Received: 9 January 2022 | Revised: 29 August 2022 | Accepted: 8 November 2022

DOI: 10.1111/sjos.12628

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Robust sure independence screening for nonpolynomial dimensional generalized linear models

Abhik Ghosh[1] | **Erica Ponzi**[2] | **Torkjel Sandanger**[3] | **Magne Thoresen**[2]

[1]Indian Statistical Institute, Kolkata, India

[2]Oslo Center for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo, Norway

[3]UiT, The Arctic University of Norway, Tromsø, Norway

Correspondence
Magne Thoresen, Oslo Center for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo, Norway.
Email: magne.thoresen@medisin.uio.no

Funding information
INSPIRE Faculty Research Grant, Department of Science and Technology, Government of India, Grant/Award Number: SRG/2020/000072; European Research Council, Grant/Award Number: ERC-2008-AdG-232997; Norwegian Research Council, Grant/Award Numbers: 248804, 262111

Abstract

We consider the problem of variable screening in ultra-high-dimensional generalized linear models (GLMs) of nonpolynomial orders. Since the popular SIS approach is extremely unstable in the presence of contamination and noise, we discuss a new robust screening procedure based on the minimum density power divergence estimator (MDPDE) of the marginal regression coefficients. Our proposed screening procedure performs well under pure and contaminated data scenarios. We provide a theoretical motivation for the use of marginal MDPDEs for variable screening from both population as well as sample aspects; in particular, we prove that the marginal MDPDEs are uniformly consistent leading to the sure screening property of our proposed algorithm. Finally, we propose an appropriate MDPDE-based extension for robust conditional screening in GLMs along with the derivation of its sure screening property. Our proposed methods are illustrated through extensive numerical studies along with an interesting real data application.

KEYWORDS

conditional screening, density power divergence, DPD-SIS, high-dimensional statistics, robustness, sure independence screening

1 | INTRODUCTION

The class of generalized linear models (GLMs) is a rich class of parametric regression models that allows to study a wide range of relationship structures for different types of response data, which makes the GLMs one of the most popular statistical tools for real-life applications across many disciplines. Let us consider the GLM in its canonical form: given a set of p predictor variables $X_1, X_2, \ldots, X_p$, the scalar response variable Y follows a distribution from the exponential family having density

$$f(y;\theta) = \exp\{y\theta - b(\theta) + c(y)\}, \tag{1}$$

for some appropriate (known) functions $b(\cdot)$ and $c(\cdot)$ and the unknown canonical parameter θ. For simplicity, we do not consider a dispersion parameter in the model (e.g., logistic or Poisson regression) although it can easily be incorporated in all our methodological discussions and the theories derived throughout the paper with slight modifications. We concentrate on the mean regression model for θ given by

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = b'(\theta) = g^{-1}\left(\boldsymbol{x}^T\boldsymbol{\beta}\right), \tag{2}$$

where $\boldsymbol{X} = (X_0 = 1, X_1, \ldots, X_p)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}$ is the vector of unknown regression coefficients and g is a monotone differentiable link function. Given independent and identically distributed (IID) data $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, our objective is to fit a GLM by efficiently estimating $\boldsymbol{\beta}$ and use it for subsequent inference.

Commonly, the regression coefficient $\boldsymbol{\beta}$ is estimated through likelihood-based approaches (or suitable extensions) under the classical low dimensional setup ($p < n$). However, recent advancements of technologies across disciplines generates data on a large number of possible covariates with limited observations leading to $p \gg n$, known as the high-dimensional set-up. In this paper, we consider ultra-high-dimensional GLMs with the number of covariates being of nonpolynomial (NP) order of n, that is, $\log(p) = O(n^l)$ for some $0 < l < 1$. However, to perform meaningful inference in such situations, we need to assume sparsity of the model — only $s \ll n$ covariates (out of the vast pool of p covariates) are actually important to explain the variability in the response. There are several statistical procedures like LASSO or other regularized approaches (Buhlmann & Van De Geer, 2011; Fan & Li, 2001; Ghosh & Majumdar, 2020; Giraud, 2014; Hastie et al., 2015) to simultaneously select these important variables and estimate the corresponding (nonzero) regression coefficients. Although they often work reasonably well in moderately high dimensions, their computation becomes highly extensive in ultra-high-dimensional setups. Therefore, it is more efficient to first reduce the set of all covariates to a sufficiently small size (maybe $< n$) through some initial screening procedure. Among these, the most popular is the sure independence screening (SIS) proposed by Fan and Lv (2008) for the linear regression model and later extended to GLMs by Fan and Song (2010). The SIS has become extremely popular for its simplicity, elegance, computational speed as well as the theoretical guarantees for sure screening of the true model, asymptotically with probability tending to one. Subsequently, SIS has been extended to different types of data and associated statistical problems; see, for example Barut et al. (2016), Luo et al. (2014), Saldana & Feng (2018), Zhao & Li (2012) among many others.

The SIS is commonly applied, besides other applications, in the context of omics data which generally include different types of noise and outliers; the same issue of data contamination also often arises in other real-life applications involving extremely large number of features. However,

the SIS procedure and its extensions are mainly based on the Pearson correlation or the maximum likelihood estimator (MLE) of the marginal regression coefficients, both of which are nonrobust against possible outliers in the data. This nonrobustness of SIS was, in fact, first noted in the discussion of the original paper itself by Gather and Guddat (2008). They proposed an alternative robust SIS using the Gnanadesikan–Kettenring correlation in place of the usual correlation while ranking the covariates in a linear regression model. Subsequently, several other robust versions of SIS, mostly nonparametric in nature, were proposed for the high-dimensional linear regression model only (Hall & Miller, 2009; Li et al., 2012b; Li, Peng, et al., 2012; Mu & Xiong, 2014; Wang et al., 2017; Zhong, 2014). Although these nonparametric versions of SIS can potentially be applied to the GLMs as well (possibly with appropriate modulation), they were never theoretically studied in the literature. Thus, there is a need for a robust variable screening procedure for the ultra-high-dimensional GLM with proper theoretical guarantees of its sure screening property. We aim to fill this gap in the literature by developing a robust sure screening procedure for the general class of GLMs.

Compared to any nonparametric robust procedure, a parametric robust approach is known to provide significantly higher efficiency when the assumed model is valid for a majority of the data except for the noise/contamination part (Hampel et al., 1986). Recently, a robust parametric version of SIS, namely the DPD-SIS, has been proposed for ultra-high-dimensional linear regression models by Ghosh and Thoresen (2021). This DPD-SIS is empirically studied and found to have significantly improved performance compared to the other existing nonparametric SIS procedures under data contamination, although no theoretical guarantees are provided in Ghosh and Thoresen (2021). They have proposed to use the marginal regression approach as in Fan and Song (2010) but to estimate the marginal regression slopes by the robust minimum density power divergence estimator (MDPDE) instead of the MLE. These MDPDEs were first proposed by Basu et al. (1998) as a robust generalization of the MLE for simple IID problems. Due to their high robustness along with their high efficiency and simple computation, the MDPDEs are subsequently extended to more complex statistical models. For linear regression models, the MDPDEs are studied by, for example, Ghosh and Basu (2013). For different GLMs as well, the MDPDEs are seen to provide highly efficient and robust parameter estimates under the classical low-dimensional setups (Basu et al., 2011, 2017, 2021; Ghosh, 2019; Ghosh & Basu, 2016). In this paper, we utilize the MDPDEs under the marginal regression approach to develop a robust variable screening procedure for ultra-high-dimensional GLMs, as an extension of the robust DPD-SIS of Ghosh and Thoresen (2021). Additionally, we prove that the proposed procedure satisfies the sure screening property for the general class of GLMs and that it can also control the selection of false positives under appropriate assumptions, which needed quite nontrivial extensions of the existing theories. To our knowledge, this is the first robust variable screening procedure for the general class of GLMs (beyond simple linear regression) with proper theoretical guarantees.

Further, we also extend our proposed DPD-SIS to develop a robust conditional screening procedure under NP-dimensional GLMs, which we will refer to as the conditional DPD-SIS. The conditional SIS (CSIS), proposed by Barut et al. (2016), has been a natural extension of the usual SIS that can take care of additional information (whenever available) about some previously chosen important variables. Among several advantages, most importantly, CSIS helps to select the hidden important variables. However, just like usual SIS, the CSIS is also extremely nonrobust under data contamination, and there is no literature available on its (parametric) robust version. Our proposed conditional DPD-SIS serves this purpose. Its population-level justifications as well as the sample-level sure screening property are also derived rigorously under reasonably practical assumptions.

We illustrate the proposed DPD-SIS through appropriate simulation studies of ultra-high-dimensional GLMs, in addition to an interesting real data application involving the search for biomarkers in lung cancer. For simplicity in presentation, all proofs are deferred to the Appendixes.

2 | THE PROPOSED DPD-SIS FOR NP-DIMENSIONAL GLMS

Let us consider the GLM described in (1) and (2) with ultra-high-dimensional covariates; for simplicity in presentation, throughout the rest of the article we will assume canonical link function so that $b' = g^{-1}$ and hence $\theta = \boldsymbol{x}^T\boldsymbol{\beta}$ in (2). Suppose that the true value of the regression coefficient $\boldsymbol{\beta}$ is denoted by $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \ldots, \beta_{0p})^T$. We assume that the true model, denoted as $\mathcal{M}_0 = \{1 \le j \le p : \beta_{0j} \ne 0\}$, is sparse with model size $s = |\mathcal{M}_0| < n$. Our aim is to perform an initial screening of the covariates in a robust manner such that it includes all the truly important variables corresponding to $\mathcal{M}_0$; this property is referred to as the sure screening property in the literature.

We follow the marginal regression approach of Fan and Song (2010) to consider the GLM for Y based on X_j (plus an intercept term) separately for each $j = 1, \ldots, p$; let us denote the associated regression coefficients for these marginal models by $\boldsymbol{\beta}_j^M = (\beta_{j0}^M, \beta_j^M)$, respectively. In this context, we also assume that the covariates are standardized so that $E(X_j) = 0$ and $E(X_j^2) = 1$ for all $j = 1, \ldots, p$. However, instead of using the MLE of $\boldsymbol{\beta}_j^M$ as in Fan and Song (2010), we propose to use their MDPDEs. Note that each marginal GLM is of low dimension, having only two parameters in $\boldsymbol{\beta}_j^M$. Hence, we can follow Ghosh and Basu (2016) to define their MDPDE as the minimizer of an appropriately defined average DPD measure between the observed data $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, and the assumed GLM density (1). After simplifications, the MDPDE of $\boldsymbol{\beta}_j^M$ with (given) tuning parameter $\alpha > 0$ is defined as

$$\widehat{\boldsymbol{\beta}}_j^{M\alpha} = \left(\widehat{\beta}_{j0}^{M\alpha}, \widehat{\beta}_j^{M\alpha}\right) = \arg\min_{\beta_{j0}, \beta_j} \frac{1}{n} \sum_{i=1}^{n} l_\alpha\left(y_i, \beta_{j0} + \beta_j x_{ij}\right), \tag{3}$$

where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^T$ for each $i = 1, \ldots, n$, and $l_\alpha(y, \theta) = \int f(s; \theta)^{1+\alpha} ds - \left(1 + \frac{1}{\alpha}\right) f(y; \theta)^\alpha + \frac{1}{\alpha}$. The tuning parameter α in the definition of the MDPDE is known to control the trade-off between the efficiency under pure data and the robustness against data contamination. In fact, $l_0(y, \theta) := \lim_{\alpha \to 0} l_\alpha(y, \theta) = -\log f(y, \theta)$ so that the MDPDE at $\alpha = 0$ (in a limiting sense) is nothing but the most efficient and highly nonrobust MLE. For $\alpha > 0$, the MDPDE provides a robust extension of the MLE having increasing robustness with a slight loss in efficiency as α increases (Ghosh & Basu, 2016). For any given $\alpha \ge 0$, each MDPDE $\widehat{\boldsymbol{\beta}}_j^{M\alpha}$ can also be obtained by solving the corresponding estimating equation

$$\sum_{i=1}^{n} \psi_\alpha(y_i, \beta_0 + \beta_j x_{ij})[1, x_{ij}]^T = \mathbf{0}_2,$$

and

$$\psi_\alpha(y, \theta) = \big(y - b'(\theta)\big) f(y; \theta)^\alpha - \xi_\alpha(\theta), \tag{4}$$

where $\xi_\alpha(\theta) = \int \big(y - b'(\theta)\big) f(y;\theta)^{1+\alpha} dy$. Note that $\xi_0(\theta) = 0$ and hence $\psi_0(y,\theta) = \big(y - b'(\theta)\big)$, the usual score function, which again leads to the MLE. The above form of ψ_α, used in the estimating equation of the MDPDE, indeed justifies the particular choice of this DPD loss function l_α for achieving robust solutions. Note that, the first part of the ψ_α function (except ξ_α, which is there to make the estimating equation unbiased at the model) is basically a weighted score function with the weight being $f(y;\theta)^\alpha$. So, for any $\alpha > 0$, observations in the region of low model probability (the outlying observations coming from data contamination) get downweighted with regard to their contributions to the MDPDE estimating equation; as a consequence, the effects of such outlying observations get reduced leading to robust parameter estimates and subsequently to robust screening results (see Section 5 for further discussions on robustness and Section 6.4 for comments on the choice of α).

Based on the MDPDEs $\widehat{\boldsymbol{\beta}}_j^{M\alpha}$ for the marginal regression coefficients, for each $j = 1,\ldots,p$, and a given $\alpha > 0$, we choose a suitable predefined threshold γ_n and select the variables in the set

$$\widehat{M}_\alpha(\gamma_n) = \left\{ 1 \le j \le p \,:\, \left|\widehat{\beta}_j^{M\alpha}\right| \ge \gamma_n \right\}. \tag{5}$$

By choosing γ_n appropriately, we can reduce the number of covariates from a large p to any smaller target, say $d < n$, so that the subsequent computation becomes feasible. With these d variables selected in $\widehat{M}_\alpha(\gamma_n)$, we can then fit any appropriate low-dimensional estimation procedure or regularized approach to get the estimated coefficient vector, say $\widehat{\boldsymbol{\beta}}_d = \big(\widehat{\beta}_{d0}, \widehat{\beta}_{d1},\ldots,\widehat{\beta}_{dd}\big)^T$ and subsequently the final model $\widehat{\mathcal{M}} = \left\{ 1 \le j \le p \,:\, \widehat{\beta}_{dj} \ne 0 \right\}$.

In our DPD-SIS, we suggest to choose γ_n from the target of attaining a fixed model size. However, in practice, it may be chosen in several other ways, for example, controlling the false positives, or prediction error, etc. (see Section 3 for an optimal rate of γ_n). Note that the case $\alpha = 0$ reduces to the ordinary SIS. Along the same lines, we will show the sure screening property of our proposed DPD-SIS, so that asymptotically $\widehat{M}_\alpha(\gamma_n)$ contains the true model $\mathcal{M}_0$ with probability tending to one, for any given $\alpha > 0$.

3 | SURE SCREENING PROPERTY OF THE DPD-SIS

We are considering the GLM in (1) and (2) with the canonical link function and the true sparse parameter value $\boldsymbol{\beta}_0$ having support $\mathcal{M}_0$ of size $s = |\mathcal{M}_0| < n$, as described in previous sections. Recall that, under the ultra-high-dimensional setup considered here, the number of covariates $p = p_n$ is assumed to grow exponentially with the sample size n; we also allow the true model size $s = s_n$ to depend on n. Further we assume that the data $(y_i, \boldsymbol{x}_i)$, for $i = 1,\ldots,n$, are IID from a true joint distribution $\Pi(dy, d\boldsymbol{x}) = F_{\beta_0}(dy|\boldsymbol{x})Q(d\boldsymbol{x})$, where F_{β_0} is the (conditional) distribution corresponding to the GLM in (1) and (2) and Q is the marginal distribution of the covariates (for which no model is assumed). Then, it is straightforward from the definition of ψ_α in (4) that, for any $\alpha \ge 0$,

$$E\big[\psi_\alpha\big(Y, \boldsymbol{x}^T\boldsymbol{\beta}_0\big)|\boldsymbol{X} = \boldsymbol{x}\big] = E\big[\psi_\alpha\big(Y, \boldsymbol{x}_1^T\boldsymbol{\beta}_{01}\big)|\boldsymbol{X} = \boldsymbol{x}\big] = \boldsymbol{0}, \tag{6}$$

where we have used the notation $\boldsymbol{x}_1 = (x_j : j \in \mathcal{M}_0)^T$ and $\boldsymbol{\beta}_{01} = (\beta_{0j} : j \in \mathcal{M}_0)^T$. Without loss of generality, let us assume $\mathcal{M}_0 = \{1, 2,\ldots, s\}$ and consider the partitions $\boldsymbol{x}^T = \big(\boldsymbol{x}_1^T, \boldsymbol{x}_2^T\big)$,

$\boldsymbol{\beta}_0^T = \left(\boldsymbol{\beta}_{01}^T, \boldsymbol{\beta}_{02}^T\right)$, $\boldsymbol{x}_i^T = \left(\boldsymbol{x}_{i1}^T, \boldsymbol{x}_{i2}^T\right)$, $\boldsymbol{\beta}^T = \left(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T\right)$ and so on for any p-vectors, where the first partitions (e.g., $\boldsymbol{x}_1, \boldsymbol{\beta}_{01}, \boldsymbol{x}_{i1}, \boldsymbol{\beta}_1$, etc.) are of length s.

3.1 | Population-level results

We first investigate the proposed DPD-SIS at population level. The population version (functional) of the marginal MDPDE $\widehat{\boldsymbol{\beta}}_j^{M\alpha}$, defined in (3), is given by

$$\boldsymbol{\beta}_j^{M\alpha} = \left(\beta_{j0}^{M\alpha}, \beta_j^{M\alpha}\right) = \arg\min_{\beta_{j0}, \beta_j} E\left[l_\alpha\left(Y, \beta_{j0} + \beta_j X_j\right)\right]. \tag{7}$$

This marginal MDPDE functional $\boldsymbol{\beta}_j^{M\alpha}$ then satisfies the estimating equations

$$E\left[\psi_\alpha\left(Y, \beta_0 + \beta_j X_j\right)\right] = 0, \quad E\left[\psi_\alpha\left(Y, \beta_0 + \beta_j X_j\right) X_j\right] = 0. \tag{8}$$

Let us denote $B_\alpha(v(\boldsymbol{x})) = b'(\boldsymbol{x}^T\boldsymbol{\beta}_0) - E[\psi_\alpha(Y, v(\boldsymbol{x}))|\boldsymbol{X} = \boldsymbol{x}]$. Clearly $B_\alpha(\boldsymbol{x}^T\boldsymbol{\beta}_0) = b'(\boldsymbol{x}^T\boldsymbol{\beta}_0)$ but $B_\alpha(\beta_{j0}^{M\alpha} + \beta_j^{M\alpha} x_j)$ do not necessarily equal $b'(\beta_{j0}^{M\alpha} + \beta_j^{M\alpha} x_j)$. However, at $\alpha = 0$ we always have $B_0(v(\boldsymbol{x})) = b'(v(\boldsymbol{x}))$ for any $v(\boldsymbol{x})$. Using Equations (6) and (8), we have proved the following two theorems. They show why the proposed DPD-SIS is expected to have the targeted sure screening property, at population level; the proofs are given in Appendix B for brevity.

Theorem 1. *For a given $\alpha \geq 0$, and for any $j = 1, \ldots, p$, the marginal MDPDE functional $\beta_j^{M\alpha} = 0$ if and only if $Cov(b'(\boldsymbol{X}^T\boldsymbol{\beta}_0), X_j) = \mathrm{Cov}(Y, X_j) = 0$.*

Theorem 2. *Given any $\alpha \geq 0$, and $j \in \mathcal{M}_0$, assume that either (B1) or (B2) holds*:

(B1) $B'_\alpha(\cdot)$ is bounded.
(B2) $B_\alpha(t)$ is strictly increasing in t and $G_\alpha(|x|) = \sup_{|u| \leq |x|} |B_\alpha(u)|$ satisfies

$$E\left[G_\alpha(a|X_j|)|X_j| I(|X_j| \geq n^\eta)\right] \leq cn^{-\kappa}, \text{ for some constants } a, c > 0, \ \eta \in (0, \kappa). \tag{9}$$

Then, whenever there exists a constant $c_1 > 0$ such that $\left|\mathrm{Cov}(b'(\boldsymbol{X}^T\boldsymbol{\beta}_0), X_j)\right| \geq c_1 n^{-\kappa}$, we have $\min_{j \in \mathcal{M}_0} |\beta_j^{M\alpha}| \geq c_2 n^{-\kappa}$, for some constant $c_2 > 0$.

The above two theorems are similar to theorems 2 and 3 of Fan and Song (2010) from the context of SIS, although the assumptions in our Theorem 2 are required on the quantity $B_\alpha(\cdot)$ instead of $b'(\cdot)$ which coincide at $\alpha = 0$. For any $\alpha \geq 0$, one can indeed show that $B'_\alpha(\cdot)$ is bounded for the normal and the logistic regression models whereas Condition (9) holds for Poisson regression with suitable covariates; see Appendix A.

In the same spirit of Fan and Song (2010), Theorem 1 implies that if the set of unimportant covariates $\{X_j : j \notin \mathcal{M}_0\}$ is independent of the set of important covariates $\{X_j : j \in \mathcal{M}_0\}$ then $\beta_j^{M\alpha} = 0$ for all $j \notin \mathcal{M}_0$ and all $\alpha \geq 0$. Further, note that, an important covariate X_j having nonzero correlation with the response has a marginal regression coefficient $\beta_j^{M\alpha} \neq 0$. These together indicate the existence of a threshold γ_n satisfying $\min_{j \in \mathcal{M}_0} |\beta_j^{M\alpha}| \geq \gamma_n$ and $\max_{j \notin \mathcal{M}_0} |\beta_j^{M\alpha}| = 0$. This forms the theoretical basis for the model selection consistency of the proposed DPD-SIS with any $\alpha \geq 0$ and justifies our proposal as an authenticate screening criterion.

On the other hand, Theorem 2 provides the conditions to yield $\min_{j \in \mathcal{M}_0} |\beta_j^{M\alpha}| \geq O(n^{-\kappa})$ for some $\kappa < 1/2$, which can be interpreted as the marginal signals being stronger than the maximum

stochastic noise level. It is an intuitive necessity for the proposed DPD-SIS, the sample version (5), to have the sure screening property. Other than the α-dependent assumption, one crucial condition in Theorem 2 is $\left|\mathrm{Cov}(b'(\boldsymbol{X}^T\boldsymbol{\beta}_0), X_j)\right| \geq c_1 n^{-\kappa}$ which is the same as required by the usual SIS in Fan and Song (2010); it can be further simplified for jointly Gaussian covariates following Proposition 1 of Fan and Song (2010) and the discussion thereafter. This theorem also provides the necessary framework to achieve sparsity in the final selected model (5).

3.2 | Sample-level results

We first show that the marginal MDPDEs $\widehat{\boldsymbol{\beta}}_j^{M\alpha}$, $j = 1,\ldots,p$, are uniformly consistent at an exponential rate which leads to the sure screening property (sample level) of our proposed DPD-SIS. In this regard, let us note that the marginal MDPDE functional $\boldsymbol{\beta}_j^{M\alpha}$ is unique and is an interior point of the parameter space by convexity of the DPD loss function $l_\alpha\left(Y, \beta_0 + \beta_j X_j\right)$ in $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_j)$ for each j. So, we can restrict the minimization of the marginal DPD loss function over the compact set $\mathcal{B} = \{|\beta_{j0}| \leq B, |\beta_j| \leq B\}$ for some large enough constant $B > 0$ such that $\boldsymbol{\beta}_j^{M\alpha}$ is also an interior point of $\mathcal{B}$. For each $j = 1,\ldots,p$, let $\boldsymbol{X}_j = (1, X_j)^T$ and define the matrices

$$\boldsymbol{J}_{j,\alpha}(\boldsymbol{\beta}_j) = E\left[\nabla^2 l_\alpha\left(Y, \beta_0 + \beta_j X_j\right)\right] = (1+\alpha)E\left[\Gamma_\alpha\left(\boldsymbol{X}_j^T\boldsymbol{\beta}_j\right)\boldsymbol{X}_j\boldsymbol{X}_j^T\right], \tag{10}$$

$$\begin{aligned}\boldsymbol{K}_{j,\alpha}(\boldsymbol{\beta}_j) &= E\left[\left(\nabla l_\alpha\left(Y, \beta_0 + \beta_j X_j\right)\right)\left(\nabla l_\alpha\left(Y, \beta_0 + \beta_j X_j\right)\right)^T\right] \\ &= (1+\alpha)^2 E\left[\left\{\Gamma_{2\alpha}\left(\boldsymbol{X}_j^T\boldsymbol{\beta}_j\right) - \xi_\alpha^2\left(\boldsymbol{X}_j^T\boldsymbol{\beta}_j\right)\right\}\boldsymbol{X}_j\boldsymbol{X}_j^T\right],\end{aligned} \tag{11}$$

where $\Gamma_\alpha(\theta) = \int (y - b'(\theta))^2 f(y;\theta)^{1+\alpha} dy$. Then, the following assumptions are needed for our subsequent theoretical investigation of the DPD-SIS; here $\alpha \geq 0$ is a fixed given tuning parameter and $\Lambda_{\min}[\cdot]$ and $\Lambda_{\max}[\cdot]$, respectively, denote the minimum and maximum eigenvalues of its argument matrix.

(A1) The GLM is such that the density f^α in (1) is bounded by some constant $L_\alpha > 0$, and $b''(\cdot)$ is continuous and positive. Also, $|\xi_\alpha(\theta)|$ is nondecreasing in θ.

(A2) For all $\boldsymbol{\beta}_j \in \mathcal{B}$, there exists some constant $V > 0$ such that $\Lambda_{\min}\left[\boldsymbol{J}_{j,\alpha}(\boldsymbol{\beta}_j)\right] \geq V$ uniformly over $j = 1,\ldots,p$.

(A3) $\boldsymbol{K}_{j,\alpha}(\boldsymbol{\beta}_j^{M\alpha})$ is finite and positive definite for each $j = 1,\ldots,p$. Also, the norm $||\boldsymbol{K}_{j,\alpha}(\boldsymbol{\beta}_j)||_{\mathcal{B}} = \sup_{\boldsymbol{\beta}_j \in \mathcal{B}, ||\boldsymbol{u}||=1} ||\boldsymbol{K}_{j,\alpha}(\boldsymbol{\beta}_j)^{1/2}\boldsymbol{u}||$ is bounded from above for each j.

(A4) There exist an $\epsilon_1 > 0$ and a large constant $K_n > 0$, such that

$$\sup_{\boldsymbol{\beta}_j \in \mathcal{B}: ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^{M\alpha}|| \leq \epsilon_1} E\left[|B_\alpha\left(\boldsymbol{X}_j^T\boldsymbol{\beta}_j\right)|||\boldsymbol{X}_j||_2 I(|X_j| > K_n)\right] \leq o\left(\frac{1}{n}\right), \quad \text{for all } j = 1, 2, \ldots, p.$$

(A5) The distribution of the covariate X_j is such that, for sufficiently large $t > 0$ and some positive constants m_0, m_1, m_2, m_3 and τ, we have $P(|X_j| > t) = (m_1 - m_2)e^{-m_0 t^\tau}$, for $j = 1, 2, \ldots, p$, and

$$E\left[\exp\left(b\left(\boldsymbol{X}^T\boldsymbol{\beta}_0 + m_3\right) - b\left(\boldsymbol{X}^T\boldsymbol{\beta}_0\right)\right)\right] + E\left[\exp\left(b\left(\boldsymbol{X}^T\boldsymbol{\beta}_0 - m_3\right) - b\left(\boldsymbol{X}^T\boldsymbol{\beta}_0\right)\right)\right] \leq m_2.$$

(A6) $\mathrm{Var}(\boldsymbol{X}^T\boldsymbol{\beta}_0)$ is bounded both from below and above by finite positive constants.

(A7) Either $b''(\cdot)$ is bounded or $\widetilde{\boldsymbol{X}} = (X_1,\ldots,X_p)^T$ follows an elliptically contoured distribution with variance Σ_1 and $\left|E\left[b'(\boldsymbol{X}^T\boldsymbol{\beta}_0)(\boldsymbol{X}^T\boldsymbol{\beta}_0 - \beta_{00})\right]\right|$ is bounded.

Note that Assumptions (A1)–(A6) are appropriate extensions of the assumptions made by Fan and Song (2010) to prove the sure screening property of the usual SIS; they coincide at $\alpha = 0$ since $L_0 = 1$, $\xi_0 \equiv 0$, $\boldsymbol{J}_{j,0} = \boldsymbol{K}_{j,0} = E\left[b''\left(\boldsymbol{X}_j^T\boldsymbol{\beta}_j\right)\boldsymbol{X}_j\boldsymbol{X}_j^T\right]$ and $B_0(v(\boldsymbol{x})) = b'(v(\boldsymbol{x}))$ for any $v(\boldsymbol{x})$. For any $\alpha > 0$, Assumption (A1) clearly holds for most common examples of GLM including the normal, Poisson and logistic regression models; other assumptions are also valid for these GLMs under mild sufficient conditions. Interestingly, Assumptions (A5)–(A7) are independent of the choice of α and are exactly the same as Assumptions (D), (F), and (G) of Fan and Song (2010), respectively. In particular, Assumption (A5) ensures that the covariates and the response variable have light tails; it implies, via lemma 1 of Fan and Song (2010), that

$$P\left(|Y| \geq \frac{m_0}{m_3}t^\tau\right) \leq m_2 e^{-m_0 t^\tau}, \quad \text{for any } t > 0. \tag{12}$$

Assumption (A6), on the other hand, implies that the variance of the response Y is bounded. In fact, denoting the variance of $\boldsymbol{X}$ by $\Sigma = \mathrm{Diag}\{0, \Sigma_1\}$, Assumption (A6) states that $\mathrm{Var}(\boldsymbol{X}^T\boldsymbol{\beta}_0) = \boldsymbol{\beta}_0^T\Sigma\boldsymbol{\beta}_0 = O(1)$. Note that the maximum eigenvalue of Σ_1 in Assumption (A7) is the same as $\Lambda_{\max}(\Sigma)$, which is a positive finite number by Assumption (A6). Further, Assumptions (A6) and (A7) along with the positiveness of $b''(\cdot)$ from Assumption (A1) imply that for any $\boldsymbol{\beta}_j$ in the interior of $\mathcal{B}$ (and hence in particular for $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^{M\alpha}$), we have

$$||\boldsymbol{\beta}_j||_2^2 = O\big(||\Sigma\boldsymbol{\beta}_0||_2^2\big) = O\big(\Lambda_{\max}(\Sigma)\big) = O\big(\Lambda_{\max}(\Sigma_1)\big). \tag{13}$$

Here, the first equality is as shown in the proof of Theorem 5 of Fan and Song (2010) while the remaining equalities are argued above.

Now, under Assumptions (A1)–(A5), we have the exponential convergence result for the marginal MDPDE as presented in the following Lemma.

Lemma 1. *Suppose that (A1)–(A5) hold for a given $\alpha \geq 0$. Then, for any $t > 0$,*

$$P\left(\sqrt{n}\left|\widehat{\beta}_j^{M\alpha} - \beta_j^{M\alpha}\right| \geq \frac{16k_n^{(\alpha)}}{V}(1+t)\right) \leq e^{-\frac{2t^2}{K_n^2}} + nm_1 e^{-m_0 K_n^\tau}, \quad j = 1,\ldots,p, \tag{14}$$

where $k_n^{(\alpha)} = (1+\alpha)\left[\frac{m_0}{m_3}K_n^\tau L_\alpha + |b'(K_nB+B)|L_\alpha + \xi_\alpha(K_nB+B)\right]$.

Note that the constant bounds involved in the above lemma are independent of the index j leading to the uniform convergence of all the marginal regression models through union bound. We will utilize this fact to derive the sure screening property of the proposed DPD-SIS along with its rate of false positive control (based on (13)), which is presented in the following theorem.

Theorem 3. *Let Assumptions (A1)–(A5) hold for a given $\alpha \geq 0$ and $\frac{n^{1-2\kappa}}{(k_nK_n)^2} \to \infty$ as $n \to \infty$, where $k_n = k_n^{(\alpha)}$ is as defined in Lemma 1. Then the following results hold.*

(a) For any given $c_3 > 0$, there exists $C > 0$ such that

$$P\left(\max_{1\le j\le p}|\widehat{\beta}_j^{M\alpha} - \beta_j^{M\alpha}| \ge c_3 n^{-\kappa}\right) \le pR_n, \tag{15}$$

where $R_n = \left[e^{-\frac{n^{1-2\kappa}C}{(k_n K_n)^2}} + nm_1 e^{-m_0 K_n^{\tau}}\right]$.

(b) If additionally the assumptions of Theorem 2 hold, then taking $\gamma_n = c_4 n^{-\kappa}$ with $c_4 \le c_2/2$, we have

$$P\left(\widehat{\mathcal{M}}(\gamma_n) \supset \mathcal{M}_0\right) \ge 1 - sR_n.$$

(c) If additionally Assumptions (A6)–(A7) hold, taking $\gamma_n = c_5 n^{-2\kappa}$, $c_5 > 0$, we get

$$P\left(|\widehat{\mathcal{M}}(\gamma_n)| \le O\left(n^{2\kappa}\Lambda_{\max}(\Sigma)\right)\right) \ge 1 - pR_n.$$

It is important to note that the bound R_n in the above theorem is exactly the same (except for the value of $k_n = k_n^{(\alpha)}$) as obtained by Fan and Song (2010) for usual SIS and it will be exponentially small for standard GLMs with appropriate choices of K_n; see Appendix A for detailed discussions. Thus, along with the additional robustness property, our proposed DPD-SIS at any $\alpha > 0$ also enjoys the same optimal rate of convergence and false discovery control as well as the similar sure screening property as the usual SIS under slightly modified assumptions. This is the most striking benefit of our proposal in the context of robust variable screening under high-dimensionality. Additionally, the sure screening property of the DPD-SIS, as stated in Theorem 3(b), does not depend on the variance and the correlation structure of the covariates for any choices of $\alpha \ge 0$. However, higher correlation among covariates may surely increase the false positive selection which can be seen by the dependence of the size of $\widehat{\mathcal{M}}(\gamma_n)$ selected via the DPD-SIS on Σ or more precisely on $\Lambda_{\max}(\Sigma)$ (Theorem 3(c)). As we have less correlation among covariates and hence smaller values of $\Lambda_{\max}(\Sigma)$, the number of variables selected via our DPD-SIS reduces, leading to less false positives due to its sure independence property. As in the usual SIS, we can also achieve model selection consistency for DPD-SIS at any $\alpha \ge 0$, that is,

$$P\left(\widehat{\mathcal{M}}(\gamma_n) = \mathcal{M}_0\right) = 1 - o(1),$$

under appropriate assumptions on $\Lambda_{\max}(\Sigma)$ along with proper control of K_n. As a particular (extreme) example, it holds with the choice of γ_n as in Theorem 3(b) if we have $\left|\mathrm{Cov}\left(b'\left(\boldsymbol{X}^T\boldsymbol{\beta}_0\right), X_j\right)\right| = o(n^{-\kappa})$ for all $j \notin \mathcal{M}_0$, along with the other necessary conditions of the theorem depending on $\alpha \ge 0$.

4 | ROBUST CONDITIONAL SCREENING: THE DPD-CSIS

Let us now extend the DPD-SIS approach to conditional screening problems in GLMs. Suppose that, along with the setup and notation of Section 2, information is available to always include a set of q covariates, say $\boldsymbol{X}_C$ (with $q < n - 1$ columns), and we need to robustly select variables from the remaining pool of $p - q$ variables (say, $\boldsymbol{X}_D$). For simplicity, in this section, we assume no intercept terms, since that can be easily incorporated within the given $\boldsymbol{X}_C$. Further,

without loss of generality, we assume that $\boldsymbol{X}_C = (X_1,\ldots,X_q)^T$ so that $\boldsymbol{X}_D = (X_{q+1},\ldots,X_p)^T$; denote $\mathcal{C} = \{1,\ldots,q\}$ and $\mathcal{D} = \{q+1,\ldots,p\}$ and hence $\boldsymbol{\beta}_C = (\beta_1,\ldots,\beta_q)^T \in \mathbb{R}^q$ and $\boldsymbol{\beta}_D = (\beta_{q+1},\ldots,\beta_p)^T \in \mathbb{R}^d$. Now, for a given $\alpha \geq 0$, we may choose the variables from $\boldsymbol{X}_D$ based on the marginal MDPDEs defined as

$$\widehat{\boldsymbol{\beta}}_{Cj}^{M\alpha} = \left(\widehat{\boldsymbol{\beta}}_{Cj1}^{M\alpha}, \widehat{\beta}_j^{M\alpha}\right) = \arg\min_{\boldsymbol{\beta}_C,\beta_j} \frac{1}{n}\sum_{i=1}^{n} l_\alpha\left(y_i, \boldsymbol{x}_{iC}^T\boldsymbol{\beta}_C + \beta_j x_{ij}\right), \quad j = q+1,\ldots,p, \tag{16}$$

where $l_\alpha(y,\theta)$ is as defined in Section 2, and $\boldsymbol{x}_{iC}$ is the ith observation on $\boldsymbol{X}_C$. Then, as in (5), given a suitable predefined threshold γ_n, we may select the variables in the set $\widehat{M}_\alpha(\gamma_n|\mathcal{D}) = \left\{q+1 \leq j \leq p : \left|\widehat{\beta}_j^{M\alpha}\right| \geq \gamma_n\right\}$. We refer to this extension as the conditional DPD-SIS, or the DPD-CSIS in short. Clearly, the DPD-CSIS again coincides with the usual CSIS of Barut et al. (2016) at $\alpha = 0$ and provides a robust generalization at $\alpha > 0$. Further, when the conditioning variable set $\boldsymbol{X}_C$ is empty (or contains only the intercept), we are back to our DPD-SIS. We here study the properties of the DPD-CSIS in line with the results derived in Section 3.

Note that, throughout this section concerning CSIS, $\mathcal{M}_0$ corresponds to the covariates from $\boldsymbol{X}_D$ having nonzero regression coefficients, and accordingly we now have $s = |\mathcal{M}_0| < n$.

4.1 | Population-level results: Justifications of DPD-CSIS

Let us continue with the notation of Section 3.1 and additionally assume that $E(X_j|\boldsymbol{X}_C) = 0$ for all $j \in \mathcal{D}$. We define the population version (functional) of $\widehat{\boldsymbol{\beta}}_{Cj}^{M\alpha}$ from (16) as

$$\boldsymbol{\beta}_{Cj}^{M\alpha} = \left(\boldsymbol{\beta}_{Cj1}^{M\alpha}, \beta_j^{M\alpha}\right) = \arg\min_{\boldsymbol{\beta}_C,\beta_j} E\left[l_\alpha\left(Y, \boldsymbol{X}_C^T\boldsymbol{\beta}_C + \beta_j X_j\right)\right], \quad j = q+1,\ldots,p. \tag{17}$$

Additionally, let us define the functional for the baseline parameter given only $\boldsymbol{X}_C$, without any additional variable, as $\boldsymbol{\beta}_C^{M\alpha} = \arg\min_{\beta_C} E\left[l_\alpha\left(Y, \boldsymbol{X}_C^T\boldsymbol{\beta}_C\right)\right]$. Then, throughout all theoretical discussions of DPD-CSIS, as in Barut et al. (2016), we need to assume that the functionals $\boldsymbol{\beta}_{Cj}^{M\alpha}$ and $\boldsymbol{\beta}_C^{M\alpha}$ are unique, that is, the associated marginal problems are fully identifiable. Now, for DPD-CSIS at any given $\alpha \geq 0$, we consider the random variables $m_{\alpha,j}$, for each $j = q+1,\ldots,p$, defined as

$$m_{\alpha,j} = \frac{B_\alpha\left(\boldsymbol{X}_{Cj}^T\boldsymbol{\beta}_{Cj}^{M\alpha}\right) - B_\alpha\left(\boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}\right)}{\boldsymbol{X}_{Cj}^T\boldsymbol{\beta}_{Cj}^{M\alpha} - \boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}}, \tag{18}$$

where $\boldsymbol{X}_{Cj} = (\boldsymbol{X}_C^T, X_j)^T$ for each j and B_α is as defined in Section 3.1. Denote by $\mathcal{M}_{0D} = \mathcal{M}_0 \cap \mathcal{D}$ the indices of the truly important variables in $\boldsymbol{X}_D$. Then, we have the following results, in analogue of Theorems 1 and 2, that justify our DPD-CSIS algorithm as a reasonable procedure for conditional screening. Here, in analogue of Barut et al. (2016), we define $\mathrm{Cov}_L(Y, X_j|\boldsymbol{X}_C) := E\left[(Y - E_L[Y|\boldsymbol{X}_C])\left(X_j - E_L[X_j|\boldsymbol{X}_C]\right)\right]$, for any $j \in \mathcal{D}$, where $E_L[\cdot|\boldsymbol{X}_C]$ denote the best linear regression fit given $\boldsymbol{X}_C$; clearly $E_L[Y|\boldsymbol{X}_C] = b'\left(\boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}\right)$.

Theorem 4. *For a given $\alpha \geq 0$ and any $j \in \mathcal{D}$, the (conditional) marginal MDPDE functional $\beta_j^{M\alpha}$ in (17) is zero if and only if* $\mathrm{Cov}_L(Y, X_j|\boldsymbol{X}_C) = 0$.

Theorem 5. *Given any $\alpha \geq 0$, suppose that $E[m_{\alpha,j}X_j^2] \leq c_2$ uniformly in $j \in \mathcal{D}$, for some constant c_2. If there exist constants $c_1 > 0, \kappa < -1/2$ such that $\left|\text{Cov}_L(Y, X_j|\boldsymbol{X}_C)\right| \geq c_1 n^{-\kappa}$ for all $j \in \mathcal{M}_{0D}$, then we have $\min_{j\in\mathcal{M}_{0D}} |\beta_j^{M\alpha}| \geq c_3 n^{-\kappa}$, for another constant $c_3 > 0$.*

4.2 | Sample-level properties: sure screening via DPD-CSIS

We now extend the results of Section 3.2 for the unconditional DPD-SIS to the case of conditional screening to show the uniform convergence of the associated (conditional) MDPDEs and the resulting sure screening property of the DPD-CSIS. We continue with the notation of Section 3.2 and assume that Assumptions (A1)–(A7) hold with $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_j^{M\alpha}$ replaced by $\boldsymbol{\beta}_{Cj} \in \mathbb{R}^{q+1}$ and $\boldsymbol{\beta}_{Cj}^{M\alpha}$, respectively, in (A2)–(A4). We also assume the following additional condition.

(A8) There exists $C > 0$ such that $\Lambda_{\min}\left(E\left[m_{\alpha,j}\boldsymbol{X}_{Cj}\boldsymbol{X}_{Cj}^T\right]\right) > C$, uniformly over $j \in \mathcal{D}$.

Note that Assumption (A8) is mild (and regular) if B_α is strictly convex implying $m_{\alpha,j} > 0$ almost surely. Further, we define $\boldsymbol{Z} = E\left(E_L[\boldsymbol{X}_D|\boldsymbol{X}_C]\left[\boldsymbol{X}^T\boldsymbol{\beta}_0 - \boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}\right]\right)$ and $\Sigma_{D|C} = E\left(\boldsymbol{X}_D - E_L[\boldsymbol{X}_D|\boldsymbol{X}_C]\right)\left(\boldsymbol{X}_D - E_L[\boldsymbol{X}_D|\boldsymbol{X}_C]\right)^T$. We can show that Assumptions (A6)–(A8) imply the following analogue of (13) for this conditional case, given by

$$||\boldsymbol{\beta}_D||_2^2 = O\left(\Sigma_{D|C} + \boldsymbol{Z}\boldsymbol{Z}^T\right). \tag{19}$$

Then, we have the desired results in analogue to Theorem 3 for the present conditional case of the DPD-CSIS which is presented in the following theorem. The proof is similar to that of Theorem 3, but using (19) instead of (13), and is hence omitted for brevity.

Theorem 6. *Suppose that, for a given $\alpha \geq 0$, Assumptions (A1)–(A5) hold with $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_j^{M\alpha}$ replaced by $\boldsymbol{\beta}_{Cj} \in \mathbb{R}^{q+1}$ and $\boldsymbol{\beta}_{Cj}^{M\alpha}$, respectively, in (A2)–(A4). Also, let $\frac{n^{1-2\kappa}}{(k_n K_n)^2} \to \infty$ as $n \to \infty$, where $k_n = k_n^{(\alpha)}$ is as defined in Lemma 1. Then, the following results hold.*

(a) For any given $c_3 > 0$, there exists $C > 0$ such that

$$P\left(\max_{q+1\leq j\leq p} |\widehat{\beta}_j^{M\alpha} - \beta_j^{M\alpha}| \geq c_3 n^{-\kappa}\right) \leq (p-q)\, R_n, \tag{20}$$

where R_n is as defined in Theorem 3.

(b) If additionally the assumptions of Theorem 5 hold, then taking $\gamma_n = c_4 n^{-\kappa}$ with $c_4 \leq c_2/2$, we have

$$P\left(\widehat{\mathcal{M}}(\gamma_n) \supset \mathcal{M}_0\right) \geq 1 - sR_n.$$

(c) If additionally Assumptions (A6)–(A8) hold, taking $\gamma_n = c_4 n^{-2\kappa}$, $c_4 > 0$, we get

$$P\left(|\widehat{\mathcal{M}}(\gamma_n)| \leq O\left(n^{2\kappa}\Lambda_{\max}\left(\Sigma_{D|C} + \boldsymbol{Z}\boldsymbol{Z}^T\right)\right)\right) \geq 1 - (p-q)\, R_n. \tag{21}$$

Note that the rate of convergence in the above theorem is exactly the same as in the unconditional case (Theorem 3) and that they are in line with the existing literature on variable screening.

In the particular case of the linear regression model, we have $\boldsymbol{Z} = \mathbf{0}$, and hence the result (21) in Theorem 6 reduces to

$$P\left(|\widehat{\mathcal{M}}(\gamma_n)| \leq O\left(n^{2\kappa}\Lambda_{\max}\left(\Sigma_{D|C}\right)\right)\right) \geq 1-(p-q)\,R_n. \tag{22}$$

In general, if we additionally assume $||\boldsymbol{Z}||_2^2 = o\left(\Lambda_{\min}\left(\Sigma_{D|C}\right)\right)$, as in condition 3(iii) of Barut et al. (2016), we can also have (22) instead of (21) in Theorem 6.

5 | ROBUSTNESS PROPERTY: THEORETICAL JUSTIFICATIONS

The robustness of the proposed DPD-SIS and DPD-CSIS under data contamination follows directly from the robustness of the associated marginal MDPDEs $\widehat{\boldsymbol{\beta}}_j^{M\alpha}$. For an intuitive understanding, recall that the MDPDE estimating equation downweights the outliers with a weight $f(y;\theta)^\alpha$ to achieve robustness. As the value of $\alpha > 0$ increases, more down-weighting takes place, reducing the contribution of outliers in the estimation process, which leads to improved robustness of the MDPDEs and the subsequent DPD-SIS or DPD-CSIS procedure. At $\alpha = 0$, there is no down-weighting of the outlying observations and so the resulting MLE or the associated SIS/CSIS are nonrobust under data contamination.

The robustness characteristic of both DPD-SIS and DPD-CSIS, that is, their increasing robustness with increasing $\alpha > 0$, compared to the usual SIS or CSIS (at $\alpha = 0$) can be theoretically justified by the classical influence function analyses under the Huber's ϵ-contamination model (Hampel et al., 1986). The influence function (IF) provides a measure of asymptotic bias, in any statistical functional, caused by infinitesimal contamination at a distant outlying point. If this IF tends to infinity (with either sign) as the contamination point moves further away, the resulting bias then increases indefinitely under contamination indicating the nonrobust nature of the associated functional (e.g., estimator). However, as long as the IF remains bounded as a function of the contamination point, the resulting functional (estimator) cannot have a value extremely far from the true value even under (infinitesimal) contamination at a very distant point, which justifies its robustness; the smaller the maximum extent of the IF (in absolute value) the greater the stability of the associated estimator. Therefore, the IF of the marginal MDPDEs would then give a theoretical justification of their robustness, and hence, the same for the proposed DPD-SIS and DPD-CSIS as well.

The existing theory of the MDPDE (Basu et al., 2011;Ghosh & Basu, 2013, 2016) has covered its IF under different parametric models including the GLMs. In particular, the IF of the marginal MDPDE $\widehat{\boldsymbol{\beta}}_j^{M\alpha}$ in the present context of DPD-SIS under a given GLM would have the form (Ghosh & Basu, 2016)

$$\mathcal{IF}_j^{(\alpha)}(y_t|x_{jt}) = \boldsymbol{\Psi}_n^{-1}\cdot\psi_\alpha(y_t, \beta_{j0}+\beta_j x_{jt})[1, x_{jt}]^T,$$

where y_t is the contamination point in the response variable with the associated covariate value being x_{jt}, β_{j0} and β_j are the assumed true parameter values, ψ_α is as defined in (4), and $\boldsymbol{\Psi}_n$ is some suitable matrix independent of the contamination point (see equation (5) of Ghosh and Basu (2016) for its exact form). For DPD-CSIS, the IF of the corresponding marginal MDPDE,

defined in (16), has the form

$$\mathcal{IF}_j^{(\alpha)}(y_t|x_{jt}) = \boldsymbol{\Psi}_n^{-1} \cdot \psi_\alpha\left(y_t, \boldsymbol{x}_{tC}^T\boldsymbol{\beta}_C + \beta_j x_{jt}\right)\left[\boldsymbol{x}_{tC}^T, x_{jt}\right]^T,$$

where $\boldsymbol{x}_{tC}$ denotes the values of the conditioning covariates associated with the contaminated response y_t. Therefore, in both cases, we can see that the form of function ψ_α determines the boundedness of the IFs of the marginal MDPDEs, and hence, the robustness of the resulting DPD-SIS and DPD-CSIS procedure. But, it can be easily noted that the function ψ_α, given in (4), is bounded for all standard GLMs at any $\alpha > 0$ and is unbounded at $\alpha = 0$ for most GLMs (specifically where the response has an unbounded support). Thus, the proposed DPD-SIS and DPD-CSIS with any $\alpha > 0$ would be robust under all GLMs. Further, the maximum value of $|\psi_\alpha|$ also decreases as α increases, indicating the increasing robustness of the associated marginal MDPDEs for increasing values of $\alpha > 0$; the same is also transferred subsequently for the proposed DPD-SIS and DPD-CSIS procedures with any $\alpha > 0$. Note that, this analysis additionally yields a theoretical justification of the non-robustness of the usual SIS at $\alpha = 0$ (with unbounded IF) for all GLMs having unbounded support for the response distribution.

The above-mentioned IFs can be further investigated for specific GLMs, by looking at the corresponding ψ_α function. For the linear regression models, for example, the associated $\psi_\alpha(y_t, \theta)$ has a form proportional to $(y_t - \theta)\exp\left(-\frac{\alpha(y_t-\theta)^2}{2(\sigma_j^{(0)})^2}\right)$; this case has been studied extensively in Ghosh and Thoresen (2021). For the case of Poisson regression models, the ψ_α function has the form

$$\psi_\alpha(y_t, \theta) = \frac{\left(y_t - e^\theta\right)}{(y_t!)^\alpha}\exp(\alpha y_t\theta + \alpha e^\theta) - \xi_\alpha(\theta).$$

Like the case of linear regression, one can also here rigorously see that, for any given covariate values (and hence any given θ) the function $\psi_\alpha(y_t, \theta)$ is bounded in y_t for all $\alpha > 0$, unbounded at $\alpha = 0$ and $\sup_{y_t}|\psi_\alpha(y_t, \theta)|$ decreases as $\alpha > 0$ increases. So, all the general robustness properties of the marginal MDPDEs, and hence, those for the DPD-SIS and DPD-CSIS, clearly hold also for Poisson regression models. For another important GLM, logistic regression, the ψ_α function, and hence, the IF of the MDPDEs would be bounded for all $\alpha \geq 0$ due to the bounded support of y_t; but its maximum value would still decrease indicating greater extent of robustness for DPD-SIS or DPD-CSIS with increasing values of α.

6 | NUMERICAL ILLUSTRATIONS

To illustrate the finite-sample performance of the proposed DPD-SIS and DPD-CSIS, we have performed extensive simulation studies for several important examples of GLMs; for brevity, a few interesting cases of the linear and logistic regression models are presented in this section. The corresponding `R` codes (for linear, logistic and also Poisson regressions) are provided in a public GitHub repository titled `dpdSIS` (available at https://github.com/abhianik/dpdSIS).

We would like to mention that prior numerical illustrations on the performance of the DPD-SIS under linear regression models, along with its comparison with several other existing robust SIS procedures, are available in Ghosh and Thoresen (2021). However, there is no literature on robust variable screening procedures for general GLMs (beyond linear regression), except for the rank-correlation-based SIS (rank-SIS) of Li et al. (2012a). So, in the present paper, we have

compared the DPD-SIS and DPD-CSIS, under the logistic regression model, with the classical SIS of Fan and Song (2010) and the only existing robust rank-SIS of Li et al. (2012a), although rank-SIS is a nonparametric method and thus philosophically different from our approach. It is also worth pointing out that even though Li et al. (2012a) did use their rank-SIS method for logistic models, the properties of their approach is proven only for linear models unlike our proposed DPD-SIS and DPD-CSIS whose sure screening properties are proven theoretically for general GLMs including the logistic regression. Finally, for consistency, the illustrations with linear regression models are also made only with the usual SIS/CSIS and rank-SIS in the present paper (which is indeed sufficient in view of the existing comparisons in Ghosh and Thoresen (2021) for linear regression models).

6.1 | Simulation settings

We simulate each sample of covariates (except intercept) from a multivariate normal distribution with mean vector $\mathbf{0}$ and a variance matrix having (i,j)th element as ρ for all $i \neq j$ and 1 for $i = j$; clearly $\rho = 0$ yields the case of independent covariates whereas a nonzero value of ρ indicates correlated covariates. The intercept term (1) is then added as the first covariate. Then the responses are generated according to a specified GLM (linear or logistic) with (true) coefficient value $\boldsymbol{\beta}_0 = (\beta_{01},\ldots,\beta_{0p})$. Considering $s = 4$ (i.e., four covariates are actually related to the response), we choose $\boldsymbol{\beta}_0$ such that four components (in addition to the first one, the intercept) are nonzero and the rest are zero. In particular, nonzero coefficients of $\boldsymbol{\beta}_0$ are considered to be sparsely distributed in positions 1 (intercept), 2, 6, 26, and 126.

Different values of $\boldsymbol{\beta}_0$ in terms of both the position and strength of these four nonzero coefficients are considered along with different values of ρ, n, and p; the results obtained for $p = 5000$, $n = 100$ and $\rho = 0, 0.3$ are presented here. For linear regression, errors are generated from the standard normal distribution and the error variance is assumed to be known (in consistence with the theoretical setup of the present paper).

For each scenario, the usual SIS and the proposed DPD-SIS are applied to the simulated sample of size $n = 100$ to identify the top $n - 1 = 99$ variables; it is then examined if the true nonzero (significant) covariates are selected. The full process is repeated 300 times and the number of true positives (number of selected variables having true nonzero coefficients) are studied as a summary measure. As mentioned previously, for comparison, we repeat the same exercise for the usual SIS of Fan and Song (2010) and the existing robust rank-SIS of Li et al. (2012a) and the corresponding results are also presented in the same figures.

Additionally, to examine the robustness, we repeat all these simulations again by contaminating 10% of the observations in 150 nonsignificant covariates (having true regression coefficient zero) by independent observations (outliers) generated from a normal distribution with mean equal -10 and variance equal 1. In order to have a clear outlier-effect, contaminations are introduced to observations having smaller responses in case of linear regression, and smaller probability of success for logistic model. Several other variations of the contamination schemes are also studied (e.g., introducing contaminations to randomly chosen observations, or using different contamination distributions, etc.) but they all produce similar results leading to the exact same conclusions about the performances of the DPD-SIS in comparison to the usual SIS and rank-SIS; so, these results are not reported here for brevity.

6.2 | Simulation results: performance of DPD-SIS

We present the box-plots of true-positives obtained by the proposed DPD-SIS procedure at different $\alpha > 0$ (along with the usual SIS and the rank-SIS) both without and with data contamination (as specified in Section 6.1) in Figures 1 and 2 for the logistic and the linear regression cases, respectively.

We observe that for independent covariates and relatively strong signal (larger coefficient values) the DPD-SIS with smaller $\alpha \leq 0.5$ performs exactly similar to the usual SIS under pure data. But, when there is contamination in the data, the performance of the usual SIS deteriorates significantly whereas the proposed DPD-SIS remains stable and yields better variable selection results, ignoring the effect of outliers (see, e.g., Figure 1a). As the signal gets weaker, the DPD-SIS fails to select all true positives but it still performs as good as the usual SIS under pure data; in case of additional contamination, the DPD-SIS remains much more stable even with weaker signal, although the usual SIS gets significantly affected (see, e.g., Figure 1b). However, the performance of the proposed DPD-SIS as well as the usual SIS becomes significantly worse when the covariates are strongly correlated which is expected from our theory as well. In such a case, the effect of outliers may not be so prominent as in the case of independent covariates, but there is still a slight decrease in the number of true positives obtained by the usual SIS under contamination and DPD-SIS again performs robustly as claimed (see, e.g., Figure 1c).

The rank-SIS method performs better than the usual SIS in terms of robustness under data contamination but worse under pure data. Additionally, the results obtained by the rank-SIS are clearly worse than the best results obtained by our proposed DPD-SIS at a suitable α under both pure and contaminated data scenarios, justifying the advantages of our proposal over the existing rank-SIS method under both linear and logistic regression models.

6.3 | Simulation results: performance of DPD-CSIS

We have also repeated our simulation exercises, as mentioned in Section 6.1, with some modifications to illustrate the performance of the proposed DPD-CSIS. In particular, for this purpose, we assume that the first four covariates are known to be important (each having true regression coefficient 5) which is used as $\mathcal{X}_C$; the remaining $p-4$ covariates are used as $\mathcal{X}_D$ from where the variable screening is performed. Among $\mathcal{X}_D$, $s = 4$ sparsely distributed covariates are again assumed to be truly significant as described in Section 6.1; the remaining settings of the simulation study are also assumed to be the same as before. Since conditional screening is more important for dependent covariates, we only present the results corresponding to the case $\rho = 0.3$ (and strong signal) in Figure 3a and b, respectively, for the logistic and the linear regression models under both pure and contaminated data.

We can again observe the advantages of the DPD-CSIS over the usual CSIS (Barut et al., 2016) in terms of robustness, producing better variable screening results under data contamination. Compared to the rank-SIS, the proposed DPD-CSIS also provides better trade-offs between the robustness and efficiency, in terms of variable screening under contaminated and pure data setups, for both linear and logistic regression models.

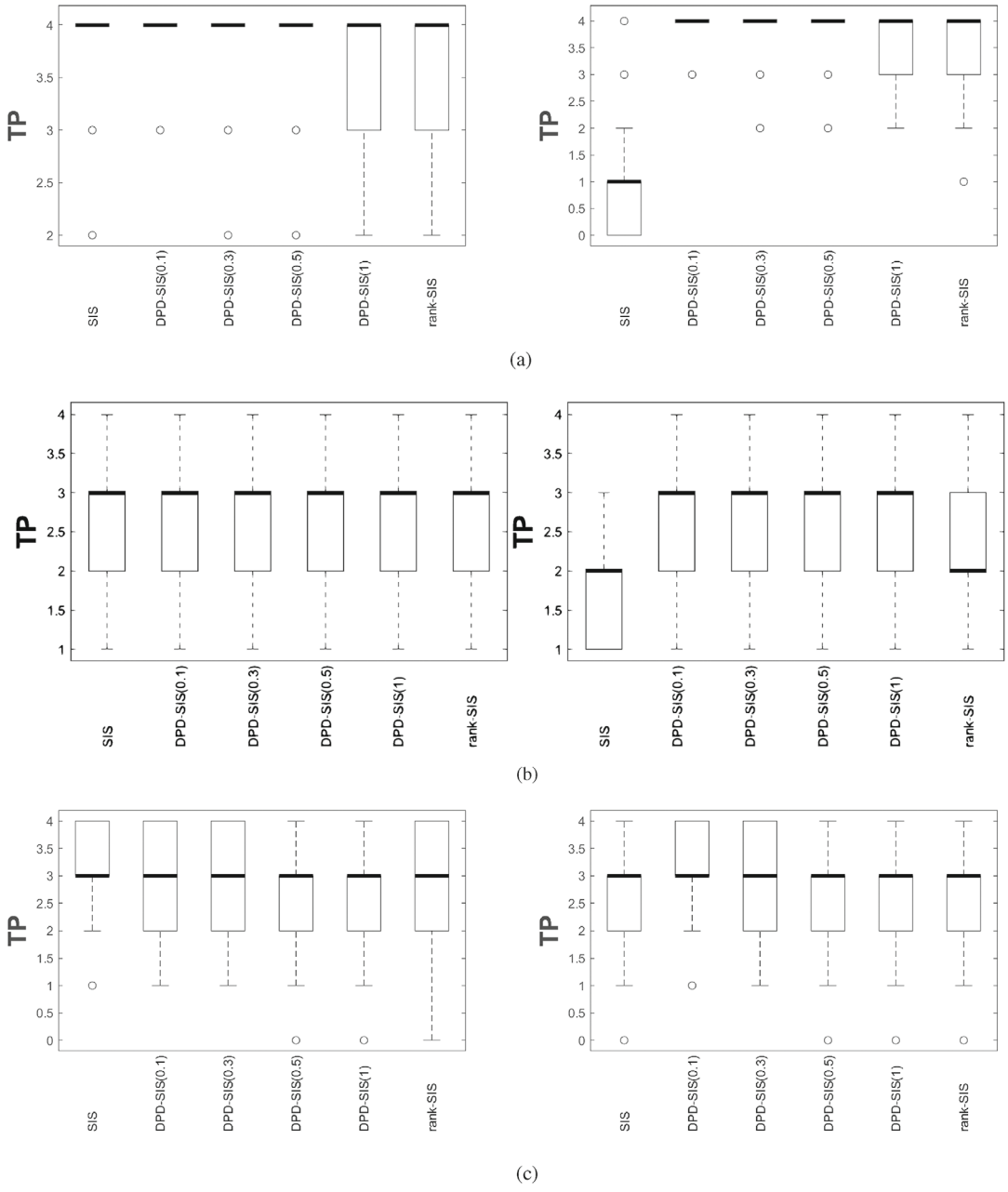

FIGURE 1 The box-plots of true-positives selected by the density power divergence-sure independence screening (SIS) at different α for different simulation setups under the logistic regression model with pure data (left panel) and 10% contaminated data (right panel). The results for usual SIS of Fan and Song (2010) and rank-SIS of Li et al. (2012a) are also presented in the same plots for comparisons. (a) Independent covariates ($\rho = 0$) strong signal (nonzero values of β_0 are all 5). (b) Independent covariates, weaker signal $(\beta_{01}, \beta_{02}, \beta_{06}, \beta_{0,26}, \beta_{0,126}) = (1, 2, 3, 1, 5)$. (c) Dependent covariates ($\rho = 0.3$) strong signal (nonzero values of β_0 are all 5)

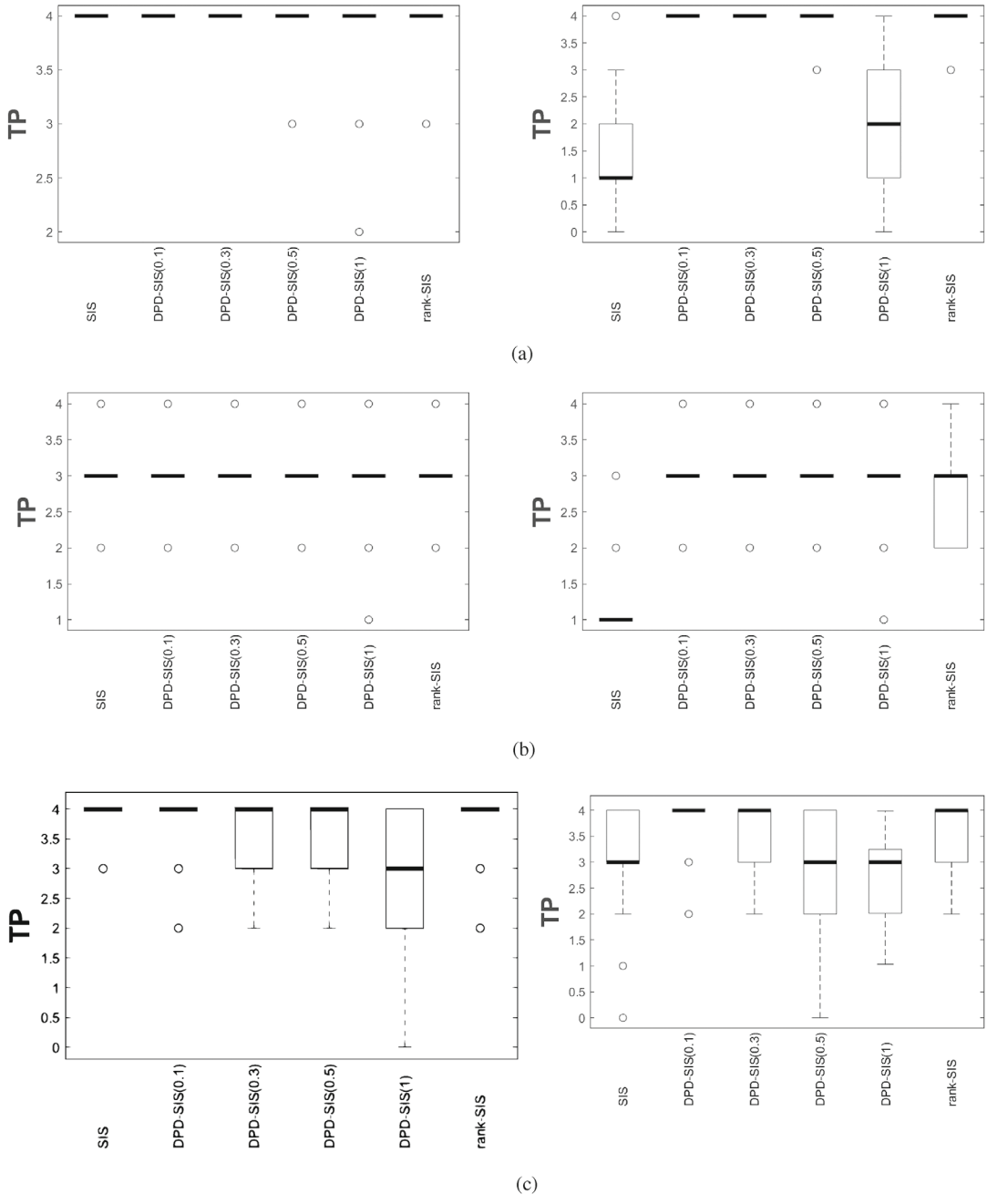

FIGURE 2 The box-plots of true-positives selected by the density power divergence-sure independence screening (SIS) at different α for different simulation setups under the linear regression model with pure data (left panel) and 10% contaminated data (right panel). The results for usual SIS of Fan and Song (2010) and rank-SIS of Li et al. (2012a) are also presented in the same plots for comparisons. (a) Independent covariates ($\rho = 0$) strong signal (nonzero values of β_0 are all 5). (b) Independent covariates, weaker signal $(\beta_{01}, \beta_{02}, \beta_{06}, \beta_{0,26}, \beta_{0,126}) = (1, 2, 3, 1, 5)$. (c) Dependent covariates ($\rho = 0.3$) strong signal (nonzero values of β_0 are all 5)

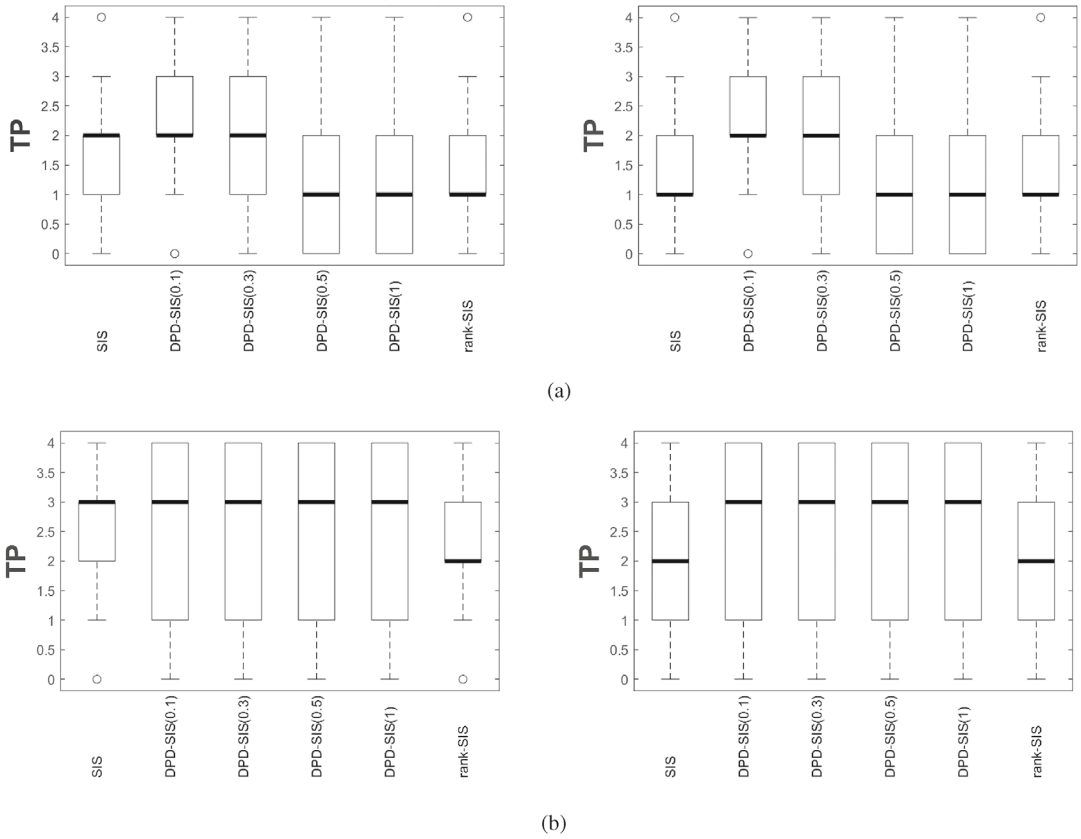

FIGURE 3 The box-plots of true-positives selected by the density power divergence-conditional sure independence screening (CSIS) at different α for simulations with dependent covariates ($\rho = 0.3$) and strong signal (nonzero values of $\boldsymbol{\beta}_0$ are all 5) under pure data (left panel) and 10% contaminated data (right panel). The results for usual CSIS of Barut et al. (2016) and rank-SIS of Li et al. (2012a) are also presented in the same plots for comparisons. (a) Logistics regression models; (b) linear regression models

6.4 | On the choices of tuning parameters α and γ_n

The proposed variable screening procedures, both DPD-SIS and DPD-CSIS, involve two tuning parameters. The first one, namely the robustness parameter α, controls the trade-off between robustness and efficiency of the marginal MDPDEs (Basu et al., 2011; Ghosh & Basu, 2013, 2016). As a consequence, as argued in Section 5, the robustness of the proposed DPD-SIS and DPD-CSIS increases with increasing $\alpha > 0$. However, since the efficiency of the marginal MDPDEs decreases as $\alpha > 0$ increases, it affects the DPD-SIS or DPD-CSIS with greater variations in the number of true positives for larger $\alpha > 0$, as seen from our extensive simulation exercises. So, we need to choose a proper α to get the optimal trade-off between the performances of DPD-SIS or DPD-CSIS under pure and contaminated data (since, in practice, the amount of contamination in a given datasets is often unknown).

As a guiding principle, we can suggest some optimal values of α (depending on the used model and correlation among variables) based on empirical investigations via extensive simulation studies. Note that, in our simulations, the performance of DPD-SIS for the cases with independent

covariates appears to be almost the same at any α in the range $[0.1, 0.5]$ both for linear and logistic regression models; so any value of α within this range should work well in practice. In case of dependent covariates though, their performances can vary significantly over α; an α value around 0.1 seems to perform the best (refer to Figures 1c, 2c, and 3). Hence, as an empirical suggestion, we recommend to apply the proposed DPD-SIS or DPD-CSIS with $\alpha = 0.1$ in any practical application with linear or logistic regression models involving small to moderate contamination (or no contamination at all). However, for datasets with higher contamination proportion, we might use a slightly larger α value (e.g., 0.3) following our theoretical robustness study presented in Section 5.

Alternatively, one might prefer a data-driven algorithm to chose the optimal α for a given dataset. For classical regression setups, there are a few existing such algorithms for choosing optimal α in the computation of the MDPDEs; most notably, the one proposed in Warwick and Jones (2005) and its recent extension by Basak et al. (2021). These algorithms have also been investigated in the context of linear regression and more general GLMs by (Ghosh & Basu, 2013, 2016) and Basak et al. (2021). In the present context of DPD-SIS or DPD-CSIS, we can directly use either of these algorithms for data-driven selection of α while computing the marginal MDPDEs and use the resulting optimal MDPDEs to perform variable screening. However, a major problem with this approach is the possibility to get different optimal α for different marginal MDPDEs (associated with different covariates) within the same dataset, leading to inconsistency in the whole variable screening procedure. Although we can bypass this obstacle by considering a summary measure from the pool of optimal α values obtained for different covariates, this exercise would be extremely time consuming for datasets with larger dimensions and would defeat the whole purpose of variable screening. So, we recommend to use the empirical suggestions for α while using DPD-SIS (or, DPD-CSIS) for initial (robust) variable screening. As an alternative, one might also consider the union of the variables selected by the DPD-SIS (or DPD-CSIS) with different possible α values in the neighborhood of the empirical suggestion (to ensure that nothing important is missed).

The second tuning parameter γ_n controls the amount of false positives in the screening procedure. The theoretically optimal rate of γ_n for controlling the false positives has been seen to be $n^{-2\kappa}$ for some $\kappa > 0$ as mentioned in Theorems 3 and 6, respectively, for the DPD-SIS and the DPD-CSIS. Note that, this is the exact same rate as the one derived in Fan and Song (2010) for the usual SIS; so any existing (data-driven) rule for the selection of γ_n for the usual SIS can also be applied for our DPD-SIS or DPD-CSIS at any $\alpha \geq 0$. As for the choice of the parameter α, one could wish for a data-driven approach to the choice of γ_n. Some efforts have been made lately to establish practical procedures to control the amount of false positives in variable screening procedures, most of them based on data splitting. Guo et al. (2022) presented a general approach to the problem that could be applied also to our DPD-SIS and DPD-CSIS procedures. However, in practice it is common to retain a fixed number of predictors, for example $[n/\log(n)]$ or $(n-1)$. Our preferred approach is to use such a hard thresholding rule followed by a regularized regression procedure where false positives can be controlled by, for example, stability selection (Meinshausen & Buhlmann, 2010).

7 | AN APPLICATION: NOWAC LUNG CANCER DATA

We will apply our DPD-SIS method to a variable selection problem related to the investigation of potential biomarkers for lung cancer. In the Norwegian Women and Cancer (NOWAC) study we have data on 125 lung cancer cases of which 97 had developed metastasis at the time of

TABLE 1 The estimated regression coefficients in the final models obtained after stability selection while using the proposed DPD-SIS ($\alpha = 0.1$) versus the usual SIS.

Probe ID	Gene	DPD-SIS ($\alpha = 0.1$)		Usual SIS	
x8tTnl6f115xQSV1X4	FER1L5	−0.4692	(0.96)	−0.4311	(0.98)
l7tROJgVSRulLNNJ18	SERHL	−0.5217	(0.95)	−0.5045	(0.92)
Hul6v6J0kV5qD3PA3U	TMEM105	−0.4952	(0.90)	−0.4631	(0.87)
9Jd97nm3QdLiQXEuzE	INVS	0.4029	(0.93)	0.3805	(0.86)
BWoVQF900KRd2rRTx8	NA	0.5487	(0.87)	0.6041	(0.84)
r15YSrezLzSP97mOnU	SCARNA14	−0.3628	(0.78)	−0.4609	(0.84)
fnRCld.v151SEwlQqk	ANO7	0.5622	(0.89)	0.5588	(0.79)
rqKhAKFSCyCDqjouuI	TBRG1	−0.3032	(0.87)	−0.3044	(0.78)
lteivVSVR8WbDmUMBQ	HBZ	—	(0.54)	−0.0209	(0.78)
Bizm6pN4NAz3jzFe3s	SNORD113-7	0.4117	(0.80)	0.4136	(0.74)
01_iIjuJFCnq.s7rqo	FLVCR1	−0.2828	(0.72)	−0.3071	(0.74)
T9XUjiuZBS6556G.gA	MGC23270	—	(0.67)	−0.3512	(0.73)
i3urBCqi6u9.0l_cB0	FAM86D	0.3580	(0.74)	0.3468	(0.70)
KA5bqKgpHd7QnntF8U	NA	−0.3219	(0.70)	−0.2756	(0.70)
EEpIrACOOgruuHCpRo	CCT6A	−0.2740	(0.81)	—	(0.69)
lldNIQXQNde_jgnE0k	ABCB4	−0.7045	(0.90)	—	
cKfyV6FSSuKdp_jz3k	MAPK7	−0.2208	(0.74)	—	

Note: The stability selection probability for each probe is given in the parenthesis.

diagnosis (Lund et al., 2008; Sandanger et al., 2018). For all these women, we have measures of mRNA in blood some time before diagnosis (ranging from 0.3 to 7.9 years before diagnosis, with a median time equal to 4.2 years). The goal is to relate the mRNA measurements to the classification of metastatic versus nonmetastatic cancer cases. The mRNA measurements are based on the microarray technology, and we have data from a total of 11,610 probes. Thus, we need to perform some sort of variable selection before running our favorite regression model. Our analysis strategy is as follows: First, we run our DPD-SIS procedure (with $\alpha = 0.1$) and select the top $n - 1$ probes. Next, we run a standard logistic regression model with elastic net penalty (with the mixing parameter fixed at 0.7 and lambda selected by cross-validation) followed by stability selection (Meinshausen & Buhlmann, 2010) on the selected probes to reach the final set. As a comparison, we run the standard SIS followed by the same standard elastic net logistic regression with stability selection. We compare the initial set of $n - 1$ probes and the final set after stability selection.

Of the 124 initially selected probes, there was an overlap of 115. After stability selection with a cut-off at a selection probability equal to 0.7 we were left with 15 probes based on the robust screening and 14 probes based on the ordinary SIS screening, which are reported in Table 1. We ran logistic regression with elastic net penalty on these 15 versus 14 probes and calculated the area under the ROC curve (AUC). The AUC values were 0.998 for the 15 robustly selected probes and 0.983 for the 14 probes selected based on ordinary SIS. Of course, these values are clearly over-optimistic due to overfitting, but they indicate that the predictions based on the robustly selected probes are no worse than those based on the nonrobust selection procedure.

Twelve probes were present on both lists; thus, we have five nonoverlapping probes worthy of further investigation (see Table 1).

The three probes not selected by the usual SIS-procedure, but selected by our DPD-SIS, are interesting. The probe EEpIrACOOgruuHCpRo is related to the CCT6A gene, which has been shown to be linked to metastatic nonsmall cell lung cancer (Zhang et al., 2020), which is the target of our analysis. Furthermore, the two remaining probes (lldNIQXQNde_jgnE0k and cKfyV6FSSuKdp_jz3k) are linked to the ABCB4 and the MAPK7 genes. Both of these are linked to several types of cancer including lung cancer (Gavine et al., 2015; Kiehl et al., 2014). Thus, all three probes seem highly relevant for our example, and they are obviously important to capture although the usual SIS fails to select them. On the other hand, two probes were selected by usual SIS but not by our DPD-SIS. They were lteivVSVR8WbDmUMBQ connected to the HBZ gene, which has been linked to leukemia but no other cancers, and T9XUjiuZBS6556G.gA connected to the MGC23270 gene that has not to our knowledge been linked to cancer at all.

This example clearly illustrate that we may miss important variables, probes in our example, by using the standard SIS procedure due to the effect of outliers in the data, and this can sometimes lead to spurious findings. Our proposed robust DPD-SIS can successfully find the most important probes, ignoring any contamination effects present in the data, also without any significant loss in the predictive performance.

8 | DISCUSSION

This paper presents a robust extension of the usual SIS and the conditional SIS for variable screening under ultra-high-dimensional GLMs using the robust minimum DPD estimators of the marginal regression parameters. We show that the proposed DPD-SIS and its conditional version (DPD-SSIS) both enjoy the sure screening property under a reasonable set of assumptions, in line of what is required by usual SIS. More importantly, the proposed DPD-SIS is extremely fast with computational times comparable to those of the usual SIS, in addition to being robust against data contaminations. It can be noted that similar robust extensions of SIS can be constructed by other low-dimensional robust estimation procedures for GLMs, but the validity of the sure screening properties is not guaranteed under the same rather simple assumptions and with such a low computational time. A real data application further reveals the advantages of our proposed DPD-SIS over the usual SIS in identifying the truly important variables from noisy data with no loss in predictive power.

The choice of variable selection criteria is controversial and represents a difficult step in the analysis of high-dimensional data, in particular different omics data. In many cases, it is common to reduce the dimensionality of the data by filtering the variables before performing any analysis. Even when performing LASSO procedures, the data is often filtered beforehand and only a portion of the variables is considered. Several approaches are used for variable filtering, most commonly based on the variance or range of the variables, or alternatively based on the association with the outcome of interest, as well as more ad hoc criteria. The chosen filtering criteria can clearly impact the analysis, and the results and conclusions might be extremely sensitive to this choice. Therefore, it is especially important to base the variable filtering on a clear, objective method that can be reproduced and justified. When using SIS-based procedures, and DPD-SIS in particular, all variables are evaluated and selected based on their marginal regression coefficient. Therefore, DPD-SIS provides a nonarbitrary, reproducible choice of the variables to be included in subsequent analysis, and can prove extremely useful in practical applications to, for example,

omics data. Additionally, the conditional extension provided in this work can prove particularly useful in the context of omics data, where it might be relevant to condition on other clinically relevant (non-omics) variables.

We will argue that robust methods, and in particular our robust DPD-based method, are particularly useful in variable filtering situations to avoid the influence of outliers, as traditional checks for outliers become infeasible due to the ultra-high dimensionality.

To control for false discoveries, it is recommendable to perform stability selection in such settings. As we illustrate in the applied example, it is quite straightforward to include a stability selection procedure in the DPD-SIS. The efficient implementation of DPD-SIS makes this possible without intense computational efforts.

However, a major problem may arise in applications of DPD-SIS when covariates are highly dependent, just like for most other (one-step) screening methods. This problem can be solved by applying the proposed DPD-SIS (or DPD-CSIS) iteratively by removing, in each step, the variables selected in the previous steps of the iteration, going through the same philosophy as the iterative SIS (Fan & Lv, 2008). The performance of such an iterative DPD-SIS in successfully addressing the dependent covariate problem has been numerically illustrated for the case of linear regression in Ghosh and Thoresen (2021); the same can also be investigated for the logistic or other GLMs in future.

Finally, due to the simplicity of the proposed DPD-SIS along with its excellent robust performance, it would be natural to extend it to variable screening under more general learning models like mixed models or censored regression models with ultra high-dimensional covariates. This may be the focus of our future work.

FUNDING INFORMATION

INSPIRE Faculty Research Grant, Department of Science and Technology, Government of India, Grant Number: SRG/2020/000072; European Research Council, Grant Number: ERC-2008-AdG-232997; Norwegian Research Council, Grant Number: 248804 and 262111.

ORCID

Magne Thoresen https://orcid.org/0000-0003-1511-5938

REFERENCES

Barut, E., Fan, J., & Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association*, *111*(515), 1266–1277.

Basak, S., Basu, A., & Jones, M. C. (2021). On the 'optimal' density power divergence tuning parameter. *Journal of Applied Statistics*, *48*(3), 536–556.

Basu, A., Ghosh, A., Mandal, A., Martin, N., & Pardo, L. (2017). A Wald-type test statistic for testing linear hypothesis in logistic regression models based on minimum density power divergence estimator. *Electronic Journal of Statistics*, *11*, 2741–2772.

Basu, A., Ghosh, A., Mandal, A., Martin, N., & Pardo, L. (2021). Robust Wald-type tests in GLM with random design based on minimum density power divergence estimators. *Statistical Methods and Applications*, *30*(3), 973–1005.

Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, *85*, 549–559.

Basu, A., Shioya, H., & Park, C. (2011). *Statistical inference: The minimum distance approach*. Chapman ς Hall/CRC Press.

Buhlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media.

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

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849–911.

Fan, J., & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, *38*(6), 3567–3604.

Gather, U., & Guddat, C. (2008). Comment on sure independence screening for ultrahigh dimensional feature space by Fan, JQ and Lv, J. *Journal of the Royal Statistical Society: Series B*, *70*, 893–895.

Gavine, P. R., Wang, M., Yu, D., Hu, E., Huang, C., Xia, J., Su, X., Fan, J., Zhang, T., Ye, Q., & Zheng, L. (2015). Identification and validation of dysregulated MAPK7 (ERK5) as a novel oncogenic target in squamous cell lung and esophageal carcinoma. *BMC Cancer*, *15*(1), 1–9.

Ghosh, A. (2019). Robust inference under the beta regression model with application to health care studies. *Statistical Methods in Medical Research*, *28*(3), 871–888.

Ghosh, A., & Basu, A. (2013). Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. *Electronic Journal of Statistics*, *7*, 2420–2456.

Ghosh, A., & Basu, A. (2016). Robust estimation in generalized linear models: The density power divergence approach. *Test*, *25*(2), 269–290.

Ghosh, A., & Majumdar, S. (2020). Ultrahigh-dimensional robust and efficient sparse regression using non-concave penalized density power divergence. *IEEE Transactions on Information Theory*, *66*(12), 7812–7827.

Ghosh, A., & Thoresen, M. (2021). A robust variable screening procedure for ultra-high dimensional data. *Statistical Methods in Medical Research*, *30*(8), 1816–1832.

Giraud, C. (2014). *Introduction to high-dimensional statistics*. Chapman & Hall/CRC Press.

Guo, X., Ren, H., Zou, C., & Li, R. (2022). Threshold selection in feature screening for error rate control. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2021.2011735

Hall, P., & Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, *18*(3), 533–550.

Hampel, F. R., Ronchetti, E., Rousseeuw, P. J., & Stahel, W. (1986). *Robust statistics: The approach based on influence functions*. John Wiley & Sons.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. CRC Press.

Kiehl, S., Herkt, S. C., Richter, A. M., Fuhrmann, L., El-Nikhely, N., Seeger, W., Savai, R., & Dammann, R. H. (2014). ABCB4 is frequently epigenetically silenced in human cancers and inhibits tumor growth. *Scientific Reports*, *4*(1), 1–9.

Li, G., Peng, H., Zhang, J., & Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics*, *40*(3), 1846–1877.

Li, R., Zhong, W., & Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, *107*(499), 1129–1139.

Lund, E., Dumeaux, V., Braaten, T., Hjartaker, A., Engeset, D., Skeie, G., & Kumle, M. (2008). Cohort profile: The Norwegian women and cancer study—NOWAC—Kvinner og kreft. *International Journal of Epidemiology*, *37*(1), 36–41.

Luo, S., Song, R., & Witten, D. (2014). Sure screening for Gaussian graphical models. *Stat*, *1050*(29), 4.

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

Mu, W., & Xiong, S. (2014). Some notes on robust sure independence screening. *Journal of Applied Statistics*, *41*(10), 2092–2102.

Saldana, D. F., & Feng, Y. (2018). SIS: An R package for sure independence screening in ultrahigh-dimensional statistical models. *Journal of Statistical Software*, *83*(2), 1–25.

Sandanger, T. M., Nøst, T. H., Guida, F., Rylander, C., Campanella, G., Muller, D. C., Van Dongen, J., Boomsma, D. I., Johansson, M., Vineis, P., & Vermeulen, R. (2018). DNA methylation and associated gene expression in blood prior to lung cancer diagnosis in the Norwegian women and cancer cohort. *Scientific Reports*, *8*(1), 1–10.

Wang, T., Zheng, L., Li, Z., & Liu, H. (2017). A robust variable screening method for high-dimensional data. *Journal of Applied Statistics*, *44*(10), 1839–1855.

Warwick, J., & Jones, M. C. (2005). Choosing a robustness tuning parameter. *Journal of Statistical Computation and Simulation*, *75*, 581–588.

Zhang, T., Shi, W., Tian, K., & Kong, Y. (2020). Chaperonin containing t-complex polypeptide 1 subunit 6A correlates with lymph node metastasis, abnormal carcinoembryonic antigen and poor survival profiles in non-small cell lung carcinoma. *World Journal of Surgical Oncology*, *18*(1), 1–10.

Zhao, S. D., & Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, *105*(1), 397–411.

Zhong, W. (2014). Robust sure independence screening for ultrahigh dimensional non-normal data. *Acta Mathematica Sinica English Series*, *30*(11), 1885–1896.

How to cite this article: Ghosh, A., Ponzi, E., Sandanger, T., & Thoresen, M. (2022). Robust sure independence screening for nonpolynomial dimensional generalized linear models. *Scandinavian Journal of Statistics*, 1–31. https://doi.org/10.1111/sjos.12628

APPENDIX A. VALIDATION OF ASSUMPTIONS FOR COMMON GLMS

A.1 Logistic regression

Let us first consider the logistic regression model, which is a prominent member of the GLM class for binary responses having Bernoulli distributions. In our notation, for the logistic regression, we have $b(\theta) = \log(1 + e^{\theta})$ and g is the logit link function, so that $\mu = E[Y|\theta] = b'(\theta) = e^{\theta}/(1 + e^{\theta})$. Therefore, for the computation of the marginal MDPDEs, we have $\xi_{\alpha}(\theta) = \frac{e^{\theta}(e^{\alpha\theta}-1)}{(1+e^{\theta})^{2+\alpha}}$, and hence, the function ψ_{α} has the form

$$\psi_{\alpha}(y, \theta) = (y - b'(\theta))\frac{e^{\alpha\theta y}}{(1 + e^{\theta})^{\alpha}} - \frac{e^{\theta}(e^{\alpha\theta} - 1)}{(1 + e^{\theta})^{2+\alpha}}.$$

Now, in the theoretical analyses of the DPD-SIS, we have first defined the function B_{α}, which can be simplified for the present case as

$$B_{\alpha}(v(\boldsymbol{x})) = b'(\boldsymbol{x}^T\boldsymbol{\beta}_0) - \frac{\left(e^{\boldsymbol{x}^T\boldsymbol{\beta}_0} - e^{v(\boldsymbol{x})}\right)\left(e^{\alpha v(\boldsymbol{x})} + e^{v(\boldsymbol{x})}\right)}{(1 + e^{\boldsymbol{x}^T\boldsymbol{\beta}_0})(1 + e^{v(\boldsymbol{x})})^{2+\alpha}}.$$

Then, one can easily check that $B'_{\alpha}(\cdot)$ is bounded so that Assumption (B1) of Theorem 2 holds and we have the theoretical justification for the DPD-SIS as a variable screening procedure under the logistic regression model. The same can also be verified easily for the DPD-CSIS.

Now, we verify Assumptions (A1)–(A7) to justify the sample-level sure screening property and the false-discovery control of the proposed DPD-SIS and DPD-CSIS. Recall that Assumptions (A5)–(A7) are exactly the same as those assumed in Fan and Song (2010) for usual SIS (independent of α), and hence, they hold under mild sufficient conditions as described in the literature of the usual SIS. Among other assumptions specific to our DPD-SIS (or DPD-CSIS), we first

note that Assumption (A1) holds for any $\alpha \geq 0$ with $L_\alpha = 1$, the bound for the Bernoulli density, and Assumptions (A2) and(A3) hold, from the standard literature on the MDPDE under logistic regression (Basu et al., 2017; Ghosh & Basu, 2016), whenever each covariate under screening is independent of the conditioning variables (which is only the intercept in DPD-SIS). Finally, Assumption (A4) requires a large constant K_n in general, and so we need to carefully choose it in order for the probability bounds in Theorems 3 and 6 to make sense. In the present case, b' and B_α are bounded and so we can take k_n to be an appropriate finite number (independent of n). So, proceeding as in Fan and Song (2010), the optimal order for the sequences in Theorems 3 and 6 are given by

$$K_n = n^{(1-2\kappa)/(\tau+2)}, \quad R_n = \exp\left(-c_4 n^{\tau(1-2\kappa)/(\tau+2)}\right),$$

and hence, the (upper) probability bounds in our theorems would make sense as long as $\log p = o\left(n^{\tau(1-2\kappa)/(\tau+2)}\right)$, covering the ultra-high-dimensional case with nonpolynomial dimensionality as desired.

A.2 Linear regression with normal error

We can also validate all the assumptions in our theoretical derivations for the usual linear regression model (with standard normal error distribution) as a special case of our GLM setup, for which we have $b(\theta) = \theta^2/2$ and g as the identity (link) function. This leads to $\xi_\alpha(\theta) = 0$, and hence, the ψ_α function in the computation of the marginal MDPDEs has the simplified form

$$\psi_\alpha(y, \theta) = \frac{(y-\theta)}{(2\pi)^{\alpha/2}} e^{-\frac{\alpha}{2}(y-\theta)^2}.$$

Therefore, for this special case, we get

$$B_\alpha(v(\boldsymbol{x})) = \boldsymbol{x}^T \boldsymbol{\beta}_0 - \frac{\left(\boldsymbol{x}^T \boldsymbol{\beta}_0 - v(\boldsymbol{x})\right)}{\sqrt{1+\alpha}(2\pi)^{\alpha/2}}.$$

Thus, B'_α is a constant and so bounded; this implies Assumption (B1) and Theorems 1 and 2 hold, justifying the DPD-SIS for the standard linear regression model at population level. Similar justifications for the DPD-CSIS is easy to observe for the present example.

Next, considering the α-specific assumptions required for the sample level properties, it is straightforward to observe that (A1) holds for any $L_\alpha > (2\pi)^{-\alpha/2}$ and Assumptions (A2) and (A3) hold whenever each screening covariate is independent of the conditioning variables. Finally, to get an idea about the choice of K_n in Assumption (A4), keeping the probability bounds in Theorems 3 and 6 meaningful, let us consider $k_n^{(\alpha)} = (1+\alpha)L_\alpha \left[B(K_n+1) + K_n^\tau/2m_3\right]$, Then, as in Fan and Song (2010), assuming $A = \max\{\tau+4, 3\tau+2\}$, we get the optimal order for the sequences in Theorems 3 and 6 as given by

$$K_n = n^{(1-2\kappa)/A}, \quad R_n = \exp\left(-c_4 n^{\tau(1-2\kappa)/A}\right).$$

With these choices, the (upper) probability bounds in our Theorems 3 and 6 would make sense whenever $\log p = o\left(n^{\tau(1-2\kappa)/A}\right)$. Thus, our theoretical results on the sure screening

property and false-discovery control of the proposed DPD-SIS and DPD-CSIS indeed hold for the example of linear regression models with ultra-high- (nonpolynomial) dimensional covariates.

A.3 Poisson regression

As our final illustration, let us consider the Poisson regression model, a special case of GLMs corresponding to count responses having Poisson distribution and the log link function. In this case, the ψ_α function, for the computation of the marginal MDPDEs, has the form

$$\psi_\alpha(y,\theta) = (y - e^\theta)\frac{e^{\alpha\theta y}}{(y!)^\alpha}e^{-\alpha e^\theta} - \xi_\alpha(\theta),$$

where $\xi_\alpha(\theta)$ has no closed form expression (an infinite sum). As a result, the deduced function $B_\alpha(v(\boldsymbol{x}))$ also does not have a closed form expression. But, one can check that $b'(\theta) = e^\theta$, and hence, G_α is of the order of an exponential function. So, as with the usual SIS in Fan and Song (2010), Assumption (B2) in Theorem 2 holds whenever the covariates are bounded or have a light tail (e.g., for sub-Gaussian covariates). For such cases, we then have the population level justification for the proposed DPD-SIS as a variable screening procedure under the Poisson regression model.

Assumptions (A1)–(A3) specific to the DPD-SIS (or, DPD-CSIS) can also be shown to hold for the Poisson regression case, as in the previous cases of linear and logistic regression models, whenever the covariates under screening are independent of the conditioning variables and the intercept. However, to get a reasonable choice of K_n in Assumption (A4), we have to make further assumptions on the covariates. For example, if the covariates are bounded, then both k_n and K_n can be taken as finite constants, and hence, we get $R_n = \exp\left(-c_4 n^{1-2\kappa}\right)$ (This is true for any GLM with bounded covariates—not only the Poisson regression). Then, the (upper) probability bounds in our Theorems 3 and 6 make sense for the ultra-high-dimensional case with $\log p = o\left(n^{1-2\kappa}\right)$ and the sample-level properties of the proposed DPD-SIS and DPD-CSIS also hold for such Poisson regression models. We conjecture that the same can be verified for sub-Gaussian covariates as well (as observed empirically) although we do not have a concrete proof at this moment.

In this respect, we must note that, there is indeed no literature on the validity of such an assumption, even for the usual SIS, under Poisson regression with more general covariate distributions. So, more research is needed to establish the validity of the required assumptions for SIS (and also for our proposed DPD-SIS or DPD-CSIS) under such general Poisson models.

APPENDIX B. PROOFS OF THEOREMS 1 AND 2

B.1 Proof of Theorem 1

This follows from the Fisher consistency of the MDPDEs (by definition) and the results of Theorem 2 in Fan and Song (2010).

B.2 Proof of Theorem 2

For $\alpha = 0$ this theorem is identical to Theorem 3 of Fan and Song (2010). We now extend their argument to prove it for any given $\alpha > 0$. Let us fix an $\alpha > 0$ and $j \in \mathcal{M}_0$.

First consider the cases where $B'_\alpha(\cdot)$ is bounded and let D be its upper bound. Then $B_\alpha(\cdot)$ is Lipschitz continuous and hence

$$\left|\left\{B_\alpha\left(\beta_{j0}^{M\alpha}+\beta_j^{M\alpha}X_j\right)-B_\alpha\left(\beta_{j0}^{M\alpha}\right)\right\}X_j\right|\le D|\beta_j^{M\alpha}|X_j^2.$$

Taking expectation, we get

$$\begin{aligned}D|\beta_j^{M\alpha}| &\ge \left|E\left[\left\{B_\alpha(\beta_{j0}^{M\alpha}+\beta_j^{M\alpha}X_j)-B_\alpha(\beta_{j0}^{M\alpha})\right\}X_j\right]\right| \\ &= \left|E\left[B_\alpha(\beta_{j0}^{M\alpha}+\beta_j^{M\alpha}X_j)X_j\right]\right| = \left|\mathrm{Cov}\left(B_\alpha(\beta_{j0}^{M\alpha}+\beta_j^{M\alpha}X_j),X_j\right)\right|.\end{aligned}$$

In the above, we have used $E\left[B_\alpha(\beta_{j0}^{M\alpha})X_j\right]=0$ which holds since $B_\alpha(\beta_{j0}^{M\alpha})$ is a constant (by definition) and $EX_j=0$ by our assumption. But, from the estimating equation in (8), we get $E[B_\alpha(\beta_{j0}^{M\alpha}+\beta_j^{M\alpha}X_j)X_j]=E[b'(\boldsymbol{X}^T\boldsymbol{\beta}_0)X_j]$, and hence, using $E(X_j)=0$,

$$\mathrm{Cov}\left(B_\alpha\left(\beta_{j0}^{M\alpha}+\beta_j^{M\alpha}X_j\right),X_j\right)=\mathrm{Cov}\left(b'\left(\boldsymbol{X}^T\boldsymbol{\beta}_0\right),X_j\right). \tag{B1}$$

Then, in view of the condition of the theorem, we get $|\beta_j^{M\alpha}|\ge D_1^{-1}c_1n^{-\kappa}$ completing the proof of the theorem.

Next, we consider the cases where the second assumption, namely Condition (9), holds. Clearly if $|\beta_j^{M\alpha}|\ge cn^{-\kappa}$ for a sufficiently large universal constant $c>0$, the result holds and we are done. So, assume $|\beta_j^{M\alpha}|\le\tilde{c}_1n^{-\kappa}$ for some $\tilde{c}>0$ and let $\beta_0^{M\alpha}$ be a constant such that $B_\alpha(\beta_0^{M\alpha})=E[Y]$. We first prove the following claim.

Claim 1: $\left|\beta_{j0}^{M\alpha}-\beta_0^{M\alpha}\right|\le\tilde{c}_2$ for all $j\in\mathcal{M}_0$ and some constant $\tilde{c}_2>0$.

To prove the claim, we fix a $j\in\mathcal{M}_0$ and consider the marginal MDPDE objective function (population version) as a function of β_0 only as $Q(\beta_0)=E[l_\alpha(Y,\beta_0+\beta_j^{M\alpha}X_j)]$ so that we get

$$Q'(\beta_0)=E\left[Y-B_\alpha\left(\beta_0+\beta_j^{M\alpha}X_j\right)\right]=B_\alpha\left(\beta_0^{M\alpha}\right)-E\left[B_\alpha\left(\beta_0+\beta_j^{M\alpha}X_j\right)\right].$$

But,

$$\begin{aligned}&\left|E\left[B_\alpha\left(\beta_0+\beta_j^{M\alpha}X_j\right)\right]-B_\alpha(\beta_0)\right| \\ &\le \sup_{|x|\le\tilde{c}_1n^{\eta-\kappa}}|B_\alpha(\beta_0+x)-B_\alpha(\beta_0)|+2E\left[G_\alpha(a|X_j|)|X_j|I(|X_j|>n^\eta)\right] \\ &= o(1)+o(1),\end{aligned}$$

by the continuity of $B_\alpha(\cdot)$ and Condition (9). Therefore, we get $Q'(\beta_0)=B_\alpha(\beta_0^{M\alpha})-B_\alpha(\beta_0)+o(1)$ and hence, for a $\tilde{c}_2>0$, we have $Q'(\beta_0^{M\alpha}-\tilde{c}_2)<0$ and $Q'(\beta_0^{M\alpha}+\tilde{c}_2)>0$ since $B_\alpha(\cdot)$ is strictly increasing. Hence $\left|\beta_{j0}^{M\alpha}-\beta_0^{M\alpha}\right|\le\tilde{c}_2$ proving our Claim 1.

Finally, to prove the theorem, we note that if $|X_j|\le n^\kappa$, then Claim 1 ensures that the points $\beta_{j0}^{M\alpha}$ and $(\beta_{j0}^{M\alpha}+\beta_j^{M\alpha}X_j)$, for all $j\in\mathcal{M}_0$, belong the interval $I=(\beta_0^{M\alpha}-h,\beta_0^{M\alpha}+h)$ independent of j, where $h=\tilde{c}-1+\tilde{c}_2$. Let $\tilde{D}=\max_{x\in I}B'_\alpha(x)$, which is finite by Lipschitz continuity of $B_\alpha(\cdot)$ in a

neighborhood of $\beta_0^{M\alpha}$ and hence, for $|X_j| \le n^\kappa$, we have

$$\left|\left\{B_\alpha\left(\beta_{j0}^{M\alpha}+\beta_j^{M\alpha}X_j\right)-B_\alpha\left(\beta_{j0}^{M\alpha}\right)\right\}\right| \le \tilde{D}|\beta_j^{M\alpha}|X_j^2.$$

Taking expectation over the region $\{|X_j| \le n^\kappa\}$, we get

$$\begin{aligned}\tilde{D}|\beta_j^{M\alpha}| &\ge \left|E\left[\left\{B_\alpha\left(\beta_{j0}^{M\alpha}+\beta_j^{M\alpha}X_j\right)-B_\alpha\left(\beta_{j0}^{M\alpha}\right)\right\}X_jI(|X_j|\le n^\kappa)\right]\right| \\ &= \left|\mathrm{Cov}(b'(\boldsymbol{X}^T\boldsymbol{\beta}_0),X_j)\right| - A_0 - A_1, \end{aligned} \tag{B2}$$

by a similar calculation leading to (B1), where $A_m = E\left[B_\alpha\left(\beta_{j0}^{M\alpha}+\beta_j^{M\alpha}X_j^m\right)X_jI(|X_j|>n^\kappa)\right]$ for $m=0,1$. But, $\left|\beta_{j0}^{M\alpha}+\beta_j^{M\alpha}X_j^m\right| \le a|X_j|$ for $|X_j|>n^\kappa$ with a sufficiently large n independent of j and m, we get from Condition (9) that $A_m \le E[G(a|X_j|)^m|X_j|I(|X_j|\ge n^\kappa)] \le dn^{-\kappa}$, for both $m=01,1$. Then, the theorem follows from (B2) using the given condition that $\left|\mathrm{Cov}(b'(\boldsymbol{X}^T\boldsymbol{\beta}_0),X_j)\right| \ge c_1n^{-\kappa}$.

APPENDIX C. PROOF OF LEMMA 1

The result in the lemma holds directly by Theorem 1 of Fan and Song (2010), provided we can show that their Conditions (A), (B) and (C) are implied by our Assumptions (A1)–(A5). In this regard, note that Assumption (A3) is indeed a reformulation of Condition (A) of Fan and Song (2010). Further, Assumption (A2) implies Condition (C) of Fan and Song (2010) via a second-order Taylor series expansion of $l_\alpha\left(Y,\boldsymbol{X}_j^T\boldsymbol{\beta}_j\right)$ with respect to $\boldsymbol{\beta}_j$ around $\boldsymbol{\beta}_j=\boldsymbol{\beta}_j^{M\alpha}$. Finally it remains to show that Condition (B) of Fan and Song (2010) holds under Assumptions (A1), (A4), and (A5).

Let us define $\Omega_n = \left\{(X_j,Y) : |X_j|\le K_n, |Y|\le K_n^*\right\}$, where K_n is as in Assumption (A4) and $K_n^* = \frac{m_0}{m_3}K_n^\tau$ with m_0, m_3 and τ being as in Assumption (A5). Then, for our present case, Condition (B) of Fan and Song (2010) becomes equivalent to

$$\begin{aligned}&\left|l_\alpha\left(Y,\boldsymbol{X}_j^T\boldsymbol{\beta}_j\right)-l_\alpha\left(Y,\boldsymbol{X}_j^T\boldsymbol{\beta}_j'\right)\right|I\left((X_j,Y)\in\Omega_n\right) \le k_n^{(\alpha)}\left|\boldsymbol{X}_j^T\boldsymbol{\beta}_j-\boldsymbol{X}_j^T\boldsymbol{\beta}_j'\right|I\left((X_j,Y)\in\Omega_n\right), \\ &\boldsymbol{\beta}_j,\boldsymbol{\beta}_j'\in\mathcal{B}, \end{aligned} \tag{C1}$$

$$\text{and}\quad \sup_{\boldsymbol{\beta}_j\in\mathcal{B}:||\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^{M\alpha}||\le\epsilon_1}\left|E\left[l_\alpha\left(Y,\boldsymbol{X}_j^T\boldsymbol{\beta}_j\right)-l_\alpha\left(Y,\boldsymbol{X}_j^T\boldsymbol{\beta}_j^{M\alpha}\right)\right]I\left((X_j,Y)\notin\Omega_n\right)\right| \le o(n^{-1}), \tag{C2}$$

where $k_n^{(\alpha)}$ is as defined in the statement of the lemma and ϵ_1 as in Assumption (A4). First, to show (C1), we use a first-order Taylor series expansion to get

$$l_\alpha\left(Y,\boldsymbol{X}_j^T\boldsymbol{\beta}_j\right)-l_\alpha\left(Y,\boldsymbol{X}_j^T\boldsymbol{\beta}_j'\right) = D(\widetilde{\boldsymbol{\beta}}_j)\left[\boldsymbol{X}_j^T\boldsymbol{\beta}_j-\boldsymbol{X}_j^T\boldsymbol{\beta}_j'\right], \tag{C3}$$

where $\widetilde{\boldsymbol{\beta}}_j\in\mathcal{B}$ lies on the line segment joining $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_j'$ and

$$D(\widetilde{\boldsymbol{\beta}}_j) = (1+\alpha)\left[\xi_\alpha\left(\boldsymbol{X}_j^T\widetilde{\boldsymbol{\beta}}_j\right)-\left(Y-b'\left(\boldsymbol{X}_j^T\widetilde{\boldsymbol{\beta}}_j\right)\right)f^\alpha\left(Y;\boldsymbol{X}_j^T\widetilde{\boldsymbol{\beta}}_j\right)\right].$$

But, on Ω_n, we have

$$\begin{aligned}\left|D(\widetilde{\boldsymbol{\beta}}_j)\right| &\le (1+\alpha)\left[\left|\xi_\alpha\left(\boldsymbol{X}_j^T\widetilde{\boldsymbol{\beta}}_j\right)\right| + \left(|Y| + |b'\left(\boldsymbol{X}_j^T\widetilde{\boldsymbol{\beta}}_j\right)|\right)\left|f^\alpha\left(Y;\boldsymbol{X}_j^T\widetilde{\boldsymbol{\beta}}_j\right)\right|\right]\\ &\le (1+\alpha)\left[|\xi_\alpha(K_nB+B)| + \left(\frac{m_0}{m_3}K_n^\tau + |b'(K_nB+B)|\right)L_\alpha\right] = k_n^{(\alpha)},\end{aligned}$$

by Assumption (A1), (A5) and the subsequent result in (12). Substituting it in (C3), we get Condition (C1).

Next, to prove (C2), we again consider the expansion (C3) with $\boldsymbol{\beta}_j' = \boldsymbol{\beta}_j^{M\alpha}$ and by taking expectation we get

$$\begin{aligned}\left|E\left[l_\alpha\left(Y,\boldsymbol{X}_j^T\boldsymbol{\beta}_j\right) - l_\alpha\left(Y,\boldsymbol{X}_j^T\boldsymbol{\beta}_j^{M\alpha}\right)\right]\right| &= E\left|D(\widetilde{\boldsymbol{\beta}}_j)\left[\boldsymbol{X}_j^T\boldsymbol{\beta}_j - \boldsymbol{X}_j^T\boldsymbol{\beta}_j^{M\alpha}\right]\right|\\ &\le (1+\alpha)||\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^{M\alpha}||_2 E\left|\left[B_\alpha\left(\boldsymbol{X}_j^T\widetilde{\boldsymbol{\beta}}_j\right) - B_\alpha\left(\boldsymbol{X}_j^T\boldsymbol{\beta}_j^{M\alpha}\right)\right]||\boldsymbol{X}_j||_2\right|,\end{aligned}$$

by an application of the Cauchy–Schwartz inequality. Therefore, we get

$$\begin{aligned}&\sup_{\boldsymbol{\beta}_j\in\mathcal{B}:||\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^{M\alpha}||\le\epsilon_1}\left|E\left[l_\alpha\left(Y,\boldsymbol{X}_j^T\boldsymbol{\beta}_j\right) - l_\alpha\left(Y,\boldsymbol{X}_j^T\boldsymbol{\beta}_j^{M\alpha}\right)\right]I\left((X_j,Y)\notin\Omega_n\right)\right|\\ &\le (1+\alpha)\epsilon_1 \sup_{\boldsymbol{\beta}_j\in\mathcal{B}:||\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^{M\alpha}||\le\epsilon_1} E\left[|B_\alpha\left(\boldsymbol{X}_j^T\widetilde{\boldsymbol{\beta}}_j\right)|||\boldsymbol{X}_j||_2 + |B_\alpha\left(\boldsymbol{X}_j^T\boldsymbol{\beta}_j^{M\alpha}\right)|||\boldsymbol{X}_j||_2\right]I\left(|X_j|>K_n\right),\\ &\le o(n^{-1}),\end{aligned}$$

by Assumption (A4), and this completes the proof.

APPENDIX D. PROOF OF THEOREM 3

D.1 Part (a)

We start with Lemma 1 and take $(1+t) = c_3Vn^{\frac{1}{2}-\kappa}\left(16k_n^{(\alpha)}\right)^{-1} > 0$ to get

$$P\left(\left|\widehat{\beta}_j^{M\alpha} - \beta_j^{M\alpha}\right| \ge c_3n^{-\kappa}\right) \le e^{-\frac{n^{1-2\kappa}}{K_n^2K_n^2}C} + nm_1e^{-m_0K_n^\tau} = R_n, \quad j=1,\ldots,p, \tag{D1}$$

Then, the uniform convergence result in Part (a) of the theorem holds from the relation (D1) via union bound of probabilities.

D.2 Part (b)

Let us consider the event $\mathcal{E}_n = \left\{\max_{j\in\mathcal{M}_0}\left|\widehat{\beta}_j^{M\alpha} - \beta_j^{M\alpha}\right| \le c_2n^{-\kappa}/2\right\}$.

By Theorem 2, on $\mathcal{E}_n$, we then have $\left|\widehat{\beta}_j^{M\alpha}\right| \ge c_2n^{-\kappa}/2$ for all $j\in\mathcal{M}_0$. Therefore, for the choice of γ_n as given in the statement of the theorem, we have $\mathcal{M}_0 \subset \widehat{\mathcal{M}}_\alpha(\gamma_n)$ on $\mathcal{E}_n$, and hence

$$P\left(\widehat{\mathcal{M}}(\gamma_n) \supset \mathcal{M}_0\right) \ge P(\mathcal{E}_n) = 1 - P(\mathcal{E}_n^c).$$

But, since $\mathcal{M}_0$ has s elements, by a union bound of probability, we get from (D1) that $P(\mathcal{E}_n^c) \le sR_n$ completing the proof of Part (b).

D.3 Part (c)

The proof is based on the result (13), evaluated at $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^{M\alpha}$, which implies that the number of variables having $|\beta_j^{M\alpha}| > \epsilon n^{-\kappa}$ cannot exceed $O(n^{2\kappa}\Lambda_{\max}(\Sigma))$ for any given $\epsilon > 0$. Now, let us consider the event

$$\widetilde{\mathcal{E}}_n = \left\{ \max_{1\le j\le p} |\widehat{\beta}_j^{M\alpha} - \beta_j^{M\alpha}| \le \epsilon n^{-\kappa} \right\}.$$

Then, on the event $\widetilde{\mathcal{E}}_n$, we have

$$\left| \left\{ j : |\widehat{\beta}_j^{M\alpha}| > 2\epsilon n^{-\kappa} \right\} \right| \le \left| \left\{ j : |\beta_j^{M\alpha}| > 2\epsilon n^{-\kappa} \right\} \right| \le O\left(n^{2\kappa}\Lambda_{\max}(\Sigma)\right).$$

Hence, taking $\epsilon = c_5/2$ for the choice of γ_n as given in the statement of the theorem, we get $P\left(|\widehat{\mathcal{M}}(\gamma_n)| \le O(n^{2\kappa}\Lambda_{\max}(\Sigma))\right) \ge P(\widetilde{\mathcal{E}}_n) = 1 - P(\widetilde{\mathcal{E}}_n^c)$. But, by Part (a) of the theorem, we have $P(\widetilde{\mathcal{E}}_n^c) \le pR_n$ completing the proof.

APPENDIX E. PROOFS OF THEOREMS 4 AND 5

We note that, for each $j \in \mathcal{D}$, the quantity $\boldsymbol{\beta}_{Cj}^{M\alpha}$, defined in (17), satisfies the estimating equations given by

$$E\left[\psi_\alpha\left(Y, \boldsymbol{X}_{Cj}^T\boldsymbol{\beta}_{Cj}^{M\alpha}\right)\boldsymbol{X}_C\right] = 0, \quad E\left[\psi_\alpha\left(Y, \boldsymbol{X}_{Cj}^T\boldsymbol{\beta}_{Cj}^{M\alpha}\right)X_j\right] = 0. \tag{E1}$$

On the other hand, the baseline quantity $\boldsymbol{\beta}_C^{M\alpha}$ satisfies

$$E\left[\psi_\alpha\left(Y, \boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}\right)\boldsymbol{X}_C\right] = 0. \tag{E2}$$

Further, for any $j \in \mathcal{D}$, using $E[X_j|\boldsymbol{X}_C] = 0$, we have

$$\begin{aligned} \mathrm{Cov}_L(Y, X_j|\boldsymbol{X}_C) &= E\left[(Y - E[Y|\boldsymbol{X}_C])X_j\right] = E\left[YX_j\right] \\ &= E\left[E(Y|\boldsymbol{X})X_j\right] = E\left[b'(\boldsymbol{X}^T\boldsymbol{\beta}_0)X_j\right], \end{aligned} \tag{E3}$$

and hence, invoking the definitions of ψ_α and B_α, we get

$$\begin{aligned} E\left[\psi_\alpha\left(Y, \boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}\right)X_j\right] &= E\left[\left(B_\alpha\left(\boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}\right) - b'\left(\boldsymbol{X}^T\boldsymbol{\beta}_0\right)\right)X_j\right] \\ &= EE\left[\left(B_\alpha\left(\boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}\right) - b'\left(\boldsymbol{X}^T\boldsymbol{\beta}_0\right)\right)X_j|\boldsymbol{X}_C\right] \\ &= -E\left[b'\left(\boldsymbol{X}^T\boldsymbol{\beta}_0\right)X_j\right] \\ &= -\mathrm{Cov}_L(Y, X_j|\boldsymbol{X}_C). \end{aligned} \tag{E4}$$

E.1 Proof of Theorem 4

Firstly, if $\beta_j^{M\alpha} = 0$ for some $j \in \mathcal{D}$, from (E1) we get

$$E\left[\psi_\alpha\left(Y, \boldsymbol{X}_C^T\boldsymbol{\beta}_{Cj1}^{M\alpha}\right)\boldsymbol{X}_C\right] = 0, \quad E\left[\psi_\alpha\left(Y, \boldsymbol{X}_C^T\boldsymbol{\beta}_{Cj1}^{M\alpha}\right)X_j\right] = 0. \tag{E5}$$

Combining the first equation with (E2) and the uniqueness of its solution we have $\boldsymbol{\beta}_{Cj1}^{M\alpha} = \boldsymbol{\beta}_{C}^{M\alpha}$ and hence the second equation in (E5) becomes

$$E\left[\psi_\alpha\left(Y, \boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}\right)X_j\right] = 0. \tag{E6}$$

This leads to the desired condition $\mathrm{Cov}_L(Y, X_j|\boldsymbol{X}_C) = 0$ by (E4).

On the other hand, if $\mathrm{Cov}_L(Y, X_j|\boldsymbol{X}_C) = 0$ for some $j \in \mathcal{D}$, then by (E4), the equation in (E6) hold. Combining (E6) with (E2), we see that $\boldsymbol{\beta}_{Cj}^{M\alpha} = \left(\boldsymbol{\beta}_{C}^{M\alpha}, 0\right)^T$ is a solution of the estimating equations in (E1), leading to $\beta_j^{M\alpha} = 0$.

E.2 Proof of Theorem 5

Fix any $j \in \mathcal{M}_{0\mathcal{D}}$ and define $\boldsymbol{\Omega}_j = E\left[m_{\alpha,j}\boldsymbol{X}_{Cj}\boldsymbol{X}_{Cj}^T\right]$ and $\boldsymbol{\beta}_{\Delta,j} = \left(\boldsymbol{\beta}_{Cj1}^{M\alpha} - \boldsymbol{\beta}_{C}^{M\alpha}\right)$. Consider a partition of $\boldsymbol{\Omega}_j$ as given by

$$\boldsymbol{\Omega}_j = \begin{bmatrix} \boldsymbol{\Omega}_{11,j} & \boldsymbol{\Omega}_{12,j} \\ \boldsymbol{\Omega}_{21,j}^T & \boldsymbol{\Omega}_{22,j} \end{bmatrix} = \begin{pmatrix} E\left[m_{\alpha,j}\boldsymbol{X}_C\boldsymbol{X}_C^T\right] & E\left[m_{\alpha,j}\boldsymbol{X}_C X_j\right] \\ E\left[m_{\alpha,j}X_j\boldsymbol{X}_C^T\right] & E\left[m_{\alpha,j}X_j^2\right] \end{pmatrix}.$$

Now, from the estimating equations (E1) and (E2), along with the definitions of B_α and $m_{\alpha,j}$, we get

$$\begin{aligned} 0 &= E\left[\left(B_\alpha\left(\boldsymbol{X}_{Cj}^T\boldsymbol{\beta}_{Cj}^{M\alpha}\right) - B_\alpha\left(\boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}\right)\right)\boldsymbol{X}_C\right] \\ &= E\left[m_{\alpha,j}\left(\boldsymbol{X}_{Cj}^T\boldsymbol{\beta}_{Cj}^{M\alpha} - \boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}\right)\boldsymbol{X}_C\right] \\ &= E\left[m_{\alpha,j}\left(\boldsymbol{X}_C^T\boldsymbol{\beta}_{\Delta,j} + X_j\beta_j^{M\alpha}\right)\boldsymbol{X}_C\right]. \end{aligned}$$

Therefore, by solving, we get $\boldsymbol{\beta}_{\Delta,j} = -\Omega_{11,j}^1\Omega_{12,j}\beta_j^{M\alpha}$. Further, note that, for any integrable function $h(\boldsymbol{X}_C)$, we have

$$E[h(\boldsymbol{X}_C)X_j] = EE[h(\boldsymbol{X}_C)X_j|\boldsymbol{X}_C] = E\left[h(\boldsymbol{X}_C)E(X_j|\boldsymbol{X}_C)\right] = 0.$$

Hence, by (E3), (E1), and the definition of $m_{\alpha,j}$, we get

$$\begin{aligned} \mathrm{Cov}_L(Y, X_j|\boldsymbol{X}_C) &= E\left[b'\left(\boldsymbol{X}^T\boldsymbol{\beta}_0\right)X_j\right] = E\left[\left(b'\left(\boldsymbol{X}^T\boldsymbol{\beta}_0\right) - B_\alpha\left(\boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}\right)\right)X_j\right] \\ &= E\left[\left(B_\alpha\left(\boldsymbol{X}_{Cj}^T\boldsymbol{\beta}_{Cj}^{M\alpha}\right) - B_\alpha\left(\boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}\right)\right)X_j\right] \\ &= E\left[m_{\alpha,j}\left(\boldsymbol{X}_{Cj}^T\boldsymbol{\beta}_{Cj}^{M\alpha} - \boldsymbol{X}_C^T\boldsymbol{\beta}_C^{M\alpha}\right)X_j\right] \\ &= \Omega_{12,j}^T\boldsymbol{\beta}_{\Delta,j} + \Omega_{22,j}\beta_j^{M\alpha} \\ &= \left[\Omega_{22,j} - \Omega_{12,j}^T\Omega_{11,j}^{-1}\Omega_{12,j}\right]\beta_j^{M\alpha}. \end{aligned}$$

Now, taking absolute value in the above and using the assumptions of the theorem, we get $c_1 n^{-\kappa} \leq c_2\left|\beta_j^{M\alpha}\right|$. Since this holds for all $j \in \mathcal{M}_{0\mathcal{D}}$, taking minimum over all such j we get the desired conclusion of the theorem with $c_3 = c_1/c_2$.