top\boldsymbol{\varepsilon}}{\mathbf{z}_j^\top\mathbf{x}_j} + \sum_{k\in\widehat{S}^c\setminus\{j\}}\frac{\mathbf{z}_j^\top\mathbf{x}_k\beta_k}{\mathbf{z}_j^\top\mathbf{x}_j}, \tag{2.11}$$

where the last equality is due to the exact orthogonalization of $\mathbf{z}_j$ to $\mathbf{X}_{\widehat{S}\setminus\{j\}}$. Furthermore, since the coefficients of order $\Omega(\sqrt{\log(p)/n})$ are typically guaranteed to be found and included in $\widehat{S}$ by the aforementioned feature screening or selection procedures under suitable conditions, the signals corresponding to the set $\widehat{S}^c\setminus\{j\}$ are weak with an aggregated magnitude no larger than the order of $s_2\sqrt{\log(p)/n}$. Thus, by taking the relaxed projection in *Step 2*, the proposed method inherits the advantage of LDPE in dealing with the weak signals. We will show that the main sparsity constraint for HOT here is that $s_2\log(p)/\sqrt{n} = o(1)$, while the allowable growth rate of s_1 is $o(n/\log p)$.

Algorithm 1: HOT

Require: $\mathbf{X}$, $\mathbf{y}$, a pre-screened set $\widehat{S}$, and the significance level α

for $j \in \{1, 2, \ldots, p\}$

  Solve the lasso regression (2.8) to obtain the estimate $\widehat{\boldsymbol{\omega}}_j$

  Compute the residual of the lasso regression (2.8) to obtain $\mathbf{z}_j$

  Compute the HOT estimator $\widehat{\beta}_j \leftarrow \frac{\mathbf{z}_j^\top \mathbf{y}}{\mathbf{z}_j^\top \mathbf{x}_j}$

  Compute the noise factor $\tau_j \leftarrow \|\mathbf{z}_j\|_2/|\mathbf{z}_j^\top \mathbf{x}_j|$

  Construct the confidence interval

$$\mathrm{CI}_j \leftarrow [\widehat{\beta}_j - \Phi^{-1}(1-\alpha/2)\widehat{\sigma}\tau_j, \widehat{\beta}_j + \Phi^{-1}(1-\alpha/2)\widehat{\sigma}\tau_j]$$

end for

Output: $\{\mathrm{CI}_j\}_{j=1}^p$ for $(1-\alpha)$-confidence intervals of $\{\beta_j\}_{j\in[p]}$

The HOT inference procedure is described in Algorithm 1, where $\widehat{\sigma}$ is a consistent estimator of σ. It is interesting to note that no initial estimate of $\boldsymbol{\beta}$ is involved in HOT. As we shall see, HOT has two main advantages over standard de-biasing techniques. First of all, HOT completely removes the impact of strong signals in equation (2.11), thereby eliminating the bias induced by the estimation errors on large coefficients. Second,

HOT accommodates at least $o(n/\log p)$ strong signals; in contrast, existing works require $s_1 = o(\sqrt{n}/\log p)$.

2.3 An equivalent form of the hybrid orthogonalization

It turns out that the aforementioned two-step hybrid orthogonalization can be formulated as a partially penalized regression procedure. The following proposition elucidates this alternative formulation.

Proposition 1. *For any j, $1 \le j \le p$, the hybrid orthogonalization residual $\mathbf{z}_j$ defined in (2.8) satisfies*

$$\mathbf{z}_j = \mathbf{x}_j - \mathbf{X}_{-j}\widehat{\boldsymbol{\pi}}_{-j}^{(j)},$$

where $\widehat{\boldsymbol{\pi}}_{-j}^{(j)}$ is the minimizer of the following partially penalized regression problem:

$$\widehat{\boldsymbol{\pi}}_{-j}^{(j)} = \underset{\mathbf{b}_{-j}}{\arg\min}\left\{(2n)^{-1}\left\|\mathbf{x}_j - \mathbf{X}_{-j}\mathbf{b}_{-j}\right\|_2^2 + \lambda_j \sum_{k\in\widehat{S}^c\setminus\{j\}} v_{jk}|b_k|\right\}.$$

Here λ_j and v_{jk} are the same as those in problem (2.8).

Proposition 1 shows that $\mathbf{z}_j$ can be viewed as the residual vector from a partially penalized regression problem. In the context of the goodness-of-fit test for high-dimensional logistic regression, Janková et al. (2020) exploited a similar partial penalty to better control the type I error of the test in its empirical study. Janková et al. (2020) argued that this technique helps with controlling the remainder term of the asymptotic expansion of the test statistic under a "beta-min condition" that all the nonzero coefficients are at least of order $\sqrt{s\log p/n}$ in absolute value. The major difference between our work and Janková et al. (2020) lies in the setups and goals: our motivation for imposing a partial penalty is to accommodate larger sparsity of the true model in inference. Besides, an important methodological difference is that we do not residualize our predictors against

the target predictor $\mathbf{x}_j$ in the exact orthogonalization step if $\mathbf{x}_j$ is among the strong predictors.

3. Theoretical properties

This section presents the theoretical guarantees of the proposed HOT estimator. Exploiting the capped L_1 sparsity condition in (2.2), we first characterize strong and weak signals using mathematical language. Specifically, we define S_1 as follows to collect all the strong signals:

$$S_1 = \{1 \le i \le p : |\beta_i| \ge \sigma\lambda_0\}. \tag{3.12}$$

Clearly, the weak signals are included in S_1^c.

3.1 Theoretical results for fixed design

We establish the asymptotic properties of the proposed estimator under fixed design $\mathbf{X}$.

Theorem 1. *Given a fixed set $\widehat{S} \supset S_1$, the HOT estimator $\widehat{\beta}_j$ satisfies*

$$\widehat{\beta}_j - \beta_j = W_j + \Delta_j,$$
$$W_j = \frac{\mathbf{z}_j^\top \boldsymbol{\varepsilon}}{\mathbf{z}_j^\top \mathbf{x}_j}, \quad |\Delta_j| \le \left(\max_{k \in \widehat{S}^c \setminus \{j\}} \left| \frac{\sqrt{n}\lambda_j \|\boldsymbol{\psi}_k^{(j)}\|_2}{\mathbf{z}_j^\top \mathbf{x}_j} \right| \right) \sum_{k \in S_1^c} |\beta_k|.$$

Suppose that the components of $\boldsymbol{\varepsilon}$ are independent and identically distributed with mean 0 and variance σ^2. If there is some $\varsigma > 0$ such that the inequality

$$\lim_{n\to\infty} \frac{1}{(\sigma\|\mathbf{z}_j\|_2)^{2+\varsigma}} \sum_{k=1}^{n} \mathbb{E}|z_{jk}\varepsilon_k|^{2+\varsigma} = 0 \tag{3.13}$$

holds, we further have

$$\lim_{n\to\infty} \mathbb{P}(W_j \le \tau_j \sigma x) = \Phi(x), \quad \forall x \in \mathbb{R},$$

where $\tau_j := \|\mathbf{z}_j\|_2 / |\mathbf{z}_j^\top \mathbf{x}_j|$, and where $\Phi(\cdot)$ is the distribution function of the standard normal distribution.

Theorem 1 presents a preliminary theoretical result of the proposed method. The first part of this theorem shows that the error of the HOT estimator $\widehat{\beta}_j$ can be decomposed into a noise term W_j and a bias term Δ_j, which is consistent with the results in van de Geer et al. (2014); Zhang and Zhang (2014). However, one interesting difference from the standard de-biasing methods here is that Δ_j is non-random, because we do not need the random initial estimator $\widehat{\boldsymbol{\beta}}^{(\text{init})}$ as shown in (2.10) and (2.11). The proposed HOT estimator can be regarded as a modified least squares estimator for high-dimensional settings, which distinguishes it from these de-biasing methods and explains the above difference. The second part of this theorem establishes the asymptotic normality of the noise term W_j based on the classical Lyapunov condition (3.13), which relaxes the Gaussian assumption in Zhang and Zhang (2014).

The conclusions in Theorem 1 are open-ended, given that neither the magnitude of $|\Delta_j|$ nor that of τ_j is specified, and that the maximum number of strong signals is not explicitly derived. In the sequel, we will solidify Theorem 1 by incorporating some random-design assumptions.

3.2 Theoretical results for random design

We provide here the theoretical results for random design $\mathbf{X}$. Before introducing the main results, we list a few technical conditions that underpin the asymptotic properties of the HOT estimator.

Condition 1. $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_n \otimes \boldsymbol{\Sigma})$, *where the eigenvalues of* $\boldsymbol{\Phi} = \boldsymbol{\Sigma}^{-1} = (\phi_{ij})_{p\times p}$ *are bounded within the interval* $[1/L, L]$ *for some constant* $L \geq 1$.

Condition 2. $s_\Phi(\log p)/n = o(1)$, *where* $s_\Phi = \max_{1\leq j\leq p} s_j^*$ *with* $s_j^* = |\{1 \leq k \leq p : k \neq j, \phi_{jk} \neq 0\}|$.

Condition 3. $s_1 = |S_1| = o\{\max(n/\log p, n/s_\Phi)\}$, *and* $\|\boldsymbol{\beta}_{S_1^c}\|_1 = o(1/\sqrt{\log p})$.

Condition 1 assumes the design matrix has independent Gaussian rows, with bounded precision matrix eigenvalues. The Gaussianity simplifies theoretical analysis, particularly for verifying identifiability conditions like the sign-restricted cone invertibility factor (Ye and Zhang, 2010), though our results extend to sub-Gaussian settings. Condition 2 restricts the sparsity of the precision matrix, ensuring relaxed projection accuracy—a common requirement in the de-biasing literature (Javanmard and Montanari, 2018; Bellec and Zhang, 2022). Notably, our sparsity constraint is weaker than the condition $s_\Phi = o(\sqrt{n}/\log p)$ imposed by both Javanmard and Montanari (2018) and Bellec and Zhang (2022).

The most appealing properties of the HOT inference are reflected in Condition 3, especially its first part. To the best of our knowledge, it is weaker than the sparsity constraints in existing high-dimensional works on statistical inference or even on estimation since it holds when either $s_1 = o(n/\log p)$ or $s_1 = o(n/s_\Phi)$. Specifically, it allows the number of strong signals to satisfy $s_1 = o(\max\{n/\log p, n/s_\Phi\})$, where $o(n/\log p)$ is a fundamental constraint on the sparsity level to guarantee consistent estimation in high dimensions. Therefore, the bound $o(n/\log p)$ is weaker than $o(\sqrt{n}/\log p)$ imposed in the aforementioned de-biasing methods or $o(n/(\log p)^2)$ in Javanmard and Montanari (2018). Remarkably, the proposed inference method even allows for $s_1 = o(n/s_\Phi)$, which can be weaker than $o(n/\log p)$ in Bellec and Zhang (2022) when $s_\Phi \ll \log p$. The second part of Condition 3 implies that the sparsity level s_2 of the weak signals can be bounded as

$$s_2 = \sum_{j \in S_1^c} \min\left\{|\beta_j|/(\sigma\lambda_0), 1\right\} \asymp \|\boldsymbol{\beta}_{S_1^c}\|_1/(\sigma\lambda_0) = o\left(\frac{1}{\sqrt{\log p}} \frac{\sqrt{n}}{\sqrt{\log p}}\right) = o(\sqrt{n}/\log p),$$

where we recall $\lambda_0 = \sqrt{2\log p/n}$. In essence, we attribute these relaxations to the presence of the acceptable set $\widehat{S}$, of which a detailed discussion is provided later.

We first analyze the statistical properties of the relaxed projection in (2.8), which is essentially a lasso regression on the residualized features $\boldsymbol{\psi}^{(j)}_{\widehat{S}^c\setminus\{j\}}$. We start by investigating

the sign-restricted cone invertibility factor (Ye and Zhang, 2010) of $\boldsymbol{\psi}^{(j)}_{\widehat{S}^c\setminus\{j\}}$, an important quantity that determines the error of the lasso estimator. Let $S_{j*} = \{1 \le k \le p : k \in \widehat{S}^c\setminus\{j\}, \phi_{jk} \neq 0\}$ and $S^c_{j*} = \widehat{S}^c\setminus(\{j\} \cup S_{j*})$. For some constant $\xi \ge 0$, the sign-restricted cone invertibility factor of the design matrix $\boldsymbol{\psi}^{(j)}_{\widehat{S}^c\setminus\{j\}}$ is defined as

$$F_2^{(j)}(\xi, S_{j*}) := \inf\left\{|S_{j*}|^{1/2}\left\|n^{-1}(\boldsymbol{\psi}^{(j)}_{\widehat{S}^c\setminus\{j\}})^\top\boldsymbol{\psi}^{(j)}_{\widehat{S}^c\setminus\{j\}}\mathbf{u}\right\|_\infty/\|\mathbf{u}\|_2 : \mathbf{u} \in \mathcal{G}_-^{(j)}(\xi, S_{j*})\right\},$$

where the sign-restricted cone

$$\mathcal{G}_-^{(j)}(\xi, S_{j*}) = \left\{\mathbf{u} : \|\mathbf{u}_{S^c_{j*}}\|_1 \le \xi\|\mathbf{u}_{S_{j*}}\|_1,\ u_k(\boldsymbol{\psi}^{(j)}_k)^\top\boldsymbol{\psi}^{(j)}_{\widehat{S}^c\setminus\{j\}}\mathbf{u}/n \le 0,\ \forall\ k \in S^c_{j*}\right\}.$$

Unlike the compatibility factor or restricted eigenvalue condition (Bickel et al., 2009), the sign-restricted cone invertibility factor (SCIF) is more flexible. The following proposition shows that the SCIF of $\boldsymbol{\psi}^{(j)}_{\widehat{S}^c\setminus\{j\}}$ is bounded away from zero with high probability, ensuring support identifiability by preventing collinearity within the sign-restricted cone.

Proposition 2. *When Conditions 1 and 2 hold and $d = |\widehat{S}| = o(n/s_\Phi)$, for any fixed constant $\delta > 1$, any j, $1 \le j \le p$, and sufficiently large n, with probability at least $1 - o(p^{1-\delta})$, there exists a positive constant c^* such that*

$$F_2^{(j)}(\xi, S_{j*}) \ge c^*.$$

Note that the conclusion of Proposition 2 is typically assumed directly in high dimensions for design matrices with independent and identically distributed (i.i.d.) rows. However, the design matrix $\boldsymbol{\psi}^{(j)}_{\widehat{S}^c\setminus\{j\}}$ here consists of projected vectors, so that the rows of $\boldsymbol{\psi}^{(j)}_{\widehat{S}^c\setminus\{j\}} = (\mathbf{I}_n - \mathbf{P}_{\widehat{S}\setminus\{j\}})\mathbf{X}_{\widehat{S}^c\setminus\{j\}}$ are neither independent nor identically distributed. We overcome these difficulties with a new technical argument that is tailored for the L_1-constrained structure of the cone. The upper bound $d = o(n/s_\Phi)$ is required to show that the difference between $\boldsymbol{\psi}^{(j)}_{\widehat{S}^c\setminus\{j\}}$ and its population projected counterpart vanishes asymptotically.

Proposition 2 lays the foundation for the following theorem on the statistical error of the relaxed projection.

Theorem 2. *Let* $\lambda_j = (1+\epsilon)\sqrt{2\delta\log(p)/(n\phi_{jj})}$ *for some constants* $\epsilon > 0$ *and* $\delta > 1$ *and any* $1 \le j \le p$. *Given a set* $\widehat{S}$, *when Conditions 1 and 2 hold and* $d = |\widehat{S}| = o(n/s_\Phi)$, *we have for any* $1 \le j \le p$ *and sufficiently large* n *that with probability at least* $1 - o(p^{1-\delta})$,

$$\|\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j\|_q = O(s_\Phi^{1/q}\lambda_j) = O\{s_\Phi^{1/q}\sqrt{\log(p)/n}\}, \quad q \in [1,2];$$
$$n^{-1/2}\|\boldsymbol{\psi}^{(j)}_{\widehat{S}^c\setminus\{j\}}(\widehat{\boldsymbol{\omega}}_j - \boldsymbol{\omega}_j)\|_2 = O(s_\Phi^{1/2}\lambda_j) = O\{\sqrt{s_\Phi\log(p)/n}\},$$

where $\boldsymbol{\omega}_j = -\boldsymbol{\Phi}^\top_{j,\widehat{S}^c\setminus\{j\}}/\phi_{jj}$ *is the population counterpart of* $\widehat{\boldsymbol{\omega}}_j$.

Theorem 2 establishes estimation and prediction error bounds for the relaxed projection in (2.8). One striking feature here is that the statistical rates no longer depend on the size of the penalization-free set $\widehat{S}$: the strong predictors have been eliminated in the exact orthogonalization step and are not involved in the relaxed projection. Next, we proceed to analyze the theoretical properties of the partially penalized regression.

Theorem 3. *Let* $\lambda_j = (1+\epsilon)\sqrt{2\delta\log(p)/(n\phi_{jj})}$ *for some constants* $\epsilon > 0$ *and* $\delta > 1$, $1 \le j \le p$. *Given any pre-screened set* $\widehat{S}$, *when Conditions 1 and 2 hold and* $d = |\widehat{S}| = o\{n/\log(p)\}$, *for any* $1 \le j \le p$ *and sufficiently large* n, *we have with probability at least* $1 - o(p^{1-\delta})$ *that*

$$\|\widehat{\boldsymbol{\pi}}^{(j)}_{-j} - \boldsymbol{\pi}^{(j)}_{-j}\|_q = O\{(s_\Phi + d)^{1/q}\lambda_j\} = O\{(s_\Phi+d)^{1/q}\sqrt{\log(p)/n}\}, \forall q \in [1,2];$$
$$n^{-1/2}\|\mathbf{X}_{-j}(\widehat{\boldsymbol{\pi}}^{(j)}_{-j} - \boldsymbol{\pi}^{(j)}_{-j})\|_2 = O\{(s_\Phi+d)^{1/2}\lambda_j\} = O\{\sqrt{(s_\Phi+d)\log(p)/n}\},$$

where $\boldsymbol{\pi}^{(j)}_{-j} = -\boldsymbol{\Phi}^\top_{j,-j}/\phi_{jj}$ *is the population counterpart of* $\widehat{\boldsymbol{\pi}}^{(j)}_{-j}$.

Theorem 3 establishes estimation and prediction error bounds for the solution of the partially penalized regression problem. It guarantees estimation and prediction consistency of $\widehat{\boldsymbol{\pi}}^{(j)}_{-j}$ when the size of the penalization-free set $\widehat{S}$ is of order $o\{n/\log(p)\}$. Such

an $\widehat{S}$ can typically be obtained through regularization methods, e.g., Bickel et al. (2009), Fan and Lv (2014), and Kong et al. (2016).

In fact, both Theorems 2 and 3 lead to the same guarantee for $\mathbf{z}_j$:

$$\mathbb{P}\left\{\|\mathbf{z}_j\|_2 \asymp n^{1/2}\right\} \geq 1 - o(p^{1-\delta});$$

see the proof of Theorem 4 for details. Perhaps more importantly, the two theorems complement each other in terms of the requirement on the size of $\widehat{S}$, so that we allow $|\widehat{S}| = o\{\max(n/\log p, n/s_\Phi)\}$. Therefore, if $\log(p) \gg s_\Phi$, we follow Theorem 2 to analyze $\mathbf{z}_j$; otherwise we follow Theorem 3.

Before presenting the main theorem on the HOT method, the following definition characterizes the properties that a pre-screened set $\widehat{S}$ should satisfy.

Definition 1 (Acceptable set)**.** The set $\widehat{S}$ is called an acceptable pre-screened set if it satisfies: (1) $\widehat{S}$ is independent of the data $(\mathbf{X}, \mathbf{y})$; (2) with probability at least $1 - \theta_{n,p}$ for some asymptotically vanishing $\theta_{n,p}$, $S_1 \subset \widehat{S}$ and $d = |\widehat{S}| = O(s_1)$.

The first independence property in Definition 1 is imposed so that the randomness of $\widehat{S}$ does not interfere with the subsequent inference procedure. The same property was suggested in Fan et al. (2013, 2015) to facilitate the theoretical analysis. Our numerical studies show, however, that it can be more of a technical assumption than a practical requisite. In practice, a standard and effective strategy to achieve such independence is the use of sample splitting, which has been widely used in existing work on statistical inference, such as Zhou et al. (2023), Chen and Zhang (2025), and Vazquez and Nan (2025). In our framework, a natural idea is to use a simple uniform sample splitting strategy to divide the available data into two equal parts: one part is used for the estimation of $\widehat{S}$, and the other part is used for the subsequent inference procedure. Clearly, although the sample size for inference is reduced from n to $n/2$, the established large-

sample theoretical results remain valid. Furthermore, we would like to mention that a single split may suffer from a loss of efficiency because only part of the data is used for inference. Recall that in the context of inference for high-dimensional generalized linear models, Vazquez and Nan (2025) proposed a multiple splitting procedure to mitigate such efficiency loss. Adapting a similar multiple splitting strategy would be useful. For more details, we refer to related sections in Vazquez and Nan (2025).

The second property is the sure screening property for the strong predictors, which assumes the size of the pre-screened set $\widehat{S}$ to be of the same order as s_1, the number of the true strong predictors. In practice, feature screening procedures such as SIS, distance correlation learning (Li et al., 2012), and HOLP can also provide such an acceptable set under certain conditions. However, the above methods may require additional restrictions on the parameter spaces of $(\boldsymbol{\beta}, \boldsymbol{\Phi})$. These interesting topics are closely related to our paper and will be left for future research. In fact, even if this assumption is violated, our theoretical results still hold as long as $d = o(\max\{n/\log p, n/s_\Phi\})$ in view of the technical arguments.

Now we are ready to present our main theorem that establishes the asymptotic normality of the HOT estimator.

Theorem 4. *Let* $\lambda_j = (1+\epsilon)\sqrt{2\delta \log(p)/(n\phi_{jj})}$ *for some constants* $\epsilon > 0$ *and* $\delta > 1$, $1 \le j \le p$. *Assume that Conditions 1-3 hold and that* $\widehat{S}$ *satisfies Definition 1. Then for any* $1 \le j \le p$ *and sufficiently large* n, *the HOT estimator* $\widehat{\beta}_j$ *satisfies*

$$\widehat{\beta}_j - \beta_j = W_j + \Delta_j,$$

where $W_j = \frac{\mathbf{z}_j^\top \boldsymbol{\varepsilon}}{\mathbf{z}_j^\top \mathbf{x}_j}$ *and* $|\Delta_j| = o_{\mathbb{P}}(n^{-1/2})$. *Moreover,*

$$n^{1/2}\tau_j \xrightarrow{\mathbb{P}} \phi_{jj}^{1/2}.$$

Conditional on $\mathbf{X}$, *suppose that the components of* $\boldsymbol{\varepsilon}$ *are independent and identically*

distributed with mean 0 and variance σ^2. *For any given* $\mathbf{X}$, *if there is some* $\varsigma > 0$ *such that the inequality*

$$\lim_{n\to\infty} \frac{1}{(\sigma\|\mathbf{z}_j\|_2)^{2+\varsigma}} \sum_{k=1}^{n} \mathbb{E}|z_{jk}\varepsilon_k|^{2+\varsigma} = 0$$

holds, we further have

$$\lim_{n\to\infty} \mathbb{P}(W_j \le \tau_j \sigma x) = \Phi(x), \quad \forall x \in \mathbb{R},$$

where $\Phi(\cdot)$ *is the distribution function of the standard normal distribution.*

Theorem 4 extends the results of Theorem 1 to the random design. Note that for any given $\mathbf{X}$, the asymptotic normality of the noise term W_j is based on a classical Lyapunov condition, a moment-based assumption that does not require unconditional independence between the errors and the predictors. In fact, the asymptotic normality in Theorem 4 is essentially established for $W_j|\mathbf{X}$, which is consistent with the result in van de Geer et al. (2014) and is precisely the key reason why the unconditional independence assumption is unnecessary in our paper.

For approximately sparse structures, Theorem 4 guarantees the asymptotic normality of the HOT estimator under a much more relaxed constraint on the number of strong signals. It is worth pointing out that $n^{1/2}\tau_j$ converges to the same constant as that of LDPE in Zhang and Zhang (2014). Given that the noise factor determines the asymptotic variance, Theorem 4 implies no efficiency loss incurred by HOT. Similarly to Javanmard and Montanari (2014) and van de Geer et al. (2014), when the noise standard deviation σ is unknown in practice, we can replace it with some consistent estimate $\widehat{\sigma}$ given by the scaled lasso for general Gaussian settings.

Finally, we state the inferential implications of the asymptotic normality results in Theorem 4. Conditional on $\mathbf{X}$, the asymptotic pointwise $(1-\alpha)$-confidence interval for

β_j is immediately given by

$$[\widehat{\beta}_j - \Phi^{-1}(1-\alpha/2)\widehat{\sigma}\tau_j, \widehat{\beta}_j + \Phi^{-1}(1-\alpha/2)\widehat{\sigma}\tau_j].$$

This result can be used to test the individual null hypothesis $H_{0,j} : \beta_j = 0$ versus the alternative $H_{A,j} : \beta_j \neq 0$ by applying the simple decision rule: reject $H_{0,j}$ if 0 is outside the confidence interval; otherwise, fail to reject $H_{0,j}$. Nevertheless, we would like to emphasize that, as in Zhang and Zhang (2014), this paper focuses primarily on the results of asymptotic confidence intervals. Therefore, the analysis concerning the significance level and statistical power of the aforementioned test is left for future work.

4. Simulation studies

In this section, we use simulated data to investigate the finite-sample performance of HOT in comparison with LDPE (Zhang and Zhang, 2014) and two other LDPE-type methods: the degrees-of-freedom adjusted LDPE (adLDPE) (Bellec and Zhang, 2022) and the small tuning parameter selector (STPS) (Shinkyu and Sueishi, 2025). We would first like to point out that the theoretical computational cost of HOT is on par with that of LDPE, adLDPE, and STPS. Specifically, the proposed HOT procedure comprises three main components: the pre-screening step, the exact orthogonalization step, and p lasso regressions. In high-dimensional settings, the p lasso calculations account for the majority of the computational cost. Note that most of the computational cost of LDPE, adLDPE, and STPS stems from p lasso calculations. Therefore, HOT exhibits a theoretical computational cost on par with these methods. Nevertheless, LDPE-type methods involve delicate parameter tuning (as described in Table 2 of Zhang and Zhang (2014)) and are typically more time-consuming in practical scenarios.

We consider two versions of HOT, denoted by H-SIS and H-HOLP, which correspond to using SIS and HOLP, respectively, to pre-screen the set $\widehat{S}$ of large coefficients. To con-

trol the number of selected identifiable coefficients in $\widehat{S}$, we first obtain the least squares estimates after restricting the covariates to $\widehat{S}$, and then apply BIC to select the optimal model. Both methods use the same data set for pre-screening and the subsequent inference procedure. For comparison, we also generate an independent data set of the same size as the original sample for the pre-screening step only; these versions are denoted by H-SIS(I) and H-HOLP(I), respectively. To calculate the hybrid orthogonalization vectors $\mathbf{z}_j$, we choose the regularization parameter λ_j by using the generalized information criterion (GIC) (Fan and Tang, 2013), which has been shown to be effective and convenient for high-dimensional penalized regressions. We estimate the noise variance using the same technique as in Shinkyu and Sueishi (2025):

$$\widehat{\sigma}^2 = \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda_{1se})\|^2/n,$$

where $\widehat{\boldsymbol{\beta}}(\lambda_{1\text{se}})$ is the lasso estimator whose tuning parameter is selected by the one-standard-error rule (the `lambda.1se` option in the `glmnet` R package).

4.1 Simulation example

We adopt simulation settings similar to those in van de Geer et al. (2014). Specifically, the simulated data sets are generated from model (2.1) with $(n, p, \sigma) = (100, 500, 1)$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}_n)$. For each data set, the rows of $\mathbf{X}$ are sampled as i.i.d. copies from $N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\rho^{|j-k|})_{p\times p}$ with $\rho = 0.9$. The true coefficient vector $\boldsymbol{\beta}$ is varied across three scenarios to examine the performance under different signal structures:

- Setting 1: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{s_1}, 0, \ldots, 0)^\top$, where $s_1 = 15$ and the nonzero β_j are independently sampled from the uniform distribution $U[0, 2]$.

- Setting 2: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, where the first 5 coefficients are equal to $3\lambda_{\text{univ}}$ with $\lambda_{\text{univ}} = \sqrt{2(\log p)/n}$, and the remaining coefficients are equal to $3\lambda_{\text{univ}}/j^2$.

- Setting 3: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ with $\beta_j = \lambda_{\text{univ}}/j^2$.

Four performance measures are used to evaluate the inference results: the average coverage probability for all coefficients (CP_all), the average coverage probability for nonzero coefficients (CP_max) in Setting 1 (or for the largest coefficients in Setting 2), the average length of confidence intervals for all coefficients, and the computation time for a single replication (Time). In Setting 3, we report the average coverage probability for the first coefficient β_1 (CP_β_1) instead of CP_max. The numerical results in the row corresponding to CP_β_1 are averages over 100 replications and are rounded to two decimal places. All simulations were performed on a Lenovo Legion Y9000P (Intel Core i9-13900HX, 32 GB RAM).

The results are summarized in Table 2. It is clear that while the averaged coverage probabilities for all coefficients (CP_all) by different methods are all around 95%, the averaged coverage probabilities for nonzero coefficients (CP_max) and the averaged coverage probabilities for the first coefficient β_1 (CP_β_1) for the different versions of HOT are significantly closer to the target than the corresponding values for LDPE in this finite-sample setting. In terms of interval length, the averaged confidence interval lengths of HOT and LDPE are comparable, and both are significantly shorter than those of adLDPE and STPS. In terms of computation time, HOT yields the shortest computational time, which is consistent with our preceding analysis: LDPE-type methods involve delicate parameter tuning and are typically more time-consuming in practice. Furthermore, the computation time of adLDPE is roughly comparable to that of LDPE, since the degrees-of-freedom adjustment imposes no additional computational burden. STPS exhibits the longest computation time, as expected, since it introduces a more computationally intensive tuning procedure on the basis of LDPE. Moreover, by comparing the results of HOT using the same or an independent data set for the pre-screening step, we find that the

independence assumption on the pre-screening set can be more of a technical assumption than a practical necessity.

Table 2: Four performance measures by different methods over 100 replications in Section 4.1.

	H-SIS	H-SIS(I)	H-HOLP	H-HOLP(I)	LDPE	adLDPE	STPS
				Setting 1			
CP_all	0.957	0.948	0.956	0.951	0.948	0.977	0.971
CP_max	0.947	0.947	0.948	0.948	0.768	0.821	0.933
Length	0.727	0.717	0.718	0.718	0.731	0.872	1.475
Time (s)	22.92	23.11	23.43	23.65	41.35	41.24	119.77
				Setting 2			
CP_all	0.963	0.954	0.965	0.961	0.949	0.966	0.969
CP_max	0.958	0.940	0.948	0.969	0.772	0.806	0.952
Length	0.707	0.700	0.702	0.705	0.737	0.815	1.439
Time (s)	23.51	23.12	23.22	23.51	41.87	42.36	120.01
				Setting 3			
CP_all	0.967	0.954	0.964	0.954	0.949	0.959	0.971
CP_β_1	0.95	0.93	0.97	0.92	0.9	0.93	0.95
Length	0.695	0.684	0.685	0.699	0.761	0.784	1.473
Time (s)	23.43	23.44	24.22	23.12	42.13	41.88	119.36

5. Application to the stock short interest and aggregate return data

We demonstrate the usefulness of the proposed methods by analyzing the stock short interest and aggregate return data. The short interest data are obtained from FINRA (`https://www.finra.org/finra-data/browse-catalog/equity-short-interest`), while the S&P 500 returns are retrieved from the FRED database using the `quantmod` R package. The raw data set includes monthly short-interest data for approximately 40,000

firms, measured as the number of shares held short for each firm, as well as the S&P 500 monthly log excess returns from January 2018 to December 2025. After removing firms with more than 5% missing observations, we select the 1,000 firms with the largest variances in short interest. As shown in Rapach et al. (2016), the short interest index, that is, the average of short interests of the firms, is arguably the strongest predictor of aggregate stock returns. Using the proposed large-scale inference technique, we can further identify which firms' short interests, among the thousands of candidates, have significant predictive associations with future S&P 500 log excess returns.

We approach this problem using the following predictive regression model

$$r_{t+1} = \mathbf{x}_t^\top \boldsymbol{\beta} + \epsilon_{t+1}, \tag{5.14}$$

where r_{t+1} is the S&P 500 log excess return of the $(t+1)$th month and $\mathbf{x}_t = (x_{1t}, \ldots, x_{pt})^\top$ with each component x_{jt} indicating the short interest of the jth firm in the tth month. Because the firms' short-interest series are highly correlated, we retain one representative from each group of firms whose pairwise correlations exceed 0.8, resulting in $p = 341$ representative firms. Moreover, the aforementioned period yields a sample size of $n = 95$, and both the predictor and response vectors are standardized to have mean zero and L_2-norm $\sqrt{n}$. We apply HOT, LDPE, adLDPE, and STPS as in Section 4 at a significance level of $\alpha = 0.001$. SIS is used in HOT to select a pre-screened set.

As summarized in Table 3, HOT identifies six significant firms. LDPE and adLDPE each select the same four firms, while STPS selects only one. It can be seen that HOT identifies all firms found by STPS and shares most findings with LDPE, while additionally identifying three firms. The predictive associations involving these firms may admit the following economic interpretations. MCSHF (Metcash) is a leading distributor of consumer staples, and its short interest may reflect shifts in consumer demand. SYRVF (Sacyr) is an international infrastructure construction company, and its short interest

Table 3: Significant firms identified by each method at $\alpha = 0.001$.

Method	Selected firms
HOT	YUXXF, MCSHF, SYRVF, NUAG, HENKY, BYLD
LDPE	YUXXF, SURVF, HENKY, BYLD
adLDPE	YUXXF, SURVF, HENKY, BYLD
STPS	HENKY

may signal market expectations regarding global investment activity. NUAG (New Pacific Metals) is a silver mining firm, and its short interest may reflect both industrial demand and market sentiment.

To further assess the practical value of these selected variables, we perform a rolling-window out-of-sample prediction exercise with a window length of 60 months. Based on the firms selected by each method, we refit model (5.14) via ordinary least squares (OLS) using only the selected firms to construct a forecasting model for each method. The out-of-sample forecasting performance is reported in Table 4. As LDPE and adLDPE select the same four firms, their out-of-sample forecasting performance is identical. We therefore only present the results of LDPE in Table 4.

Table 4: Out-of-sample forecasting performance.

Method	Variables	RMSE	MAE	Correlation
HOT	6	0.0352	0.0305	0.447
LDPE	4	0.0360	0.0318	0.328
STPS	1	0.0387	0.0334	0.163

As we can see, the HOT method yields the lowest RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error), and notably achieves a PCC (Pearson Correlation

Coefficient) of 0.447 between predicted and actual returns—substantially higher than LDPE (0.328) and STPS (0.163). This indicates that the additional firms identified by HOT carry incremental predictive information that is not captured by the alternative approaches.

Supplementary Material

The extension of HOT to sub-Gaussian designs and all technical details are relegated to the Supplementary Material for this paper.

Acknowledgments

This research has been supported by the National Key Research and Development Program of China (2022YFA1008000), the Natural Science Foundation of China (72571258; 12526560), the Anhui Provincial Natural Science Foundation (JZ2025AKZR0533), the Natural Science Foundation of Hefei University of Technology (JZ2025HGTA0143), and the Fundamental Research Funds for the Central Universities (WK2040000114).

References

Bellec, P. C. and Zhang, C.-H. (2022). De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli 28*(2), 713–743.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist. 37*(4), 1705–1732.

Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist. 45*(2), 615–646.

Chen, K. and Zhang, Y. (2025). Semi-supervised linear regression: enhancing efficiency and robustness in high dimensions. *Biometrics* **81**(3), ujaf113.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 70*(5), 849–911.

Fan, Y., Demirkaya, E. and Lv, J. (2019). Nonuniformity of p-values can occur early in diverging dimensions. *J. Mach. Learn. Res. 20*(77), 1–33.

Fan, Y., Jin, J. and Yao, Z. (2013). Optimal classification in sparse Gaussian graphic model. *Ann. Statist. 41*(5), 2537–2571.

Fan, Y., Kong, Y., Li, D. and Zheng, Z. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *Ann. Statist. 43*(3), 1243–1272.

Fan, Y. and Lv, J. (2014). Asymptotic properties for combined L_1 and concave regularization. *Biometrika 101*(1), 57–70.

Fan, Y. and Tang, C. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 75*(3), 531–552.

Janková, J., Shah, R. D., Bühlmann, P. and Samworth, R. J. (2020). Goodness-of-fit testing in high dimensional generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 82*(3), 773–795.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res. 15*(82), 2869–2909.

Javanmard, A. and Montanari, A. (2018). Debiasing the Lasso: Optimal sample size for Gaussian designs. *Ann. Statist. 46*(6A), 2593–2622.

Kong, Y., Zheng, Z. and Lv, J. (2016). The constrained Dantzig selector with enhanced consistency. *J. Mach. Learn. Res. 17*(123), 1–22.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc. 107*(499), 1129–1139.

Rapach, D. E., Ringgenberg, M. C. and Zhou, G. (2016). Short interest and aggregate stock returns. *J. Financ. Econ. 121*(1), 46–65.

Shinkyu, A. and Sueishi, N. (2025). Small tuning parameter selection for the debiased lasso. *J. Bus. Econom. Statist. 43*(4), 872–883.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika 99*(4), 879–898.

Sur, P., Chen, Y. and Candès, E. (2020). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields, 175*(1–2), 487–558.

Tian, X. and Taylor, J. (2018). Selective inference with a randomized response. *Ann. Statist. 46*(2), 679–710.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 58*(1), 267–288.

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc. 111*(514), 600–620.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist. 42*(3), 1166–1202.

Vazquez, O. and Nan, B. (2025). Debiased lasso after sample splitting for estimation and inference in high-dimensional generalized linear models. *Can J Stat.* **53**(1), e11827.

Wainwright, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory 55*(12), 5728–5741.

Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 78*(3), 589–611.

Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res. 11*(114), 3519–3540.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist. 38*(2), 894–942.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 76*(1), 217–242.

Zhou, K., Li, K.-C. and Zhou, Q. (2023). Honest confidence sets for high-dimensional regression by projection and shrinkage. *J. Amer. Statist. Assoc.* **118**(541), 469–488.

School of Management, University of Science and Technology of China

E-mail: (tjly@ustc.edu.cn)

School of Management, University of Science and Technology of China

E-mail: (zhengzm@ustc.edu.cn)

School of Economics, Hefei University of Technology

E-mail: (tszhjia@mail.ustc.edu.cn)

Radix Trading, Chicago

E-mail: (zzw9348ustc@gmail.com)