Estimation and Inference of Semiparametric Single-Index Models with High-Dimensional Covariates

Ruixuan Liu and Jing Tao

Emory University and University of Washington

Abstract

This paper develops new estimation and inference methods of high-dimensional single-index models. We propose a simple two-stage estimation method based on the average derivative estimator (ADE). This ADE is composed of weighted score functions of covariates that can easily be estimated under a semiparametric Gaussian copula structure. In the first stage, we plug in standard nonparametric estimates for marginal features and a regularized estimator for the precision matrix of the Gaussian copula to obtain high-dimensional score functions. In the second stage, we conduct LASSO-type thresholding to get sparse estimates of the regression coefficients in single-index models. Both stages involve only convex minimization problems. We derive the non-asymptotic bound of our estimator. For inference, we prove the asymptotic normality of a de-biased estimator using the one-step Newton-Raphson update. Our inferential tools do not rely on the Gaussian copula restriction and are more generally applicable with other pilot estimators.

Keywords: average derivatives; Gaussian copula; single-index models; Newton-Raphson update

1 Introduction

The single-index model (Ichimura, 1993; Klein and Spady, 1993) is one of the most important semiparametric models since its introduction in econometrics and statistics. Its prototypical version postulates the following relationship between a scalar dependent variable Y_i and p-dimensional covariates X_i:

$$Y_i = m\left(X_i^\top \beta_0\right) + \varepsilon_i, \quad i = 1, \cdots, n, \tag{1.1}$$

in which the error term ε_i is independent and identically distributed ($i.i.d.$) with conditional mean equal to zero given X. The regression coefficient β_0 is a unit vector of p unknown parameters that provides a concise measure of the covariates' effect. The unspecified link function $m(\cdot)$ offers great modeling flexibility, because (1.1) nests many parametric generalized linear models (GLMs) as special cases (McCullagh and Nelder, 1983; Horowitz, 2009).

There have been extensive studies on root-n consistent estimators for β_0, which can be classified into two categories in the low-dimensional regime where p is fixed and relatively small. The first well-known estimator for the index vector is the semiparametric least squares (SLS) or semiparametric maximum likelihood estimator (SMLE) (Ichimura, 1993; Klein and Spady, 1993), which solves the full-blown M-estimation problem by finding the function m and coefficient β pair that minimizes (maximizes) the least squares (likelihood) criterion function. The SLS or SMLE is cumbersome to compute as the solution to a nonlinear optimization problem whose objective function may be non-convex or multimodal. The average derivative estimator (ADE) stands as an alternative strategy that does not require solving a hard optimization problem. Let $\nabla m(x^\top \beta)$ denote the

vector of partial derivatives of the conditional mean with respect to the covariates. Then the average derivative identifies the regression coefficient β_0 up to some scale normalization, i.e., $\beta_0 \propto \mathbb{E}\left[\nabla m(X^\top \beta_0)\right]$. It is commonly referred to as a direct estimation method (Powell et al., 1989; Hardle and Stoker, 1989) by averaging nonparametric derivative estimation over the observed samples, and it can also be employed as a pilot estimator in an iteration procedure for obtaining the SLS or its variants (Hristache et al., 2001; Xia et al., 2002; Horowitz, 2009).

Both routes face significant challenges in the high-dimensional (H-D) regime where the ambient dimension p can be much larger than the sample size n. The tremendous success of LASSO (Tibshirani, 1996) and related estimators with ℓ_1-type penalties enabled the exploration of H-D single-index models. Several existing proposals combine the SLS with an ℓ_1 penalty for the coefficient β under the sparsity assumption of β_0. However, the main difficulty is the non-convexity of the criterion function in Ichimura (1993), which makes the well-developed machinery for H-D (generalized) linear models inapplicable, besides the significant computation burden.[1] The theoretical consistency is stated with respect to a local stationary point, which also requires a sufficiently accurate pilot estimator; see Loh (2017) for a rigorous treatment. On the other hand, the plain version of the ADE, which averages over slope estimates from local polynomial fitting (Chaudhuri et al., 1997; Li et al., 2003), still suffers from the curse of dimensionality and is not feasible in the H-D regime.

Motivated by modern empirical studies in which many different covariates per obser-

[1] A number of modifications are available. A local linear approximation of the profiled criterion function is suggested in Peng and Huang (2011) and Ma and He (2016). Nonetheless, it depends crucially on an accurate pilot estimator, which is often selected in an ad-hoc way. Radchenko (2015) suggested a constrained iterative least squares estimation for a fixed L_1 norm of β, s.t. $|\beta|_1 = t$ and then selected the resulting estimates over a set of grid points for $t \in \mathcal{T}$. We do not attempt to survey all related methods in this area, but refer interested readers to Su and Zhang (2014) for a comprehensive review.

vation are available, we develop new estimation and inference methods for H-D single-index models with $p \gg n$. We provide a H-D version of the one-step update approach suggested by Horowitz (2009). Due to the lack of a proper pilot estimator such as the ADE, we first propose a simple two-stage semiparametric estimator for the weighted average derivative (1.2) that leverages on the Gaussian copula structure on X. To fix the idea, the following representation is fundamental for our estimation:

$$\beta_0^* = \mathbb{E}\left[Y\dot{l}(X)w(X)\right], \tag{1.2}$$

with some non-negative weighting or trimming function $w(X)$ and the score function $\dot{l}(\cdot) := \frac{\nabla f(X)}{f(X)}$ of covariates, where $f(\cdot)$ is the joint density function of covariates X and $\nabla f(\cdot)$ is the vector of its first-order derivative. In the first stage of estimating $\dot{l}(\cdot)$, we plug in nonparametric kernel-type estimates for marginal features, such as the density and its derivative functions, and adopt the CLIME from Cai et al. (2011) for the precision matrix of the Gaussian copula function. In the second stage, we conduct soft thresholding to get a sparse estimator of β_0^* and then normalize it to obtain the estimated regression coefficient. The key steps of the implementation involve only convex optimization and enable fast computation with existing R packages. We also present the non-asymptotic bound of our estimator. For inference, we establish the asymptotic normality of the one-step updated or de-biased estimator utilizing the estimating equation and Hessian matrix from Section 2.6.4 in Horowitz (2009). Our proposal can be viewed as a feasible scheme of Horowitz (2009) targeted at H-D single-index models. It is worth emphasizing that our inference procedure does not rely on the Gaussian copula assumption and is generally applicable with other pilot estimators as well.

The main contributions of our paper are as follows. First, we design a new estimator that combines the ADE with the Gaussian copula structure of covariates for H-D single-index models. This allows us to construct a pilot estimator through convex minimization problems that can be executed in a computationally efficient manner, in contrast with several existing proposals that entail non-convex minimization. Second, our methodol-

ogy accommodates H-D cases where $p \gg n$ with explicit non-asymptotic bounds. We rigorously characterize the contribution of the estimated marginal features and the precision matrix from the Gaussian copula. Third, we combine the insight from Jankova and Van De Geer (2018) on the Newton-Raphson update with classical arguments in Ichimura (1993) in order to develop a valid inference procedure. Unlike the partial linear model in Jankova and Van De Geer (2018) with a separable structure between parametric and nonparametric components, the single-index model is more complicated because the finite-dimensional and infinite-dimensional parameters are bundled together. Aiming to address this challenge, we carry out a careful expansion of the kernel estimator that accounts for the estimation uncertainty of pilot estimators.

Related Literature. Our estimation method is inspired by researchers from the machine learning community (Plan and Vershynin, 2016; Yang et al., 2017) who utilized Stein's identity in single-index models. Plan and Vershynin (2016) first showed that $1/n \sum_{i=1}^{n} X_i Y_i$ estimates a vector proportional to β_0 when $X \sim \mathbb{N}(0, \mathbb{I}_{p\times p})$. Yang et al. (2017) generalized this framework using $1/n \sum_{i=1}^{n} \dot{l}(X_i) Y_i$ (upon proper trimming) for a *known* score function $\dot{l}(\cdot)$, which essentially requires a *known joint density function* of covariates X. Stein's identity is equivalent to the average derivative (Hardle and Stoker, 1989), because both are just integration by parts. We believe that it is helpful to formalize this complementary notion, as the ADE is an important semiparametric estimand in its own right (Powell et al., 1989; Li et al., 2003; Cattaneo et al., 2013). Compared with Plan and Vershynin (2016) and Yang et al. (2017), who assumed a known covariate density function, our setup is more realistic for empirical applications. Moreover, our assumptions on the Gaussian copula structure are related to H-D nonparanormal graphical models (Liu et al., 2009). This nonparanormal graph (Liu et al., 2009) extends the popular Gaussian graphical model by imposing only the Gaussianity on the copula, while leaving the marginals completely unspecified. In addition, the sparsity pattern of the precision matrix is interpretable, because it encodes the conditional independence

structure of covariates.

The additional restriction on X might seem odd at first sight; however, it shares motivation with several recent works that exploit additional information from covariates to obtain more tractable estimation procedures. For example, there is an increased interest in applying sliced inverse regression (SIR) to H-D single-index models (Lin et al., 2018, 2019, 2021), which leads to a convex minimization problem with an ℓ_1 type penalty. Note that its validity relies on the following linearity condition: $\mathbb{E}[X^\top b|X^\top\beta_0]$ is a linear function of b for any b, which is satisfied if the distribution of X is elliptic symmetric.[2] Neither the elliptic symmetric distribution nor the Gaussian copula family nests the other. Furthermore, for the monotone single-index model, Dai et al. (2021) developed an iterative approach that alternates between sparse-thresholded gradient-like steps and isotonic regressions to estimate the regression coefficient and the monotone link function simultaneously. They derived an non-asymptotic bound with a cubic-root rate for β_0, when X follows a finite normal mixture.

Our work also contributes to the literature on de-biased inference for H-D models. Within the framework of H-D (generalized) linear models, a number of pioneering works, including Javanmard and Montanari (2014); Van de Geer et al. (2014); Zhang and Zhang (2014); Belloni et al. (2015); Ning and Liu (2017), have laid down the general foundation. The extension to semiparametric models constitutes an active research area (Jankova and Van De Geer, 2018; Chernozhukov et al., 2020). Hirshberg and Wager (2020) adapted the approach from Chernozhukov et al. (2020) to analyze H-D single-index models with a *known* link function. A related approach in Nekipelov et al. (2020) also allows for a *known* nonlinear link function and additional nonparametric components within this link function; see their Section 2.2 for the precise formulation and Section 2.3 for motivating examples. Chakrabortty et al. (2019) developed de-biased inference for H-D

[2]As such, its applications to economics can be limited as covariates often exhibit high skewness and heavy tail (Horowitz, 2009).

semiparametric models with missing dependent variables. Their framework covers H-D single-index models when the covariates are elliptical symmetric. Our paper fills the gap and develops valid inference for single-index models with an *unknown* link function and general covariates without strong distributional assumptions.

Notation. We summarize the notation to be used throughout the paper. Let $\Phi(\cdot)$ be the cumulative distribution function of a standard Gaussian random variable, and let $\Phi^{-1}(\cdot)$ denote the quantile function. A Gaussian copula is parameterized as follows:

$$C(u_1, \cdots, u_p; \Sigma) = \Phi_p(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_p); \Sigma),$$

where $\Phi_p(\cdot; \Sigma)$ is the joint CDF of standard Gaussian random variables with a correlation matrix Σ. For any vector $\beta = (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^p$, we define the ℓ_0, ℓ_q and ℓ_∞ vector norms as $|\beta|_0 = \sum_{j=1}^p \mathbb{I}(\beta_j \neq 0)$, $|\beta|_q = \left(\sum_{j=1}^p |\beta_j|^q\right)^{1/q}$, and $|\beta|_\infty = \max_{1\leq j\leq p} |\beta|_j$. The true coefficient β_0 satisfies $|\beta_0|_2 = 1$ and its first coordinate is positive. We also use the following L_q and L_∞ matrix norms: $\|A\|_q = \max_{|b|=1} |Ab|_q$, and $\| A \|_\infty = \max_{1\leq i,j\leq p} |A_{ij}|$. We are particularly interested in the L_1 operator norm $\|A\|_1 = \max_{|b|_1=1} |Ab|_1 = \max_{1\leq j\leq p} \sum_{k=1}^p |a_{kj}|$, for a $p \times p$ matrix A. For a generic vector β, let $s_\beta = |\beta_0|_0$, and for a generic symmetric matrix A, let s_A be the number of non-zero off-diagonal terms.

The data consist of i.i.d. observations $(Y_i, X_i; i = 1, \cdots, n)$. We use the standard empirical process notation as follows. For a function $f(\cdot)$ of a random vector $Z = (Y, X)$ that follows distribution P, $Pf = \int f(z) dP(z)$, $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(Z_i)$, and $\mathbb{G}_n f = n^{1/2} (\mathbb{P}_n - P) f$. Let $\| \cdot \|_{\psi_\alpha}$ be the ψ_α-Orlicz norm with $\alpha \geq 1$. Recall that the Orlicz norm of a random variable Z is given by $\| Z \|_{\psi_\alpha} := \inf\{\eta > 0 : \mathbb{E}[\psi_\alpha(|Z|/\eta)] \leq 1\}$. If a random variable Z is a sub-exponential random variable or sub-Gaussian, it has a finite ψ_α-Orlicz norm with $\alpha = 1$ or $\alpha = 2$, respectively. Our analysis requires a trimming function $w_j(x) := \mathbb{I}\{x_j \in \mathcal{X}_j\}, \quad j = 1, \cdots, p$, and $w(x) := (w_1(x), \cdots, w_p(x))^\top$.

Accordingly, we define $\mathcal{X} = \prod_{j=1}^{p} \mathcal{X}_j$. The soft thresholding operator is denoted by

$$S(y; \lambda) = sign(y)[|y| - \lambda]_+. \tag{1.3}$$

We denote a standard kernel function and its first-order derivative by $K(\cdot)$ and $K'(\cdot)$, respectively.

Organization. The rest of our paper is organized as follows. Section 2 introduces the semiparametric estimator utilizing the ADE and the Gaussian copula structure of covariates. The de-biased estimator, which leads to valid inference, is also proposed in this section. Section 3 establishes the non-asymptotic bound for our estimator and shows asymptotic normality of the de-biased estimator. Section 4 conducts Monte Carlo simulations, and applies our method to a real dataset. Section 5 concludes. All proofs are relegated to the online supplementary material.

2 Semiparametric Estimation and Inference

We first introduce our two-stage semiparametric estimation method of H-D single-index models exploring the ADE under a Gaussian copula design of covariates. We highlight its computational convenience. Second, we describe our de-biased estimator, which is built on the Newton-Raphson update using the estimating equation and Hessian matrix from Section 2.6.4 of Horowitz (2009). In the low-dimensional regime, this type of one-step update can also be used to obtain the SLS of Ichimura (1993), which usually improves the efficiency of ADE; see the discussion leading to Theorem 4.1 of Xia (2006). Because we only require only a certain convergence rate of the pilot estimator, one can opt for other pilot estimators, such as those proposed in Lin et al. (2019) or Dai et al. (2021), depending on different modeling assumptions.

2.1 Two-Stage Semiparametric Estimation

Without additional restrictions on the underlying dependence structure, it remains a daunting exercise to deal with a H-D score function $\dot{l}(x)$. It is natural to exploit the copula function, as it has been a powerful tool for capturing the dependence structure and separating the dependence modeling from the marginal specification. In other words, the joint distribution of X is modeled by a copula $C(u_1, \cdots, u_p)$ with unrestricted marginal distributions $F_j(\cdot)$, for $j = 1, \cdots, p$. The Gaussian copula, denoted by $C(u_1, \cdots, u_p; \Sigma)$ with its correlation matrix Σ, is central in the study of semiparametric copula models (Klaassen and Wellner, 1997; Chen et al., 2006; Segers et al., 2014). In our context, the crucial feature of a Gaussian copula model is the closed-form score function $\dot{l}(\mathbf{x}) := \Big(\dot{l}_1(\mathbf{x}), \cdots, \dot{l}_p(\mathbf{x})\Big)^\top$:

$$\dot{l}_j(\mathbf{x}) := \frac{f_j'(x_j)}{f_j(x_j)} + \left[\frac{\Phi^{-1}(F_j(x_j))}{\phi(\Phi^{-1}(F_j(x_j)))} - \sum_{k=1}^{p} \Omega_{kj} \frac{\Phi^{-1}(F_k(x_k))}{\phi(\Phi^{-1}(F_j(x_j)))}\right] f_j(x_j), \tag{2.1}$$

for $j = 1, \cdots, p$, in which $\Omega := (\Omega_{j,k})_{1 \le j,k \le p}$ is the precision matrix, i.e., the inverse of Σ (Segers et al., 2014).

The semiparametric Gaussian copula model is equivalent to the nonparanormal model (Liu et al., 2009) in H-D data analysis, which relaxes the full parametric assumption in the Gaussian graphic model. The literature has developed rank-based estimation of the correlation matrix (Liu et al., 2012; Xue and Zou, 2012) to uncover the dependence structure. The (j, k)-th component in the correlation matrix Σ is estimated by

$$\widehat{\Sigma}_{jk} := sin\left(\frac{\pi}{2}\hat{\tau}_{jk}\right), \quad \text{for} \quad j \neq k, \tag{2.2}$$

where τ_{jk} stands for the rank correlation (Kendall's tau) between covariates $X_{i,j}$ and $X_{i,k}$, and its estimator is

$$\hat{\tau}_{jk} := \frac{2}{n(n-1)} \sum_{i \neq i'} \mathbb{I}\{X_{i,j} > X_{i',j}\}\mathbb{I}\{X_{i,k} > X_{i',k}\}. \tag{2.3}$$

Now we are ready to describe the implementation of our semiparametric estimation.

Step E1 (i). Nonparametric Estimation of Marginal Features. Each marginal density and its derivative functions are estimated by kernel smoothing methods:

$$\hat{f}_j(x_j) := \frac{1}{nh_j}\sum_{i=1}^{n} K\left(\frac{X_{i,j}-x_j}{h_j}\right), \quad \hat{f}_j'(x_j) := \frac{1}{nh_j^2}\sum_{i=1}^{n} K'\left(\frac{X_{i,j}-x_j}{h_j}\right). \tag{2.4}$$

Also, let $\hat{F}_j(\cdot)$ be the empirical distribution for X_j, $j = 1, \cdots, p$.

Step E1(ii). Regularized Estimation of the Precision Matrix. The estimated matrix solves the following problem:

$$\min \parallel \Omega \parallel_1, \text{ subject to } \parallel \widehat{\Sigma}\Omega - \mathbf{I}_{p\times p} \parallel_\infty \leq \lambda_\Omega. \tag{2.5}$$

We take the symmetrized solution as $\hat{\Omega}$, which is also the CLIME proposed by Cai et al. (2011).

Step E2. Thresholding the ADE. We obtain the plug-in estimator as $\hat{\dot{l}}_n := (\hat{\dot{l}}_{n,1}, \ldots, \hat{\dot{l}}_{n,p})^\top$:

$$\hat{\dot{l}}_{n,j}(\mathbf{x}) = \frac{\hat{f}_j'(x_j)}{\hat{f}_j(x_j)} + \left[\frac{\Phi^{-1}(\hat{F}_j(x_j))}{\phi(\Phi^{-1}(\hat{F}_j(x_j)))} - \sum_{k=1}^{p}\hat{\Omega}_{kj}\frac{\Phi^{-1}(\hat{F}_k(x_k))}{\phi(\Phi^{-1}(\hat{F}_j(x_j)))}\right]\hat{f}_j(x_j).$$

Then we apply the soft thresholding operator to each coordinate of $\hat{\beta}_n^* = (\hat{\beta}_{n,1}^*, \cdots, \hat{\beta}_{n,p}^*)^\top$:

$$\hat{\beta}_{n,j}^* := S(n^{-1}\sum_{i=1}^{n} Y_i w_j(X_i)\hat{\dot{l}}_{n,j}(X_i); \lambda_\beta), \quad j = 1, \cdots, p. \tag{2.6}$$

Finally, we normalize the estimated $\hat{\beta}_n$ as

$$\hat{\beta}_n := \hat{\beta}_n^* sign(\hat{\beta}_{n,1}^*)/|\hat{\beta}_n^*|_2. \tag{2.7}$$

We elaborate on the computational simplicity of the above procedure. First, the nonparametric kernel estimation is applied to the marginal density function and its derivative for each coordinate separately. Second, the H-D Gaussian copula is parameterized by the correlation matrix, which is estimated in a pairwise fashion by rank-based

methods (Liu et al., 2012; Xue and Zou, 2012). To obtain an accurate estimation of its inverse, i.e., the precision matrix, we leverage on the sparsity restriction, which can be motivated by the sparsity of the corresponding nonparanormal graph (Liu et al., 2009). The CLIME of Cai et al. (2011) can be solved via p linear programming problems. The final soft thresholding is performed on the ADE coordinate-wise. As noted by Yang et al. (2017), the regression coefficient can also be expressed as the solution of

$$\hat{\beta}_n^* = \arg\min_{\beta} \left[\beta\beta^\top - 2n^{-1} \sum_{i=1}^{n} Y_i w(X_i) \hat{\dot{l}}_n^\top(X_i)\beta + \lambda_\beta |\beta|_1 \right]. \tag{2.8}$$

Other coordinate-wise thresholding, such as SCAD (Fan and Li, 2001) or hard thresholding (Bühlmann and van de Geer, 2011) can also be employed herein. We define the adaptive LASSO version by taking

$$\hat{\beta}_{n,j}^{*adap} := S(n^{-1} \sum_{i=1}^{n} Y_i \hat{w}_{n,j}(X_i) \hat{\dot{l}}_{n,j}(X_i); \lambda_\beta / |\hat{\beta}_{n,j}^*|), \tag{2.9}$$

and we denote the normalized version by $\hat{\beta}_{n,j}^{adap}$ for $j = 1, \cdots, p$.

Remark 2.1. The ADE cannot directly be used to estimate coefficients associated with discrete covariates. Nonetheless, one can adapt the method proposed by Horowitz and Hardle (1996). To illustrate the idea, let D denote a binary regressor with its coefficient α, so the single-index model is $\mathbb{E}[Y|X = x, D = d] = m_0(x^\top \beta_0 + d\alpha_0)$. We can obtain the estimates $\hat{\beta}_n^{(0)}$ and $\hat{\beta}_n^{(1)}$ based on subsamples for which $D_i = 0$ or 1 separately and then take a weighted average denoted by $\hat{\beta}_n$. When D is evaluated at d, define

$$\begin{aligned} J(d) := \int_{v_0}^{v_1} & [c_0 \mathbb{I}\{m_0(u + d\alpha) < c_0\} + c_1 \mathbb{I}\{m_0(u + d\alpha) > c_1\} \\ & + m_0(u + d\alpha) \mathbb{I}\{c_0 \leq m_0(u + d\alpha) \leq c_1\}] \, du, \end{aligned}$$

with some given constant terms (c_0, c_1) and (v_0, v_1) that satisfy Assumption G on page 38 of Horowitz (2009). Given the consistency of $\hat{\beta}_n$ for the continuous part, one can

construct a kernel-type estimator $J_n(d)$ for $J(d)$. The estimator of α is then $\hat{\alpha}_n = \frac{J_n(1)-J_n(0)}{c_1-c_0}$.[3]

2.2 One-Step Newton-Raphson Update

We propose an asymptotically normal non-sparse estimator that leads to confidence regions or testing for subvectors of β_0. We adapt the one-step Newton-Raphson update procedure, which is a general method for constructing efficient estimators for parametric and semiparametric models; see Sections 2.5 and 7.3 in Bickel et al. (1993). This idea has resurged to develop inference for H-D models based on the so-called de-biased estimator, considered by Javanmard and Montanari (2014); Van de Geer et al. (2014); Zhang and Zhang (2014), among many others. Existing work has focused mostly on parameters in (generalized) linear models or partial linear models (Jankova and Van De Geer, 2018), where the finite-dimensional parameter of interest is separable from the nuisance nonparametric component. This is no longer the case for the single-index model. We build on the outline given in Section 2.6 in Horowitz (2009) and offer a feasible plan to the H-D single-index model.

We introduce additional notations to facilitate our presentation. For identification purpose, we construct the de-biased estimator only for $\beta_{0,-1} := (\beta_{0,2}, \cdots, \beta_{0,p})$, because the first coordinate is determined by $\beta_{0,1} = \sqrt{1 - \sum_{j=2}^{p} \beta_{0,j}^2}$ under the normalization scheme. Henceforth, we partition our estimators as $\hat{\beta}_n = (\hat{\beta}_{n,1}, \hat{\beta}_{n,-1}^\top)^\top$ and $\tilde{\beta}_n = (\tilde{\beta}_{n,1}, \tilde{\beta}_{n,-1}^\top)^\top$. We write $X = (X_1, X_{-1}^\top)$ accordingly. Denote the conditional mean of Y, given the linear index $X^\top\beta$, by

$$m(u;\beta) := E[Y|X^\top\beta = u]. \tag{2.10}$$

Let $K_h(\cdot) = K(\cdot/h)$. Given a generic pilot estimator $\hat{\beta}_n$, consider the following leave-

[3]To clarify the notation, our dummy variable D is z in Horowitz (2009). The vector W therein simplifies to a scalar equal to 1, pertaining to the single dummy variable D.

one-out kernel estimators:

$$\hat{\mathbf{r}}_{\phi,n}(X_i^\top\hat{\beta}_n) := \frac{1}{nh}\sum_{l\neq i}\phi_l K_h\left((X_l - X_i)^\top\hat{\beta}_n\right),\ \hat{\mathbf{r}}'_{\phi,n}(X_i^\top\hat{\beta}_n) := \frac{1}{nh^2}\sum_{l\neq i}\phi_l K'_h\left((X_l - X_i)^\top\hat{\beta}_n\right),$$
$$\hat{\mathbf{f}}_n(X_i^\top\hat{\beta}_n) := \frac{1}{nh}\sum_{l\neq i}K_h\left((X_l - X_i)^\top\hat{\beta}_n\right),\ \hat{\mathbf{f}}'_n(X_i^\top\hat{\beta}_n) := \frac{1}{nh^2}\sum_{l\neq i}K'_h\left((X_l - X_i)^\top\hat{\beta}_n\right).$$

Let $\hat{\mu}_{n,-1}(X_i) = (\hat{\mu}_{n,2}(X_i), \cdots, \hat{\mu}_{n,p}(X_i))^\top$, where

$$\hat{\mu}_{n,j}(X_i) := \frac{\hat{\mathbf{r}}_{X_j,n}(X_i^\top\hat{\beta}_n)}{\hat{\mathbf{f}}_n(X_i^\top\hat{\beta}_n)},\quad j = 2, \cdots, p,$$

and

$$\hat{\varpi}_n(X_i) := \frac{\hat{\mathbf{r}}'_{Y,n}(X_i^\top\hat{\beta}_n)}{\hat{\mathbf{f}}_n(X_i^\top\hat{\beta}_n)} - \frac{\hat{\mathbf{r}}_{Y,n}(X_i^\top\hat{\beta}_n)\hat{\mathbf{f}}'_n(X_i^\top\hat{\beta}_n)}{\hat{\mathbf{f}}_n^2(X_i^\top\hat{\beta}_n)}.$$

We can estimate the conditional mean function by $\widehat{m}_n(X_i^\top\hat{\beta}_n;\hat{\beta}_n) = \frac{\hat{\mathbf{r}}_{Y,n}(X_i^\top\hat{\beta}_n)}{\hat{\mathbf{f}}_n(X_i^\top\hat{\beta}_n)}$. Denote its derivative w.r.t. the regression coefficient by $\frac{\partial\widehat{m}_n(X_i^\top\hat{\beta}_n;\hat{\beta}_n)}{\partial\beta_{-1}} = (\frac{\partial\widehat{m}_n(X_i^\top\hat{\beta}_n;\hat{\beta}_n)}{\partial\beta_2}, \cdots, \frac{\partial\widehat{m}_n(X_i^\top\hat{\beta}_n;\hat{\beta}_n)}{\partial\beta_p})^\top$.

We describe how to implement the de-biased estimator as follows:

Step I1. Estimating $\hat{\Psi}_n(\hat{\beta}_n)$ and $\hat{\dot{\Psi}}_n(\hat{\beta}_n)$. We estimate the score function and Hessian matrix in Ichimura (1993) by

$$\hat{\Psi}_n(\hat{\beta}_n) := \frac{1}{n}\sum_{i=1}^n w(X_i)(Y_i - \widehat{m}_n(X_i^\top\hat{\beta}_n;\hat{\beta}_n))\frac{\partial\widehat{m}_n(X_i^\top\hat{\beta}_n;\hat{\beta}_n)}{\partial\beta_{-1}}, \tag{2.11}$$

$$\hat{\dot{\Psi}}_n(\hat{\beta}_n) := \frac{1}{n}\sum_{i=1}^n w(X_i)\hat{\varpi}_n^2(X_i)(X_{i,-1} - \hat{\mu}_{n,-1}(X_i))^{\otimes 2}. \tag{2.12}$$

Step I2. Estimating Θ. The matrix $\widehat{\Theta}$ is a regularized inverse by CLIME as the solution to the following problem:

$$\widehat{\Theta} := \arg\min \parallel \Theta \parallel_1,\quad \text{subject to}\quad \parallel \hat{\dot{\Psi}}_n(\hat{\beta}_n)\Theta - \mathbb{I}_{(p-1)\times(p-1)} \parallel_\infty \leq \lambda_\Psi. \tag{2.13}$$

Step I3. N-R update. Obtain the de-biased estimator $\tilde{\beta}_n$ such that

$$\tilde{\beta}_{n,-1} := \hat{\beta}_{n,-1} - \widehat{\Theta}\hat{\Psi}_n(\hat{\beta}_n). \tag{2.14}$$

The probability limit of the sample Hessian matrix $\hat{\dot{\Psi}}_n(\hat{\beta}_n)$ is

$$\dot{\Psi}_0 := \mathbb{E}[w(X)m'(X^\top\beta_0)(X_{-1} - \mathbb{E}[X_{-1}|X^\top\beta_0])^{\otimes 2}]. \tag{2.15}$$

The regularized estimation in Step I2 requires the sparsity of its inverse $\Theta_0 := \dot{\Psi}_0^{-1}$. We explain why we did not construct a de-biased version of our ADE directly by undoing any thresholding or regularization. The reason is that our estimated scores depend on a linear combination of columns from the precision matrix; see expression (2.1). One cannot generally show the asymptotic normality for a non-sparse linear combination of each column or row of this estimated precision matrix, even if one applies a de-biasing procedure for the precision matrix as in Gu et al. (2015). In addition, we need to bound the reminder terms by controlling the estimation error in terms of its L_1 operator norm. Unfortunately, there is no guarantee that the de-biased estimator can achieve this.

Remark 2.2. There has been considerable interest in applying the double or de-biased machine learning approach to semiparametric models (Chernozhukov et al., 2020; Hirshberg and Wager, 2020; Nekipelov et al., 2020; Chakrabortty et al., 2019). The classical notion of Neyman orthogonality has emerged as a natural and flexible condition. The general method developed by Chernozhukov et al. (2020) applies to a given coordinate $\beta^*_{0,j}$ of the average derivative (see their Example 4) in our notation, which is not the same as $\beta_{0,j}$ in the single-index model per se. Hirshberg and Wager (2020) and Nekipelov et al. (2020) extended the approach from Chernozhukov et al. (2020) to H-D single-index models with a *known* link function. The setup in Nekipelov et al. (2020) also allows for nuisance functions within this known link function; see their Equation (2.4) and their discussion of it. Chakrabortty et al. (2019) developed a double robust inference procedure for H-D M-estimation involving missing dependent variables. Their framework covers single-index models with elliptically symmetric covariates X that can be handled by SIR. Because their M-estimation begins with a convex loss function in terms of the parameter of interest, it does not directly apply to single-index models without the elliptical symmetry assumption on X. It is well known that the criterion function in

Ichimura (1993) is not convex in β. Referring to the basis expansion $\Phi(X)^\top\beta$ therein, the proper extension to SLS should allow the basis function $\Phi(X;\beta)$ to depend on β as well; see the formulation in Radchenko (2015) or Ma and He (2016).

3 Theoretical Results

3.1 Rate of Convergence

This section starts with a non-asymptotic bound of our estimator. The aim is to show that the estimator has good theoretical properties. Under reasonable assumptions on the marginal densities and the precision matrix in the Gaussian copula, the estimation error is of the same order of magnitude as the oracle upper bound that one would have if one knew the joint density of X. Once the non-asymptotic bound is established, it is clear how to obtain the rate of convergence. Recall that the trimmed support induced by $w(\cdot)$ is $\mathcal{X} := \prod_{j=1}^{p} \mathcal{X}_j$. The following regularity conditions are imposed throughout.

Assumption 1. We observe i.i.d. copies of (Y_i, X_i) generated from the single-index model (1.1). We normalize $|\beta_0|_2 = 1$ with its first coordinate $\beta_{0,1}$ to be strictly positive. The error term $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is sub-Gaussian with sub-Gaussian norm $\tau_\varepsilon := \|\varepsilon\|_{\psi_2}$ for a constant $0 < \tau_\varepsilon < \infty$.

Assumption 2. The copula function of $(X_1, \cdots, X_p)$ is Gaussian and is denoted by $C(\cdot;\Sigma)$. Its precision matrix $\Omega := \Sigma^{-1}$ belongs to the family $\mathcal{G}_{q_1}(c_\Omega, M_\Omega)$ of semidefinite positive matrices, as follows:

$$\mathcal{G}_{q_1}(c_\Omega, M_\Omega) := \left\{ \Omega = (\omega_{kj})_{1\le j,k\le p} : \max_j \sum_{k=1}^{p} |\omega_{kj}|^{q_1} \le c_\Omega, \| \Omega \|_1 \le M_\Omega \right\}. \tag{3.1}$$

Moreover, the largest and smallest eigenvalues of Ω, $\lambda_{\max}(\Omega)$ and $\lambda_{\min}(\Omega)$, are bounded away from infinity and zero, respectively.

Assumption 3. Each marginal density function f_j of X_j belongs to the uniform Hölder class with order $(r+1)$. That is, $f_j(\cdot)$ is r-th times continuously differentiable and $\forall x, x' \in \mathcal{X}_j$,

$$|f_j^{(r)}(x) - f_j^{(r)}(x')| \leq L|x - x'| \tag{3.2}$$

with a Lipschitz constant $L > 0$. Also, we have $0 < c_0 \leq f_j(x_j) \leq C_0 < \infty$, $\forall x_j$ in the compact set $\mathcal{X}_j$.

Assumption 4. The link function m is bounded on $\mathcal{X}$ and is continuously differentiable. In addition, $\mathbb{E}[\nabla m(X^\top \beta_0)] \neq 0$. The second-order derivative of the link function $m''(\cdot)$ is continuous and bounded in the support of $x^\top \beta$ for $x \in \mathcal{X}$.

Assumption 5. (i) The bandwidth h_j satisfies the following restrictions: $h_j \to 0$, $nh_j \to \infty$, and $nh_j^{-3}/\log p \to \infty$ as $n \to \infty$ and $c \log n/n \leq h_j \leq 1$ with probability 1, for $1 \leq j \leq p$. (ii) The tuning parameter in the CLIME satisfies $\lambda_\Omega \asymp O(\sqrt{\log p/n})$.

Assumption 6. The kernel function $K(\cdot)$ is a bounded r-order kernel function with compact support, i.e.,

$$\int K(u)\, du = 1 \text{ and } \int u^j K(u)\, du = 0 \text{ for } j = 1, ..., r-1.$$

Moreover, it can be written as $K(x) = \Psi(\mathbf{p}(x))$, with $\Psi(\cdot)$ being of bounded variation and $\mathbf{p}(x)$ being a real polynomial on $\mathcal{R}$. The first-order derivative $K'(\cdot)$ satisfies the same set of restrictions with possibly different choices of Ψ and $\mathbf{p}$.

Some comments on the conditions are in order regarding estimation of marginal features and the precision matrix.

Remark 3.1. The smoothness condition on the marginal density and its derivative is standard (Giné and Nickl, 2016). Under the restriction of the kernel and its first derivative functions, we can invoke maximal inequalities for the VC-type functional class. It is common to assume the restricted eigenvalue condition in a (generalized) linear model

using ℓ_1-type penalized estimation. Heuristically speaking, this enforces weak dependence among covariates; see Bühlmann and van de Geer (2011) and Wainwright (2019). Referring to a proper extension to the SLS (Ichimura, 1993) with an ℓ_1 penalty, the analog is stated in Assumption (A.7) from Radchenko (2015). Under the framework of SIR, Lin et al. (2018) also impose sparsity conditions on the covariance matrix (not the precision matrix) of X; see their Assumption (A.6). In comparison, we restrict the dependence among covariates by means of the sparsity and bandedness (Cai et al., 2011) of its precision matrix of the Gaussian copula. A sparse Ω directly represents the conditional independence relationship among covariates. That is, $\Omega_{jk} = 0$ if and only if $X_j \perp X_k | X_{\setminus\{j,k\}}$, where $X_{\setminus\{j,k\}}$ is the vector of X after excluding X_j and X_k. The bandedness condition is indispensable if we want to obtain a fast enough rate in the L_1 operator norm; see Cai et al. (2016).

Theorem 1 (Non-asymptotic Bounds)**.** Let Assumptions 1 to 6 hold. If the regularization parameter $\lambda_\beta \geq C \left(s_\Omega \sqrt{\frac{\log p}{n}} + \max_j h_j^r + M_\Omega^{1-q} c_\Omega \left(\frac{\log p}{n}\right)^{(1-q_1)/2} \right)$, for some large constant C, then with probability $1 - O(p^{-1})$, we have

$$|\hat{\beta}_n^* - \beta_0^*|_2 \leq \sqrt{s_\beta}\lambda_\beta \quad \text{and} \quad |\hat{\beta}_n^* - \beta_0^*|_1 \leq s_\beta \lambda_\beta. \tag{3.3}$$

Corollary 1 (Rate of Convergence)**.** The rate of convergence for the normalized estimator $\hat{\beta}_n$ is as follows:

$$|\hat{\beta}_n - \beta_0|_\ell = O_p \left(s_\beta^{1/\ell} \left(\sqrt{\frac{\log p}{n}} + \max_j h_j^r + M_\Omega^{1-q} c_\Omega \left(\frac{\log p}{n}\right)^{(1-q_1)/2} \right) \right), \quad \ell = 1, 2.$$

Considering the adaptive-LASSO version of our estimation, we get the same order of convergence rate for $|\hat{\beta}_n^{adap} - \beta_0|_\ell$ for $\ell = 1, 2$ by Theorem 7.9 from Bühlmann and van de Geer (2011). Compared with the convergence rate in Yang et al. (2017) under a known joint density assumption on covariates, the additional term

$$A_n := s_\beta \left(\max_{1 \leq j \leq p} h_j^r + M_\Omega^{1-q} c_{n\Omega} \left(\frac{\log p}{n}\right)^{(1-q_1)/2} \right)$$

represents the cost for estimating the marginal density functions and the Gaussian copula structure. It is straightforward to see that A_n can be of $o\left(s_\beta\sqrt{\log p/n}\right)$ for sufficiently smooth marginal densities and a relatively sparse Ω.

Remark 3.2. Besides the convolution-type kernel estimator, one can also work with wavelet- or spline-based estimators (Chen et al., 2006). For those estimators that can be explicitly expressed as multi-resolution or projection kernel estimators, the corresponding maximal inequalities are available from Giné and Nickl (2016) for the marginal density (Proposition 5.1.12) and the derivative of the density (Proposition 5.1.9). Because the score function $\dot{l}(x)$ depends on the linear combination of columns in Ω, it is essential to adopt the CLIME (Cai et al., 2011) or ACLIME (Cai et al., 2016) procedure to estimate the precision matrix when $p \gg n$. In comparison, the graphic LASSO used in Xue and Zou (2012) entails a convergence rate in terms of the L_1 matrix norm that requires $p/n \to 0$.

3.2 Asymptotic Normality

Let $\mathbf{e}_j$ be a $(p-1)$-dimensional selection vector with its $(j-1)$-th coordinate being 1 and the rest being 0 such that $\tilde{\beta}_{n,j} = \mathbf{e}_j^\top \tilde{\beta}_{n,-1}$ for $2 \le j \le p$. This section derives the asymptotic normality of $\tilde{\beta}_{n,j}$, as well as a linear combination $\xi^\top \tilde{\beta}_{n,-1}$ for a given constant vector $\xi \in \mathbb{R}^{p-1}$. We denote $[a,b]$ as the interval support of $x^\top\beta_0$ where $x \in \mathcal{X}$. Let c_{1n} be the convergence rate of a pilot estimator. For example, $c_{1n} = s_\beta\left(s_\Omega\sqrt{\frac{\log p}{n}} + \max_{1\le j\le p} h_j^r + M_\Omega^{1-q_1} c_\Omega \left(\frac{\log p}{n}\right)^{(1-q_1)/2}\right)$ is for our ADE derived from Corollary 1. Denote the conditional mean of each covariate given the linear index by $\mu_j(x^\top\beta_0) := \mathbb{E}[X_{ij}|X_i^\top\beta_0 = x^\top\beta_0]$ for $j = 2, \cdots, p$. We make the following assumptions to facilitate our analysis.

Assumption 7. The marginal density function $g_{\beta_0}(\cdot)$ of $X^\top\beta_0$ is continuous and positive on the interval (a,b). For any compact subinterval $[a_0,b_0] \subset (a,b)$, there exist constants

c and C such that $0 < c \leq g_{\beta_0}(x_\beta) \leq C < \infty$ with $x_\beta \in [a_0, b_0]$.

Assumption 8. (i) $\|m(\cdot)\|_\infty \leq C_m$, and the second-order derivative of the link function $m''(\cdot)$ is continuous and bounded on (a, b). (ii) The density function g_{β_0}; (iii) $\forall 1 \leq j \neq j' \leq p$, the joint density functions $g_j(x_\beta, x_j)$ of $(X^\top \beta_0, X_j)$ and $g_{j,j'}(x_\beta, x_j, x_{j'})$ of $(X^\top \beta_0, X_j, X_{j'})$ are continuous and have continuous partial derivatives of order one with respect to x_β on $(a, b) \times \mathcal{X}$ and $(a, b) \times \mathcal{X}^2$. (iv) The conditional mean $\mu_j(\cdot)$ belongs to the uniform Hölder class with order γ for some $\gamma \geq 2$ for all $j = 2, \cdots, p$.

Assumption 9. The kernel function K that is used to estimate $\hat{m}_n(\cdot)$ is a symmetric probability function with compact support, whose second order derivative K'' is Lipschitz continuous on $\mathbb{R}$. That is, there exists a constant $L > 0$ such that for all $u, v \in \mathbb{R}$ with $|u - v| \leq L$, $|K''(u) - K''(v)| \leq c_K(u)|u - v|$ for some bounded and integrable function $c_K(\cdot) : \mathbb{R} \to \mathbb{R}^+$ with $\int_{\mathbb{R}} c_K(u)du \leq A_K$ and $\|c_K(\cdot)\|_\infty \leq A'_K$ for some non-negative constants A_K, A'_K.

Assumption 10. The sparsity s_β satisfies $\sqrt{n} s_\beta c_{1n}^2 = o(1)$.

Assumption 11. The covariate $\tilde{X}_{ij} := X_{ij} - \mathbb{E}\left[X_{ij} | X_i^\top \beta_0\right]$ is sub-Gaussian for each $i = 1, \ldots, n$, and $j = 2, \ldots, p$.

Assumption 12. The inverse of the Hessian matrix $\Theta_o := \dot{\Psi}_0^{-1}$ belongs to the family

$$\mathcal{G}_{q_2}(c_\Theta, M_\Theta) := \left\{ \Theta_o = (\theta_{kj})_{1 \leq kj \leq p} : \max_j \sum_{k=1}^{p} |\theta_{kj}|^{q_2} \leq c_\Theta, \|\Theta_o\|_1 \leq M_\Theta, \Theta \succ 0 \right\},$$

where the largest and smallest eigenvalues of Θ_o, $\lambda_{\max}(\Theta_o)$ and $\lambda_{\min}(\Theta_o)$, are bounded away from infinity and zero, respectively.

Assumption 13. The tuning parameter defined in Algorithm (2.13) satisfies that $\lambda_\Psi c_{1n} = o(n^{-1/2})$ and $\lambda_\Psi \gtrsim M_\Theta \left(\frac{c_{1n}^2}{h^3} + \sqrt{\frac{\log np}{nh^2}} + h^2 + c_{1n} \right)$.

Assumption 14. The bandwidth $h = h_n$ satisfies that (i) $h = o(1)$, $nh^8 \to 0$, $nh^{2\gamma} \to 0$, and $nh^4/(\log n \log p) \to \infty$; and (ii) for M_Θ and c_Θ defined in Assumption 12, $(M_\Theta^2 \vee c_\Theta^2) \left(\frac{c_{1n}^2}{h^3} + \sqrt{\frac{\log np}{nh^2}} + h^2 + c_{1n} \right) = o(1)$.

Remark 3.3. Assumptions 7, 8, and 9 are mild smoothness conditions for the nonparametric functions and kernel function. Those standard conditions have been introduced in the semiparametric single-index literature when the dimension of X is fixed (e.g., Assumptions (A2)-(A5) in Gu and Yang (2015)). Based on the result in Corollary 1, Assumption 10 guarantees that $|\hat{\beta} - \beta_0|_1^2 = o_p(n^{-1/2})$, which controls the higher-order terms when we derive the asymptotic destribution of the de-biased estimator $\tilde{\beta}_n$. Assumption 11 is stronger than needed, which is introduced to simplify our analysis. One can relax this sub-Gaussian condition following the argument in, for instance, Chernozhukov et al. (2017) and Kuchibhotla and Chakrabortty (2018). In Assumption 12, the constants q_2, c_Θ, and M_Θ are positive constants that are allowed to grow as n and p grow. Even though this uniformity class of matrices restricts the Hessian matrix in a certain way, it is widely used in the H-D estimation and inference literature. Consider an example with $X \sim \mathbb{N}(0, I_{p\times p})$. It is easy to observe that

$$\Theta_0 = \frac{1}{\mathbb{E}[m_0'(X^\top\beta_0)]} \times \left[I_{(p-1)\times(p-1)} + \frac{1}{1+\beta_{0,-1}^\top\beta_{0,-1}}\beta_{0,-1}\beta_{0,-1}^\top \right].$$

Under the sparsity assumption of β_0, the above matrix is also sparse. The condition in Assumption 13 is slightly stronger than, for example, the $\sqrt{\log p/n}$ rate in Cai et al. (2011), where they apply a similar procedure to estimate the sparse precision matrix. This is due to the two-stage procedure and plug-in estimation where the variations of first-stage's nonparametric estimation of the marginal density functions, the Gaussian copula, and the LASSO estimation of $\hat{\beta}_n$ affect the convergence rates of $\hat{\dot{\Psi}}_n(\hat{\beta}_n)$ and $\widehat{\Theta}$. Finally, Assumption 14 is about the kernel bandwidth rate. It is slightly stronger than the condition in Ichimura (1993), because we need to quantify the additional price for estimating the H-D single-index parameter β besides the error rate of a standard univariate kernel regression.

The following notation is used to construct the asymptotic variance of the de-biased estimator. Let $\Sigma_{\beta_0} = \mathbb{E}\left[w(X)\sigma^2(X)\left(m_0'\left(X^\top\beta_0\right)\right)^2\left(X_{-1} - \mathbb{E}\left[X_{-1}|X^\top\beta_0\right]\right)^{\otimes 2}\right]$ for $\sigma^2(X) =$

$\mathbb{E}[\varepsilon^2|X]$. Let $\widehat{\Sigma}_{\hat{\beta}_n} = \mathbb{P}_n\left[w(X_i)\hat{\varepsilon}_i^2\left(\hat{m}_n'\left(X_i^\top\hat{\beta}_n\right)\right)^2\left(X_{i,-1} - \hat{\mu}_n(X_{i,-1};\hat{\beta}_n)\right)^{\otimes 2}\right]$ with $\hat{\varepsilon}_i = Y_i - \widehat{m}_n(X_i^\top\hat{\beta}_n)$. Suppose that $\lambda_{\max}(\Sigma_{\beta_0})$ is bounded from above and away from zero.

Theorem 2 (Asymptotic Normality)**.** Under the assumptions in Theorem 1 and Assumptions 10-14, we have the asymptotic normality for any given j-th coordinate of the regression coefficient:

$$\sqrt{n}\left(\tilde{\beta}_{n,j} - \beta_{0,j}\right) \Rightarrow \mathbb{N}(0, \sigma_{\beta,j}^2), \tag{3.4}$$

where

$$\sigma_{\beta,j}^2 = \mathbf{e}_j^\top\Theta_o\mathbb{E}[w(X)\sigma^2(X)(m_0'(X^\top\beta_0))^2(X_{-1} - \mathbb{E}[X_{-1}|X^\top\beta_0])^{\otimes 2}]\Theta_o\mathbf{e}_j. \tag{3.5}$$

Moreover, for $\hat{\sigma}_{\beta,j}^2 := \mathbf{e}_j^\top\widehat{\Theta}\widehat{\Sigma}_{\hat{\beta}_n}\widehat{\Theta}\mathbf{e}_j$, we have $\hat{\sigma}_{\beta,j}^2/\sigma_{\beta,j}^2 = o_p(1)$ for each $j = 2, \ldots, p$.

Remark 3.4. Theorem 2 allows the number of non-zero $\beta_{0,j}$ to increase as the sample size grows to infinity as long as the growth rate is not too fast. Moreover, these results are still valid when the dimension of the parametric parameters is bigger than the sample size. Under the conditions provided in Theorem 2, when the unknown function $m_0(\cdot)$ is estimated by the local polynomial estimator or the kernel smoothing estimator, the asymptotic variance of the de-biased estimator is identical to the one obtained in Ichimura (1993) when p is fixed and much smaller than n. When $\varepsilon_i = Y_i - m_0(X^\top\beta_0)$ has conditionally homoskedastic errors, i.e., $\mathbb{E}[\varepsilon_i^2|X_i] = \sigma^2$, $\sigma_{\beta_j}^2$ coincides with the semiparametric efficiency bound derived in Ichimura (1993).

We next turn to a uniform limit theory. Let $B(s_\beta, p)$ be the set for p-dimensional vectors with at most s_0 non-zero coordinates for β_0. Let $\Phi(t)$ be the CDF of the standard normal distribution evaluated at t.

Theorem 3. Under the assumptions in Theorem 2, let $\xi = \xi_n \in \mathbb{R}^{p-1}$ be any fixed sequence of vectors satisfying $\|\xi\|_2 = 1$. We have

$$\sup_{t\in\mathbb{R}}\sup_{\beta_0\in B(s_\beta,p)}\left|\mathbb{P}\left(\frac{\sqrt{n}\xi^\top\left(\tilde{\beta}_{n,-1} - \beta_{0,-1}\right)}{\sqrt{\xi^\top\widehat{\Theta}\widehat{\Sigma}_{\hat{\beta}_n}\widehat{\Theta}\xi}} \le t\right) - \Phi(t)\right| \to 0. \tag{3.6}$$

Moreover, letting $z_{1-\tau/2}$ be the $1-\tau/2$ percentile of the standard normal distribution, one has for all $j = 2, \ldots, p$ and for $\hat{\sigma}^2_{\beta_j}$ defined in Theorem 2,

$$\lim_{n\to\infty} \inf_{\beta_0 \in B(s_\beta, p)} \mathbb{P}\left(\beta_{0,j} \in \left[\tilde{\beta}_{n,j} - z_{1-\tau/2}\frac{\hat{\sigma}_{\beta_j}}{\sqrt{n}}, \tilde{\beta}_{n,j} + z_{1-\tau/2}\frac{\hat{\sigma}_{\beta_j}}{\sqrt{n}}\right]\right) = 1-\tau. \tag{3.7}$$

Remark 3.5. The first result (3.6) in Theorem 3 illustrates that convergence to the standard normal distribution is valid uniformly over the ℓ_0 ball radius at most s_β. As a consequence, as is shown in (3.7), the confidence band is asymptotically honest for $\beta_{0,j}$ uniformly over $B(s_\beta, p)$. We stress that this does not contradict Theorem 2 of Pötscher (2009), who showed that honest confidence bands based on sparse estimators must be large because unlike the LASSO estimator, the de-biased estimator is not a sparse estimator. We also emphasize that the above results add to the previous literature by giving conditions for uniform inference of H-D single-index models. We extend the uniform convergence results in Theorem 3 of Caner and Kock (2018) from H-D linear regression to single-index models.

4 Numeric Results

4.1 Monte Carlo

We conduct a small-scale Monte Carlo simulation to evaluate the finite-sample performance of our estimator. We consider the following two conditional mean specifications:

$$\begin{aligned} \text{Model I:}\ \ \mathbb{E}[Y_i|X_i] &= 3(X_i^\top \beta_0) + 10\sin(X_i^\top \beta_0), \\ \text{Model II:}\ \ \mathbb{E}[Y_i|X_i] &= (X_i^\top \beta_0) + 4\sqrt{|X_i^\top \beta_0 + 1|}. \end{aligned}$$

The first one is from Yang et al. (2017) and the second one comes from Zeng et al. (2011). For both DGPs, we draw the error term from a standard normal distribution independent of all covariates. As the covariates' distribution plays an important role in the estimation

strategy of our approach and that of Yang et al. (2017), we consider deviations from the independent normal design. In particular, we demonstrate the deviation by changing the marginal distribution or by allowing correlation among regressors. We generate 200 random datasets with sample size $n = 500$, and we vary the dimensionality of covariates $p \in \{25, 50, 100, 200\}$. The first five covariates are relevant ones throughout the exercise. The true regression coefficient is generated through the p-dimensional vector $(1, 1, 2, -1, -1, 0, \cdots, 0)$ with its Euclidean norm normalized to be 1.

The coefficient β_0 is estimated based on one of the following four methods: (I) our estimator with the ℓ_1 penalty (AD-Lasso), (II) our estimator with the adaptive ℓ_1 penalty (AD-AdapLasso), (III) the ADE using the oracle score function *without* thresholding (OracleScore), (II) the ADE based on the oracle score function with the ℓ_1 penalty (OracleScore-Lasso). The last one is exactly the LASSO estimator in Yang et al. (2017), so that we use the set of trimming and thresholding parameters suggested therein. To put everything on a comparable basis, we normalize each ADE by taking its norm to be 1 and fixing its first coordinate to be positive under our identification restriction. The reason that we include (III) is that this plain ADE (without any thresholding) is still a legitimate estimator in terms of the ℓ_∞ norm. As a matter of fact, the maximum error of any individual coordinate is of the order $O_p(\sqrt{\log p/n})$ under our sub-Gaussian tail restriction. However, it is not guaranteed to control the overall error for the entire vector in the ℓ_2 norm.

The thresholding parameter λ_β is selected to minimize a fivefold cross-validated empirical risk discussed in Appendix D. We take $\mathcal{X}_j = (F_j^{-1}(0.025), F_j^{-1}(0.975))$ in the trimming function $w_j(\cdot)$ for $j = 1, \cdots, p$. Regarding oracle scores, when the covariates are standard normal, $\dot{l}_j(x_j) = -x_j$ (Plan and Vershynin, 2016). When the marginal covariates follows a t-distribution with five degrees of freedom, i.e., $X_j \sim t(5)$, we have $\dot{l}_j(x_j) = -\frac{6x_j}{5+x_j^2}$ (Yang et al., 2017). For the correlated normal covariates, we consider

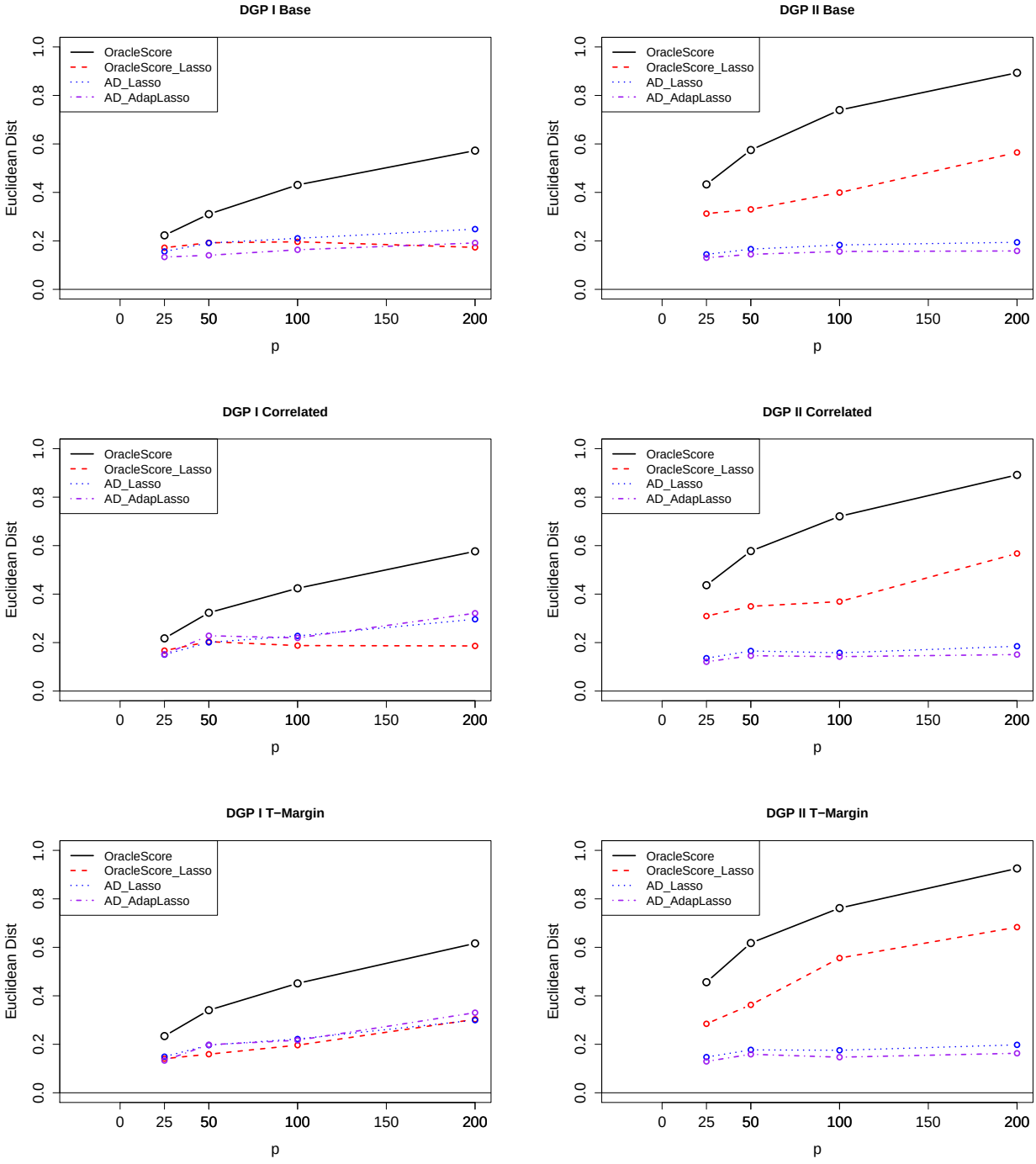
the case with a sparse and banded precision matrix, as follows:

$$\Omega = \begin{bmatrix} 1 & -\rho_0 & 0 & 0 & \cdots & \cdots & 0 \\ -\rho_0 & 1+\rho_0^2 & -\rho_0 & 0 & \cdots & \cdots & 0 \\ 0 & -\rho_0 & 1+\rho_0^2 & -\rho_0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & 0 & -\rho_0 & 1 \end{bmatrix},$$

where $\rho_0 = 0.5$. Its sparsity is self-evident, whereas the banded properties refers to the fact that the sum of the absolute values in each row (or column) is also bounded. According to formula (2.1), it is easy to check that $\dot{l}_j(\boldsymbol{x}) = -\sum_{k=1}^{p} \Omega_{kj} x_k$.

Figure **??** displays the results in terms of $|\hat{\beta}_n - \beta_0|_2$ (Euclidean distance) averaged over simulated samples for four estimators. Of our two estimators, the adaptive Lasso version works better than the plain Lasso one. Nonetheless, the difference is relatively small. In general, OracleScore-Lasso is slightly better than our estimator AD-Lasso and comparable to our estimator AD-AdapLasso in DGP I. Note that all differences are not distinguishable except for the correlated case when $p = 200$. This is not surprising, given the additional errors incurred by estimating the score function in a H-D setup. However, the performance of OracleScore-Lasso deteriorates significantly for DGP II, which is from the simulation design in Zeng et al. (2011). One plausible explanation is that the rule-of-thumb choice for $\lambda_\beta = 4\sqrt{\log p/n}$ in Yang et al. (2017) does not work well for DGP II. A closer examination reveals that it leads to an overly sparse estimate in our simulation. In contrast, the cross-validation proposed in our paper demonstrates robust performance for both DGPs. We conjecture that the better performance of our estimators is most likely due to the cross-validated choice of λ_β instead of the rule-of-thumb choice in Yang et al. (2017). Another notable point is that as the dimensionality scales up, the estimation error measured by the ℓ_2 norm increases only modestly except for the unthresholded estimator, which is again consistent with the theory.

Figure 1: Comparisons of Different ADEs.



4.2 A Real Data Example

As an empirical illustration, we revisit the Oregon health experiment data to explore the effect of health insurance on healthcare use and health outcomes. In 2008, the state of Oregon conducted eight lottery drawings to randomly select names for its Medicaid program from a waiting list of almost 90,000 uninsured, low-income adults. This created a rare opportunity to study the effects of Medicaid coverage for the uninsured on different health outcomes. Approximately two years after the experiment, researchers obtained interview data from adults who were not selected and adults who were selected for the program. This dataset is particularly advantageous for collecting the demographic characteristics of each individual, including age, gender, education, income, and location.

In this study, we focus on the interview data from elderly individuals whose ages range from 50 to 65 and who were eligible for the program. We have 1793 observations in total. Among them, 767 individuals finished the application and got enrolled into Medicaid ($D = 1$), while 1026 individuals did not ($D = 0$). A statistical summary of key variables is reported in Table 1 for the treated group ($D = 1$) and the control group ($D = 0$). Overall the sample is well balanced. For example, the average age was about 56 years for both treated and control groups, and the average number of years of education is about 12.5. For the outcome variables we are interested in, the treated group had less out-of-pocket spending (Spend). Happiness was measured on a scale from 1 (unhappiest) to 3 (happiest). Mental and physical health composite scores were denoted as MCS and PCS, respectively. The control group had on average higher scores for happiness, MCS, and PCS compared with the treatment group.

Prior analyses of this dataset estimated the treatment effect of medical insurance on various clinical outcomes (Finkelstein et al., 2012; Baicker et al., 2013, 2014; Finkelstein et al., 2016). However, previous studies controlled only a small set of covariates. We believe that it is beneficial to control for detailed medical histories and personal charac-

Table 1: *Some descriptive statistics of the survey respondents in OHIE.*

	Control($D=0$)				Treatment ($D=1$)			
	Mean	Std. Error	Min	Max	Mean	Std. Error	Min	Max
				X				
Age	55.798	4.054	50	64	55.597	3.94	50	64
Edu	12.503	2.151	9	16	12.409	2.114	9	16
Weight	1.174	0.44	0.897	7.323	1.146	0.36	0.895	4.026
Health Index	3.692	1.117	1	6	3.408	1.183	1	6
BMI	30.079	7.129	16.896	61.779	30.412	7.856	12.089	79.167
				Y				
Spend	774.161	1755.998	0	32700	406.5268	1495.627	0	27640
Happiness	1.868	0.655	1	3	1.773	0.626	1	3
MCS8	44.972	11.1	12.706	64.922	42.857	12.005	11.466	66.87
PCS8	43.979	10.553	15.139	63.841	40.581	10.711	12.295	65.351
E.D. Visits	0.727	1.808	0	30	1.126	2.407	0	43
Any Doctor Visits	0.693	0.461	0	1	0.847	0.36	0	1

teristics, as well as quadratic and interaction terms of these covariates. We generated 92 control variables based on a second-order polynomial (excluding the constant) from a set of covariates that include medical history and demographic variables, such as income, family size, education level, location, catastrophic expenditures, and existing borrowed or skipped bills.

To analyze the data, we consider the single-index model

$$Y = m\left(\alpha D + \beta_1^\top X_1 + \beta_2^\top D X_1 + \beta_3^\top X_2\right) + \varepsilon, \tag{4.1}$$

in which Y represents the different outcomes of interest, X_1 contains age and education, and X_2 contains the rest of the covariates. The treatment effects of Medicaid coverage on various outcomes of interest are reported in Table 2. In this study, we are interested in the heterogeneous effect of health insurance on the different outcomes of interest. We are particularly interested in the coefficients of D, $D * age$, and $D * education$. To estimate the coefficient of the treatment variable D, we apply the method discussed in Remark

Table 2: *Heterogeneous treatment effects of Medicaid coverage on outcomes.*

	Spend	Happiness	MCS8	PCS8	Hospital Visits	E.D. Visits	Any Doctor Visits
D	-9.887	0.230	102.600	3.073	1.212	0.205	0.000
	(0.010)	(0.032)	(8.208)	(12.759)	(0.191)	(0.688)	(0.001)
D*age	0.198	0.195	0.488	0.154	0.001	0.313	-0.132
	(0.001)	(0.053)	(3.126)	(1.356)	(0.032)	(5.764)	(1.491)
D*edu	0.016	0.339	14.691	2.575	0.832	0.056	0.046
	(3.215)	(0.038)	(0.106)	(11.831)	(1.281)	(0.939)	(0.032)
age	0.166	0.005	0.002	0.413	0.323	0.055	0.003
	(0.030)	(0.022)	(0.466)	(0.989)	(3.141)	(0.442)	(0.001)
edu	0.010	0.009	0.020	0.010	3.093	-0.011	0.018
	(0.014)	(0.031)	(0.751)	(0.127)	(0.121)	(0.002)	(0.011)
$\hat{s}_\beta$	31	37	32	36	32	39	33

2.2 . The standard errors are presented in parentheses, which are constructed from the de-biased estimator. Furthermore, $\hat{s}_\beta$ represents the number of significant covariates selected by the Benjamini-Hochberg multiple testing procedure on the p-values of each coefficient from the proposed de-biased estimator (4.10) at the 10% level for FDR. From Table 2, it can be seen that Medicaid coverage led to lower out-of-pocket spending, a higher level of happiness, and higher MCS scores, indicating overall improvements in the economic well-being and mental health of the participants. Notably, we also find that the effect on spending decreases with an increase in age, while the effect on happiness increases with an increase in age and education. Note that for the self-reported levels of happiness, previous studies using the same data failed to find a significant overall effect of Medicaid coverage, which was arguably a measure of overall subjective well-being (Baicker et al., 2013), so our findings are new to the literature.

5 Conclusion

The single-index model provides a useful middle ground between the fully parametric and nonparametric models, as it combines the strengths of traditional parametric regression with the enhanced flexibility of nonparametric techniques. Despite its original purpose of achieving sufficient dimension reduction, the empirical applications of a single-index model with moderate- or high-dimensional covariates are rare. Addressing this challenge, we propose a simple two-stage semiparametric estimation method of H-D single-index models exploring the average derivatives under a Gaussian copula design of covariates. Its implementation involves only convex minimization problems with straightforward computation. We present a non-asymptotic bound of our estimator that carefully accounts for the estimation uncertainty of marginal features and a H-D precision matrix. Furthermore, we develop de-biased inference of an one-step updated estimator using a Newton-Raphson correction that also allows for general pilot estimators. A small-scale Monte Carlo simulation compares our estimator with that of Yang et al. (2017) and demonstrates its competitive performance. We also illustrate the inferential procedure with a real data application.

References

Baicker, K., Finkelstein, A., Song, J., and Taubman, S. (2014). The impact of medicaid on labor market activity and program participation: evidence from the oregon health insurance experiment. *American Economic Review*, 104(5):322–328.

Baicker, K., Taubman, S. L., Allen, H. L., Bernstein, M., Gruber, J. H., Newhouse, J. P., Schneider, E. C., Wright, B. J., Zaslavsky, A. M., and Finkelstein, A. N. (2013). The oregon experiment—effects of medicaid on clinical outcomes. *New England Journal of Medicine*, 368(18):1713–1722.

Belloni, A., Chernozhukov, V., and Kato, K. (2015). Uniform post-selection inference for leaset absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Cai, Tony, T., Liu, W., and Zhou, H. (2016). Estimation sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Annals of Statistics*, 44:455–488.

Caner, M. and Kock, A. B. (2018). Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. *Journal of Econometrics*, 203(1):143–168.

Cattaneo, M. D., Crump, R. K., and Jansson, M. (2013). Generalized jackknife estimators of weighted average derivatives. *Journal of the American statistical Association*, 108:1243–1256.

Chakrabortty, A., Lu, J., Cai, T. T., and Li, H. (2019). High dimensional m-estimation with missing outcomes: A semi-parametric framework. *arXiv preprint arXiv:1911.11345*.

Chaudhuri, P., Doksum, K., and Samarov, A. (1997). On average derivative quantile regression. *Annals of Statistics*, 25:715–744.

Chen, X., Fan, Y., and Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association*, 101(475):1228–1240.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352.

Chernozhukov, V., Newey, W., and Singh, R. (2020). De-biased machine learning of global and local parameters using regularized riesz representers. *arXiv preprint arXiv:1802.08667*.

Dai, R., Song, H., Barber, R., and Raskutti, G. (2021). Convergence guarantee for the sparse monotone single index model. *arxiv preprint*, arxiv:2105.07587v1.

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

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and Group, O. H. S. (2012). The oregon health insurance experiment: evidence from the first year. *Quarterly Journal of Economics*, 127(3):1057–1106.

Finkelstein, A. N., Taubman, S. L., Allen, H. L., Wright, B. J., and Baicker, K. (2016). Effect of medicaid coverage on ed use—further evidence from oregon's experiment. *New England Journal of Medicine*, 375(16):1505–1507.

Giné, E. and Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press.

Gu, L. and Yang, L. (2015). Oracally efficient estimation for single-index link function with simultaneous confidence band. *Electronic Journal of Statistics*, 9:1540–1561.

Gu, Q., Cao, Y., Ning, Y., and Liu, H. (2015). Local and global inference for high dimensional gaussian copula graphical models. *arXiv preprint*, arXiv:1502.02347.

Hardle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association*, 84:986–995.

Hirshberg, D. A. and Wager, S. (2020). Debiased inference of average partial effects in single-index models: comment on wooldridge and zhu. *Journal of Business & Economic Statistics*, 38:19–24.

Horowitz, J. and Hardle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American statistical Association*, 91:1632–1640.

Horowitz, J. L. (2009). *Semiparametric and nonparametric methods in econometrics*. Springer.

Hristache, M., Juditsky, A., and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *The Annals of Statistics*, 29:595–623.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58.

Jankova, J. and Van De Geer, S. (2018). Semiparametric efficiency bounds for high-dimensional models. *The Annals of Statistics*, 46(5):2336–2359.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

Klaassen, C. A. and Wellner, J. A. (1997). Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, 3(1):55–77.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, 61(2):387–421.

Kuchibhotla, A. K. and Chakrabortty, A. (2018). Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*.

Li, Q., Lu, X., and Ullah, A. (2003). Multivariate local polynomial regression for estimating average derivatives. *Journal of Nonparametric Statistics*, 15:607–624.

Lin, Q., Li, X., Huang, D., and Liu, J. (2021). On the optimality of sliced inverse regression in high dimensions. *Annals of Statistics*, 49:1–20.

Lin, Q., Zhao, Z., and Liu, J. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *Annals of Statistics*, 46:580–610.

Lin, Q., Zhao, Z., and Liu, J. (2019). Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association*, 114:1726–1739.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric gaussian copula graphical models. *Annals of Statistics*, 40:2293–2326.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *ournal of Machine Learning Research*, 10:2295–2328.

Loh, P. L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *Annals of Statistics*, 45:866–896.

Ma, S. and He, X. (2016). Inference for single-index quantile regression models with profile optimization. *The Annals of Statistics*, 44(3):1234–1268.

McCullagh, P. and Nelder, J. (1983). *Generalized linear models*. CRC.

Nekipelov, D., Semenova, V., and Syrgkanis, V. (2020). Regularized orthogonal machine learning for nonlinear semiparametric models. *arXiv preprint*, 1806.04823v6.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195.

Peng, H. and Huang, T. (2011). Penalized least squares for single index models. *Journal of Statistical Planning and Inference*, 141:1362–1379.

Plan, Y. and Vershynin, R. (2016). The generalized lasso with non-linear observations. *IEEE Transactions on Information Theory*, 62:1528–1537.

Pötscher, B. M. (2009). Confidence sets based on sparse estimators are necessarily large. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, pages 1–18.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57:1403–1430.

Radchenko, P. (2015). High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282.

Segers, J., van den Akker, R., and Werker, B. J. (2014). Semiparametric gaussian copula models: Geometry and efficient rank-based estimation. *Annals of Statistics*, 42(5):1911–1940.

Su, L. and Zhang, Y. (2014). Variable selection in nonparametric and semiparametric regression models. In *The Oxford handbook of applied nonparametric and semiparametric econometrics and statistics*, pages 249–307. Oxford University Press.

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

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, 22:1112–1137.

Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64:363–410.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Annals of Statistics*, 40:2541–2571.

Yang, Z., Balasubramanian, K., and Liu, H. (2017). On stein's identity and near-optimal estimation in high-dimensional index models. *International Conference on Machine Learning*, pages 3851–3860.

Zeng, P., He, T., and Zhu, Y. (2011). A lasso-type approach for estimation and variable selection in single index models. *Journal of Computational and Graphical Statistics*, 21:92–109.

Zhang, C.-H. and Zhang, 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.

Supplementary Material to "Estimation and Inference of Semiparametric Single-index Models with High-Dimensional Covariates"

Ruixuan Liu and Jing Tao
Emory University and University of Washington

Appendix A: Proofs of Main Results

In this section, we first prove Theorem 1 about the non-asymptotic bound of our ADE. Then we establish the asymptotic normality of the de-biased estimator in the proof of Theorem 2. For any vector $z \in \mathbb{R}^p$ and an index set $\mathcal{A}$, we define the restriction of z to $\mathcal{A}$ by letting $[z_{\mathcal{A}}]_j = z_j \quad \text{if} \quad j \in \mathcal{A}, \quad [z_{\mathcal{A}}]_j = 0 \quad \text{otherwise}$. To simplify the exposition, we define

$$\chi_j(X_i) := \Phi^{-1}(F_j(X_{i,j})) - \sum_{k=1}^{p} \Omega_{kj}\Phi^{-1}(F_k(X_{i,k})),$$
$$\upsilon_j(X_i) := \frac{\chi_j(X_i)}{\phi(\Phi^{-1}(F_j(X_{i,j})))}, \text{ and } \zeta_j(X_{i,j}) := \frac{f_j(X_{i,j})}{\phi(\Phi^{-1}(F_j(X_{i,j})))}.$$

Proof of Theorem 1. Our ADE is the solution of the following minimization problem:

$$\hat{\beta}_n^* = \arg\min_{\beta}[L(\beta) + \lambda_\beta|\beta|_1] := \arg\min_{\beta}\left[\beta\beta^\top - 2n^{-1}\sum_{i=1}^{n} Y_i w(X_i)\hat{\dot{l}}_n^\top(X_i)\beta + \lambda_\beta|\beta|_1\right],$$

with the derivative vector $\nabla L(\beta_0^*) = \check{\beta}_n^* - \beta_0^*$, where the non-sparse estimator $\check{\beta}_n^* = (\check{\beta}_{n,1}^*, \cdots, \check{\beta}_{n,p}^*)^\top$ is defined by

$$\check{\beta}_{n,j}^* := n^{-1}\sum_{i=1}^{n} Y_i w_j(X_i)\hat{\dot{l}}_j(X_i), \ j = 1, \cdots, p. \tag{S.0.1}$$

Because the Hessian matrix of the sample criterion function $L(\cdot)$ is $\mathbb{I}_{p\times p}$, the restricted eigenvalue condition is automatically satisfied, cf. Section 7.3.1 in Wainwright (2019).

We seek a non-asymptotic bound for $\nabla L(\beta_0^*) = \check{\beta}_n^* - \beta_0^*$. For this purpose, we consider the following decomposition:

$$\check{\beta}_{n,j}^* - \beta_{0,j}^* = T_{n,j} + U_{n,j} + V_{n,j} + W_{n,j} + R_{n,j}, \tag{S.0.2}$$

where

$$\begin{aligned}
T_{n,j} :=& n^{-1}\sum_{i=1}^{n} Y_i w_j(X_i)\dot{l}_j(X_i) - \mathbb{E}[Y w_j(X)\dot{l}_j(X)],\\
U_{n,j} :=& \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i w_j(X_i)}{f_j(X_{i,j})}\left[\hat{f}_j'(X_{i,j}) - f_j'(X_{i,j})\right] - \sum_{i=1}^{n}\frac{Y_i w_j(X_i) f_j'(X_{i,j})}{f_j^2(X_{i,j})}\left[\hat{f}_j(X_{i,j}) - f_j(X_{i,j})\right]\\
&+ \frac{2}{n}\sum_{i=1}^{n} Y_i w_j(X_i) v_j(X_i)\left[\hat{f}_j(X_{i,j}) - f_j(X_{i,j})\right],\\
V_{n,j} :=& \frac{1}{n}\sum_{i=1}^{n} Y_i w_j(X_i) f_j(X_{i,j})\left[\frac{\Phi^{-1}(F_j(X_{i,j}))}{\phi(\Phi^{-1}(F_j(X_{i,j})))} - \frac{\Phi^{-1}(\hat{F}_j(X_{i,j}))}{\phi(\Phi^{-1}(\hat{F}_j(X_{i,j})))}\right]\\
&+ \frac{1}{n}\sum_{i=1}^{n} Y_i w_j(X_i)\zeta_j(X_{i,j})\left[\sum_{k=1}^{p}\Omega_{kj}\left(\Phi^{-1}(F_k(X_{i,k})) - \Phi^{-1}(\hat{F}_k(X_{i,k}))\right)\right],\\
W_{n,j} :=& \frac{1}{n}\sum_{i=1}^{n} Y_i w_j(X_i)\zeta_j(X_{i,j})\left[\sum_{k=1}^{p}\left(\hat{\Omega}_{kj} - \Omega_{kj}\right)\Phi^{-1}(F_k(X_{i,k}))\right],
\end{aligned}$$

whereas the negligible remainder term $R_{n,j}$ is presented in Section 0.2.

We provide an outline here and delegate detailed proofs to Section 0.2. First, $T_{n,j}$ is an sample average of *i.i.d.* terms which can be handled by a standard maximal inequality given the sub-Gaussian assumption. Second, we apply the Hoeffding decomposition to U-statistics derived from $U_{n,j}$ and then the maximal inequalities from Kuchibhotla and Chakrabortty (2018) and Major (2006) to bound the linear and quadratic terms. We also bound the bias term uniformly as in Hardle and Stoker (1989). The order of $V_{n,j}$ is determined by the supnorm bound for the marginal empirical distribution. We also make use of the sparsity of (Ω_{kj}) in handling the second term of $V_{n,j}$. $W_{n,j}$ is a linear

combination involving the precision matrix, so that we rely on the sharp bound in terms of the L_1 operator norm for the CLIME (Cai et al., 2011). To that end, we prove that with probability $1 - O(p^{-1})$, the following holds

$$\max_{1\leq j\leq p} |\check{\beta}^*_{n,j} - \beta^*_{0,j}| \lesssim s_\Omega \sqrt{\frac{\log p}{n}} + \max_{1\leq j\leq p} h^r_j + M_\Omega^{1-q_1} c_\Omega \left(\frac{\log p}{n}\right)^{(1-q_1)/2}. \tag{S.0.3}$$

Given our bound for $\max_{1\leq j\leq p} |\check{\beta}^*_{n,j} - \beta^*_{0,j}|$, the rest proof essentially follows from an analogous argument in Theorem 3.2 of Yang et al. (2017). We include the proof for completeness. The first-order condition of the optimization problem leads to

$$\nabla L(\hat{\beta}^*_n) + \lambda_\beta z = 0, \quad \text{where} \quad z \in \partial |\hat{\beta}^*_n|_1. \tag{S.0.4}$$

Each coordinate of z is given by

$$z_j = sign(\hat{\beta}^*_{n,j}), \quad \forall j \in supp(\hat{\beta}^*_n); \quad z_j \in [-1, 1], \quad \forall j \notin supp(\hat{\beta}^*_n).$$

Let $\mathcal{S} = supp(\beta^*_0)$ for the true β^*_0 and we write $\beta = \beta_\mathcal{S} + \beta_{\mathcal{S}^c}$ for any vector β. Then it holds

$$\langle \nabla L(\hat{\beta}^*_n) - \nabla L(\beta^*_0), \theta \rangle \leq -\lambda \langle z_\mathcal{S} + z_\mathcal{S}, \theta \rangle + |\nabla L(\beta^*_0)|_\infty |\theta|_1, \tag{S.0.5}$$

where $\theta = \hat{\beta}^*_n - \beta^*_0$. By the characterizing property of z, we get

$$-\lambda_\beta \langle z_{\mathcal{S}^c}, \theta \rangle = |\theta_\mathcal{S}|_1 \quad \text{and} \quad -\lambda_\beta \langle z_\mathcal{S}, \theta \rangle \leq |\theta_\mathcal{S}|_1,$$

given that $|z|_\infty \leq 1$. Combining previous results together, we get

$$|\theta|_2^2 = \langle \nabla L(\hat{\beta}^*_n) - \nabla L(\beta^*_0), \theta \rangle \leq \lambda_\beta |\theta_{\mathcal{S}^c}|_1 + \lambda_\beta |\theta_\mathcal{S}|_1 + |\nabla L(\beta^*_0)|_\infty |\theta|_1.$$

Given our assumption that $\lambda_\beta > |\nabla L(\beta^*_0)|_\infty$, one arrives at

$$|\theta|_2^2 \leq -\lambda_\beta/2 |\theta_{\mathcal{S}^c}|_1 + 3\lambda_\beta/2 |\theta_\mathcal{S}|_1 \leq 2\lambda_\beta |\theta_\mathcal{S}|_1.$$

Given $|\theta|_2^2 \geq 0$, we first get $|\theta_{\mathcal{S}^c}|_1 \leq |\theta_\mathcal{S}|_1$. Furthermore, we obtain $|\theta_\mathcal{S}|_1 \leq \sqrt{s_\beta} |\theta_\mathcal{S}|_2$ and $|\theta_\mathcal{S}|_2 \leq |\theta|_2$, which allows us to conclude that $|\theta|_2 \leq \sqrt{s_\beta} \lambda_\beta$. The bound in terms of the l_1 norm can be shown in a similar way; see e.g. Theorem 7.13 of Wainwright (2019). $\square$

Proof of Corollary 1. Under our normalization scheme, we proceed with

$$\hat{\beta}_n - \beta_0 = \frac{\hat{\beta}_n^* sign(\hat{\beta}_{n,1}^*)}{|\hat{\beta}_n^*|_2} - \frac{\hat{\beta}_n^* sign(\beta_{0,1}^*)}{|\hat{\beta}_n^*|_2} + \frac{\hat{\beta}_n^* sign(\beta_{0,1}^*)}{|\hat{\beta}_n^*|_2} - \frac{\beta_0^*}{|\beta_0^*|_2}. \tag{S.0.6}$$

Because $\beta_{0,1}^*$ is strictly positive, then for some positive constant terms c_0 and c_1, the events $\{\hat{\beta}_{n,1}^* > c_0\}$ and $\{|\hat{\beta}_{n,1}^*|_2 > c_1\}$ hold simultaneously with probability at least $1 - O(p^{-1})$, given the conclusion from Theorem 1. Thus, the first term on the r.h.s. of (S.0.6) is exactly zero and the second term is of $O_p\left(\sqrt{s_\beta}\left(\sqrt{\frac{\log p}{n}} + \max_j h_j^r + M_\Omega^{1-q_1} c_\Omega \left(\frac{\log p}{n}\right)^{(1-q_1)/2}\right)\right)$, with probability at least $1 - O(p^{-1})$. □

Proof of Theorem 2. We first establish the following decomposition:

$$\sqrt{n}\left(\tilde{\beta}_{n,j} - \beta_{0,j}\right)/\sigma_{\beta,j} = \mathbb{Z}_n + rem_j, \tag{S.0.7}$$

where $\mathbb{Z}_n \Longrightarrow \mathbb{N}(0,1)$ and $|rem_j| = o_p(1)$. Then we prove the plug-in estimation $\hat{\sigma}_{\beta,j}$ is consistent.

By definition of the de-biased estimator, we have

$$\tilde{\beta}_{n,-1} - \beta_{0,-1} = \widehat{\Theta}\hat{\Psi}_n(\beta_0) + (\hat{\beta}_{n,-1} - \beta_{0,-1})\left[\mathbb{I}_{(p-1)\times(p-1)} - \widehat{\Theta}\hat{\dot{\Psi}}_n(\beta_0)\right] + O_p(|\hat{\beta}_n - \beta_0|_2^2),$$

where $|\hat{\beta}_n - \beta_0|_2^2 = o_p(n^{-1/2})$ from Corollary 1 and Assumption 10. The second term on the r.h.s. of the above equality satisfies

$$\begin{aligned}
&\| \left(\mathbb{I}_{(p-1)\times(p-1)} - \widehat{\Theta}\hat{\dot{\Psi}}_n(\beta_0)\right)\left(\hat{\beta}_{n,-1} - \beta_{0,-1}\right) \|_\infty \leq \|\mathbb{I}_{(p-1)\times(p-1)} - \widehat{\Theta}\hat{\dot{\Psi}}_n(\beta_0)\|_\infty |\hat{\beta}_n - \beta_0|_1 \\
&= O_p\left(\lambda_\Psi |\hat{\beta}_n - \beta_0|_1\right) = o_p(n^{-1/2}),
\end{aligned}$$

where the second equality follows from Lemma 10. Referring to the first term, Lemma 14 in Appendix S.2 proves the following result:

$$\hat{\Psi}_n(\beta_0) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i m_0'(X_i^\top \beta_0)(X_{i,-1} - \mathbb{E}[X_{i,-1}|X_i^\top \beta_0]) + o_p(n^{-1/2}). \tag{S.0.8}$$

It suffices to verify that

$$\mathbf{e}_j^\top \Theta_o \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i m_0'(X_i^\top \beta_0)(X_i - \mathbb{E}[X_i|X_i^\top \beta_0]) \Longrightarrow \mathbb{N}(0, \sigma_{\beta,j}^2). \tag{S.0.9}$$

Note that for $r_\varepsilon > 4$,

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{E}\left[\mathbf{e}_j^\top m_0'(X_i^\top \beta_0)\tilde{X}_i\varepsilon_i/\sqrt{n}\right]^{r_\varepsilon/2} \leq \sum_{i=1}^n \mathbb{E}\left[\|\mathbf{e}_j\|_1 \|m_0'(X_i^\top \beta_0)\tilde{X}_i\varepsilon_i/n\|_\infty\right]^{r_\varepsilon/2} \\
&\leq \sum_{i=1}^n \left(\frac{s_\beta}{\sqrt{n}}\right)^{r_\varepsilon/2} \max_{1\leq k\leq p} \mathbb{E}\left[|m_0'(X_i^\top \beta_0)|^{r_\varepsilon/2}\right] = O_p\left(\frac{s_\beta^{r_\varepsilon/2}}{n^{r_\varepsilon/4-1}}\right)
\end{aligned}$$

and

$$\left(\mathbf{e}_j^\top \Theta_o \Sigma_{\beta_0} \Theta_o \mathbf{e}_j\right)^{r_\varepsilon/4} \geq \left[\|\mathbf{e}_j\|_2^2 \lambda_{\min}(\Omega_\beta)\lambda_{\min}^2(\Theta_o)\right]^{r_\varepsilon/4} = p^{r_\varepsilon/2}[\lambda_{\min}(\Sigma_{\beta_0})\lambda_{\max}(\dot{\Psi}(\beta_0))^{-2}]^{r_\varepsilon/4} > 0$$

which is bounded away from zero since $\lambda_{\min}(\Sigma_{\beta_0})$ is bounded away from zero and $\lambda_{\max}(\dot{\Psi}(\beta_0))$ is bounded from above. By invoking the Lyapunov central limit theorem, we establish (S.0.9).

Because $\Theta_o \in \mathcal{G}_q(c_\Theta, M_\Theta)$, we proceed as follows

$$\begin{aligned}
&\|\widehat{\Theta} - \Theta_o\|_\infty \leq \|\Theta_o\|_1 \|\mathbb{I}_{(p-1)\times(p-1)} - \widehat{\Theta}\dot{\Psi}_0\|_\infty \\
\leq &\|\Theta_o\|_1 \|\mathbb{I}_{(p-1)\times(p-1)} - \widehat{\Theta}\hat{\dot{\Psi}}_n(\hat{\beta}_n)\|_\infty + \|\Theta_o\|_1 \|\hat{\Theta}\|_1 \|\hat{\dot{\Psi}}_n(\hat{\beta}_n) - \dot{\Psi}_0\| \\
= &O_p\left(M_\Theta \lambda_\Psi + M_\Theta^2\left(\frac{c_{1n}^2}{h^3} + \sqrt{\frac{\log np}{nh^2}} + h^2 + c_{1n}\right)\right),
\end{aligned}$$

where in the last inequality we use Lemma 10.

Moreover, by the fact that Σ_{β_0} and Θ_o are symmetric and $\mathbf{e}_j^\top \mathbf{e}_j = 1$, then

$$|\mathbf{e}_j^\top \widehat{\Theta}\Sigma_{\beta_0}\widehat{\Theta}^\top \mathbf{e}_j - \mathbf{e}_j^\top \Theta_o \Sigma_{\beta_0}\Theta_o^\top \mathbf{e}_j| \leq \lambda_{\max}^2(\Sigma_{\beta_0})\|\left(\widehat{\Theta} - \Theta_o\right)^\top \mathbf{e}_j\|_2^2 + 2\|\Sigma_{\beta_0}\Theta_o^\top \mathbf{e}_j\|_2 \|\left(\widehat{\Theta} - \Theta_o\right)\mathbf{e}_j\|_2.$$

Because

$$\|\Sigma_{\beta_0}^\top \Theta_o \mathbf{e}_j\|_2 \leq \sqrt{\lambda_{\max}(\Sigma_{\beta_0})^2 \mathbf{e}_j^\top \Theta_o \Theta_o \mathbf{e}_j} = O(1)$$

from Assumption 12 and

$$\| \left(\widehat{\Theta} - \Theta_o \right) \mathbf{e}_j \|_2 \leq \| \widehat{\Theta} - \Theta_o \|_\infty = O_p(M_\Theta \lambda_\Psi),$$

we have

$$|\mathbf{e}_j^\top \widehat{\Theta} \Sigma_{\beta_0} \widehat{\Theta}^\top \mathbf{e}_j - \mathbf{e}_j^\top \Theta_o \Sigma_{\beta_0} \Theta_o \mathbf{e}_j| = O_p \left(M_\Theta \lambda_\Psi \right).$$

Furthermore,

$$\left| \mathbf{e}_j^\top \widehat{\Theta} \widehat{\Sigma}_{\hat{\beta}_n} \widehat{\Theta} \mathbf{e}_j - \mathbf{e}_j^\top \widehat{\Theta} \Sigma_{\beta_0} \widehat{\Theta} \mathbf{e}_j \right| \leq \left| \mathbf{e}_j^\top \widehat{\Theta} \right|_1^2 \| \widehat{\Sigma}_{\hat{\beta}_n} - \Sigma_{\beta_0} \|_\infty$$

where $\left| \mathbf{e}_j^\top \widehat{\Theta} \right|_1^2 \leq \left| \mathbf{e}_j^\top \Theta_o \right|_1^2 \leq c_\Theta^2$ with probability approaching one because $\widehat{\Theta}$ is the optimal solution to Algorithm 2.13 and

$$\| \widehat{\Sigma}_{\hat{\beta}_n} - \Sigma_{\beta_0} \|_\infty = O_p \left(\frac{c_{1n}^2}{h^3} + \sqrt{\frac{\log np}{nh^2}} + h^2 + c_{1n} \right).$$

from Lemma 16 in Appendix S.2. Thus,

$$|\mathbf{e}_j^\top \widehat{\Theta} V_\beta \widehat{\Theta}^\top \mathbf{e}_j - \mathbf{e}_j^\top \Theta V_\beta \Theta^\top \mathbf{e}_j| = O_p(M_\Theta \lambda_\Psi + c_\Theta^2 \| \widehat{\Sigma}_{\hat{\beta}_n} - \Sigma_{\beta_0} \|_\infty) = o_p(1),$$

where the last equality follows from Assumption 14. □

Proof of Theorem 3. Let

$$\begin{aligned} \Delta =& \sqrt{n}(\hat{\beta}_{n,-1} - \beta_{0,-1}) \left[\mathbb{I}_{(p-1)\times(p-1)} - \widehat{\Theta} \hat{\dot{\Psi}}_n(\beta_0) \right] \\ &+ \widehat{\Theta} \left(\hat{\Psi}_n(\beta_0) - \mathbb{E}_n[m_0'(X_i^\top \beta_0) \tilde{X}_i \varepsilon_i] \right) + R(\hat{\beta}_n, \beta_0), \end{aligned}$$

where $R(\hat{\beta}_n, \beta_0)$ is the remainder term of order $O(|\hat{\beta}_n - \beta_0|_2^2)$.

Let $\hat{w} = \widehat{\Theta} \xi$ and $w_0 = \Theta \xi$. For $\epsilon > 0$, we consider the following events:

$$\mathcal{U}_{1n} := \left\{ \sup_{\beta \in B(s_\beta, p)} |\Delta| < \epsilon \right\}, \quad \mathcal{U}_{2n} := \left\{ \sup_{\beta \in B(s_\beta, p)} \left| \frac{\sqrt{\hat{w}^\top \widehat{\Sigma}_{\hat{\beta}_n} \hat{w}}}{\sqrt{w_0^\top \Sigma_{\beta_0} w_0}} - 1 \right| < \epsilon \right\}.$$

Note that

$$
\begin{aligned}
&\left|\Pr\left(\frac{\sqrt{n}\left(\xi^\top\tilde{\beta}_{n,-1}-\xi^\top\beta_{0,-1}\right)}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}}\le t\right)-\Phi(t)\right| \\
&=\left|\Pr\left(\frac{\sqrt{n}w_0^\top\mathbb{E}_n\left[m_0'(X_i^\top\beta_0)\tilde{X}_i\varepsilon_i\right]}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}}-\frac{\Delta}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}}\le t\right)-\Phi(t)\right| \\
&=\left|\Pr\left(\frac{\sqrt{n}w_0^\top\mathbb{E}_n\left[m_0'(X_i^\top\beta_0)\tilde{X}_i\varepsilon_i\right]}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}}-\frac{\Delta}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}}\le t,\mathcal{U}_{1n},\mathcal{U}_{2n}\right)-\Phi(t)\right|+\mathbb{P}\left(\mathcal{U}_{1n}^c\cup\mathcal{U}_{2n}^c\right) \\
&:=P_1(\beta_0)+P_2(\beta_0),
\end{aligned}
$$

where for a constant $D>0$,

$$
\begin{aligned}
&\sup_{\beta_0\in B(s_\beta,p)}P_1(\beta_0) \\
&=\sup_{\beta_0\in B(s_\beta,p)}\left|\Pr\left(\frac{\sqrt{n}w_0^\top\mathbb{E}_n\left[m_0'(X_i^\top\beta_0)\tilde{X}_i\varepsilon_i\right]}{\sqrt{w_0^\top\Sigma_{\beta_0}w_0}}-\frac{\Delta}{\sqrt{w_0^\top\Sigma_{\beta_0}w_0}}\le\frac{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}}{\sqrt{w_0^\top\Sigma_{\beta_0}w_0}}t,\mathcal{U}_{1n},\mathcal{U}_{2n}\right)-\Phi(t)\right| \\
&\le\sup_{\beta_0\in B(s_\beta,p)}\left|\Pr\left(\frac{\sqrt{n}w_0^\top\mathbb{E}_n\left[m_0'(X_i^\top\beta_0)\tilde{X}_i\varepsilon_i\right]}{\sqrt{w_0^\top\Sigma_{\beta_0}w_0}}\le t(1+\epsilon)+D\epsilon\right)-\Phi(t)\right|,
\end{aligned}
$$

where the second inequality follows because the definition of $\mathcal{U}_{1n}$, $\mathcal{U}_{2n}$ and the fact that $\sqrt{w_0^\top\Sigma_{\beta_0}w_0}$ does not depend on β_0 and is bounded away from zero.

We have established the asymptotic normality of $\frac{\sqrt{n}w_0^\top\mathbb{E}_n\left[\varepsilon_i m_0'(X_i^\top\beta_0)\tilde{X}_i\right]}{\sqrt{w_0^\top\Sigma_{\beta_0}w_0}}$ in Theorem 2. Hence for n sufficiently large, we can choose ϵ sufficiently small to conclude that for any $\iota>0$,

$$
\inf_{\beta_0\in B(s_0,p)}\Pr\left(\frac{\sqrt{n}w_0^\top\mathbb{E}_n\left[m_0'(X_i^\top\beta_0)\tilde{X}_i\varepsilon_i\right]}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}}-\frac{\Delta}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}}\le t,\mathcal{U}_{1n},\mathcal{U}_{2n}\right)\ge\Phi(t)-2\epsilon-\iota. \tag{S.0.10}
$$

A similar argument implies that there exists a positive constant D' such that

$$\Pr\left(\frac{\sqrt{n}w_0^\top \mathbb{E}_n\left[m_0'(X_i^\top\beta_0)\tilde{X}_i\varepsilon_i\right]}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}} - \frac{\Delta}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}} \le t, \mathcal{U}_{1n}, \mathcal{U}_{2n}\right)$$
$$\ge \Pr\left(\frac{\sqrt{n}w_0^\top \mathbb{E}_n\left[m_0'(X_i^\top\beta_0)\tilde{X}_i\varepsilon_i\right]}{\sqrt{w_0^\top\Sigma_{\beta_0}w_0}} \le t(1-\epsilon) - D'\epsilon, \mathcal{U}_{1n}, \mathcal{U}_{2n}\right) + \mathbb{P}\left(\mathcal{U}_{1n}\cap\mathcal{U}_{2n}\right) - 1.$$

Since $\Pr\left(\mathcal{U}_{1n}\cap\mathcal{U}_{2n}\right)$ can be made arbitrarily close to one when n is sufficiently large by the argument in the proof of Theorem 2, we have

$$\inf_{\beta_0\in B(s_0,p)} \Pr\left(\frac{\sqrt{n}w_0^\top \mathbb{E}_n\left[m_0'(X_i^\top\beta_0)\tilde{X}_i\varepsilon_i\right]}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}} - \frac{\Delta}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}} \le t, \mathcal{U}_{1n}, \mathcal{U}_{2n}\right)$$
$$\ge \Pr\left(\frac{\sqrt{n}w_0^\top \mathbb{E}_n\left[m_0'(X_i^\top\beta_0)\tilde{X}_i\varepsilon_i\right]}{\sqrt{w_0^\top\Sigma_{\beta_0}w_0}} \le t(1-\epsilon) - D'\epsilon\right) - \epsilon.$$

Because of the asymptotic normality established in Theorem 2, for n sufficiently large and any $\iota > 0$, we can choose ϵ sufficiently small to have that

$$\inf_{\beta_0\in B(s_0,p)} \Pr\left(\frac{\sqrt{n}w_0^\top \mathbb{E}_n\left[m_0'(X_i^\top\beta_0)\tilde{X}_i\varepsilon_i\right]}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}} - \frac{\Delta}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}} \le t, \mathcal{U}_{1n}, \mathcal{U}_{2n}\right) \ge \Phi(t) - 2\epsilon - \iota.$$

Because $\sup_{\beta_0\in B(s_\beta,p)} \Pr\left(\mathcal{U}_{1n}^c \cup \mathcal{U}_{2n}^c\right) \to 0$, it implies that

$$\sup_{\beta_0\in B(s_\beta,p)} \left|\Pr\left(\frac{\sqrt{n}\left(\xi^\top\tilde{\beta}_{n,-1} - \xi^\top\beta_{0,-1}\right)}{\sqrt{\hat{w}^\top\widehat{\Sigma}_{\hat{\beta}_n}\hat{w}}} \le t\right) - \Phi(t)\right| \to 0,$$

which is (3.6).

Next, we show (3.7). Observe that

$$\Pr\left(\beta_{0,j} \notin \left[\tilde{\beta}_{n,j} - z_{1-\tau/2}\frac{\hat{\sigma}_{\beta,j}}{\sqrt{n}}, \tilde{\beta}_{n,j} + z_{1-\tau/2}\frac{\hat{\sigma}_{\beta,j}}{\sqrt{n}}\right]\right)$$
$$\le 1 - \Pr\left(\frac{\sqrt{n}(\tilde{\beta}_{n,j} - \beta_{0,j})}{\hat{\sigma}_{\beta,j}} > z_{1-\tau/2}\right) + \Pr\left(\frac{\sqrt{n}(\tilde{\beta}_{n,j} - \beta_{0,j})}{\hat{\sigma}_{\beta,j}} \le -z_{1-\tau/2}\right).$$

Then taking the supremum over $\beta_0 \in B(s_\beta, p)$ and letting n go to infinity yields that

$$\lim_{n\to\infty} \inf_{\beta_0 \in \mathcal{B}(s_\beta,p)} \Pr\left(\beta_{0,j} \in \left[\tilde{\beta}_{n,j} - z_{1-\tau/2}\frac{\hat{\sigma}_{\beta,j}}{\sqrt{n}}, \tilde{\beta}_{n,j} + z_{1-\tau/2}\frac{\hat{\sigma}_{\beta,j}}{\sqrt{n}}\right]\right) \leq 1-\tau.$$

The reverse inequality follows upon noting that for any $c > 0$,

$$\begin{aligned}
&\Pr\left(\beta_{0,j} \notin \left[\tilde{\beta}_{n,j} - z_{1-\tau/2}\frac{\hat{\sigma}_{\beta,j}}{\sqrt{n}}, \tilde{\beta}_{n,j} + z_{1-\tau/2}\frac{\hat{\sigma}_{\beta,j}}{\sqrt{n}}\right]\right) \\
&\geq 1 - \Pr\left(\frac{\sqrt{n}(\tilde{\beta}_{n,j} - \beta_{0,j})}{\hat{\sigma}_{\beta,j}} > z_{1-\alpha/2}\right) + \Pr\left(\frac{\sqrt{n}(\tilde{\beta}_{n,j} - \beta_{0,j})}{\hat{\sigma}_{\beta,j}} \leq -z_{1-\tau/2}\right).
\end{aligned}$$

So the conclusion follows.

□

Appendix B: Technical Proofs Related to Results in Section 3.1

0.1 Auxiliary Lemmas

Let $\|\cdot\|_{\mathcal{X}_j,\infty}$ be the supnorm for any function defined on $\mathcal{X}_j$, where $j = 1, \cdots, p$. We abbreviate it as $\|\cdot\|_\infty$ when there is no confusion. We first list key maximal inequalities for kernel type estimators from Prop. 5.1.9 and 5.1.12 in Giné and Nickl (2016).

Lemma 1 (Maximal Inequalities of Marginal Kernel Estimators)**.** The kernel type estimates for f_j, and f_j' satisfy the following bounds for any $t > 0$:

$$\mathbb{P}\left\{\| \hat{f}_{n,j} - \mathbb{E}\hat{f}_{n,j} \|_\infty \geq \frac{C\sqrt{\log h_j^{-1}}}{\sqrt{nh_j}} + C\sqrt{\frac{t\log h_j^{-1}}{nh_j}}\right\} \leq e^{-t}, \tag{S.0.11}$$

and

$$\mathbb{P}\left\{\| \hat{f}_{n,j}' - \mathbb{E}\hat{f}_{n,j}' \|_\infty \geq \frac{C\sqrt{\log h_j^{-1}}}{\sqrt{nh_j^2}} + C\sqrt{\frac{t\log h_j^{-1}}{nh_j^2}}\right\} \leq e^{-t}. \tag{S.0.12}$$

To analyze the linear term in the Hoeffding decomposition, we use Theorem 3.1 of Kuchibhotla and Chakrabortty (2018), which we state below for convenience. This allows us to handle independent but not necessarily identically distributed random variables.

Lemma 2. Considering independent r.v.s $(\mathbf{Z}_i)_{i=1,\cdots,n}$ in $\mathbb{R}^p$ such that for some $K_n > 0$,

$$\max_{1\le i\le n}\max_{1\le j\le p} \| \mathbf{Z}_{i,j} \|_{\psi_2} \le K_n,$$

then for any $t \ge 0$, with probability at least $1 - 3e^{-t}$,

$$\max_{1\le j\le p}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_{i,j}\right| \le 7\sqrt{\frac{\Gamma_n(t+\log p)}{n}} + C\frac{K_n\sqrt{\log n}(t+\log p)}{n},$$

where $\Gamma_n := \max_{1\le j\le p} n^{-1}\sum_{i=1}^{n}\mathbb{E}[\mathbf{Z}_{i,j}^2]$.

We recall some standard notion related to U-statistics (Giné and Nickl, 2016). For a kernel function f of k variables, we denote

$$U_n^{(k)}(f) = \frac{(n-k)!}{n!}\sum_{i\in I_n^k} f\left(X_{i_1},\cdots,X_{i_k}\right),$$

where $I_n^m = \{(i_1,\cdots,i_m) : 1 \le i_j \le n, i_j \ne i_k \text{ if } j \ne k\}$. Now suppose f is symmetric in its entries, we have the well-known Hoeffding decomposition:

$$U_n^{(m)}(f) - Ef = \sum_{k=1}^{m} U_n^{(k)}(\pi_k f),$$

where

$$\pi_k f = (\delta_{x_1} - P)\times\cdots\times(\delta_{x_k} - P)\times P^{m-k} f.$$

Moreover let σ^2 (which we call maximal variance) be any number satisfying

$$\sup_{f\in\mathcal{F}}\left|P^m f^2\right| \le \sigma^2 \le M^2.$$

We say a class of functions $\mathcal{F}$ is of VC type with respect to an envelope F if the covering number $N(\mathcal{F}, L_2(Q), \varepsilon)$, the smallest number of $L_2(Q)$ open balls of radius ε required to cover $\mathcal{F}$, satisfies

$$N(\mathcal{F}, L_2(Q), \varepsilon) \le \left(\frac{M\|F\|_{L_2(Q)}}{\varepsilon}\right)^v \text{ for } 0 < \varepsilon \le 2\|F\|_{L_2(Q)},$$

for some universal positive constants M, v and for every probability measure Q on the underlying space (Giné and Nickl, 2016).

Lemma 3 (Theorem 2 of Major (2006))**.** Let $\mathcal{F}$ be a collection of measurable symmetric functions $f : S^m \to \mathcal{R}$ uniformly bounded up by M in absolute values. Assume $\mathcal{F}$ is of VC type with envelope function F and with characteristics A and ν. Then we have

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\left|n^{k/2}U_n^{(k)}\left(\pi_k f\right)\right| \geq t\right\} \leq M\exp\left[-M\left(\frac{t}{\sigma}\right)^{2/k}\right], \tag{S.0.13}$$

if $n\sigma^2 \geq \left(\frac{t}{\sigma}\right)^{2/k} \geq M\left(\frac{1}{\log n}\right)^{3/2}\log\left(\frac{2}{\sigma}\right)$.

Lemma 4. Consider the following bias terms defined by

$$B_{n,j}^{(1)} := \mathbb{E}\left[\left(\hat{f}_j'(X_{i,j}) - f_j'(X_{i,j})\right)Y_i w_j(X_i)/f_j(X_{i,j})\right], \tag{S.0.14}$$

and

$$B_{n,j}^{(2)} := \mathbb{E}\left[\left(\hat{f}_j(X_{i,j}) - f_j(X_{i,j})\right)Y_i w_j(X_i)(\upsilon_j(X_i) + f_j'(X_{i,j})/f^2(X_{i,j}))\right]. \tag{S.0.15}$$

Then there exists some constant terms $C_1 > 0$ and $C_2 > 0$ such that

$$|B_{n,j}^{(1)}| \leq C_1 h_j^r \quad \text{and} \quad |B_{n,j}^{(2)}| \leq C_2 h_j^{r+1} \tag{S.0.16}$$

Proof. The proof essentially follows from Step 3 in the proof of Theorem 3.1 of Hardle and Stoker (1989). We integrate by parts to get

$$\begin{aligned} B_{n,j}^{(1)} &= \int_{\mathcal{X}_j}\mathbb{E}[Y|X_j = x_j]\int K(u)[f_j'(x_j + uh_j) - f_j'(x_j)]dudx_j \\ &= \int_{\mathcal{X}_j}\mathbb{E}[Y|X_j = x_j]K(u)\left[\sum_{2\leq\iota\leq r-1}\int h_j^\iota f_j^{(\iota+1)}(x_j)u^\iota + h_j^r f_j^{(r+1)}(x_j^*)\right]dudx_j, \end{aligned}$$

where x_j^* lies on the line segment between x_j and $x_j + uh_j$. Given our assumptions on the kernel order and the smoothness of marginal densities, the desired conclusion follows. The same analysis can be done for $B_{n,j}^{(2)}$. □

Lemma 5. The derivative of a normal density satisfies the following two identities

$$\phi'(z) = -z\phi(z), \quad \text{and} \quad (\phi \circ \Phi^{-1})'(z) = \Phi^{-1}(z), \quad \forall z. \tag{S.0.17}$$

Also, it holds that

$$\left(\frac{\Phi^{-1}(z)}{\phi \circ \Phi^{-1}(z)}\right)' = \frac{(\Phi^{-1}(z))^2 - 1}{(\phi \circ \Phi^{-1}(z))^2}. \tag{S.0.18}$$

Lemma 6. For given matrices A, $\hat{A}$, and B, we have

$$\|(\hat{A} - A)B\|_\infty \leq \|\hat{A} - A\|_\infty \parallel B \parallel_1 . \tag{S.0.19}$$

Lemma 7. (Theorem 4.2 in Liu et al. (2012).) For any $n > 1$, with probability at least $1 - 1/p$, we have

$$\|\widehat{\Sigma} - \Sigma\|_\infty \leq 2.45\pi\sqrt{\frac{\log p}{n}}. \tag{S.0.20}$$

Lemma 8. Let $\lambda_\Omega = C_0 M_\Omega \sqrt{\log p/n}$, where $C_0 > 0$. Then with probability at least $1 - 1/p$, we have

$$\|\widehat{\Omega} - \Omega_o\|_1 \leq c_\Omega M_\Omega^{2(1-q_1)} \left(\frac{\log p}{n}\right)^{(1-q_1)/2} \tag{S.0.21}$$

Proof. To bound $\|\widehat{\Omega} - \Omega_o\|_1$, we use the result from Lemma 7.1 of Cai et al. (2016). Essentially, we need to check

$$|\hat{\omega}_{\cdot j}|_1 \leq |\omega^0_{\cdot j}|_1, \quad \text{for } j = 1, \cdots, p, \tag{S.0.22}$$

where $\omega^0_{\cdot j}$ stands for j-th column of the true precision matrix Ω_0. We argue that Ω_0 satisfies the constraint (with high probability) in the optimization problem defining CLIME, i.e., whether $\parallel \widehat{\Sigma}\Omega_0 - \mathbb{I}_{p\times p} \parallel_\infty \leq \tau_n$. This is indeed the case with probability at least $1 - 1/p$:

$$\parallel \widehat{\Sigma}\Omega_0 - \mathbb{I}_{p\times p} \parallel_\infty \leq \parallel \widehat{\Sigma} - \Sigma_0 \parallel_\infty \parallel \Omega_0 \parallel_1 \leq C s_\Omega \sqrt{\frac{\log p}{n}},$$

by applying Lemma 6 and then Lemma 7. Because the CLIME $\hat{\Omega}$ is the minimizer of the constrained optimization problem, we know that the inequalities specified in (S.0.22) are satisfied, which concludes the proof.

□

0.2 Non-asymptotic Bounds for Terms in ADE

Proposition 1 (*Bounds for Linear Functionals of Kernel Estimates*)**.** For n large enough, we have

$$\Pr\left\{\max_{1\leq j\leq p}|U_{n,j}|\leq C\sqrt{\frac{\log p}{n}}+C\max_{1\leq j\leq p}h_j^r\right\}\geq 1-O(p^{-1}). \tag{S.0.23}$$

Proof. By the result from Lemma 4, we have $\max_{1\leq j\leq p}|\mathbb{E}[U_{n,j}]|\lesssim h_j^r$ for any $j=1,\cdots,p$. Once we center $U_{n,j}$ by its expectation $\mathbb{E}[U_{n,j}]$, we can employ the theory of canonical U-statistics:

$$U_{n,j}-\mathbb{E}[U_{n,j}]=\sum_{k=1}^{2}U_n^{(k)}(\pi_k g_{1,j}+\pi_k g_{2,j}),$$

where

$$g_{1,j}(Z_i,Z_k)=h_j^{-2}K'\left(\frac{X_{ij}-X_{kj}}{h_j}\right)Y_i w_j(X_i)/f_j(X_{i,j}),$$
$$g_{2,j}(Z_i,Z_k)=h_j^{-1}K\left(\frac{X_{ij}-X_{kj}}{h_j}\right)Y_i w_j(X_i)\left[f_j'(X_{i,j})/f_j(X_{i,j})+2\chi_j(X_i)\right].$$

We shall apply the maximal inequality from Major (2006) to the second-order term in the Hoeffding decomposition. This concerns a VC type class of functions $\mathcal{F}_j$ where $j=1,\cdots,p$. Then, we use the union bound to obtain the uniform results over the classes $\mathcal{F}_j$ for $j=1,\cdots,p$. Note that the second moments of the kernel functions are bounded by

$$\mathbb{E}[g_{1,j}^2(Z_i,Z_k)]\lesssim\underbrace{h_j^{-3}}_{\sigma_{1,j}^2}\quad\text{and}\quad\mathbb{E}[g_{2,j}^2(Z_i,Z_k)]\lesssim\underbrace{h_j^{-1}}_{\sigma_{2,j}^2}.$$

Specifically, we get

$$\Pr\left\{\max_{1\le j\le p}\sup_{g_{1,j},g_{2,j}}\left|nU_n^{(2)}\left(\pi_2[g_{1,j}+g_{2,j}]\right)\right|\ge\frac{(t+\log p)}{\sigma}\right\}\le M\exp\left[-Mt\right], \tag{S.0.24}$$

where $\sigma := \max_{1\le j\le p}\sigma_{1,j}\vee\sigma_{2,j}$. By taking $t=n\sigma\frac{\sqrt{\log p}}{\sqrt{n}}$, we get

$$\Pr\left\{\max_{1\le j\le p}\sup_{g_{1,j},g_{2,j}}\left|nU_n^{(2)}\left(\pi_2[g_{1,j}+g_{2,j}]\right)\right|\ge\frac{\sqrt{\log p}}{\sqrt{n}}+\frac{\log p}{n\sigma}\right\}\le M\exp\left[-M(\sigma\sqrt{n\log p})\right], \tag{S.0.25}$$

which is of order $o(p^{-1})$ under our assumption on the bandwidth. In other words, the second-order terms $U_n^{(2)}(\pi_2 g_{1,j}+\pi_2 g_{2,j})$ are of negligible orders.

Considering the linear term in the Hoeffding expansion[1], we have

$$\pi_1 g_{1,j} = Y_i w_j(X_i)/f_j(X_{i,j})\int K(u)f'(X_{i,j}-uh_j)du.$$

In order to apply the maximal inequality in Lemma 2, note that $\pi_1 g_{1,j}$ has a bounded Orliz norm for any j under our assumption. In addition, its second moment can be bounded by

$$\mathbb{E}[(\pi_1 g_{1,j})^2]\le(1+o(1))\mathbb{E}[(Y_i w_j(X_i)f_j'(X_{i,j})/f_j(X_{i,j}))^2],$$

Thus, we get

$$\Pr\left\{\max_{1\le j\le p}\left|U_n^{(1)}\left(\pi_1 g_{1,j}\right)\right|\ge M\sqrt{\frac{\log p}{n}}\right\}=O(p^{-1}), \tag{S.0.26}$$

as $p\to\infty$ when we take $t=\log p$. A similar bound holds for $\pi_1 g_{2,j}$. The desired conclusion follows from combining the linear and second order terms in U-statistics, as well as their bias terms. □

Proposition 2 (*Bounds for Nonlinear Functional of Marginal Distributions*)**.** For large enough n, we have

$$\Pr\left\{\max_{1\le j\le p}|V_{n,j}|\le Cs_\Omega\frac{\sqrt{\log p}}{\sqrt{n}}\right\}\ge 1-O(p^{-1}). \tag{S.0.27}$$

[1]For the linear term ($k=1$), the inequality from Major (2006) is not sharp, because the variance term σ^2 is too rough for $\|\ P[\pi_1 g]^2\ \|_{\mathcal{F}}$.

Proof. We decompose $V_{n,j}$ by

$$\begin{aligned} V_{n,j} &:= \frac{1}{n}\sum_{i=1}^{n} Y_i w_j(X_i) f_j(X_{i,j}) \left[\frac{\Phi^{-1}(F_j(X_{i,j}))}{\phi(\Phi^{-1}(F_j(X_{i,j})))} - \frac{\Phi^{-1}(\hat{F}_j(X_{i,j}))}{\phi(\Phi^{-1}(\hat{F}_j(X_{i,j})))} \right] \\ &+ \frac{1}{n}\sum_{i=1}^{n} Y_i w_j(X_i) \zeta_j(X_{i,j}) \left[\sum_{k=1}^{p} \Omega_{kj} \left(\Phi^{-1}(F_k(X_{i,k})) - \Phi^{-1}(\hat{F}_k(X_{i,k})) \right) \right] \\ &:= V_{n,j}^{(1)} + V_{n,j}^{(2)}. \end{aligned}$$

By Lemma 5, the function $\frac{\Phi^{-1}(z)}{\phi \circ \Phi^{-1}(z)}$ is Lipschitz continuous over any compact support. Thus, the stochastic order of $V_{n,j}^{(1)}$ is driven by marginal empirical distribution functions. Consider the sample average $\mathbb{P}_n[Y\chi_j(X_i)]$ in which Y has a bounded Orliz norm ψ_2 and $\max_{1\leq j\leq p} |\chi_j|$ is uniformly bounded. Then for any $t > 0$, we have

$$\max_{1\leq j\leq p} |(\mathbb{P}_n - P)[Y\chi_j(X)]| \leq C\sqrt{\frac{(t+\log p)}{n}} + C\frac{\sqrt{\log(2n)(t+\log p)}}{n}, \tag{S.0.28}$$

with probability at least $1 - 3e^{-t}$. It also holds that

$$\max_{1\leq j\leq p} |\mathbb{P}_n Y\chi_j(X)| \leq \| \max_{1\leq j\leq p} P|Y\chi_j(X)| \|_\infty + C\sqrt{\frac{(t+\log p)}{n}} + C\frac{\sqrt{\log(2n)(t+\log p)}}{n}, \tag{S.0.29}$$

with probability at least $1 - 3e^{-t}$.

In particular, we bound it by

$$V_{n,j}^{(1)} \leq C \max_{1\leq j\leq p} \left(\| \hat{F}_j - F_j \|_\infty \left[\max_{1\leq j\leq p} P\left|Y w_j(X) f_j(X_j)\right| + \frac{\sqrt{\log p}}{\sqrt{n}} \right] \right).$$

Note that the second term $V_{n,j}^{(2)}$ involves the summation of all estimated marginal distributions in the bracket so that we need the sparsity condition on the precision matrix Ω to deliver the following bound:

$$V_{n,j}^{(2)} \leq C \max_{1\leq j\leq p} \left(s_\Omega \| \hat{F}_j - F_j \|_\infty \left[\max_{1\leq j\leq p} \mathbb{E}\left|Y w_j(X) \zeta_j(X_j)\right| + \frac{\sqrt{\log p}}{\sqrt{n}} \right] \right).$$

In sum,

$$\mathbb{P}\left\{\max_{1\leq j\leq p}|V_{n,j}|\geq s_\Omega\sqrt{\frac{\log p}{n}}\right\}\leq O(p^{-1}).$$

by the DKW maximal inequality:

$$\mathbb{P}\left\{\sqrt{n}\parallel\hat{F}_j-F_j\parallel_\infty\geq t\right\}\leq Ce^{-Ct^2}, \tag{S.0.30}$$

where C is some finite constant[2]. □

Proposition 3. [*Bounds for Linear Combination of the Precision Matrix*] Under our assumptions, we have

$$\Pr\left\{\max_{1\leq j\leq p}|W_{n,j}|\leq C_1c_\Omega M_\Omega^{2(1-q_1)}\left(\frac{\log p}{n}\right)^{(1-q_1)/2}\right\}\geq 1-O(p^{-1}). \tag{S.0.31}$$

Proof. To bound $\max_{1\leq j\leq p}|W_{n,j}|$, let A be a $p\times p$ matrix with element

$$A_{lk}:=\frac{1}{n}\sum_{i=1}^{n}\Phi^{-1}(F_k(X_{i,k}))\zeta_l(X_{i,l})Y_i. \tag{S.0.32}$$

Let $\tilde{W}_n$ be a $p\times p$ matrix with element $\tilde{W}_{n,lj}=\sum_{k=1}^{p}A_{lk}\left(\widehat{\Omega}_{kj}-\Omega_{kj}\right)$ for $l,j=1,\ldots,p$. Then $W_{n,j}:=\tilde{W}_{n,jj}$ is the jth element on the diagonal of $\tilde{W}_n$. Thus,

$$\begin{aligned}\max_{1\leq j\leq p}|W_{n,j}|&=\max_{1\leq j\leq p}\left|\sum_{k=1}^{p}a_{jk}\left(\widehat{\Omega}_{kj}-\Omega_{kj}\right)\right|=\max_{1\leq j\leq p}|\tilde{W}_{n,jj}|\leq\|\tilde{W}_n\|_\infty\\&=\|A(\widehat{\Omega}-\Omega)\|_\infty\leq\|A\|_\infty\|\widehat{\Omega}-\Omega\|_1.\end{aligned}$$

Since $\|A\|_\infty<C_1+\sqrt{\frac{2\log p}{n}}$ holds with probability larger than $1-O(p^{-1})$, it suffices to bound $\|\widehat{\Omega}-\Omega\|_1$. Then Equation (S.0.31) follows from the conclusion in Lemma 8. □

[2]One can also resort to the kernel smoothed distribution function by the following DKW type exponential inequality from Giné and Nickl (2009).

A routine calculation shows that the remainder term is equal to

$$
\begin{aligned}
R_{n,j} =& \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i w_j(X_i)}{f_j(X_{i,j})\hat{f}_j(X_{i,j})} \left[\hat{f}_j(X_{i,j}) - f_j(X_{i,j})\right]\left[\hat{f}'_j(X_{i,j}) - f'_j(X_{i,j})\right] \\
& - \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i w_j(X_i) f'_j(X_{i,j})}{f_j^2(X_{i,j})\hat{f}_j(X_{i,j})} \left[\hat{f}_j(X_{i,j}) - f_j(X_{i,j})\right]^2 \\
& + \frac{1}{n}\sum_{i=1}^{n} Y_i w_j(X_i) \left[\frac{\Phi^{-1}(\hat{F}_j(X_{i,j}))}{\phi(\Phi^{-1}(\hat{F}_j(X_{i,j})))} - \frac{\Phi^{-1}(F_j(X_{i,j}))}{\phi(\Phi^{-1}(F_j(X_{i,j})))}\right]\left[\hat{f}_j(X_{i,j}) - f_j(X_{i,j})\right] \\
& + \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i w_j(X_i)}{\phi(\Phi^{-1}(\hat{F}_j(X_{i,j})))} \left[\sum_{k=1}^{p} (\widehat{\Omega}_{kj} - \Omega_{kj})\Phi^{-1}(\hat{F}_k(X_{i,k}))\right]\left[\hat{f}_j(X_{i,j}) - f_j(X_{i,j})\right] \\
& - \frac{1}{n}\sum_{i=1}^{n} Y_i w_j(X_i) \left[\sum_{k=1}^{p} \Omega_{kj}\left(\frac{\Phi^{-1}(\hat{F}_k(X_{i,k}))}{\phi(\Phi^{-1}(\hat{F}_j(X_{i,j})))} - \frac{\Phi^{-1}(F_k(X_{i,k}))}{\phi(\Phi^{-1}(F_j(X_{i,j})))}\right)\right]\left[\hat{f}_j(X_{i,j}) - f_j(X_{i,j})\right] \\
:=& R_{n,j}^{(1)} + R_{n,j}^{(2)} + R_{n,j}^{(3)} + R_{n,j}^{(4)} + R_{n,j}^{(5)}.
\end{aligned}
$$

Essentially, the remainder consists of second order terms involving the product of errors arising from terms $U_{n,j}, W_{n,j}, V_{n,j}$. For those terms related to $U_{n,j}$, we do not need further linearization to keep them as negligible. Under our assumptions on bandwidths, a straightforward analysis with supnorm bounds on the kernel density or its derivative suffices. To that end, we combine Lemma 1 with standard bias calculation for kernel estimators to arrive at the following maximal inequalities:

$$
\Pr\left\{\max_{1\le j\le p} \parallel \hat{f}_{n,j} - f_j \parallel_\infty \ge C h_j^{r+1} + \frac{C\sqrt{\log h_j^{-1}}}{\sqrt{nh_j}} + \sqrt{\frac{Ct(\log h_j^{-1} + \log p)}{nh_j}}\right\} \le e^{-t}, \tag{S.0.33}
$$

and

$$
\Pr\left\{\max_{1\le j\le p} \parallel \hat{f}'_{n,j} - f'_j \parallel_\infty \ge C h_j^{r} + \frac{C\sqrt{\log h_j^{-2}}}{\sqrt{nh_j^2}} + \sqrt{\frac{Ct(\log h_j^{-2} + \log p)}{nh_j^2}}\right\} \le e^{-t}. \tag{S.0.34}
$$

Proposition 4 (*Bounds for the Remainder Term*)**.** Under our assumptions, we have

$$\Pr\left\{\max_{1\leq j\leq p}|R_{n,j}|\leq\left(C_1c_\Omega M_\Omega^{2(1-q_1)}\left(\frac{\log p}{n}\right)^{(1-q_1)/2}\sqrt{\frac{\log p}{nh_j}}\right)\bigvee\left(C_2\frac{\log p}{nh_j^3}\right)\right\}\geq 1-O(p^{-1}). \tag{S.0.35}$$

Proof. We first take care of random denominators. Given our assumptions on the marginal density's behavior over the compact set $\mathcal{X}_j$ and the uniform convergence of $\hat{f}_j$, we have the following

$$\min_{1\leq j\leq p}\inf_{x_j\in\mathcal{X}_j}\hat{f}_j(x_j)\geq c_0/2$$

hold with probability at least $1-O(p^{-1})$. In the same vein, we also get

$$\min_{1\leq j\leq p}\inf_{x_j\in\mathcal{X}_j}\phi(\Phi^{-1}(\hat{F}_j(x_j)))\geq c_0/2.$$

We bound each term by

$$\begin{aligned}
R_{n,j}^{(1)} &\leq C\max_{1\leq j\leq p}\left(\parallel\hat{f}_j-f_j\parallel_\infty\parallel\hat{f}_j'-f_j'\parallel_\infty\left[\mathbb{P}_n\left|Yw_j(X)/f_j(X)\right|\right]\right)\\
&\leq C\max_{1\leq j\leq p}\left(\parallel\hat{f}_j-f_j\parallel_\infty\parallel\hat{f}_j'-f_j'\parallel_\infty\left[\max_{1\leq j\leq p}P\left|Yw_j(X)/f_j(X)\right|+\frac{\sqrt{\log p}}{\sqrt{n}}\right]\right),
\end{aligned}$$

$$R_{n,j}^{(2)}\leq C\max_{1\leq j\leq p}\left(\parallel\hat{f}_j-f_j\parallel_\infty^2\left[\max_{1\leq j\leq p}P\left|Yw_j(X)f_j'(X_j)/f_j^2(X_j)\right|+\frac{\sqrt{\log p}}{\sqrt{n}}\right]\right),$$

$$R_{n,j}^{(3)}\leq C\max_{1\leq j\leq p}\left(\parallel\hat{f}_j-f_j\parallel_\infty\parallel\hat{F}_j-F_j\parallel_\infty\left[\max_{1\leq j\leq p}P\left|Yw_j(X)\right|+\frac{\sqrt{\log p}}{\sqrt{n}}\right]\right),$$

$$R_{n,j}^{(4)}\leq C\max_{1\leq j\leq p}\left(\parallel\hat{f}_j-f_j\parallel_\infty\left[\max_{1\leq j\leq p}P\left|Yw_j(X)\right|+\frac{\sqrt{\log p}}{\sqrt{n}}\right]\right)\parallel\widehat{\Omega}-\Omega_0\parallel_1,$$

and

$$R_{n,j}^{(5)}\leq C\max_{1\leq j\leq p}\left(\parallel\hat{f}_j-f_j\parallel_\infty s_\Omega\parallel\hat{F}_j-F_j\parallel_\infty\left[\max_{1\leq j\leq p}P\left|Yw_j(X)\right|+\frac{\sqrt{\log p}}{\sqrt{n}}\right]\right).$$

It is straightforward to see the dominating terms are $R_{n,j}^{(1)}$ and $R_{n,j}^{(4)}$. Thus, the conclusion follows from Lemma 7, as well as inequalities (S.0.33) and (S.0.34). □

Appendix C: Technical Proofs Related to Results in Section 3.2

We introduce some notation that facilitate our presentation in this section. Let $\mathcal{A}_n$ be the event $\left|\hat{\beta}_n - \beta_0\right|_1 \leq c_{1n}$ and $\mathcal{A}_n^c$ be the complement of $\mathcal{A}_n$. Let $K_h(\cdot) = K(\cdot/h)$. Given a generic pilot estimator $\hat{\beta}_n$, consider the following leave-one-out kernel estimators:

$$\hat{\mathbf{r}}_{\phi,n}(X_i^\top \hat{\beta}_n) := \frac{1}{nh} \sum_{l \neq i} \phi_l K_h \left((X_l - X_i)^\top \hat{\beta}_n \right), \qquad \hat{\mathbf{r}}'_{\phi,n}(X_i^\top \hat{\beta}_n) := \frac{1}{nh^2} \sum_{l \neq i} \phi_l K'_h \left((X_l - X_i)^\top \hat{\beta}_n \right),$$

$$\hat{\mathbf{f}}_n(X_i^\top \hat{\beta}_n) := \frac{1}{nh} \sum_{l \neq i} K_h \left((X_l - X_i)^\top \hat{\beta}_n \right), \qquad \hat{\mathbf{f}}'_n(X_i^\top \hat{\beta}_n) := \frac{1}{nh^2} \sum_{l \neq i} K'_h \left((X_l - X_i)^\top \hat{\beta}_n \right).$$

Let $\hat{\mu}_{n,-1}(X_i^\top \hat{\beta}_n) = (\hat{\mu}_{n,2}(X_i^\top \hat{\beta}_n), \cdots, \hat{\mu}_{n,p}(X_i^\top \hat{\beta}_n))^\top$ where

$$\hat{\mu}_{n,j}(X_i^\top \hat{\beta}_n) = \frac{\hat{\mathbf{r}}_{X_j,n}(X_i^\top \hat{\beta}_n)}{\hat{\mathbf{f}}_n(X_i^\top \hat{\beta}_n)}, \quad j = 2, \cdots, p,$$

and

$$\hat{\varpi}_n(X_i) = \frac{\hat{\mathbf{r}}'_{Y,n}(X_i^\top \hat{\beta}_n)}{\hat{\mathbf{f}}_n(X_i^\top \hat{\beta}_n)} - \frac{\hat{\mathbf{r}}_{Y,n}(X_i^\top \hat{\beta}_n)\hat{\mathbf{f}}'_n(X_i^\top \hat{\beta}_n)}{\hat{\mathbf{f}}_n^2(X_i^\top \hat{\beta}_n)}.$$

We estimate the conditional mean function by $\widehat{m}_n(X_i^\top \hat{\beta}_n; \hat{\beta}_n) = \frac{\hat{\mathbf{r}}_{Y,n}(X_i^\top \hat{\beta}_n)}{\hat{\mathbf{f}}_n(X_i^\top \hat{\beta}_n)}$.

Suppose that $\hat{\beta}_n$ satisfies a basic high-level L_1 error guarantee such that

$$\mathbb{P}\left(\left| \hat{\beta}_n - \beta_0 \right|_1 > c_{1n} \right) \leq q_n$$

for some $c_{1n} \geq 0$ and $q_n = o(1)$ from Corollary 1. Let $\mu_j(X_i'\beta_0) = \mathbb{E}\left[X_{ij} | X_i^\top \beta_0 \right], \hat{\mu}_{n,j}(X_i^\top \hat{\beta}_n)$ be its kernel estimator, $\hat{X}_{i,j} = (X_{i,j} - \hat{\mu}_{n,j}(X_i^\top \hat{\beta}_n))$, $\tilde{X}_{i,j} = (X_{i,j} - \mu_j(X_i^\top \beta_0))$. We denote $\hat{X}_i := (\hat{X}_{i,2}, \cdots, \hat{X}_{i,p})^\top$ and $\tilde{X}_i := (\tilde{X}_{i,2}, \cdots, \tilde{X}_{i,p})^\top$. For notational convenience, we drop the trimming function $w(X_i)$ when there is no confusion. For instance, we simply write

$$\dot{\Psi}_n(\hat{\beta}_n) = \frac{1}{n} \sum_{i=1}^n \hat{\varpi}_n^2(X_i; \hat{\beta}_n) \hat{X}_i \hat{X}_i^\top.$$

Lemma 9. Under Assumptions in Theorem 2, we have that

$$\hat{\Psi}_n(\beta_0) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i m_0'(X_i^\top \beta_0)\tilde{X}_i + o_p(n^{-1/2}). \tag{S.0.36}$$

Proof. We take the following decomposition of the score function as in Ichimura (1993):

$$\begin{aligned}
\hat{\Psi}_n(\beta_0) :=& \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_n(X_i^\top\beta_0;\beta_0))\frac{\partial \widehat{m}_n(X_i^\top\beta_0;\beta_0)}{\partial \beta_{-1}} \\
=& \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \frac{\partial m_0(X_i^\top\beta_0;\beta_0)}{\partial \beta_{-1}} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \left(\frac{\partial \widehat{m}_n(X_i^\top\beta_0;\beta_0)}{\partial \beta_{-1}} - \frac{\partial m_0(X_i^\top\beta_0;\beta_0)}{\partial \beta_{-1}}\right) \\
&+ \frac{1}{n}\sum_{i=1}^{n}(\widehat{m}_n(X_i^\top\beta_0;\beta_0) - m_0(X_i^\top\beta_0))\frac{\partial m_0(X_i^\top\beta_0;\beta_0)}{\partial \beta_{-1}} \\
&+ \frac{1}{n}\sum_{i=1}^{n}(\widehat{m}_n(X_i^\top\beta_0;\beta_0) - m_0(X_i^\top\beta_0))\left(\frac{\partial \widehat{m}_n(X_i^\top\beta_0;\beta_0)}{\partial \beta_{-1}} - \frac{\partial m_0(X_i^\top\beta_0;\beta_0)}{\partial \beta_{-1}}\right),
\end{aligned}$$

where

$$\frac{\partial m_0(X_i^\top\beta_0;\beta_0)}{\partial \beta_{-1}} = m_0'(X_i^\top\beta_0)(X_{i,-1} - \mathbb{E}[X_{i,-1}|X_i^\top\beta_0]) = m_0'(X_i^\top\beta_0)\tilde{X}_i.$$

Thus, the leading term is

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \frac{\partial m_0(X_i^\top\beta_0;\beta_0)}{\partial \beta_{-1}} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i m_0'(X_i^\top\beta_0)\tilde{X}_i$$

and the rest are all asymptotically negligible terms because of the results in Lemma 10. □

The following lemma requires an expansion as in the proof of Lemma 5.8 on Page 114 in Ichimura (1993). In particular, we strengthen his arguments to accommodate the presence of high-dimensional covariates by the maximal inequality of Major (2006).

Lemma 10. Under Assumptions in Theorem 2, we have that

(i)
$$\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(\frac{\partial\widehat{m}_n(X_i^\top\beta_0;\beta_0)}{\partial\beta_{-1}}-\frac{\partial m_0(X_i^\top\beta_0;\beta_0)}{\partial\beta_{-1}}\right)\right|_\infty=o_p(n^{-1/2}),$$

(ii)
$$\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{m}_n(X_i^\top\beta_0;\beta_0)-m_0(X_i^\top\beta_0)\right\}m'(X_i^\top\beta_0)\tilde{X}_i\right|_\infty=o_p\left(n^{-1/2}\right),$$

(iii)
$$\left|\frac{1}{n}\sum_{i=1}^{n}(\widehat{m}_n(X_i^\top\beta_0;\beta_0)-m_0(X_i^\top\beta_0))\left(\frac{\partial\widehat{m}_n(X_i^\top\beta_0;\beta_0)}{\partial\beta_{-1}}-\frac{\partial m_0(X_i^\top\beta_0;\beta_0)}{\partial\beta_{-1}}\right)\right|_\infty=o_p\left(n^{-1/2}\right).$$

Proof. Due to the similarity, we only prove Part (i). Since the bias term is of order h^γ, we get

$$\begin{aligned}&\frac{\partial\hat{m}_n(X_i^\top\beta_0)}{\partial\beta_{-1}}-\frac{\partial m(X_i^\top\beta_0;\beta_0)}{\partial\beta_{-1}}\\=&\frac{\partial\hat{\mathbf{r}}_{Y,n}(X_i^\top\beta_0)/\partial\beta_{-1}}{\hat{\mathbf{f}}_n(X_i^\top\beta_0)}-\frac{\partial\mathbf{r}_Y(X_i^\top\beta_0)/\partial\beta_{-1}}{\mathbf{f}(X_i^\top\beta_0)}\\&-\frac{\hat{\mathbf{r}}_{Y,n}(X_i^\top\beta_0)}{\hat{\mathbf{f}}_n^2(X_i^\top\beta_0)}\times\hat{\mathbf{f}}_n(X_i^\top\beta)/\partial\beta_{-1}+\frac{\mathbf{r}_Y(X_i^\top\beta_0)}{\mathbf{f}^2(X_i^\top\beta_0)}\times\mathbf{f}(X_i^\top\beta_0)/\partial\beta_{-1}+O(h^\gamma).\end{aligned}$$

Following the same linearization from Ichimura (1993), e.g., the long display between equations (27) and (28), it suffices to prove

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\varepsilon_i}{\mathbf{f}(X_i^\top\beta_0)}\left[\partial\hat{\mathbf{r}}_{Y,n}(X_i^\top\beta_0)/\partial\beta_{-1}-\partial\mathbf{r}_Y(X_i^\top\beta_0)/\partial\beta_{-1}\right]\right|_\infty=o_p(1/\sqrt{n}).$$

This can be written as a second order degenerate U-statistics:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\varepsilon_i}{\mathbf{f}(X_i^\top\beta_0)}\left[\partial\hat{\mathbf{r}}_{Y,n}(X_i^\top\beta_0)/\partial\beta_{-1}-\partial\mathbf{r}_Y(X_i^\top\beta_0)/\partial\beta_{-1}\right]=\frac{1}{n(n-1)}\sum_{i\neq l}\psi_{il},$$

in which

$$\psi_{il}=\frac{\varepsilon_i}{h^2\mathbf{f}(X_i^\top\beta_0)}\left[Y_l(X_{i,-1}-X_{l,-1})K'\left(\frac{(X_i-X_l)^\top\beta_0}{h}\right)-\partial\mathbf{r}_Y(X_i^\top\beta_0)/\partial\beta_{-1}\right].$$

As shown on page 116 of Ichimura (1993), the second order moment of ψ_{il} can be bounded up by $O(h^4)$. By first applying the maximal inequality in Major (2006) and then the union bound, we get

$$\mathbb{E}\max_{2\leq j\leq p}\left|U_n^{(2)}\left(\pi_2\psi\right)\right| \leq C\frac{\sigma\log p}{n}\left(\log\left(\frac{A}{\sigma}\right)\right) \asymp \frac{h^2\log(h^{-2})\log p}{n}, \tag{S.0.37}$$

Under our assumption that $nh^4/(\log n\log p)\to\infty$, we get the desired result.

□

Lemma 11. Under Assumptions in Theorem 2, we have

$$\|\hat{\dot{\Psi}}_n(\hat{\beta}_n)-\dot{\Psi}_0\|_\infty = O_p\left(\frac{c_{1n}^2}{h^3}+\sqrt{\frac{\log np}{nh^2}}+h^2+c_{1n}\right).$$

Proof. We consider the following decomposition such that

$$\begin{aligned}
&\left\|\hat{\dot{\Psi}}_n(\hat{\beta}_n)-\dot{\Psi}_0\right\|_\infty \qquad\qquad (\text{S.0.38})\\
&= \|\frac{1}{n}\sum_{i=1}^n \hat{\varpi}_n^2(X_i;\hat{\beta}_n)\hat{X}_i\hat{X}_i^\top - \underbrace{\frac{1}{n}\sum_{i=1}^n m_0'(X_i^\top\beta_0)^2\tilde{X}_i\tilde{X}_i^\top}_{\dot{\Psi}_n(\beta_0)}\|_\infty + \left\|\left(\mathbb{E}_n-\mathbb{E}\right)m_0'(X_i^\top\beta_0)^2\tilde{X}_i\tilde{X}_i^\top\right\|_\infty\\
&\leq \left\|\frac{1}{n}\sum_{i=1}^n\left\{\hat{\varpi}_n^2(X_i^\top\hat{\beta}_n)-m_0'(X_i^\top\beta_0)^2\right\}\hat{X}_i\hat{X}_i^\top\right\|_\infty + \left\|\frac{1}{n}\sum_{i=1}^n m_0'(X_i^\top\beta_0)\left(\hat{X}_i\hat{X}_i^\top-\tilde{X}_i\tilde{X}_i^\top\right)\right\|_\infty\\
&\quad + \left\|\left(\mathbb{P}_n-\mathbb{P}\right)m_0'(X_i^\top\beta_0)^2\tilde{X}_i\tilde{X}_i^\top\right\|_\infty\\
&\leq \left\|\left(\mathbb{P}_n-\mathbb{P}\right)m_0'(X_i^\top\beta_0)^2\tilde{X}_i\tilde{X}_i^\top\right\|_\infty + \max_{1\leq i\leq n}\left|m_0'(X_i^\top\beta_0)^2\right|\left\|\frac{1}{n}\sum_{i=1}^n\hat{X}_i\hat{X}_i^\top-\tilde{X}_i\tilde{X}_i^\top\right\|_\infty.\\
&\quad + \max_{1\leq i\leq n}\left|\hat{\varpi}_n^2(X_i^\top\hat{\beta}_n)-m_0'(X_i^\top\beta_0)^2\right|\left\|\frac{1}{n}\sum_{i=1}^n\hat{X}_i\hat{X}_i\right\|_\infty\\
&:= I_1+I_2+I_3.
\end{aligned}$$

First, because $m'(X_i^\top\beta_0)$ is bounded from Assumption 10 (i) , $m'(X_i^\top\beta_0)\tilde{X}_i$ is sub-Gaussian with $\|m'(X_i^T\beta_0)\tilde{X}_i\|_{\psi_2}\leq K_X$ for a constant $0<K_X<\infty$,

$$I_1 = \|\left(\mathbb{P}_n-\mathbb{P}\right)m_0'(X_i^\top\beta_0)^2\tilde{X}_i\tilde{X}_i^\top\|_\infty = O_p\left(\sqrt{\frac{\log 2p}{n}}\right). \tag{S.0.39}$$

We next consider how to bound I_2. We first find the upper bound for $\left\|\frac{1}{n}\sum_{i=1}^n\left(\hat{X}_i\hat{X}_i^\top-\tilde{X}_i\tilde{X}_i^\top\right)\right\|_\infty$.

Notice that

$$\left\|\frac{1}{n}\sum_{i=1}^n\left(\hat{X}_i\hat{X}_i^\top-\tilde{X}_i\tilde{X}_i^\top\right)\right\|_\infty\leq\left\|\frac{1}{n}\sum_{i=1}^n\left(\hat{X}_i-\tilde{X}_i\right)\left(\hat{X}_i-\tilde{X}_i\right)^\top\right\|_\infty+2\left\|\frac{1}{n}\sum_{i=1}^n(\hat{X}_i-\tilde{X}_i)\tilde{X}_i^\top\right\|_\infty.$$

Lemma 13 and the followed discussion implies that

$$\left\|\frac{1}{n}\sum_{i=1}^n\left(\hat{X}_i-\tilde{X}_i\right)\left(\hat{X}_i-\tilde{X}_i\right)^\top\right\|_\infty=O_p\left(\left(c_{1n}+\frac{c_{1n}^2}{h^2}+\sqrt{\frac{\log np}{nh}}+h^2\right)^2\right). \quad \text{(S.0.40)}$$

Next, notice that for any constant $t>0$,

$$\mathbb{P}\left(\max_{1\leq j\leq p}\max_{1\leq k\leq p}\left|\frac{1}{n}\sum_{i=1}^n\left(\hat{X}_{ik}-\tilde{X}_{ik}\right)\tilde{X}_{ij}\right|\geq\max_{1\leq i\leq n}\max_{1\leq k\leq p}\left|\hat{X}_{ik}-\tilde{X}_{ik}\right|\max_{1\leq j\leq p}\sqrt{\mathbb{E}_n\tilde{X}_{ij}^2}\sqrt{2\left(t^2+\frac{\log 2p}{n}\right)}\right)$$
$$\leq\exp(-nt^2),$$

because of Lemma 14.15 of Bühlmann and van de Geer (2011) and the sub-Gaussian property of $\tilde{X}_{ij}$. Choosing $t=\sqrt{\frac{\log 2p}{n}}$ implies that

$$\left\|\frac{1}{n}\sum_{i=1}^n(\hat{X}_i-\tilde{X}_i)\tilde{X}_i^\top\right\|_\infty=O_p\left(\left(c_{1n}+\frac{c_{1n}^2}{h^2}+\sqrt{\frac{\log np}{nh}}+h^2\right)\sqrt{\frac{\log 2p}{n}}\right). \quad \text{(S.0.41)}$$

Thus, (S.0.40) and (S.0.41) and the boundedness of $\max_{1\leq i\leq n}|m_0'(X_i^\top\beta_0)|$ implies that

$$I_2=O_p\left(\left(c_{1n}+\frac{c_{1n}^2}{h^2}+\sqrt{\frac{\log np}{nh}}+h^2\right)^2+\left(c_{1n}+\frac{c_{1n}^2}{h^2}+\sqrt{\frac{\log np}{nh}}+h^2\right)\sqrt{\frac{\log 2p}{n}}\right). \quad \text{(S.0.42)}$$

From the above analysis, we know that $\left\|\frac{1}{n}\sum_{i=1}^n\hat{X}_i\hat{X}_i\right\|_\infty=O_p(1)$. Moreover, because

$$\max_{1\leq i\leq n}\left|\hat{\varpi}_n^2(X_i^\top\hat{\beta}_n)-m_0'(X_i^\top\beta_0)^2\right|=O_p\left(\frac{c_{1n}^2}{h^3}+\sqrt{\frac{\log np}{nh^2}}+h^2+c_{1n}\right) \quad \text{(S.0.43)}$$

from Lemma 14, it follows that

$$I_3=O_p\left(\frac{c_{1n}^2}{h^3}+\sqrt{\frac{\log np}{nh^2}}+h^2+c_{1n}\right). \quad \text{(S.0.44)}$$

Then (S.0.39), (S.0.42) and (S.0.44) combined imply that

$$\|\hat{\dot{\Psi}}_n(\hat{\beta}_n) - \dot{\Psi}_n(\beta_0)\|_\infty = O_p\left(\frac{c_{1n}^2}{h^3} + \sqrt{\frac{\log np}{nh^2}} + h^2 + c_{1n}\right).$$

□

Lemma 12. Under Assumptions in Theorem 2,

$$\left\|\mathbb{I}_{(p-1)\times(p-1)} - \widehat{\Theta}\hat{\dot{\Psi}}_n(\beta_0)\right\|_\infty = O_p\left(M_\Theta\left(\frac{c_{1n}^2}{h^3} + \sqrt{\frac{\log np}{nh^2}} + h^2 + c_{1n}\right)\right).$$

Proof. Lemma 11 gives the convergence rate for $\|\hat{\dot{\Psi}}_n(\hat{\beta}_n) - \dot{\Psi}_0\|_\infty$ and it also induces the same rate for $\|\hat{\dot{\Psi}}_n(\beta_0) - \dot{\Psi}_0\|_\infty$ by simply taking $\hat{\beta}_n$ to be β_0. Considering the fact that Θ_o is a feasible solution to Algorithm (2.25) w.p.1, we get $\|\widehat{\Theta}\|_1 \leq \|\Theta_o\|_1 \leq M_\Theta$. Then we proceed as follows:

$$\begin{aligned}\left\|\mathbb{I}_{(p-1)\times(p-1)} - \widehat{\Theta}\hat{\dot{\Psi}}_n(\beta_0)\right\|_\infty &\leq \left\|\mathbb{I}_{(p-1)\times(p-1)} - \widehat{\Theta}\hat{\dot{\Psi}}_n(\hat{\beta}_n)\right\|_\infty + \left\|\widehat{\Theta}\right\|_1 \left\|\hat{\dot{\Psi}}_n(\hat{\beta}_n) - \hat{\dot{\Psi}}_n(\beta_0)\right\|_\infty \\ &\leq \lambda_\Psi + M_\Theta \left\|\hat{\dot{\Psi}}_n(\hat{\beta}_n) - \hat{\dot{\Psi}}_n(\beta_0)\right\|_\infty\end{aligned}$$

The result follows from Assumption 15 and Lemma 10. □

Consider a generic value x_β in the support of $x^\top\beta$ for which $x \in \mathcal{X}$ and any β in the parameter space.

Lemma 13. Under Assumptions in Theorem 2, we have that

$$\sup_{x_\beta} |\hat{m}_n(x_\beta) - m_0(x_\beta)| = O_p\left(c_{1n} + \frac{c_{1n}^2}{h^2} + \sqrt{\frac{\log np}{nh}} + h^2\right).$$

In addition, we get

$$\max_{2\leq j\leq p} \sup_{x_\beta} |\hat{\mu}_{n,j}(x_\beta) - \mu_j(x_\beta)| = O_p\left(c_{1n} + \frac{c_{1n}^2}{h^2} + \sqrt{\frac{\log np}{nh}} + h^2\right).$$

Proof. Let

$$\tilde{m}_n(x_\beta) = \frac{\sum_{i=1}^n Y_i K_h\left(X_i^\top \beta_0 - x_\beta\right)}{\sum_{i=1}^n K_h\left(X_i^\top \beta_0 - x_\beta\right)}, \quad \hat{m}_n(x_\beta) = \frac{\sum_{i=1}^n Y_i K_h\left(X_i^\top \hat{\beta}_n - x_\beta\right)}{\sum_{i=1}^n K_h\left(X_i^\top \hat{\beta}_n - x_\beta\right)}.$$

For $m_0(x_\beta)$, we consider the following decompositions:

$$\hat{m}_n\left(x_\beta\right) - m_0(x_\beta) = \frac{\sum_{i=1}^n K_h(X_i^\top \hat{\beta}_n - x_\beta)\left\{Y_i - m_0(x_\beta)\right\}}{\sum_{i=1}^n K_h(X_i^\top \hat{\beta}_n - x_\beta)} := \frac{\hat{L}_1(x_\beta) + \hat{L}_2(x_\beta)}{\hat{g}_{\hat{\beta}_n}(x_\beta)},$$
$$\tilde{m}_n\left(x_\beta\right) - m_0(x_\beta) = \frac{\sum_{i=1}^n K_h(X_i^\top \beta_0 - x_\beta)\left\{Y_i - m_0(x_\beta)\right\}}{\sum_{i=1}^n K_h(X_i^\top \beta_0 - x_\beta)} := \frac{L_1(x_\beta) + L_2(x_\beta)}{\hat{g}_{\hat{\beta}_n}(x_\beta)},$$

where

$$\hat{L}_1(x_\beta) = \frac{1}{n}\sum_{i=1}^n K_h(X_i^\top \hat{\beta}_n - x_\beta)\left\{m_0(X_i^\top \beta_0) - m_0(x_\beta)\right\},$$
$$\hat{L}_2(x_\beta) = \frac{1}{n}\sum_{i=1}^n K_h(X_i^\top \hat{\beta}_n - x_\beta)\varepsilon_i,$$
$$\hat{g}_{\hat{\beta}_n}(x_\beta) = \frac{1}{n}\sum_{i=1}^n K_h\left(X_i^\top \hat{\beta}_n - x_\beta\right),$$
$$L_1(x_\beta) = \frac{1}{n}\sum_{i=1}^n K_h(X_i^\top \beta_0 - x_\beta)\left\{m_0(X_i^\top \beta_0) - m_0(x_\beta)\right\},$$
$$L_2(x_\beta) = \frac{1}{n}\sum_{i=1}^n K_h(X_i^\top \beta_0 - x_\beta)\varepsilon_i.$$

Let

$$B\left(b_{1n}\right) = \mathbb{P}\left\{\sup_{x_\beta}\left|\hat{L}_1(x_\beta) - L_1(x_\beta)\right| > b_{1n}\right\} \leq a_{1n},$$
$$B(b_{2n}) = \mathbb{P}\left\{\sup_{x_\beta}\left|\hat{L}_2(x_\beta) - L_2(x_\beta)\right| > b_{2n}\right\} \leq a_{2n},$$
$$B(b_{3n}) = \mathbb{P}\left\{\sup_{x_\beta}\left|\hat{g}_{\hat{\beta}_n}(x_\beta) - g_\beta(x_\beta)\right| > b_{3n}\right\} \leq a_{3n}.$$

For a_{kn} and $b_{kn}, k = 1, 2, 3$ derived from Lemmas 15-17 respectively, we denote $a_n := a_{1n} + a_{2n} + a_{3n}$ and $b_n := b_{1n} + b_{2n} + b_{3n}$. Let $\hat{L}(x_\beta) = \hat{L}_1(x_\beta) + \hat{L}_2(x_\beta)$, $L(x_\beta) = L_1(x_\beta) + L_2(x_\beta)$, $\hat{m}_n(\cdot) = \hat{L}(\cdot)/\hat{g}_{\hat{\beta}_n}(\cdot)$, and $\tilde{m}_n(\cdot) = L(\cdot)/\hat{g}_{\hat{\beta}_n}(\cdot)$. Note that

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{x_\beta \in \mathcal{W}} \left|\hat{g}_{\hat{\beta}_n}(x_\beta)\right| |\hat{m}_n(x_\beta) - m_0(x_\beta)| > (1 + C_m) b_n\right) \\
&\leq \mathbb{P}\left(\sup_{x_\beta \in \mathcal{W}} \left|\hat{g}_{\hat{\beta}_n}(x_\beta) \{\hat{m}_n(x_\beta) - \tilde{m}(x_\beta)\}\right| > (1 + C_m) b_n\right) \\
&\quad + \mathbb{P}\left(\sup_{x_\beta \in \mathcal{W}} \left|\hat{g}_{\hat{\beta}_n}(x_\beta) \{\tilde{m}_n(x_\beta) - m_0(x_\beta)\}\right| > (1 + C_m) b_n\right) \\
&\leq \mathbb{P}\left(\sup_{x_\beta \in \mathcal{W}} \left|\hat{L}(x_\beta) - L(x_\beta)\right| > (1 + C_m) b_n\right) \\
&\quad + \mathbb{P}\left(\sup_{x_\beta \in \mathcal{W}} \left|\hat{g}_{\hat{\beta}_n}(x_\beta) \{\tilde{m}_n(x_\beta) - m_0(x_\beta)\}\right| > (1 + C_m) b_n\right)
\end{aligned}
$$

Moreover, by Assumption 9, $\sup_{x_\beta \in \mathcal{W}} g_\beta(x_\beta) \geq c_f > 0$, and with probability $1 - c/n$ for some constant $c > 0$,

$$
\sup_{x_\beta \in \mathcal{W}} |\tilde{m}_n(x_\beta) - m_0(x_\beta)| \lesssim \sqrt{\frac{\log n}{nh}} + h^2 \lesssim b_n,
$$

we conclude that with probability approaching one we have

$$
\sup_{x_\beta \in \mathcal{W}} \left|\hat{g}_{\hat{\beta}_n}(x_\beta) \{\tilde{m}_n(x_\beta) - m_0(x_\beta)\}\right| \lesssim b_n,
$$

where $b_n = O\left(c_{1n}^2/h^2 + \sqrt{\log np/nh} + h^2\right)$. Then using a union bound argument and the results in Lemma 15-16, we have

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{x_\beta \in \mathcal{W}} |\hat{m}_n(x_\beta) - m_0(x_\beta)| > \frac{2(1 + C_m)}{c_f} b_n\right) \\
&\leq \mathbb{P}\left(\sup_{x_\beta \in \mathcal{W}} \left|\hat{g}_{\hat{\beta}_n}(x_\beta)\right| |\hat{m}_n(x_\beta) - m_0(x_\beta)| > (1 + C_m) b_n\right) + \mathbb{P}\left(\left|\hat{g}_{\hat{\beta}_n}(x_\beta)\right| < c_f/2\right) \\
&\lesssim a_n.
\end{aligned}
$$

The proof for $\hat{\mu}_{n,j}$ follows an analogous argument upon some notational changes. It is clear that all upper bounds involved can be made independent of the coordinate value of covariates. □

We apply the above uniform convergence result with $x_\beta = X_i^\top \hat{\beta}_n$. Under the Lipschitz continuity of the true link function $m_0(\cdot)$ and the uniform boundedness of covariates in the trimmed support, we can easily bound

$$\left| m_0(X_i^\top \hat{\beta}_n) - m_0(X_i^\top \beta_0) \right| \lesssim |\hat{\beta}_n - \beta_0| = O_p(c_{1n}),$$

and

$$\max_{2 \le j \le p} \left| \mu_j(X_i^\top \hat{\beta}_n) - \mu_j(X_i^\top \beta_0) \right| \lesssim |\hat{\beta}_n - \beta_0| = O_p(c_{1n}),$$

Lemma 14. Under Assumptions in Theorem 2, we have that

$$\sup_{x_\beta} |\hat{m}_n'(x_\beta) - m_0'(x_\beta)| = O_p(c_{1n} + \tilde{b}_n),$$

where $\tilde{b}_n := O\left(c_{1n}^2/h^3 + \sqrt{\log np/(nh^2)} + h^2\right)$.

Proof. Note that

$$\begin{aligned}
&\sup_{x_\beta} |\hat{m}_n'(x_\beta) - m_0'(x_\beta)| \\
&\lesssim \sum_{0 \le \mu \le 1} \sup_{x_\beta} \left| D^{1-\mu} \hat{L}(x_\beta) D^\mu \left[\hat{g}_{\hat{\beta}_n}^{-1}(x_\beta) \right] - D^{1-\mu} L(x_\beta) D^\mu \left[g_{\beta_0}^{-1}(x_\beta) \right] \right|.
\end{aligned}$$

Thus, to save space the details are skipped because the rest applies almost the same strategy we used in Lemma 13. □

Lemma 15. Under Assumptions in Theorem 2, we have that

$$\mathbb{P}\left(\sup_{x_\beta \in \mathcal{W}} \left| \hat{L}_1(x_\beta) - L_1(x_\beta) \right| > b_{1n} \right) \le a_{1n},$$

where

$$a_{1n} \lesssim 1/p + (1/np) + q_n, \tag{S.0.45}$$

$$b_{1n} \lesssim c_{1n} + c_{1n}\left(\sqrt{\frac{\log np}{nh}} + \frac{(\log np + \log p)\sqrt{\log n}}{nh^2} + \frac{1}{n^4h^3}\sqrt{\frac{16\log\{(1+p)p\}}{np}}\right) + c_{1n}^2 h^{-2}, \tag{S.0.46}$$

Proof. Since

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x_\beta}\left|\hat{L}_1(x_\beta) - L_1(x_\beta)\right| > b_{1n}\right) \\
&\leq \mathbb{P}\left(\sup_{x_\beta}\left|\hat{L}_1(x_\beta) - L_1(x_\beta)\right| > b_{1n}, \mathcal{A}_n\right) + \mathbb{P}\left(\mathcal{A}_n^c\right) \\
&\leq \mathbb{P}\left(\sup_{x_\beta}\left|\hat{L}_1(x_\beta) - L_1(x_\beta)\right| > b_{1n}, \mathcal{A}_n\right) + q_n,
\end{aligned}$$

it suffices to find an upper bound for

$$\mathbb{P}\left(\sup_{x_\beta}\left|\hat{L}_1(x_\beta) - L_1(x_\beta)\right| > b_{1n}, \mathcal{A}_n\right).$$

Consider the following decomposition:

$$\begin{aligned}
\hat{L}_1(x_\beta) =& L_1(x_\beta) + \frac{1}{n}\sum_{i=1}^n \frac{1}{h^2}K'\left(\frac{X_i^\top\beta_0 - x_\beta}{h}\right)\left\{m_0(X_i^\top\beta_0) - m_0(x_\beta)\right\}X_i^\top\left(\hat{\beta}_n - \beta_0\right) \\
&+ \frac{1}{n}\sum_{i=1}^n h^{-1}R_{i,\beta_0}(x_\beta)\left\{m_0(X_i^\top\beta_0) - m_0(x_\beta)\right\} \\
:=& L_1(x_\beta) + \hat{L}_{11}(x_\beta) + \hat{L}_{12}(x_\beta),
\end{aligned}$$

where

$$R_{i,\beta_0}(x_\beta) := \int_{(X_i^\top\beta_0 - x_\beta)/h}^{(X_i^\top\hat{\beta}_n - x_\beta)/h} K''(t)\left(\frac{X_i^\top\hat{\beta}_n - x_\beta}{h} - t\right)dt, 1 \leq i \leq n.$$

Thus,

$$\mathbb{P}\left(\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_1(x_\beta)-L_1(x_\beta)\right|>\epsilon_1(t),\mathcal{A}_n\right)$$
$$\leq\mathbb{P}\left(\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_{11}(x_\beta)>\epsilon_1(t),\mathcal{A}_n\right|\right)+\mathbb{P}\left(\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_{12}(x_\beta)\right|>\epsilon_1(t),\mathcal{A}_n\right).$$

We first find an upper bound for $\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_{11}(x_\beta)\right|$ on $\mathcal{A}_n$.

Let $\xi_{in,j}(x_\beta)=h^{-2}K'\left(h^{-1}\left(X_i^\top\beta-x_\beta\right)\right)\left\{m\left(X_i^\top\beta\right)-m(x_\beta)\right\}X_{ij}$ and $\xi^*_{in,j}(x_\beta)=\xi_{in,j}(x_\beta)-\mathbb{E}\left[\xi_{in,j}(x_\beta)\right]$. Let $\xi_{in}(x_\beta)=\left(\xi_{in,1}(x_\beta),\ldots,\xi_{in,p}(x_\beta)\right)^\top$ be a $p\times 1$ vector. Then

$$\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_{11}(x_\beta)\right|\leq\left|\hat{\beta}_n-\beta_0\right|_1\left[\sup_{x_\beta\in\mathcal{W}}\left\|\frac{1}{n}\sum_{i=1}^n\xi^*_{in}(x_\beta)\right\|_\infty+\sup_{x_\beta\in\mathcal{W}}\left\|\mathbb{E}\left[\xi_{in}(x_\beta)\right]\right\|_\infty\right]$$
$$:=\left|\hat{\beta}_n-\beta_0\right|_1\left[\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_{111}(x_\beta)\right|+\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_{112}(x_\beta)\right|\right].$$

We discretize $\mathcal{W}$ by equally spaced $a_0=x_{\beta,0}<x_{\beta,1}<...<x_{\beta,M_n}=b_0$, $M_n=n^4-1$. Then

$$\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_{111}(x_\beta)\right|=\sup_{x_\beta\in\mathcal{W}}\left\|\frac{1}{n}\sum_{i=1}^n\xi^*_{in}(x_\beta)\right\|_\infty=\max_{1\leq j\leq p}\sup_{x_\beta\in\mathcal{W}}\left|\frac{1}{n}\sum_{i=1}^n\xi^*_{in,j}(x_\beta)\right|$$
$$\leq\max_{1\leq j\leq p}\max_{0\leq\tau\leq M_n}\left|\frac{1}{n}\sum_{i=1}^n\xi^*_{in,j}(x_{\beta,\tau})\right|+\max_{1\leq j\leq p}\max_{0\leq\tau\leq M_n-1}\sup_{x_\beta\in[x_{\beta,\tau},x_{\beta,\tau+1}]}\left|\frac{1}{n}\sum_{i=1}^n\left[\xi^*_{in,j}(x_\beta)-\xi^*_{in,j}(x_{\beta,\tau})\right]\right|.$$

To bound $\max_{1\leq j\leq p}\max_{0\leq\tau\leq M_n}\left|\frac{1}{n}\sum_{i=1}^n\xi^*_{in,j}(x_{\beta,\tau})\right|$, first notice that

$$\max_{1\leq i\leq n}\max_{1\leq j\leq p}\sup_{x_\beta}\left|\xi_{in,j}(x_\beta)\right|\leq h^{-2}\|K'\|_\infty\|m'\|_\infty h\max_{1\leq j\leq p}|X_{ij}|\leq ch^{-1}\max_{1\leq i\leq n}\max_{1\leq j\leq p}|X_{ij}|.$$

Moreover, for some constants $C,C'>0$ and $\mu_{j,k}(x_\beta)=\mathbb{E}[X_j^k|X^\top\beta_0=x_\beta]$,

$$\mathbb{E}\left[\xi_{in,j}\right]\leq Cm'(x_\beta)g_{\beta_0}(x_\beta)\mu_{j,1}(x_\beta)\int vK'(v)dv+o(1),\tag{S.0.47}$$

where $\int v^2K'(v)dv=0$. We also have that

$$\mathbb{E}\left[\xi^2_{in,k}\right]=h^{-1}m'(x_\beta)^2g_{\beta_0}(x_\beta)\mu_{j,2}(x_\beta)\int v^2\left(K''(v)\right)^2dv+o(h^{-1})$$
$$\leq C'h^{-1}m'(x_\beta)^2g_{\beta_0}(x_\beta)\mu_{j,2}(x_\beta)\int v^2\left(K''(v)\right)^2dv$$

Then Theorem 3.4 of Kuchibhotla and Chakrabortty (2018) implies that for any $t \geq 0$

$$\begin{aligned}
&\mathbb{P}\left(\max_{0\leq\tau\leq M_n}\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\xi^*_{in,j}(x_{\beta,\tau})\right| \geq \left\{C_1 + C_2\frac{(t+\sqrt{\log p})}{\sqrt{nh}} + \frac{C_3\sqrt{\log n}(t+\log p)}{nh}\right\}\right) \\
&\leq \sum_{\tau=0}^{M_n}\mathbb{P}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\xi^*_{in,j}(x_{\beta,\tau})\right| \geq \left\{C_1 + C_2\frac{(t+\sqrt{\log p})}{\sqrt{nh}} + \frac{C_3\sqrt{\log n}(t+\log p)}{nh}\right\}\right) \\
&\leq 3M_n\exp(-t).
\end{aligned}$$

Next, consider any $x_\beta \in [x_{\beta,\tau}, x_{\beta,\tau+1}], \tau = 0, 1, \ldots, M_n - 1$. We have

$$\begin{aligned}
&|x_{in,j}(x_\beta) - x_{in,j}(x_{\beta,\tau+1})| \\
&= h^{-2}|X_{ij}|\left|K'\left(\frac{X_i^\top\beta_0 - x_\beta}{h}\right)\left\{m_0(X_i^\top\beta_0;\beta_0) - m_0(x_\beta)\right\}\right. \\
&\left. -K'\left(\frac{X_i^\top\beta_0 - x_{\beta,\tau}}{h}\right)\left\{m_0(X_i^\top\beta_0;\beta_0) - m_0(x_{\beta,\tau})\right\}\right| \\
&\leq h^{-2}\left\{\|K'\|_\infty\|m_0'\|_\infty + 2\|K''\|_\infty\|m_0\|_\infty h^{-1}\right\}|x_\beta - x_{\beta,\tau+1}||X_{ij}| \\
&\lesssim h^{-3}(b_0-a_0)(2p)^{-1/2}|X_{ij}| \lesssim h^{-3}n^{-4}(2p)^{-1/2}|X_{ij}|,
\end{aligned}$$

which yields that for $(\varrho_1, \ldots, \varrho_n)$ being independent Rademacher variables with $\mathbb{P}(\varrho_i = 1) = \mathbb{P}(\varrho_i = -1) = 1/2$ for each i,

$$\begin{aligned}
&\mathbb{E}\left[\max_{1\leq j\leq p}|(\mathbb{E}_n - \mathbb{E})\left\{\xi_{in,j}(x_\beta) - \xi_{in,j}(x_{\beta,\tau+1})\right\}|\right] \leq 2\mathbb{E}\left[\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\varrho_i\left\{\xi_{in,j}(x_\beta) - \xi_{in,j}(x_{\beta,\tau+1})\right\}\right|\right] \\
&\lesssim 2h^{-3}n^{-4}(2p)^{-1/2}\mathbb{E}\max_{1\leq j\leq p}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\varrho_i X_{ij}\right|\right] \lesssim h^{-3}n^{-4}(2p)^{-1/2}\sqrt{\frac{2\log(2+2p)}{n}}
\end{aligned}$$

where the first and second inequalities follow by, respectively, the symmetrization theorem (Theorem 14.3 of Bühlmann and van de Geer (2011)) and the contraction theorem (Theorem 14.4 of Bühlmann and van de Geer (2011)), and the third inequality follows by Hoeffding's moment inequality (Lemma 14.14 of Bühlmann and van de Geer (2011)). Then Massart's inequality (Theorem 14.2 of Bühlmann and van de Geer (2011)) implies

that with probability at least $1-2t'$ for any $t' \geq 0$,

$$\begin{aligned}
&\max_{1\leq j\leq p} \left|\left(\mathbb{E}_n - \mathbb{E}\right)\left\{\xi_{in,j}(x_\beta) - \xi_{in,j}(x_{\beta,\tau+1})\right\}\right| \\
&\lesssim h^{-3}n^{-4}(2p)^{-1/2}\left\{2\sqrt{\frac{2\log(2+2p)}{n}} + \sqrt{\frac{8\log(1/2t')}{n}}\right\} \\
&\leq h^{-3}n^{-4}(2p)^{-1/2}\sqrt{\frac{16\log\left\{(1+p)/t'\right\}}{n}},
\end{aligned}$$

with probability at least $1-2t'$. Therefore, with probability $1 - M_n\exp(-t^2) - 2t'$, on $\mathcal{A}_n$

$$\begin{aligned}
&\max_{1\leq j\leq p}\sup_{x_\beta\in\mathcal{W}} \left|\left(\mathbb{E}_n - \mathbb{E}\right)\xi_{in,j}(x_\beta)\right| \\
&\leq \max_{0\leq\tau\leq M_n}\max_{1\leq j\leq p}\left|\sum_{i=1}^n \xi^*_{in,j}(x_{\beta,\tau})\right| + \max_{0\leq\tau\leq M_n-1}\max_{1\leq j\leq p}\sup_{x_\beta\in[x_{\beta,\tau},x_{\beta,\tau+1}]}\left|\mathbb{E}_n\xi^*_{in,j}(x_\beta) - \xi^*_{in,j}(x_{\beta,\tau})\right| \\
&\lesssim \left(\frac{(t+\sqrt{\log p})}{\sqrt{nh}} + \frac{(t^2+\log p)\sqrt{\log n}}{nh^2}\right) + \frac{1}{h^3n^4}\sqrt{\frac{16\log\left\{(1+p)/t'\right\}}{np}}.
\end{aligned}$$

Combine (S.0.47), on $\mathcal{A}_n$, we have with probability at least $1 - M_n\exp(-t^2) - 2t'$,

$$\begin{aligned}
&\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_{11}(x_\beta)\right| \\
&\lesssim c_{1n} + c_{1n}\left(\left(\frac{(t+\sqrt{\log p})}{\sqrt{nh}} + \frac{(t^2+\log p)\sqrt{\log n}}{nh^2}\right) + \frac{1}{h^3n^4}\sqrt{\frac{16\log\left\{(1+p)/t'\right\}}{np}}\right).
\end{aligned}$$

Next, we consider $\hat{L}_{12}(x_\beta)$ on $\mathcal{A}_n$. Because

$$\sup_{x_\beta\in\mathcal{W}}|R_{i,\beta_0}(x_\beta)| \leq \|K''\|_\infty h^{-2}\left|X_i^\top(\hat{\beta}_n - \beta_0)\right|^2 \leq \|K''\|_\infty \max_{1\leq j\leq p} h^{-1}|X_{ij}|^2\left|\hat{\beta}_{0,n} - \beta_0\right|_1^2,$$

then for constant $C_m < \infty$ such that $\|m'\|_\infty < C_m$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x_\beta \in \mathcal{W}} \left|\hat{L}_{12}(x_\beta)\right| > c_{1n}^2 h^{-2} C_K C_m, \mathcal{A}_n\right) \\
&= \mathbb{P}\left(\sup_{x_\beta \in \mathcal{W}} \left|\frac{1}{nh}\sum_{i=1}^n R_{i,\beta_0}(x_\beta)\left\{m_0\left(X_i^\top \beta_0\right) - m_0(x_\beta)\right\}\right| > c_{1n}^2 h^{-2} C_K C_m, \mathcal{A}_n\right) \\
&\leq \mathbb{P}\left(\max_{1\leq j\leq p}\left|\frac{c_{1n}^2 C_m}{nh^2}\sum_{i=1}^n \left\{|X_{ij}|^2\right\}\right| > \epsilon(t)\right),
\end{aligned}$$

where by Bernstein's inequality for a constant $C_X > 0$,

$$\mathbb{P}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^n |X_{ij}|^2 - \mathbb{E}|X_{ij}|^2\right| \geq C_X\sqrt{\frac{\log(2p)}{n}}\right) \leq 1/(2p),$$

which further implies that

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq j\leq p}\left|\frac{c_{1n}^2 C_m}{nh^2}\sum_{i=1}^n \left\{|X_{ij}|^2\right\}\right| > 2c_{1n}^2 h^{-2} C_K C_m\right) \\
&\leq \mathbb{P}\left(\max_{1\leq j\leq p}\frac{1}{n}\sum_{i=1}^n |X_{ij}|^2 \geq \mathbb{E}|X_{ij}|^2 + C_K\sqrt{\frac{\log(2p)}{n}}\right) \leq 1/(2p),
\end{aligned}$$

i.e.,

$$\mathbb{P}\left(\sup_{x_\beta \in \mathcal{W}} \left|\hat{L}_{12}(x_\beta)\right| > c_{1n}^2 h^{-2} C_K C_m, \mathcal{A}_n\right) \leq 1/(2p).$$

To summarize, with probability at least $1 - M_n \exp(-t^2) - 2t' - 1/(2p) - q_n$,

$$\begin{aligned}
&\sup_{x_\beta \in \mathcal{W}} \left|\hat{L}_1(x_\beta) - L_1(x_\beta)\right| \\
&\lesssim c_{1n} + c_{1n}\left(\left(\frac{(t+\sqrt{\log p})}{\sqrt{nh}} + \frac{(t^2+\log p)\sqrt{\log n}}{nh^2}\right) + \frac{1}{h^3 n^4}\sqrt{\frac{16\log\left\{(1+p)/t'\right\}}{np}}\right) + c_{1n}^2 h^{-2}.
\end{aligned}$$

Let $t = 3\sqrt{\log(np)}$ and $t' = p^{-1}$, then the conclusion follows for a_{1n} and b_{1n} defined in (S.0.45) and (S.0.46) respectively. □

Lemma 16. Under Assumptions in Theorem 2, we have that

$$\mathbb{P}\left(\sup_{x_\beta \in \mathcal{W}} \left|\hat{L}_2(x_\beta) - L_2(x_\beta)\right| > b_{2n}\right) \leq a_{2n},$$

where

$$a_{2n} \lesssim 1/(np) + p + q_n,$$

$$\begin{aligned} b_{2n} \lesssim\, & c_{1n} + c_{1n}\left(\left(\frac{\sqrt{\log np}}{\sqrt{nh}} + \frac{(\log np)\sqrt{\log n}}{nh^2} + \frac{1}{n^4h^3}\sqrt{\frac{\log\{(1+p)p\}}{np}}\right)\right) \\ & + c_{1n}^2\left(\frac{\sqrt{\log np}}{\sqrt{nh^2}} + \frac{(\log np)\sqrt{\log n}}{nh^3} + \frac{1}{n^4h^4}\sqrt{\frac{\log\{(1+p)p\}}{np}}\right) \\ & + c_{1n}^3\left(\frac{1}{h^4} + \frac{\sqrt{\log(np)}}{nh^4} + \frac{(\log(np))\sqrt{n}}{h^4}\right). \end{aligned}$$

Proof. We write $\hat{L}_2(x_\beta)$ as

$$\begin{aligned} \hat{L}_2(x_\beta) =& \frac{1}{n}\sum_{i=1}^n K_h(X_i^\top\hat{\beta}_n - x_\beta)\varepsilon_i \\ =& \frac{1}{n}\sum_{i=1}^n K_h(X_i^\top\beta_0 - x_\beta)\varepsilon_i + \frac{1}{nh^2}\sum_{i=1}^n K'\left(\frac{X_i^\top\beta_0 - x_\beta}{h}\right)\varepsilon_i X_i^\top\left(\hat{\beta}_n - \beta_0\right) \\ & + (\hat{\beta}_n - \beta_0)^\top\frac{1}{2nh^3}\sum_{i=1}^n K''\left(\frac{X_i^\top\beta_0 - x_\beta}{h}\right)\varepsilon_i X_i X_i^\top\left(\hat{\beta}_n - \beta_0\right) \\ & + \frac{1}{nh}\sum_{i=1}^n \varepsilon_i\int_{(X_i^\top\beta_0 - x_\beta)/h}^{(X_i^\top\hat{\beta}_{0,n} - x_\beta)/h}\left\{K''(t) - K''\left(\frac{X_i^\top\beta_0 - x_\beta}{h}\right)\right\}\left(\frac{X_i^\top\hat{\beta}_n - x_\beta}{h} - t\right)dt \\ :=& L_2(x_\beta) + \hat{L}_{21}(x_\beta) + \hat{L}_{22}(x_\beta) + \hat{L}_{23}(x_\beta). \end{aligned}$$

Similar to the argument for $\hat{L}_{11}(x_\beta)$, one can show that with probability at least $1 - M_n\exp(-t^2) - 2t'$, on the set $\mathcal{A}_n$,

$$\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_{21}(x_\beta)\right| \lesssim c_{1n}\left(\frac{t + \sqrt{\log p}}{\sqrt{nh}} + \frac{(t^2 + \log p)\sqrt{\log n}}{nh^2} + \frac{1}{n^4h^3}\sqrt{\frac{\log p\{(1+p)/t'\}}{np}}\right).$$

and with probability at least $1 - M_n\exp(-t^2) - 2t'$, on the set $\mathcal{A}_n$,

$$\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_{22}(x_\beta)\right| \lesssim c_{1n}^2\left(\left(\frac{(t + \sqrt{\log p})}{\sqrt{nh^2}} + \frac{(t^2 + \log p)\sqrt{\log n}}{nh^3}\right) + \frac{1}{n^4h^4}\sqrt{\frac{\log\{(1+p)/t'\}}{np}}\right).$$

We next consider $\hat{L}_{23}(x_\beta)$. For

$$\tilde{R}_{i,\beta}(x_\beta) := \left| \int_{(X_i^\top \beta_0 - x_\beta)/h}^{(X_i^\top \hat{\beta}_n - x_\beta)/h} \left\{ K''(t) - K'' \left(\frac{X_i^\top \beta_0 - x_\beta}{h} \right) \right\} \left(\frac{X_i^\top \hat{\beta}_n - x_\beta}{h} - t \right) dt \right|, \tag{S.0.48}$$

we find the bound such that

$$\sup_{x_\beta} \max_{1 \le i \le n} \left| \tilde{R}_{i,\beta}(x_\beta) \right| \lesssim \max_{1 \le i \le n} h^{-3} \left| X_i^\top (\hat{\beta}_n - \beta_0) \right|^3 \lesssim \max_{1 \le i \le n} \max_{1 \le j \le p} h^{-3} |X_{ij}|^3 \left| \hat{\beta}_n - \beta_0 \right|_1^3.$$

so on the set $\mathcal{A}_n$, with probability $1 - M_n \exp(-t^2)$,

$$\sup_{x_\beta \in \mathcal{W}} \left| \hat{L}_{23}(x_\beta) \right| = \sup_{x_\beta \in \mathcal{W}} \left| \frac{1}{nh} \sum_{i=1}^n \varepsilon_i \tilde{R}_{i,\beta}(x_\beta) \right| \lesssim c_{1n}^3 \left(h^{-4} + \frac{t}{nh^4} + \frac{t^2 \sqrt{\log n}}{h^4} \right).$$

Hence, with probability at least $1 - 3M_n \exp(-t^2) - 2t' - q_n$,

$$\begin{aligned}
&\sup_{x_\beta \in \mathcal{W}} \left| \hat{L}_2(x_\beta) - L_2(x_\beta) \right| \\
&\lesssim c_{1n} + c_{1n} \left(\frac{t + \sqrt{\log p}}{\sqrt{nh}} + \frac{(t^2 + \log p)\sqrt{\log n}}{nh^2} + \frac{1}{n^4 h^3} \sqrt{\frac{\log p \{(1+p)/t'\}}{np}} \right) \\
&\quad + c_{1n}^2 \left(\frac{(t + \sqrt{\log p})}{\sqrt{nh^2}} + \frac{(t^2 + \log p)\sqrt{\log n}}{nh^3} + \frac{1}{n^4 h^4} \sqrt{\frac{\log p \{(1+p)/t'\}}{np}} \right) \\
&\quad + c_{1n}^3 \left(h^{-4} + \frac{t}{nh^4} + \frac{t^2 \sqrt{\log n}}{h^4} \right).
\end{aligned}$$

Thus, the conclusion follows by setting $t = 3\sqrt{\log(np)}$ and $t' = 1/2p$. □

Lemma 17. Under Assumptions in Theorem 2, we have that

$$\mathbb{P} \left\{ \sup_{x_\beta} \left| \hat{g}_{\hat{\beta}_n}(x_\beta) - g_\beta(x_\beta) \right| > b_{3n} \right\} \le a_{3n},$$

where

$$a_{3n} \lesssim 1/(np) + 1/p + q_n,$$

$$
\begin{aligned}
b_{3n} &\lesssim c_{1n} + c_{1n}\left(\frac{\sqrt{\log np}}{\sqrt{nh}} + \frac{(\log np)\sqrt{\log n}}{nh^2} + \frac{1}{n^4h^3}\sqrt{\frac{\log\{(1+p)p\}}{np}}\right) \\
&\lesssim c_{1n}^2\left(\frac{\sqrt{\log(np)}}{nh^5} + \frac{(\log(np))\sqrt{\log n}}{nh^3} + \frac{1}{n^4h^4}\sqrt{\frac{\log\{(1+p)p\}}{np}}\right) \\
&\quad + c_{1n}^3\left(\frac{1}{h^4} + \frac{\sqrt{\log(np)}}{nh^4} + \frac{\log(np)\sqrt{\log n}}{h^4}\right) + \sqrt{\frac{\log(np)}{nh}} + h^2.
\end{aligned}
$$

Proof. Consider $\hat{g}_{\hat{\beta}_n}(x_\beta)$ such that

$$
\begin{aligned}
\hat{g}_{\hat{\beta}_n}(x_\beta) =& \hat{g}_{\beta_0}(x_\beta) + \frac{1}{nh^2}\sum_{i=1}^n K'\left(\frac{X_i^\top\beta_0 - x_\beta}{h}\right)X_i^\top\left(\hat{\beta}_n - \beta_0\right) \\
&+ \left(\hat{\beta}_n - \beta_0\right)^\top\frac{1}{2nh^3}\sum_{i=1}^n K''\left(\frac{X_i^\top\theta_0 - x_\beta}{h}\right)X_iX_i^\top\left(\hat{\beta}_n - \beta_0\right) + \frac{1}{nh}\sum_{i=1}^n \tilde{R}_{i,\beta_0}(x_\beta) \\
:=& L_3(x_\beta) + \hat{L}_{31}(x_\beta) + \hat{L}_{32}(x_\beta) + \hat{L}_{33}(x_\beta).
\end{aligned}
$$

where $\tilde{R}_{i,\beta_0}(x_\beta)$ is the same defined in (S.0.48) so similar argument yields that on $\mathcal{A}_n$,

$$
\mathbb{P}\left(\sup_{x_\beta}\left|\hat{L}_{33}(x_\beta)\right| > c_{1n}^3\left(\frac{1}{h^4} + \frac{t}{nh^4} + \frac{t^2\sqrt{\log n}}{h^4}\right)\right) \lesssim 1 - 3\exp(-t^2).
$$

For $\hat{L}_{32}(x_\beta)$ note that

$$
\max_{1\le j,k\le p}\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_{32}(x_\beta)\right| \le \max_{1\le j,k\le p}\sup_{x_\beta}\left|\frac{1}{2nh^3}\sum_{i=1}^n K''\left(\frac{X_i^\top\beta_0 - x_\beta}{h}\right)X_{ij}X_{ik}\right|\left|\hat{\beta}_n - \beta_0\right|_1^2,
$$

and

$$
\max_{1\le j,k\le p}\mathbb{E}\left[\left|h^{-3}K''\left(\frac{X_i^\top\beta_0 - x_\beta}{h}\right)X_{ij}X_{ik}\right|\right] = O(h^{-2}).
$$

Because X_{ij} is a bounded random variable, i.e., $|X_{ij}| < M$ a.s. for some constant $M < \infty$, then $\|X_{ij}^2\|_{\psi_2} \le M\|X_{ij}\|_{\psi_2} < \infty$, and $\|X_{ij}^2X_{ik}^2\|_{\psi_1} \le \|X_{ij}^2\|_{\psi_2}\|X_{ik}^2\|_{\psi_2} < \infty, 1 \le j,k \le p$. Then Theorem 3.4 in Kuchibhotla and Chakrabortty (2018) implies that with

probability $1-3\exp(-t^2)$,

$$\begin{aligned}
&\frac{1}{nh^3}\sum_{i=1}^{n}\max_{1\le j,k\le p}\left|K''\left(\frac{X_i^\top\beta_0-x_\beta}{h}\right)X_{ij}X_{ik}\right| \\
&\lesssim (\mathbb{E}_n-\mathbb{E})\max_{1\le j,k\le p}\left|h^{-3}K''\left(\frac{X_i^\top\beta_0-x_\beta}{h}\right)X_{ij}X_{ik}\right|+\max_{1\le j,k\le p}\mathbb{E}\left|h^{-3}K''\left(\frac{X_i^\top\beta_0-x_\beta}{h}\right)X_{ij}X_{ik}\right| \\
&\lesssim \frac{t}{nh^5}+\frac{t^2\sqrt{\log n}}{nh^3}+h^{-2},
\end{aligned}$$

Moreover, by the same discretization technique, similar to the way to bound $\hat{L}_{11}(x_\beta)$, we conclude that with probability $1-M_n\exp(-t^2)-2t'-3/p$ for $t,t'>0$, on the set $\mathcal{A}_n$ we have

$$\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_{32}(x_\beta)\right|\lesssim c_{1n}^2h^{-2}+c_{1n}^2\left(\frac{t}{nh^5}+\frac{t^2\sqrt{\log n}}{nh^3}+\frac{1}{n^4h^4}\sqrt{\frac{\log\{(1+p)/t'\}}{np}}\right)$$

Similar to the argument for $\hat{L}_{11}(x_\beta)$, we can also show that on $\mathcal{A}_n$ with probability at least $1-M_n\exp(-t^2)-2t'$,

$$\sup_{x_\beta\in\mathcal{W}}\left|\hat{L}_{31}(x_\beta)\right|\lesssim c_{1n}+c_{1n}\left(\left(\frac{(t+\sqrt{\log p})}{\sqrt{nh}}+\frac{(t^2+\log p)\sqrt{\log n}}{nh^2}\right)+\frac{1}{n^4h^3}\sqrt{\frac{\log\{(1+p)/t'\}}{np}}\right)$$

Thus, with probability at least $1-M_n\exp(-t^2)-M_n\exp(-n\tilde{t}^2)-4t'-q_n-3/p$,

$$\begin{aligned}
&\sup_{x_\beta\in\mathcal{W}}\left|\hat{g}_{\hat{\beta}_n}(x_\beta)-\hat{g}_{\beta_0}(x_\beta)\right| \\
&\lesssim c_{1n}+c_{1n}\left(\left(\frac{(t+\sqrt{\log p})}{\sqrt{nh}}+\frac{(t^2+\log p)\sqrt{\log n}}{nh^2}\right)+\frac{1}{n^4h^3}\sqrt{\frac{\log\{(1+p)/t'\}}{np}}\right) \\
&\quad+c_{1n}^2h^{-2}+c_{1n}^2\left(\left(\frac{t}{nh^5}+\frac{t^2\sqrt{\log n}}{nh^3}\right)+\frac{1}{n^4h^4}\sqrt{\frac{\log\{(1+p)/t'\}}{np}}\right) \\
&\quad+c_{1n}^3\left(\frac{1}{h^4}+\frac{t}{nh^4}+\frac{t^2\sqrt{\log n}}{h^4}\right).
\end{aligned}$$

Moreover, with probability at least $1-3\exp(-t^2)$,

$$\sup_{x_\beta}|\hat{g}_{\beta_0}(x_\beta)-g_{\beta_0}(x_\beta)|\lesssim\frac{t}{\sqrt{nh}}+\frac{t^2\sqrt{\log n}}{nh}+h^2,$$

so with probability at least $1-M_n\exp(-t^2)-M_n\exp(-n\tilde{t}^2)-4t'-2q_n-3/p-3\exp(-t^2)$, we have

$$\begin{aligned}
&\sup_{x_\beta}\left|\hat{g}_{\hat{\beta}_n}(x_\beta)-g_{\beta_0}(x_\beta)\right| \\
&\lesssim c_{1n}+c_{1n}\left(\left(\frac{(t+\sqrt{\log p})}{\sqrt{nh}}+\frac{(t^2+\log p)\sqrt{\log n}}{nh^2}\right)+\frac{1}{n^4h^3}\sqrt{\frac{\log\{(1+p)/t'\}}{np}}\right) \\
&\quad+c_{1n}^2h^{-2}+c_{1n}^2\left(\left(\frac{t}{nh^5}+\frac{t^2\sqrt{\log n}}{nh^3}\right)+\frac{1}{n^4h^4}\sqrt{\frac{\log\{(1+p)/t'\}}{np}}\right) \\
&\quad+c_{1n}^3\left(\frac{1}{h^4}+\frac{t}{nh^4}+\frac{t^2\sqrt{\log n}}{h^4}\right)+\frac{t}{\sqrt{nh}}+\frac{t^2\sqrt{\log n}}{nh}+h^2.
\end{aligned}$$

The conclusion follows by letting $t=3\sqrt{\log(np)}$ and $t'=1/2p$. □

Appendix D: Selection of Tuning Parameters

There are several tuning parameters that need to be determined in our pilot estimator. We use plug-in type bandwidth selectors h_j for $1\le j\le p$ from Sheather and Jones (1991) for all marginal densities and their derivatives. This choice is made because it is a data-driven algorithm with an efficient implementation and it satisfies the rate condition in Section 3.1. The tuning parameter λ_Ω for estimating the precision matrix is fixed at $\sqrt{\log p/n}$.

The key tuning parameter λ_β in the thresholding step is chosen by the $V-$fold cross-validation as follows. Let a kernel estimator for the conditional mean be denoted by $\widehat{m}_n(\cdot;\hat{\beta}_n)$ for the given $\hat{\beta}_n$. We choose its bandwidth for fitting this local constant function by a plug-in rule proposed by Ruppert et al. (1995). Let $I_1\cup\cdots\cup I_V$ be a partition of $\{1,\cdots,n\}$. Referring to a given set I_k, we adopt the estimator $\widehat{m}^{[-I_k]}(X_i^\top\hat{\beta}_{n,\lambda_\beta};\hat{\beta}_{n,\lambda_\beta})$ built by applying the estimation scheme with the tuning parameter corresponding to λ_β

only on the subsample $(Y_i, X_i)_{i \notin I_k}$. The V-fold cross-validation risk is defined as

$$R_{CV}(\lambda_\beta) = \frac{1}{V} \sum_{k=1}^{V} \sum_{i \in I_k} (Y_i - \widehat{m}^{[-I_k]}(X_i^\top \hat{\beta}_{n,\lambda_\beta}; \hat{\beta}_{n,\lambda_\beta}))^2 \tag{S.0.49}$$

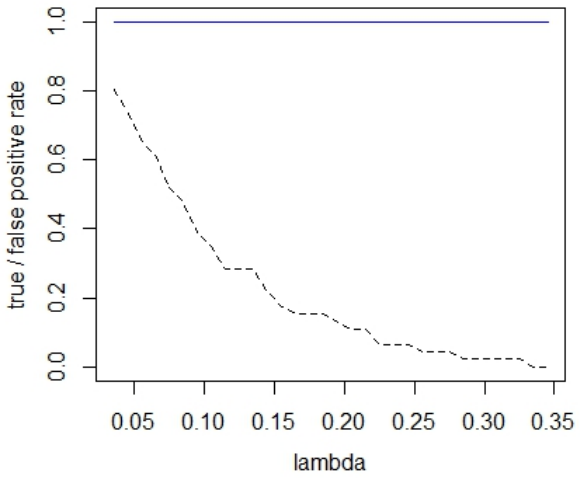
The V-fold CV proceeds by selecting the tuning parameter $\hat{\lambda}_\beta$ with the smallest V-fold $R_{CV}(\lambda_\beta)$.

We demonstrate the influence of λ_β on the performance of our estimators with LASSO and adaptive LASSO versions through a simple numerical exercise. We generate a random sample of size 500 using the specification of DGP I in Section 4.1 with independent normal covariates. The dimensionality of covariates is $p = 50$. We plot the path of $|\hat{\beta}_n - \beta_0|_2$ and the true/false positive rates[3] as λ_β varies. The U-shaped pattern of the estimation error suggests that an appropriate choice of λ_β matters a lot in practice. For both estimation methods, the true positive rates stay at 1, which means that all relevant regressors are included. The false positive rate drops steadily for LASSO penalty as λ_β gradually increases, whereas it quickly reaches zero for the adaptive LASSO version. This phenomenon agrees with the conventional wisdom that LASSO tends to over-select the variables. Note that once λ_β passes the turning point, a high percentage of zeros in the coefficient can deteriorate the estimation precision. In practice, a proper choice of λ_β is desirable that balances the estimation accuracy and variable selection.

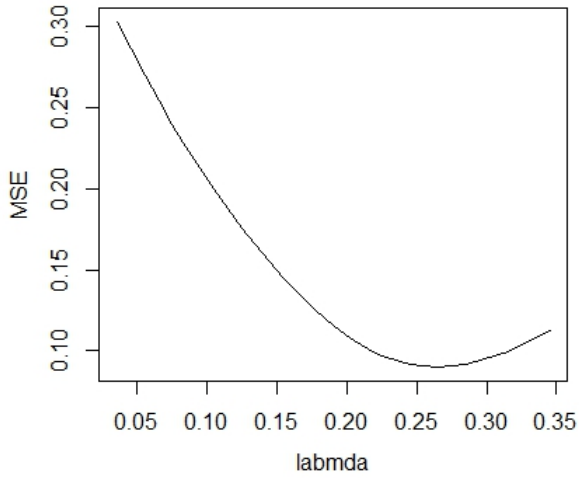
[3]The true positive rate is measured by $\sum_{j=1}^{s} \mathbb{I}\{\hat{\beta}_j \neq 0\}/s$ and the false positive rate is $\sum_{j=s+1}^{p} \mathbb{I}\{\hat{\beta}_j \neq 0\}/(p-s)$.

Figure 1: Plots of True/False Positive Rates and MSE versus λ

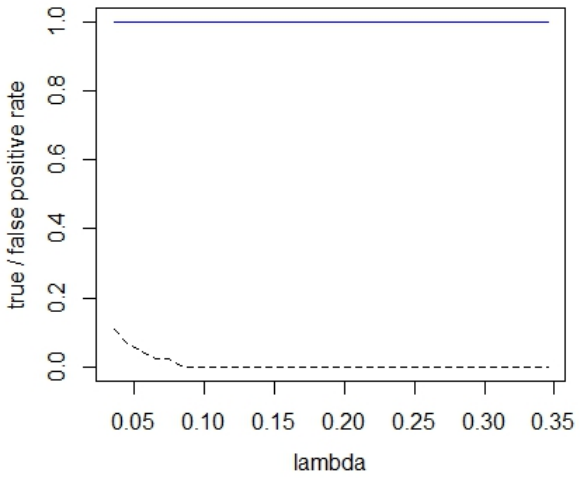
(a) T/F positive rates for LASSO.



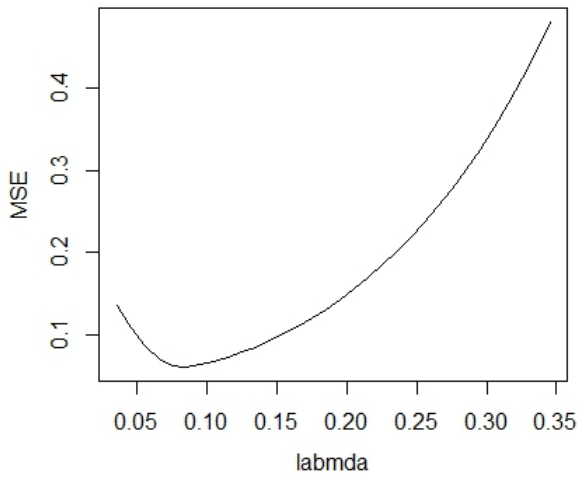
(b) MSE for LASSO.



(c) T/F positive rates for Adaptive LASSO.



(d) MSE for Adaptive LASSO.



References

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Cai, Tony, T., Liu, W., and Zhou, H. (2016). Estimation sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Annals of Statistics*, 44:455–488.

Giné, E. and Nickl, R. (2009). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probability Theory and Related Fields*, 143:569–596.

Giné, E. and Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press.

Hardle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association*, 84:986–995.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58.

Kuchibhotla, A. K. and Chakrabortty, A. (2018). Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric gaussian copula graphical models. *Annals of Statistics*, 40:2293–2326.

Major, P. (2006). An estimate on the supremum of a nice class of stochastic integrals and u-statistics. *Probability Theory and Related Fields*, 134:489–537.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American statistical Association*, 90:1257–1270.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 53:683–690.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.

Yang, Z., Balasubramanian, K., and Liu, H. (2017). On stein's identity and near-optimal estimation in high-dimensional index models. *International Conference on Machine Learning*, pages 3851–3860.