Inference In General Single-Index Models Under High-dimensional Symmetric Designs

Hamid Eftekhari hamidef@umich.edu
Moulinath Banerjee moulib@umich.edu
Ya'acov Ritov yritov@umich.edu
Department of Statistics
University of Michigan

Editor:

arXiv:1909.03540v1 [math.ST] 8 Sep 2019

Abstract

We consider the problem of statistical inference for a finite number of covariates in a generalized single-index model with $p > n$ covariates and unknown (potentially random) link function under an elliptically symmetric design. Under elliptical symmetry, the problem can be reformulated as a proxy linear model in terms of an identifiable parameter, which characterization is then used to construct estimates of the regression coefficients of interest that are similar to the de-biased lasso estimates in the standard linear model and exhibit similar properties: square-root consistency and asymptotic normality. The procedure is *agnostic* in the sense that it completely bypasses the estimation of the link function, which can be extremely challenging depending on the underlying structure of the problem. Our method allows testing for the importance of pre-fixed covariates in the single-index model, as well as testing for the relative importance of coefficients via straightforward application of the delta method. Furthermore, under Gaussianity, we extend our approach to prescribe improved, i.e., more efficient estimates of the coefficients using a sieved strategy that involves an expansion of the true regression function in terms of Hermite polynomials. Finally, we illustrate our approach via carefully designed simulation experiments.

Keywords:

1. Introduction and Background

The single-index model has been the subject of extensive investigation in both the statistics and econometric literatures over the last few decades. Its utility lies in that it generalizes the linear model to scenarios where the regression function $E(Y|X)$ is not necessarily linear in the covariates; rather it is connected to the covariates by an unknown transformation of $X^T\tau_0$, τ_0 being the regression co-efficients, i.e. $E(Y|X) = G_0(X^T\tau_0)$. While allowing broad generality in the structure of the mean response, the single-index model also circumvents the curse of dimensionality by modeling the mean response in terms of a low-dimensional functional of X. An excellent review of single-index models appears, for example, in the work of Horowitz (2009).

It is well-known that the parameter τ_0 is in general unidentifiable in the single-index model, since any scaling of τ_0 can always be absorbed into the function G, and therefore, some identifiability constraints are imposed for statistical estimation and inference, with the most popular choice being to set $\|\tau\| = 1$. General schemes of estimation of the parameter of interest, τ_0, involve optimizing an appropriate loss function (likelihood/pseudo-likelihood/least squares) in (τ, G) (generic parameter values) by alternately updating estimates of τ and G (Carroll et al., 1997), or devising a profile estimating equation for τ that involves some estimate of G computed under the assumption that the true parameter is indeed τ: see, e.g. the WNLS estimator in Section 2.5 of Horowitz (2009). In any case, the estimation of G figures critically in most estimation schemes. Inference on τ_0 requires appropriate regularity assumptions on G, typically involving smoothness constraints, and while G is only estimable at rate slower than $\sqrt{n}$, τ_0 possesses $\sqrt{n}$ consistent asymptotically normal and efficient estimates, under certain regularity conditions.

High-dimensional single index models have also attracted interest with various authors studying variable selection, estimation and inference using penalization schemes. For example, Ganti et al. (2015) proposed ℓ_1 penalized estimates for learning high-dimensional index models and provided theoretical guarantees on excess risk for bounded responses; Foster et al. (2013) proposed an algorithm for variable selection in a monotone single index model via the adaptive lasso but without any rigorous proofs; Luo and Ghosal (2016) used a penalized forward selection technique for high-dimensional single index models for a monotone link function, but again without any theoretical guarantees; Cheng et al. (2017) used cubic B-splines for estimating the single index model in conjunction with a SICA (smooth integration of counting and absolute deviation) penalty function for variable selection; and Radchenko (2015) also studied simultaneous variable selection and estimation in high-dimensional SIMs using a penalized least-squares criterion, with the link function estimated via B-splines, and provided theoretical results on the rate of convergence. Very recently, Hirshberg and Wager (2018) have proposed a method for average partial effect estimation in high-dimensional single-index models that is root-n-consistent and asymptotically unbiased under sparsity assumptions on the regression, and to the best of our knowledge, this is the only work that provides asymptotic distributions in the high-dimensional setting. However, their method critically uses the form of the link function ψ and also requires it to be adequately differentiable.

In this paper, we develop an inference scheme for the regression coefficients (up to scale) of a high-dimensional single-index model with minimal restrictions on the (potentially random) link function (indeed, even discontinuous link functions are allowed) that *completely bypasses* the estimation of the link. Thus, by dispensing with most regularity conditions on the link function, our approach can accommodate diverse underlying model structures. The price one pays for the feasibility of such an agnostic approach is an elliptically symmetric restriction on X. This assumption may be overly restrictive in applications where the entries of X are determined by nature. However, in some applications one has the luxury of designing the measurement matrix. For example, Gaussian random matrices have been used as the measurement matrix in compressed sensing with applications to radar and ultrasound imaging(see e.g., Baraniuk and Steeghs, 2007; Achim et al., 2014).

Next we introduce our model. Consider the semiparametric single-index model :

$$y_i = f_i(\langle x_i, \tau \rangle), \quad i \in \{1, 2, \ldots, n\}, \tag{1}$$

where $f_i : \mathbf{R} \to \mathbf{R}$ are iid realizations of an unknown random function f, independent of x_i, and $\tau \in \mathbf{R}^p$ is an unknown parameter whose direction is the object of estimation. *For starters*, assume that $x_i \sim_{iid} N(0_p, \Sigma)$ for a positive definite matrix Σ. While τ is not identifiable in this model [1], an appropriate scalar multiple: $\beta = \mu\tau$ (where μ will be defined shortly) is. The new parameter β turns out to be the vector of average partial derivatives of the regression function with respect to the covariates in the model, under certain conditions.[2]

We will write $y = (y_1, \ldots, y_n)^T$, and let X denote the matrix with $x_1^T, \ldots, x_n^T$ in its rows. For subsets of indices $I \subset \{1, \ldots, n\}, J \subset \{1, \ldots, p\}$ we let $X_{I,J}$ be the submatrix of X containing the rows with indices in I and columns with indices in J. For Gaussian covariates, there is a specific feature of this model that obviates the need to estimate the link function that we now describe. Define:

$$\mu := \mathbf{E} f(\zeta)\zeta, \tag{2}$$

$$\beta := \mu\tau, \tag{3}$$

$$z := y - X\beta, \tag{4}$$

where ζ is a standard normal variable independent of f. We also assume $\|\Sigma^{\frac{1}{2}}\tau\|_2 = 1$, as otherwise we can rescale τ and f appropriately without changing β. Also, as we seek to make inference on the importance/relative importance of the components of τ via estimates of β, we assume henceforth that $\mu \neq 0$. It can be shown (see B.1 in the appendix) that

$$\mathbf{E} X^T z = 0, \tag{5}$$

which is equivalent to $\mathbf{E} z_i x_i = 0$ for all $i = 1, \ldots, n$. We therefore have the representation $Y = X\beta + z$ with X and z uncorrelated, which implies that $\beta \in \arg\min_{\beta'} \mathbf{E}\|y - X\beta'\|_2^2$, thus motivating the use of (penalized) least squares methods for estimation. The above equation will be referred to as the orthogonality property of X and z and will be used subsequently at several places.

The orthogonality property appears to have been first noted in the work of Brillinger (1982) who used it to study the properties of the least squares estimator $\hat{\beta}_{ls}$ in the classical fixed p setting, and in particular, the estimator was shown to be asymptotically normal. Plan and Vershynin (2016) studied the estimation of β in the $p > n$ setting using a generalized constrained lasso. While their results are quite general, they require knowing the constraint set K over which least squares is performed. Under a different set of assumptions, Thrampoulidis et al. (2015) obtained an asymptotically precise expression for the estimation error of the regularized generalized lasso when X has iid $N(0, 1)$ entries and β is generated according to a density in $\mathbf{R}^p$ with marginals that are independent of p.

Recall that a random vector V is called spherically symmetric if its distribution is invariant to all possible rotations, i.e. $V \equiv_d PV$ for all orthogonal matrices P. Say that X has an elliptically symmetric distribution if for some fixed vector μ and p.d. matrix Σ, $\Sigma^{-1/2}(X - \mu)$ is spherically symmetric. It turns out that the proxy linear model representation also holds *more generally* when X follows an elliptically symmetric distribution; see Section 3 for the details. This fact appears in the work of Li et al. (1989), who then use the linear model representation to construct asymptotically normal and unbiased estimators of the regression coefficients in a fixed dimension parameter setting. More recently, Goldstein et al. (2018)

1. Identifiability is discussed in more detail in Appendix B (section B.3).
2. See section B.2 in the appendix for more details on this connection.

have studied structured signal recovery (in high dimensions) from a single-index model with elliptically symmetric X, which can be viewed as an extension of Plan and Vershynin (2016).

Our approach in this paper is to construct asymptotically normal and unbiased estimates of pre-fixed finite dimensional components of β, by using de-biasing techniques (van de Geer et al., 2014; Zhang and Zhang, 2014; Javanmard and Montanari, 2014) on a consistent pilot estimator $\hat{\beta}$ in both the Gaussian and the general elliptically symmetric cases. As will be shown below, we have considerable flexibility in the choice of a pilot estimator, so long as it has a sufficiently fast rate of convergence in terms of ℓ_1 error. In the interests of concrete implementation, we prescribe using an ℓ_1 regularized lasso estimator of β that has the desired convergence rate (at least when the residual z is sub-Gaussian). To fix ideas, we initially develop detailed results for Gaussian design in Section 2 and then present results under general elliptical symmetry in Section 3, using the linear model representation. Section 4 extends the approach to construct improved (as in more efficient) estimates of these parameters under Gaussianity via a sieved estimation strategy for g (under appropriate regularity conditions) which involves its expansion in terms of the first m components of the orthonormal Hermite polynomial basis, with m slowly varying in n, the sample size.

2. Inference under Gaussian design

In this section we apply the debiasing technique to obtain $\sqrt{n}$-consistent estimators of individual coordinates of β under the assumption of Gaussian design. Our main theorems assume the existence of pilot estimators of β that possess sufficiently fast (ℓ^1-norm) rates of convergence. Subsection 2.3 discusses the existence and construction of these pilot estimators. The extension of the results in this section to the case of elliptically symmetric design is considered in Section 3.

2.1 Inference on β when Σ is known

We first consider the case where Σ is known. In this case, the distribution of $x_i \sim N(0, \Sigma)$ is fully known, and we can therefore compute the L_2 projection of any covariate, $x_{n,1}$, over the remaining ones, $x_{n,-1}$. Specifically, define

$$\gamma := (\mathbf{E} x_{n,-1} x_{n,-1}^T)^{-1} \mathbf{E} x_{n,-1} x_{n,1}.$$

In words, γ is the vector of coefficients when regressing the first covariate on the rest, at the population level. The resulting residuals

$$r_i := x_{i,1} - \langle \gamma, x_{i,-1} \rangle \tag{6}$$

satisfy

$$\begin{aligned} \mathbf{E} r_i x_{i,-1}^T &= \mathbf{E} x_{i,1} x_{i,-1}^T - \gamma^T \mathbf{E} x_{i,-1} x_{i,-1}^T \\ &= 0. \end{aligned}$$

Using these residuals and the pilot estimator $\hat{\beta}$, define a de-biased estimator of β_1 as

$$\tilde{\beta}_1 := \hat{\beta}_1 + \frac{\sum_{i=1}^n r_i (y_i - \langle x_i, \hat{\beta} \rangle)}{\sum_{i=1}^n r_i x_{i,1}}. \tag{7}$$

The following theorem characterizes the asymptotic distribution of $\tilde{\beta}_1$.

Theorem 1 *Suppose* $(x_i, y_i)_{i=1}^n$ *follow the model (1) with* $x_i \sim N(0, \Sigma)$ *and let*

$$\nu^2 = \frac{\mathbf{E} r_n^2 z_n^2}{(\mathbf{E} r_n^2)^2}.$$

Assume also that the following conditions are satisfied:

1. $\mathbf{E} r_n^2 z_n^2$ *is bounded away from zero.*
2. $\hat{\beta}$ *is an estimate of* β *satisfying*
$$\sqrt{\log(p)} \|\hat{\beta} - \beta\|_1 \to_p 0.$$
3. *There exist* $0 < c, C < \infty$ *such that* $c \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C$.
4. $\mathbf{E}|y_n|^{2+\alpha} \leq M < \infty$ *for some* $\alpha, M > 0$ *and all* $n \geq 1$.

Then

$$\frac{\sqrt{n}(\tilde{\beta}_1 - \beta_1)}{\nu} \to_d N(0, 1).$$

Remark 2 *In the special case of a standard linear model, i.e. when* $y_i = \mu \cdot \langle x_i, \tau \rangle + \epsilon_i$ *with* $\epsilon_i \perp\!\!\!\perp x_i$*, the asymptotic variance of* $\sqrt{n}\tilde{\beta}_1$ *reduces to* $\nu^2 = \sigma^2 / \mathbf{E}\mathbf{r}_n^2$ *which is the variance of the OLS estimator of* β_k *in the low-dimensional case and the debiased estimator (van de Geer et al., 2014) in the high-dimensional case. (In fact, in this case* $\tilde{\beta}_1$ *itself reduces to the de-biased estimator.)*

Remark 3 *Using the* $(2 + \alpha)$ *moment bound on* y_n*, the proof of Theorem 1 shows that* $\mathbf{E} r_n^2 z_n^2$ *is uniformly bounded above. Together with* $\mathbf{E} r_n^2 \geq \lambda_{\min}(\Sigma)$*, this shows that* ν^2 *is uniformly bounded above and thus the rate of convergence of* $\tilde{\beta}_1$ *is indeed* $\sqrt{n}$.

Remark 4 *If the design matrix* X *is not centered, but* $\mathbf{E} x_n$ *is known, then one can first center* X *by using* $\tilde{X} = X - \mathbf{E}X$ *and redefining the link function:*

$$\tilde{f}_i(\cdot) = f_i(\cdot + \langle \mathbf{E} x_i, \tau \rangle).$$

Even though f_i *now depends on the mean of* x_i*, it is independent of* x_i*, and hence all the theoretical results in this paper still continue to hold. In simulations we consider a case where* X *has nonzero mean and is centered on the sample, prior to constructing the estimator (since, in reality,* EX *will not be known). While this, strictly speaking, does not fall within the purview of our approach (since centering around the sample mean induces some dependence between* x_i *and* f_i*), our inference procedure is still seen to produce satisfactory results.*

2.2 Inference when Σ is unknown

In this section we consider the problem of de-biasing an estimate $\hat{\beta}_1$ of β_1 when the precision matrix Σ^{-1} is unknown but estimable. Suppose that we have a sample $\mathcal{S} = (x_i, y_i)_{i=1}^{2n}$ of size $2n$, and that we use the second sub-sample $\mathcal{S}_2 = (x_i, y_i)_{i=n+1}^{2n}$ to find an estimate $\hat{\gamma}$ of γ, which is subsequently used to find estimates $\hat{r}_i$ of r_i on the first sub-sample $\mathcal{S}_1 = (x_i, y_i)_{i=1}^{n}$:

$$\hat{r}_i := x_{i,1} - \langle x_{i,-1}, \hat{\gamma} \rangle, \quad i \in \{1, \dots, n\}. \tag{8}$$

The debiased estimator of β_1 on the first subsample is defined by

$$\tilde{\beta}_1 := \hat{\beta}_1 + \frac{\sum_{i=1}^n \hat{r}_i (y_i - \langle x_i \hat{\beta} \rangle)}{\sum_{i=1}^n \hat{r}_i x_{i,1}}. \tag{9}$$

Theorem 5 *Suppose $(x_i, y_i)_{i=1}^{2n}$ follow the model (1) with $x_i \sim N(0, \Sigma)$ and let*

$$\begin{aligned}
\gamma :&= (\mathbf{E} x_{n,-1} x_{n,-1}^T)^{-1} \mathbf{E} x_{n,-1} x_{n,1}, \\
r_i :&= x_{i,1} - \langle \gamma, x_{i,-1} \rangle, \\
\nu^2 :&= \frac{\mathbf{E} r_n^2 z_n^2}{(\mathbf{E} r_n^2)^2}.
\end{aligned}$$

Assume also that the following conditions are satisfied:

1. *$\mathbf{E} r_n^2 z_n^2$ is bounded away from zero.*
2. *$\hat{\beta}$ is an estimate of β satisfying*

$$s\sqrt{\log(p)} \|\hat{\beta} - \beta\|_1 \to_p 0.$$

3. *There exists an estimate $\hat{\gamma}$ of γ, depending on data in the second sub-sample $\mathcal{S}_2$, satisfying*

$$\mathbf{P}\left(\|\hat{\gamma} - \gamma\|_1 \le c_\gamma s \sqrt{\frac{\log p}{n}} \right) \to 1,$$

   for a constant c_γ not dependent on n.

4. *There exist $0 < c, C < \infty$ such that $c \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le C$.*
5. *$\mathbf{E}|y_n|^{2+\alpha} \le M < \infty$ for some $\alpha, M > 0$ and all $n \ge 1$.*

Then

$$\frac{\sqrt{n}(\tilde{\beta}_1 - \beta_1)}{\nu} \to_d N(0, 1).$$

Remark 6 *1. Our approach to approximately de-correlating the design matrix uses sample splitting for estimation of γ, and the supporting argument is somewhat different from the ones in Zhang and Zhang (2014), Javanmard and Montanari (2014) and van de Geer et al. (2014). This is because in our linear model $y_i = \langle x_i, \beta \rangle + z_i$, the error z_i is not independent of, but only uncorrelated with, x_i. It is not clear if the argument could be changed to justify the use of standard de-biasing techniques (introduced in the aforementioned papers) in this case.*

2. Under the assumption of Gaussian design, the estimator $\hat{\gamma}$ of γ exists if (for example) the first row of $\Omega = \Sigma^{-1}$ is s-sparse. Denoting by Ω_1 the first row of Ω, it can be shown that $(1, -\gamma^T) = (\mathbf{E}r_i^2)\Omega_1$ where $r_i = x_{i1} - \gamma^T x_{i,-1}$. Thus γ is s-sparse as well, and can be estimated via the linear model

$$x_{i,1} = \langle \gamma, x_{i,-1} \rangle + r_i, \quad i = 1, \ldots, n,$$

using node-wise lasso on the m-th subsample:

$$\hat{\gamma} = \arg\min_{\gamma'} \left\{ \frac{1}{2n} \|X_{\cdot,1} - X_{\cdot,-1}\gamma'\|_2^2 + \lambda_k \|\gamma'\|_1 \right\}, \tag{10}$$

Note that removing a column of X does not decrease its restricted eigenvalues. So as long as the RE condition holds for X, it holds for $X_{\cdot,-1}$ with the same constants. See also Lemma 5.4 in (van de Geer et al., 2014). Obviously, all considerations here go through if instead of estimating β_1, we are interested in estimating β_k for some other fixed k: we replace 1 by k at the pertinent places.

Remark 7 *In order to avoid loss of efficiency due to sample splitting, one can change the roles of the two sub-samples in the theorem to compute two estimates $\tilde{\beta}_1^1, \tilde{\beta}_1^2$, and use the average of $\tilde{\beta}_1^1$ and $\tilde{\beta}_1^2$ as the final estimator. The proof of Theorem 5 shows that*

$$\frac{\sqrt{n}(\tilde{\beta}_1^m - \beta_1)}{\nu} = G^m + o_p(1), \quad m = 1, 2,$$

where

$$G^m = \frac{1}{\sqrt{n}} \frac{\sum_{i \in I_m} r_i z_i}{\nu \mathbf{E} r_n^2}$$

depends only on the m-th subsample so that G^0 and G^1 are independent. Moreover, we have $G^m \to_d N(0,1)$ as $n \to \infty$, and so by independence

$$G^0 + G^1 \to_d N(0, 2).$$

Consequently, for the average estimator $\tilde{\beta}_1^{avg} = (\tilde{\beta}_1^1 + \tilde{\beta}_1^2)/2$ we have

$$\begin{aligned}
\sqrt{2n} \frac{(\tilde{\beta}_1^{avg} - \beta_1)}{\nu} &= \frac{1}{\sqrt{2}} \left(\sqrt{n} \frac{(\tilde{\beta}_1^0 - \beta_1)}{\nu} + \sqrt{n} \frac{(\tilde{\beta}_1^1 - \beta_1)}{\nu} \right) \\
&= \frac{1}{\sqrt{2}} (G^0 + G^1 + o_p(1)) \\
&\to_d N(0, 1),
\end{aligned}$$

showing that the lost efficiency due to sample splitting is regained by switching the roles of sub-samples.

Remark 8 *Inference on multivariate components of β can be done similarly. To simplify notation, suppose a sample of size $(x_i, y_i)_{i=1}^{2n}$ is given, $I = \{1, \dots, n\}$, and assume $K \subset \{1, \dots, p\}$ is a fixed set of indices, which without loss of generality, can be taken to be $\{1, \dots, |K|\}$. For $k \in K$ and $i \in I$, suppose r_{ik} is computed as in (6) and $\hat{r}_{ik}$ is computed as in (8), where the corresponding node-wise lasso coefficient is computed on the second subsample $X_{I^c} = (x_{i,k}, x_{i,-k})_{i=n+1}^{2n}$. Let $R = (r_{ik})_{i,k} \in \mathbf{R}^{n \times |K|}$ and $\hat{R} = (\hat{r}_{ik})_{i,k} \in \mathbf{R}^{n \times |K|}$. The debiased estimate of β_K is then defined as*

$$\tilde{\beta}_K = \hat{\beta}_K + (\hat{R}^T X_{I,K})^{-1}(\hat{R}^T(y_I - X_I\hat{\beta})).$$

Under assumptions similar to those in Theorem (5), one can show that

$$\sqrt{n}\Theta^{-\frac{1}{2}}(\frac{\hat{R}^T X_{I,K}}{n})(\tilde{\beta}_K - \beta_K) \to_d N_{|K|}(0, I),$$

where $\Theta = \mathbf{E}[z_n^2 R_n R_n^T]$ and R_n^T denotes the n-th row of R.

2.3 A pilot estimator of β

In this subsection the construction of a pilot estimator for β is discussed. Consistency and rates of convergence of (generalized) constrained lasso estimators for single-index models under Gaussian or elliptically symmetric design have been established in other works (Plan and Vershynin, 2016; Goldstein et al., 2018). In what follows we state and prove the consistency of penalized lasso under assumptions similar to the ones typically used in high dimensional linear models.

Let $\hat{\beta}$ be a solution to the penalized lasso problem:

$$\hat{\beta} = \arg\min_{\beta'}\{\frac{1}{2n}\|Y - X\beta'\|_2^2 + \lambda\|\beta'\|_1\}. \tag{11}$$

The following proposition provides sufficient conditions for consistency of $\hat{\beta}$, following the arguments in (Bickel et al., 2009). Before we state the proposition, we review the concept of restricted eigenvalues. A matrix A is said to satisfy the restricted eigenvalue condition with parameters (s, κ, α), if for all $S \subset \{1, \dots, p\}$ with $|S| \leq s$ and all $\theta \in \mathbf{R}^p$ with $\|\theta_{S^c}\|_1 \leq \kappa\|\theta_S\|_1$ we have

$$\|A\theta\|_2 \geq \alpha\|\theta_S\|_2.$$

Proposition 9 *Suppose that model (1) holds with $x \sim N(0, \Sigma)$ and let $\sigma_x^2 = 4\max_j \Sigma_{jj}$. Assume that*

1. *$z_i = y_i - \langle x_i, \beta\rangle$ is subgaussian with $\|z_i\|_{\psi_2} \leq \sigma_z$,*
2. *β is a s-sparse vector, i.e. $|\{j : \beta_j \neq 0\}| \leq s$,*
3. *$\Sigma^{\frac{1}{2}}$ satisfies the restricted eigenvalue condition with parameters $(s, 9, 2\alpha)$ for some $\alpha > 0$, and that α and $\lambda_{\max}(\Sigma)$ are bounded away from $0, \infty$.*

Then there exists an absolute constant $c_0 > 0$ such that for $\lambda > c_0\sigma_z\sigma_x\sqrt{\log(p)/n}$ and $n \geq c_0(1 \vee \sigma_x^4)s\log(p/s)$ we have

$$\|\hat{\beta} - \beta\|_1 \leq \frac{16s\lambda}{\alpha^2} \quad \text{and} \quad \frac{\|X(\hat{\beta} - \beta)\|_2}{\sqrt{n}} \leq \frac{4\sqrt{s}\lambda}{\alpha},$$

with probability no less than $1 - 2p^{-1} - \exp(-c_0 n^2/\sigma_x^4)$.

Remark 10 *1. Assumption (1) is the analogue of subgaussian errors in linear models. This is a strong assumption on the approximation error z_i, but is nevertheless satisfied in some interesting cases such as when y_i is bounded almost surely by a constant or when the model can be written as $y_i = g(\langle x_i, \tau\rangle) + e_i$, where g is an unknown Lipschitz function and e_i is a mean-zero subgaussian error independent of x_i.*

2. Restricted eigenvalue conditions on the design matrix are standard in the analysis of high-dimensional linear models (Bickel et al., 2009). With random designs one can show that the RE condition for the sample design matrix follows with high probability from the corresponding condition on the population covariance matrix (Rudelson and Zhou, 2012). It is clear that the RE condition on $\Sigma^{\frac{1}{2}}$ is weaker than the assumption that Σ has eigenvalues that are bounded away from $0, \infty$ (as is assumed in theorems 1, 5).

3. Inference for a General Elliptically Symmetric design

In this section, extensions of Proposition 9, Theorem 1 and Theorem 5 to the more general setting of elliptically symmetric design are considered. We start by reviewing the definitions of elliptically symmetric and sub-gaussian vectors.

Definition 11 *A centered random vector $x \in \mathbf{R}^p$ follows an elliptically symmetric distribution with parameters Σ and F_v if*

$$x \stackrel{d}{=} vBU,$$

where the random variable $v \in \mathbf{R}$ has distribution F_v, the random vector $U \in \mathbf{R}^p$ is uniformly distributed over the unit sphere $\mathbb{S}^{p-1}$ and is independent of v, and B is a matrix satisfying $\Sigma = BB^T$. In this case we write $x \sim \mathcal{E}(0, \Sigma, F_v)$.

Note that the matrix B and the random variable v in the above definition are not uniquely determined. In particular, for any orthogonal matrix Q and $t > 0$, if the pair (B, v) satisfies the definition then so does the pair $(tBQ, v/t)$. For comparability with the case of Gaussian random vectors, in this work we assume that in this representation $\mathbf{E}v^2 = p$, so that the variance-covariance matrix of x is equal to Σ, i.e. $\mathbf{E}xx^T = \Sigma$.

It is well-known that elliptically symmetric distributions generalize the multivariate normal distribution, and in particular, include distributions that have heavier or lighter tails than the normal distribution. More precisely, in the above definition, if $v = \sqrt{u}$ where $u \sim \chi_p^2$, then $\mathcal{E}(0, \Sigma, F_v) = N(0, \Sigma)$.

Definition 12 *A centered random vector $x \in \mathbf{R}^p$ is subgaussian with subgaussian constant σ if for all unit vectors $u \in \mathbb{S}^{p-1}$ we have that $\langle u, x\rangle$ is a subgaussian random variable with $\|\langle u, x\rangle\|_{\psi_2} \leq \sigma$. In this case we write $\|x\|_{\psi_2} \leq \sigma$.*

Under an elliptically symmetric design $x_i \sim \mathcal{E}(0, \Sigma, F_v)$, the linear representation $y_i = \mu\langle x_i, \tau\rangle + z_i$ is still valid with $\mathbf{E} z_i x_i = 0$, when μ and z_i are defined by

$$\begin{aligned}\mu &= \mathbf{E} y_i \langle x_i, \tau\rangle, \\ z_i &= y_i - \mu\langle x_i, \tau\rangle,\end{aligned}$$

and we use the normalization $\|\Sigma^{\frac{1}{2}}\tau\|_2 = 1$. The argument for $\mathbf{E} z_i x_i = 0$ is exactly as in the case of Gaussian design, since, as far as the distribution of x_i is concerned, the proof in Section B.1 only requires the normalization $\|\Sigma^{\frac{1}{2}}\tau\|_2 = 1$ and the fact that the conditional expectation of x given $\langle x, \tau\rangle$ is linear in $\langle x, \tau\rangle$, that is, there exists a (non-random) vector b such that $\mathbf{E}[x \mid \langle x, \tau\rangle] = \langle x, \tau\rangle b$. The latter property also holds for elliptically symmetric random vectors (see Goldstein et al., 2018, Corollary 2.1). Besides the orthogonality property $\mathbf{E} z_i x_i = 0$, our proofs rely on controlling the tail probabilities of certain random variables, such as $\|z^T X\|_\infty$ in Proposition 9 and $\max_{j\neq k} |\sum_{i=1}^n r_i z_i|$ in Theorems 1 and 5. In addition, in the case of unknown Σ, subgaussian tails of x_i were used to control the moments of $r_i - \hat{r}_i = \langle x_{i,-1}, \hat{\gamma} - \gamma\rangle$. The assumption of sub-gaussianity of x_i allows the same proofs go through in the case of elliptically symmetric designs.

Remark 13 *A sufficient condition for $x \sim \mathcal{E}(0, \Sigma, F_v)$ to be subgaussian is that in the representation $x = vBU$ the random variable v is subgaussian. This follows because for all unit vectors $w \in \mathbb{S}^{p-1}$,*

$$\begin{aligned}|\langle w, x\rangle| &= |\langle w, BU\rangle| \cdot |v| \\ &\leq \|B^T w\|_2 \cdot |v| \\ &\leq \sqrt{\lambda_{\max}(BB^T)}|v|, \quad a.s..\end{aligned}$$

Thus $\langle w, x\rangle$ is subgaussian if v is subgaussian. Moreover, in this work we assume that $\lambda_{\max}(\Sigma)$ is uniformly (in n, p) bounded above, which implies that up to an absolute constant x and v have the same subgaussian constant.

With these definitions, Proposition 9 and Theorems 1, 5 can be extended as follows.

Proposition 14 *Let $\hat{\beta}$ be the penalized lasso estimator defined in (11). Suppose that model (1) holds with $x \overset{iid}{\sim} \mathcal{E}(0, \Sigma, F_v)$ and assume that*

1. *v is subgaussian with $\|v\|_{\psi_2} \leq \sigma_x$*
2. *$z_i = y_i - \langle x_i, \beta\rangle$ is subgaussian with $\|z_i\|_{\psi_2} \leq \sigma_z$ for all $1 \leq i \leq n$,*
3. *β is a s-sparse vector, i.e. $|\{j : \beta_j \neq 0\}| \leq s$,*
4. *$\Sigma^{\frac{1}{2}}$ satisfies the restricted eigenvalue condition with parameters $(s, 9, 2\alpha)$ for some $\alpha > 0$, and that α and $\lambda_{\max}(\Sigma)$ are bounded away from $0, \infty$.*

Then there exists an absolute constant $c_0 > 0$ such that for $\lambda > c_0\sigma_z\sigma_x\sqrt{\log(p)/n}$ and $n \geq c_0(1 \vee \sigma_x^4)s\log(p/s)$ we have

$$\|\hat{\beta} - \beta\|_1 \leq \frac{16s\lambda}{\alpha^2} \quad \text{and} \quad \frac{\|X(\hat{\beta} - \beta)\|_2}{\sqrt{n}} \leq \frac{4\sqrt{s}\lambda}{\alpha},$$

with probability no less than $1 - 2p^{-1} - \exp(-c_0 n^2/\sigma_x^4)$.

Theorem 15 *Let $\tilde{\beta}_1$ be the estimator defined by (7). Suppose $(x_i, y_i)_{i=1}^n$ follow the model (1) with $x_i \overset{iid}{\sim} \mathcal{E}(0, \Sigma, F_\mu)$ and let*

$$\nu^2 = \frac{\mathbf{E}r_n^2 z_n^2}{(\mathbf{E}r_n^2)^2}.$$

Assume also that the following conditions are satisfied:

1. $\|x_i\|_{\psi_2} \leq \sigma_x$ *with* σ_x *uniformly (over* n*) bounded above.*
2. $\mathbf{E}r_n^2 z_n^2$ *is bounded away from zero.*
3. $\hat{\beta}$ *is an estimate of* β *satisfying*

$$\sqrt{\log(p)}\|\hat{\beta} - \beta\|_1 \to_p 0.$$

4. *There exist* $0 < c, C < \infty$ *such that* $c \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C$.
5. $\mathbf{E}|y_n|^{2+\alpha} \leq M < \infty$ *for some* $\alpha, M > 0$ *and all* $n \geq 1$.

Then

$$\frac{\sqrt{n}(\tilde{\beta}_1 - \beta_1)}{\nu} \to_d N(0, 1).$$

Theorem 16 *(Unknown Σ) Let $\tilde{\beta}_1$ be the estimator defined in (9). Suppose $(x_i, y_i)_{i=1}^{2n}$ follow the model (1) with $x_i \overset{iid}{\sim} \mathcal{E}(0, \Sigma, F_\mu)$ and let*

$$\begin{aligned}
\gamma &:= (\mathbf{E}x_{n,-1}x_{n,-1}^T)^{-1}\mathbf{E}x_{n,-1}x_{n,1}, \\
r_i &:= x_{i,1} - \langle \gamma, x_{i,-1} \rangle, \\
\nu^2 &:= \frac{\mathbf{E}r_n^2 z_n^2}{(\mathbf{E}r_n^2)^2}.
\end{aligned}$$

Assume also that the following conditions are satisfied:

1. $\|x_i\|_{\psi_2} \leq \sigma_x$ *with* σ_x *uniformly (over* n*) bounded above.*
2. $\mathbf{E}r_n^2 z_n^2$ *is bounded away from zero.*
3. $\hat{\beta}$ *is an estimate of* β *satisfying*

$$s\sqrt{\log(p)}\|\hat{\beta} - \beta\|_1 \to_p 0.$$

4. *There exists an estimate* $\hat{\gamma}$ *of* γ *that is independent of* $(x_i, y_i)_{i=1}^n$ *and satisfies*

$$\mathbf{P}(\|\hat{\gamma} - \gamma\|_1 \leq c_\gamma s\sqrt{\frac{\log p}{n}}) \to 1,$$

   for a constant c_γ *not dependent on* n.

5. *There exist* $0 < c, C < \infty$ *such that* $c \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C$.

6. $\mathbf{E}|y_n|^{2+\alpha} \leq M < \infty$ *for some* $\alpha, M > 0$ *and all* $n \geq 1$.

Then the estimator $\tilde{\beta}_1$ *defined by*

$$\frac{\sqrt{n}(\tilde{\beta}_1 - \beta_1)}{\nu} \to_d N(0,1).$$

Remark 17 *The remarks that followed Proposition 9 and Theorems 1, 5 remain valid in the more general context of Proposition 14 and Theorems 15 and 16. For the sake of brevity they are not restated here.*

4. Towards More Efficient Inference

In previous sections, a linear approximation of the link function was used to obtain estimates of β_1. While this approach avoids the estimation of the link function, the variance of the resulting estimator, $\nu^2 = \mathbf{E}r_n^2(y_n - \langle x_n, \beta\rangle)^2/(\mathbf{E}r_n^2)^2$, depends heavily on the quality of this linear approximation.

In this section we show how, under smoothness assumptions on the link function, we can go beyond a linear approximation and obtain more efficient estimators of β_1. To this end, we use an expansion of the link function in terms of Hermite polynomials, as the latter form an orthonormal basis of the Hilbert space $L^2(\mathbf{R}, N(0,1))$ and are thus particularly useful in our setting.

Let us write $\mathbf{E}(y_i|x_i = x) = g(\langle x, \tau\rangle)$ and $e_i = y_i - g(\langle x_i, \tau\rangle)$, so that $\mathbf{E}(e_i|x_i) = 0$. Assume that g can be expanded as $g(\xi) = \sum_{j=0}^{\infty} \mu_j h_j(\xi)$, where h_j is the normalized Hermite polynomial of j-th degree:

$$h_j(\xi) = \frac{(-1)^j}{\sqrt{j!}} e^{\frac{\xi^2}{2}} \frac{d^j}{d\xi^j} e^{-\frac{\xi^2}{2}}. \tag{12}$$

In order to simplify notation, assume that we have a sample $(x_i, y_i)_{i=1}^{2n}$ of size $2n$. For a given m, we compute $\{\hat{\mu}_j\}_{j=1}^m$ and $\hat{\tau}$ on $\mathcal{S}_2 = (x_i, y_i)_{i=n+1}^{2n}$ as follows:

1. Compute a pilot estimate $\hat{\beta}$ of β using only $\mathcal{S}_{21} = (x_i, y_i)_{i=n+1}^{n+\lfloor n/2\rfloor}$.

2. Define $\hat{\mu}_1 := \|\Sigma^{\frac{1}{2}}\hat{\beta}\|_2$ and $\hat{\tau} := \hat{\mu}_1^{-1}\hat{\beta}$.

3. For $0 \leq j \leq m$ and $j \neq 1$ define

$$\hat{\mu}_j := \frac{1}{\lceil n/2\rceil} \sum_{i=n+\lfloor n/2\rfloor+1}^{2n} y_i h_j(\langle x_i, \hat{\tau}\rangle).$$

4. Finally, define the de-biased estimate of β_1 by

$$\tilde{\beta}_1 = \hat{\beta}_1 + \frac{\sum_{i=1}^n r_i(y_i - \sum_{j=0}^m \hat{\mu}_j h_j(\langle x_i, \hat{\tau}\rangle))}{\sum_{i=1}^n r_i x_{i1}}. \tag{13}$$

Theorem 18 *Suppose that $(x_i, y_i)_{i=1}^{2n}$ are i.i.d. observations from the model*

$$y_i = g(\langle \tau, x_i \rangle) + e_i, \quad x_i \sim N_p(0, \Sigma), \quad \mathbf{E}[e_i | x_i] = 0.$$

Let $m = \lfloor \log^{\frac{2}{3}}(n) \rfloor$ and suppose that $\hat{\beta}, \hat{\tau}, \{\hat{\mu}_j\}_{j=0}^m$ are computed as in the above procedure. Assume also that the following conditions are satisfied:

1. *There exist $0 < c, C < \infty$ such that $c \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C$.*
2. *There exists $c' > 0$ such that $\mu_1 = \mathbf{E} y_n \langle x_n, \tau \rangle > c'$ for all $n \geq 1$.*
3. *$\mathbf{E} r_n^2 e_n^2$ is bounded away from zero.*
4. *y_n has a finite fourth moment: $\mathbf{E} y_n^4 < C_y^4$.*
5. *The link function g is differentiable with*

$$\|g'\|_{L_2}^2 = \mathop{\mathbf{E}}_{\xi \sim N(0,1)} |g'(\xi)|^2 < L^2 < \infty,$$

for a constant L not depending on n.

6. *The pilot estimator $\hat{\beta}$ satisfies*

$$\sqrt{\log(p)} \|\hat{\beta} - \beta\|_1 \to_p 0.$$

7. *$\mathbf{E}|e_n|^{2+\alpha} < M < \infty$ for some $\alpha, M > 0$ and all $n \geq 1$.*

Then we have

$$\frac{\sqrt{n}(\tilde{\beta}_1 - \beta_1)}{(\sqrt{\mathbf{E} r_n^2 e_n^2} / \mathbf{E} r_n^2)} \to_d N(0, 1).$$

Remark 19 *1. Inspecting the proof of Theorem (18) shows that the argument in remark (7) applies here as well, so that changing the role of the two subsamples $\mathcal{S}_1$ and $\mathcal{S}_2$ leads to efficient use of the full sample.*

2. From the proof it can be seen that if $\mathbf{E} r_n^2 e_n^2 \to 0$, then

$$\sqrt{n}(\tilde{\beta}_1 - \beta_1) \to_p 0,$$

showing that in this case $\hat{\beta}$ has a rate of convergence faster than $\sqrt{n}$. Assumption (3) rules out such degenerate asymptotic distributions and is readily satisfied for interesting statistical models.

Remark 20 (Efficiency) *1. Using $y_n = g(\langle x_n, \tau \rangle) + e_n$ and $\mathbf{E}[e_n | x_n] = 0$ we can write*

$$\begin{aligned} \mathbf{E} r_n^2 z_n^2 \equiv \mathbf{E} r_n^2 (y_n - \langle x_n, \beta \rangle)^2 &= \mathbf{E} r_n^2 (g(\langle x_n, \tau \rangle) - \langle x_n, \beta \rangle)^2 + \mathbf{E} r_n^2 e_n^2 \\ &\geq \mathbf{E} r_n^2 e_n^2, \end{aligned}$$

showing that the estimator in Theorem 18 is indeed more efficient than the one in Theorem 1.

2. Assume that $e_i \sim N(0, \sigma^2)$ and that $e_i \perp\!\!\!\perp x_i$. Then the asymptotic variance of our (more efficient) estimator reduces to $\sigma^2/\mathbf{E}r_n^2$, which is precisely the (asymptotic) variance of the OLS estimator (or in the high-dimensional case, the debiased lasso estimator) applied to linear observations $y_i = \langle x_i, \beta \rangle + e_i$, and has certain optimality properties(van de Geer et al., 2014, Section 2.3.3.) for a discussion of optimality and semiparametric efficiency of the debiased lasso in high dimensional linear models. We do not, however, expect our estimator to be semiparametrically efficient, as we are not *in a linear model setting, and the construction of such estimators, while interesting, is beyond the scope of this work.*

Remark 21 *In Theorem (18) we let the number of basis functions $\{h_j\}_{j=0}^m$ grow with n at the rate $m = \lfloor \log^{\frac{2}{3}}(n) \rfloor$. The theorem would still be valid for slower rates of growth of m, as long as $m \to \infty$. The slow growth rate of m is used to control the variance of $\hat{\mu}_j$'s in the proof of the theorem.*

5. Conclusions

We have shown that in the generic single-index model, it is easy to obtain $\sqrt{n}$-consistent estimators of finite-dimensional components of β in the high-dimensional setting using a procedure that is perfectly agnostic to the link function, provided we have a Gaussian (or more generally, elliptically symmetric) design. Even though this rate can be achieved under minimal assumptions on the link function, we also showed that using an estimate of the link function to refine the debiased estimator enhances efficiency. Some words of caveat are in order. First, the the independence of f and x is critical to our development. Indeed, if f depends upon x. there is no guarantee that one can estimate individual co-efficients at $\sqrt{n}$ rate. As an example, consider the binary choice model $\Delta = 1(\beta^T X + \epsilon > 0)$ where ϵ given X depends non-trivially on X (Manski, 1975, 1985). The recent results of Mukherjee et al. (2019, see Theorem 3.4) imply that when p is fixed, the co-efficients of β can be estimated at a rate no faster than $n^{1/3}$, with the maximum score estimator of Manski attaining this rate (Kim et al., 1990). It is clear that we should not expect the de-biasing approach of our paper to work in this model. Second, if f is independent of X and discontinuous, e.g. $f(t) = 1(t > 0) + \epsilon$ where $\epsilon > 0$, this becomes a multi-dimensional change-point problem (a change-plane problem to be precise) and the work of Wei and Kosorok (2018) implies that the co-efficients are estimable even at rate n (for the fixed p-case), and the $\sqrt{n}$ rate derived in this paper is sub-optimal. These two examples serve to illustrate the fact that while the de-biased agnostic scheme is attractive, it can fail under model-misspecification, and may not produce optimal convergence rates in certain cases.
Of course the $\sqrt{n}$ will be typically optimal when f is sufficiently smooth, e.g. $f(t) = P(t) + \epsilon$ where ϵ is independent of X and P is a polynomial of fixed degree. In this case, the debiased estimator has to be rate-optimal since even if we knew P there is no way we can estimate the co-efficients at a rate faster than $\sqrt{n}$.

Acknowledgments

This work was done under the auspices of NSF Grant DMS-1712962.

Appendix A. (Tail bounds)

We collect in this appendix some facts about sub-Gaussian and sub-exponential random variables. Proofs can be found in the book by Vershynin (2018, chapter 2).

Denote by $\|\cdot\|_{\psi_2}$ and $\|\cdot\|_{\psi_1}$ the sub-gaussian and the sub-exponential norms, respectively.

Proposition 22 *There exists an absolute constant $C > 0$ such that the following are true:*

1. *If $X \sim N(0, \sigma^2)$, then $\|X\|_{\psi_2} \leq C\sigma$.*
2. *If X is sub-guassian, then X^2 is sub-exponential and $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$.*
3. *If X, Y are sub-gaussian, then XY is sub-exponential and $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \cdot \|Y\|_{\psi_2}$.*
4. *If X is sub-exponential then $\|X - \mathbf{E}X\|_{\psi_1} \leq C\|X\|_{\psi_1}$.*
5. *(Bernstein's Inequality). Let $x_1, \ldots, x_n$ be independent, mean zero sub-exponential random variables and $a = (a_1, \ldots, a_n) \in \mathbf{R}^n$. Then for every $t > 0$ we have*

$$\mathbf{P}\{|\sum_{i=1} a_i x_i| \geq t\} \leq 2\exp\left[-c \min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right],$$

where $K = \max_i \|x_i\|_{\psi_1}$ and c is an absolute constant.

The following corollary will be used multiple times in the text.

Corollary 23 *Suppose that $x_i \in \mathbf{R}^p$ are i.i.d. random vectors with $\|x_{ij}\|_{\psi_1} \leq \rho_x$ for $1 \leq i \leq n$ and $1 \leq j \leq p$. Assume also that $\log(p)/n \to 0$. Then*

1. *for any $1 \leq i \leq n$ and $1 \leq j \leq p$, the variable $x_{ij} - \mathbf{E}x_{ij}$ is sub-exponential with $\|x_{ij} - \mathbf{E}x_{ij}\|_{\psi_1} \leq C\rho_x$, for some absolute constant C.*
2. *We have $\mathbf{P}(\max_{1\leq j\leq p} |\frac{1}{\sqrt{n}}\sum_{i=1}^n x_{ij} - \mathbf{E}x_{ij}| < C\rho_x\sqrt{\log p}) \to 1$ for an absolute constant $C > 0$*

Proof 1. This follows immediately from proposition (22).

2. Apply Bernstein's inequality with $a_i = 1/\sqrt{n}$ and $t = \kappa\rho_x\sqrt{\log p}$ for a constant κ that will be determined shortly. We obtain for each $1 \leq j \leq p$,

$$\mathbf{P}\left\{\frac{1}{\sqrt{n}}\left|\sum_{i=1}^n x_{ij} - \mathbf{E}x_{ij}\right| \geq \kappa\rho_x\sqrt{\log p}\right\} \leq 2e^{-c\min\left(\frac{\kappa^2\rho_x^2\log p}{C^2\rho_x^2}, \frac{\kappa\rho_x\sqrt{\log p}\sqrt{n}}{C\rho_x}\right)}.$$

In order to get sub-gaussian tail in Bernstein's inequality, we need

$$\frac{\kappa^2\log p}{C^2} \leq \frac{\kappa\sqrt{n}\sqrt{\log p}}{C},$$

which is equivalent to $\sqrt{\log p}/\sqrt{n} \leq C/\kappa$, and holds for large enough n (and any fixed value of κ) as in our asymptotic regime, $\sqrt{\log p}/\sqrt{n} \to 0$. For such large n, p, apply a union bound to the above inequality to get

$$\mathbf{P}\{\frac{1}{\sqrt{n}}\max_{1\le j\le p}|\sum_{i=1}^{n}x_{ij}-\mathbf{E}x_{ij}|\ge\kappa\rho_x\sqrt{\log p}\}\le 2p\exp\left[-c\left(\frac{\kappa^2\log p}{C^2}\right)\right]$$
$$=2p\cdot p^{-\frac{c\kappa^2}{C^2}}$$
$$=2p^{1-\frac{c\kappa^2}{C^2}}\to 0,$$

as long as $\kappa^2 > C^2/c$, where C, c are absolute constants.

■

Appendix B.

B.1 Orthogonality of X and z

Let $\xi = \langle x_i, \tau\rangle \sim N(0,1)$. Gaussianity of x_i implies that $\mathbf{E}[x \mid \xi] = \xi b$ for some (non-random) $b \in \mathbf{R}^p$. Using the definition of z_i and the tower property of conditional expectations,

$$\begin{aligned}
\mathbf{E}z_i x_i &= \mathbf{E}[(f_i(\xi)-\mu\xi)x_i] \\
&= \mathbf{E}[\mathbf{E}[(f_i(\xi)-\mu\xi)x_i \mid \xi, f_i]] \\
&= \mathbf{E}[(f_i(\xi)-\mu\xi)\mathbf{E}[x_i \mid \xi]] && (f_i \perp\!\!\!\perp x_i) \\
&= (\mathbf{E}[f_i(\xi)\xi]-\mu\mathbf{E}\xi^2)b^T \\
&= (\mu-\mu\cdot 1)b \\
&= 0.
\end{aligned}$$

B.2 Average partial effect interpretation

Write $\mathbf{E}(y|X=x) = g(\langle x,\tau\rangle)$ and assume that g is differentiable. The average partial effect with respect to the j-th covariate is then defined as:

$$\mathbf{E}_x\left(\frac{\partial}{\partial x_j}\mathbf{E}(y|x)\right) = \tau_j\mathbf{E}_x g'(\langle x,\tau\rangle).$$

By Stein's lemma (Stein, 1981), $\mathbf{E}g'(\langle x,\tau\rangle) = \mathbf{E}\langle x,\tau\rangle g(\langle x,\tau\rangle)$, assuming both expectations exist and using the fact that $\langle x,\tau\rangle \sim N(0,1)$. Thus we have

$$\begin{aligned}
\mathbf{E}_x\left(\frac{\partial}{\partial x_j}\mathbf{E}(y|x)\right) &= \tau_j\mathbf{E}_x\left(\langle x,\tau\rangle\mathbf{E}(y|x)\right) \\
&= \tau_j\mathbf{E}y\langle x,\tau\rangle \\
&= \mu\tau_j = \beta_j.
\end{aligned}$$

B.3 Identifiability

In this appendix we discuss the identifiability of the parameters the model (1). Even though the parameter τ is not identifiable (since its norm can be absorbed into the link function

f), the parameter $\beta = \mu\tau$ is in fact identifiable when Σ is non-singular. This follows since as shown before, X and z are uncorrelated, and so

$$\begin{aligned}\mathbf{E}\|Y - X\beta'\|_2^2 &= \mathbf{E}\|X\beta + z - X\beta'\|_2^2 \\ &= \mathbf{E}\|X(\beta - \beta')\|_2^2 + \mathbf{E}\|z\|_2^2 \\ &= \|\Sigma^{\frac{1}{2}}(\beta - \beta')\|_2^2 + \mathbf{E}\|z\|_2^2.\end{aligned}$$

Thus when Σ is non-singular, β is the unique minimizer of $\mathbf{E}\|Y - X\beta'\|_2^2$. Since the latter only depends on the distribution of (X, Y), identifiability follows.

Remark 24 *The proofs of Proposition 9 and Theorems 1, 5 only use the facts that x_i is a subgaussian vector (with a uniformly bounded subgaussian constant) and that $\mathbf{E}[x_i \mid \langle x_i, \tau\rangle]$ is linear in $\langle x_i, \tau\rangle$. It is well-known and easy to verify that both of these conditions are satisfied for Gaussian vectors $x_i \sim N(0, \Sigma)$ when the extreme eigenvalues of Σ are uniformly bounded away from zero and ∞. In particular, for any unit vector $u \in \mathbf{R}^p$ we have $\|\langle u, x_i\rangle\|_{\psi_2} \leq C\lambda_{\max}(\Sigma) < \infty$. The validity of these properties for elliptically symmetric random vectors has been discussed in section 3.*

B.4 Proof of proposition 9

Lemma 25 *Suppose that $X \in \mathbf{R}^{n\times p}$ is a random matrix with rows x_i^T that are iid samples from $\mathcal{E}(0, \Sigma, F_v)$ with $\|v\|_{\psi_2} \leq \sigma_x$ for $i = 1, \ldots, n$. Assume that $\Sigma^{\frac{1}{2}}$ satisfies the RE condition with parameters $(s, 9, 2\alpha)$ and that $0 < c \leq \alpha, \lambda_{\max}(\Sigma) \leq C < \infty$ for some c, C not depending on n, p. Then, as long as $n \geq c'\sigma_x^4 s\log(p/s)$, there exist constants $c', C' > 0$ not depending on n such that with probability at least $1 - \exp(-c'n^2/\sigma_x^4)$ the matrix X satisfies the RE condition with parameters $(s, 3, \alpha)$.*

Proof By elliptical symmetry, x_i^T can be decomposed as

$$x_i^T = v_i u_i^T \Sigma^{\frac{1}{2}},$$

where $v_i \perp\!\!\!\perp u_i$, the random vector u_i is uniformly distributed on the sphere and $\mathbf{E}v_i^2 = p$. It is easy to verify that for any unit vector $a \in \mathbb{S}^{p-1}$ we have

- $\mathbf{E}\langle a, v_i u_i\rangle^2 = 1$, and,
- $\|\langle a, v_i u_i\rangle\|_{\psi_2} \leq \sigma_x$.

In other words, the random vector $v_i u_i$ is isotropic and subgaussian with constant σ_x. Since $\Sigma^{\frac{1}{2}}$ satisfies the RE condition with parameters $(s, 9, 2\alpha)$ and $\lambda_{\max}(\Sigma) \leq C$, it follows from Theorem 6 in Rudelson and Zhou (2012) that for some constants c', C' depending only on c, C and all $n \geq c'\sigma_x^4 s\log(p/s)$, the matrix $X/\sqrt{n}$ satisfies the restricted eigenvalue condition with parameters $(s, 3, \alpha)$ with probability at least $1 - \exp(-c'n^2/\sigma_x^4)$. ∎

Proof [Proposition 9] Using the definition of $\hat\beta$ we can write

$$\begin{aligned}\frac{1}{2n}\|Y - X\hat\beta\|_2^2 + \lambda\|\hat\beta\|_1 &\leq \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \\ &= \frac{1}{2n}\|z\|_2^2 + \lambda\|\beta\|_1.\end{aligned}$$

Expand $\|Y - X\hat{\beta}\|_2^2$ in the above inequality and rearrange to get

$$\frac{1}{n}\|X(\hat{\beta}-\beta)\|_2^2 \le \frac{2}{n} z^T X(\hat{\beta}-\beta) + 2\lambda(\|\beta\|_1 - \|\hat{\beta}\|_1). \tag{14}$$

Using Hölder's inequality we can write

$$|z^T X(\hat{\beta}-\beta)| \le \|z^T X\|_\infty \cdot \|\hat{\beta}-\beta\|_1.$$

For any i, j, $z_i X_{ij}$ is a subexponential random variable with

$$\|z_i X_{ij}\|_{\psi_1} \le \|z_i\|_{\psi_2} \cdot \|X_{ij}\|_{\psi_2} \le \sigma_x \sigma_z$$

So for each $1 \le j \le p$, the variable $(z^T X)_j$ is a sum of iid, mean-zero sub-exponential random variables, and thus Bernstein's inequality implies

$$\mathbf{P}\{|\sum_{i=1}^n z_i X_{ij}| \ge t\} \le 2\exp\left[-c\min\left(\frac{t^2}{n\sigma_x^2\sigma_z^2}, \frac{t}{\sigma_x\sigma_z}\right)\right],$$

for an absolute constant $c > 0$. As long as $t \le n\sigma_x\sigma_z$, the subgaussian tail bound prevails. For such t and using a union bound, we get

$$\mathbf{P}\{\|z^T X\|_\infty \ge t\} \le 2p\exp\left(-c\frac{t^2}{n\sigma_x^2\sigma_z^2}\right).$$

Setting $t = \sigma_x\sigma_z\sqrt{2c^{-1}n\log p}$, the last inequality reads

$$\mathbf{P}\{\|z^T X\|_\infty \ge t\} \le 2p \cdot p^{-2} = 2p^{-1},$$

As long as $\log(p)/n \le c/2$. Choosing $\lambda \ge 2\sigma_x\sigma_z\sqrt{2c^{-1}\log(p)/n}$, we have shown that the event

$$\mathcal{T}_1 := \left[\frac{2}{n}\|z^T X\|_\infty \le \lambda\right]$$

has probability no less than $1 - 2p^{-1}$. On this event $\mathcal{T}_1$ we can continue with inequality (14) to get

$$\frac{1}{n}\|X(\hat{\beta}-\beta)\|_2^2 \le \lambda\|\hat{\beta}-\beta\|_1 + 2\lambda(\|\beta\|_1 - \|\hat{\beta}\|_1).$$

Adding $\lambda\|\hat{\beta}-\beta\|_1$ to both sides yields

$$\frac{1}{n}\|X(\hat{\beta}-\beta)\|_2^2 + \lambda\|\hat{\beta}-\beta\|_1 \le 2\lambda\left(\|\hat{\beta}-\beta\|_1 + \|\beta\|_1 - \|\hat{\beta}\|_1\right) \tag{15}$$

$$\le 4\lambda\|\hat{\beta}_S - \beta\|_1, \tag{16}$$

where S is the support of β. It follows from the last inequality that

$$\hat{\beta}-\beta \in \mathcal{C}(S,3) = \{\delta \in \mathbf{R}^p : \|\delta_{S^c}\|_1 \le 3\|\delta_S\|_1\}.$$

Let $\mathcal{T}_2$ be the event that $X/\sqrt{n}$ satisfies the RE condition with parameters $(s, 3, \alpha)$. By Lemma 25, there exist constants $c', C' > 0$ such that for all $n \geq C'\sigma_x^4 s \log(p/s)$ we have $\mathbf{P}(\mathcal{T}_2) \geq 1 - \exp(-c'n^2/\sigma_x^4)$. Using the Cauchy-Schwartz inequality and the RE condition on $T_1 \cap T_2$,

$$\begin{aligned}\|\hat{\beta}_S - \beta\|_1 &\leq \sqrt{s}\|\hat{\beta}_S - \beta\|_2 \\ &\leq \frac{\sqrt{s}}{\alpha}\frac{\|X(\hat{\beta} - \beta)\|_2}{\sqrt{n}}.\end{aligned}$$

Comparing this bound with (16) and some algebra gives

$$\frac{1}{\sqrt{n}}\|X(\hat{\beta} - \beta)\|_2 \leq \frac{4\lambda\sqrt{s}}{\alpha} \quad \text{and} \quad \|\hat{\beta} - \beta\|_1 \leq \frac{16s\lambda}{\alpha^2}.$$

Note that $\mathbf{P}(\mathcal{T}_1 \cap \mathcal{T}_2) \geq 1 - 2p^{-2} - \exp(-c_0 n^2/\sigma_x^4)$ as long as $n \geq c_0(1 \vee \sigma_x^4)s\log(p/s)$, where $c_0 := c' \vee C' \vee (2/c) \vee 2\sqrt{2/c}$. ■

Lemma 26 *Suppose that x_n is a subgaussian vector with variance proxy σ_x. The projection r_n of $x_{n,1}$ on the ortho-complement of the span of $x_{n,-1}$ satisfies*

1. $\lambda_{\min}(\Sigma) \leq \mathbf{E}r_n^2 \leq \lambda_{\max}(\Sigma)$
2. $\|r_n\|_{\psi_2} \leq \frac{\mathbf{E}r_n^2}{\lambda_{\min}(\Sigma)} \cdot \sigma_x \leq \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \cdot \sigma_x$

Proof 1. We have

$$\mathbf{E}r_n^2 \leq \mathbf{E}x_{n,1}^2 = \Sigma_{11} \leq \lambda_{\max}(\Sigma),$$

which proves the upper bound. For the lower bound, note that by definition, $r_n = (1, -\gamma^T)x_n$, and that $\mathbf{E}r_n x_n^T = (\mathbf{E}r_n^2)\mathbf{e}_1^T$, where $\mathbf{e}_k$ is that k-th standard basis vector in $\mathbf{R}^p$. From the last equality and $\Sigma = \mathbf{E}x_n x_n^T$, we get

$$\begin{aligned}\mathbf{E}r_n^2 &= \|(\mathbf{E}r_n^2)\mathbf{e}_1^T\|_2 \\ &= \|(1, -\gamma^T) \cdot \Sigma\|_2 \\ &\geq \|(1, -\gamma^T)\|_2 \cdot \lambda_{\min}(\Sigma) \qquad (\star) \\ &\geq \lambda_{\min}(\Sigma),\end{aligned}$$

proving the lower bound.

2. From the definition of subgaussian vectors and the inequality $(\star)$ in the proof of part (1) we have

$$\begin{aligned}\|r_n\|_{\psi_2} &= \|(1, -\gamma^T)\|_2 \cdot \left\|\frac{(1, -\gamma^T)x_n}{\|(1, -\gamma^T)\|_2}\right\|_{\psi_2} \\ &\leq \|(1, -\gamma^T)\|_2 \cdot \sigma_x \\ &\leq \frac{\mathbf{E}r_n^2}{\lambda_{\min}(\Sigma)}\sigma_x,\end{aligned}$$

where we used $(\star)$ in the last inequality. Using the upper bound $\mathbf{E}r_n^2 \leq \lambda_{\max}(\Sigma)$ obtained in part (1) completes the proof. ■

B.5 Proof of Theorem (1)

Proof Without loss of generality, assume that $k = 1$. Use the representation $y_i = \langle x_i, \beta\rangle + z_i$ to rewrite $\tilde{\beta}_1$ as

$$\begin{aligned}\tilde{\beta}_1 &= \hat{\beta}_1 + \frac{\sum_{i=1}^n r_i z_i + r_i\langle x_i, \beta - \hat{\beta}\rangle}{\sum_{i=1}^n r_i x_{i,1}} \\ &= \beta_1 + \frac{\sum_{i=1}^n r_i z_i + r_i\langle x_{i,-1}, \beta_{-1} - \hat{\beta}_{-1}\rangle}{\sum_{i=1}^n r_i x_{i,1}}\end{aligned}$$

Subtracting β_1 from both sides and multiplying by $\sqrt{n}$ yields

$$\sqrt{n}(\tilde{\beta}_1 - \beta_1) = \frac{\overbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n r_i z_i}^{B} + \overbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n r_i\langle x_{i,-1}, \beta_{-1} - \hat{\beta}_{-1}\rangle}^{C}}{\underbrace{\frac{1}{n}\sum_{i=1}^n r_i x_{i,1}}_{A}}.$$

We start by showing $A/\mathbf{E}r_n^2 \to_p 1$. That $\mathbf{E}A/\mathbf{E}r_n^2 = 1$ follows from the definition of r_n. By the second part of Lemma 26, $\|r_i/\mathbf{E}r_n^2\|_{\psi_2} \le 1/\lambda_{\min}(\Sigma)$. Also, we have by assumption that x_n is subgaussian with variance proxy σ_x, implying that $\|x_{i,1}\|_{\psi_2} \le \sigma_x$. Thus $(r_i x_{i,1})/\mathbf{E}r_n^2$ is subexponential with

$$\begin{aligned}\|\frac{r_i x_{i,1}}{\mathbf{E}r_n^2}\|_{\psi_1} &\le \|\frac{r_i}{\mathbf{E}r_n^2}\|_{\psi_2} \cdot \|x_{i,1}\|_{\psi_2} \\ &\le \frac{\sigma_x}{\lambda_{\min}(\Sigma)},\end{aligned}$$

which is uniformly (in n) bounded above by assumption. Bernstein's inequality now implies that $\mathbf{E}A/\mathbf{E}r_n^2 \to_p 1$.

Next, we consider bounding C. By Hölder's inequality,

$$\left|\frac{C}{\mathbf{E}r_n^2}\right| \le \max_{j\neq 1}\frac{1}{\sqrt{n}}\left|\sum_{i=1}^n \frac{r_i x_{i,j}}{\mathbf{E}r_n^2}\right| \cdot \|\beta - \hat{\beta}\|_1.$$

By the construction of r_i, we have $\mathbf{E}r_i x_{i,j} = 0$ for $j \neq 1$. An argument similar to the previous part shows that $r_i x_{i,j}$ is subexponential with $\|r_i x_{i,j}/\mathbf{E}r_n^2\|_{\psi_1} \le \sigma_x/\lambda_{\min}(\Sigma)$. Thus by the second part of Corollary 23, for an absolute constant $c > 0$ we have

$$\mathbf{P}\left\{\max_{j\neq 1}\frac{1}{\sqrt{n}}\left|\sum_{i=1}^n \frac{r_i x_{ij}}{\mathbf{E}r_n^2}\right| \ge c\frac{\sigma_x}{\lambda_{\min}(\Sigma)}\sqrt{\log p}\right\} \to 0.$$

Combining the last two inequalities, we have

$$\mathbf{P}\left\{\frac{|C|}{\mathbf{E}r_n^2} \ge c\frac{\sigma_x}{\lambda_{\min}(\Sigma)}\sqrt{\log p}\|\beta - \hat{\beta}\|_1\right\} \to 0.$$

Noting that by assumption, $\sigma_x/\lambda_{\min}(\Sigma)$ is bounded away from ∞ and $\sqrt{\log p}\|\beta-\hat{\beta}\|_1 \to_p 0$, it follows that

$$\frac{|C|}{\mathbf{E}r_n^2} \to_p 0, \text{ as } n \to \infty.$$

Finally, we show that for $\nu^2 := \mathbf{E}r_n^2 z_n^2$, the term B/ν converges to $N(0,1)$ in distribution. In light of the Lyapunov condition for the central limit theorem, it is sufficient to show that $|r_n z_n|$ has a finite and bounded $(2+\delta)$-th moment for some $\delta > 0$. Let us argue that this follows from $\mathbf{E}|y_n|^{2+\alpha} < M < \infty$.

For $q \geq 1$, denote the L^q norm of random variables by $\|\cdot\|_{L^q} := \sqrt[q]{\mathbf{E}|\cdot|^q}$. Let $q = 2+\alpha$ and use the triangle inequality to write

$$\begin{aligned}\|z_n\|_{L^q} &= \|y_n - \mu\langle x_n, \tau\rangle\|_{L^q} \\ &\leq \|y_n\|_{L^q} + |\mu| \cdot \|\langle x_n, \tau\rangle\|_{L^q}.\end{aligned}$$

By assymption, $\|y_n\|_{L^q} \leq M^{1/q}$. Using the Cauchy-Shwartz inequality,

$$\begin{aligned}|\mu| &= |\mathbf{E}y_n\langle x_n, \tau\rangle| \\ &\leq \sqrt{\mathbf{E}|y_n|^2} \cdot \sqrt{\mathbf{E}\langle x_n, \tau\rangle^2} \\ &\leq \sqrt[q]{M},\end{aligned}$$

where the last inequality uses the normalization of τ and the fact that $\|y_n\|_2 \leq \|y_n\|_q$ for $q = 2+\alpha \geq 2$. Next, note that $\langle x_n, \tau\rangle$ is a subgaussian random variable with $\|\langle x_n, \tau\rangle\|_{\psi_2} \leq \sigma_x \cdot \|\tau\|_2$. By the properties of subgaussian random variables (Vershynin, 2018, Proposition 2.5.2), the L^q norms of subgaussian random variables are bounded by their ψ_2 norms, so we have

$$\|\langle x_n, \tau\rangle\|_{L^q} \leq c\sigma_x \cdot \|\tau\|_2 \cdot \sqrt{q},$$

for an absolute constant $c > 0$. Noting that $\|\Sigma^{1/2}\tau\|_2 = 1$ implies $\|\tau\|_2 \leq 1/\sqrt{\lambda_{\min}(\Sigma)}$, we obtain

$$\|z_n\|_{L^q} \leq M^{1/q}(1 + c\sigma_x\sqrt{q/\lambda_{\min}(\Sigma)}), \tag{17}$$

■

proving that $\|z_n\|_{L^q}$ is bounded (uniformly in n) away from infinity. Lemma 23 shows that $\|r_n\|_{\psi_2} \leq \sigma_x\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$. Using Hölder's inequality and the bound on the moments of subgaussian random variables, for $q' = 2+\alpha/2 < q$,

$$\begin{aligned}\mathbf{E}|r_n z_n|^{q'} &\leq \||z_n|^{q'}\|_{L^{q/q'}} \cdot \||r_n|^{q'}\|_{L^{q/(q-q')}} \\ &\leq \|z_n\|_{L^q}^{q'} \cdot \|r_n\|_{L^{q''}}^{q'} && (q'' = (q-q')/(qq')) \\ &\leq M^{q'/q}\left(1 + c\sigma_x\sqrt{\frac{q}{\lambda_{\min}(\Sigma)}}\right)^{q'} \cdot \left(c\sigma_x\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\sqrt{q''}\right)^{q'} \\ &\leq c'\end{aligned}$$

for some $c' < \infty$ that does not depend on n. It follows that for $q' = 2 + \alpha/2$ the Lyapunov condition is satisfied:

$$\frac{n\mathbf{E}|r_n z_n|^{2+\alpha/2}}{(\sqrt{n\mathbf{E}r_n^2 z_n^2})^{2+\alpha/2}} \leq \frac{c'}{n^{\alpha/4}c_{\mathrm{rz}}} \to 0,$$

and the proof is complete.

B.6 Proof of Theorem (5)

Proof Using $y_i = \langle x_i, \beta\rangle + z_i$ in the the definition of $\tilde{\beta}_1$ and rearranging yields

$$\begin{aligned}
\tilde{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^n \hat{r}_i(y_i - \langle\hat{\beta}_{-1}, x_{i,-1}\rangle)}{\sum_{i=1}^n \hat{r}_i x_{i,1}} \\
&= \beta_1 + \frac{\sum_{i=1}^n \hat{r}_i z_i + \sum_{i=1}^n \hat{r}_i(\langle\beta_{-1} - \hat{\beta}_{-1}, x_{i,-1}\rangle)}{\sum_{i=1}^n \hat{r}_i x_{i,1}} \\
&= \beta_1 + \left(\frac{n\mathbf{E}r_n^2}{\sum_{i=1}^n \hat{r}_i x_{i,1}}\right)\frac{\sum_{i=1}^n r_i z_i + \langle\beta_{-1} - \hat{\beta}_{-1}, \sum_{i=1}^n r_i x_{i,-1}\rangle}{n\mathbf{E}r_n^2} \\
&+ \left(\frac{n\mathbf{E}r_n^2}{\sum_{i=1}^n \hat{r}_i x_{i,1}}\right)\frac{\sum_{i=1}^n (r_i - \hat{r}_i) z_i + \langle\beta_{-1} - \hat{\beta}_{-1}, \sum_{i=1}^n (r_i - \hat{r}_i) x_{i,-1}\rangle}{n\mathbf{E}r_n^2}
\end{aligned}$$

Subtracting β_1 from both sides and multiplying by $\sqrt{n}$,

$$\sqrt{n}(\tilde{\beta}_1 - \beta_1) = A^{-1}(B + C + D)$$

where

$$\begin{aligned}
A &= \frac{1}{n\mathbf{E}r_n^2}\sum_{i=1}^n \hat{r}_i x_{i,1}, \\
B &= \frac{\sum_{i=1}^n r_i z_i + \langle\beta_{-1} - \hat{\beta}_{-1}, \sum_{i=1}^n r_i x_{i,-1}\rangle}{\sqrt{n}\mathbf{E}r_n^2}, \\
C &= \frac{\sum_{i=1}^n (r_i - \hat{r}_i) z_i}{\sqrt{n}\mathbf{E}r_n^2}, \\
D &= \frac{\langle\beta_{-1} - \hat{\beta}_{-1}, \sum_{i=1}^n (r_i - \hat{r}_i) x_{i,-1}\rangle}{\sqrt{n}\mathbf{E}r_n^2}.
\end{aligned}$$

The proof of Theorem 1 shows that

$$\frac{\mathbf{E}r_n^2}{\sqrt{\mathbf{E}r_n^2 z_n^2}} B \to_d N(0,1).$$

The rest of the proof shows that $A \to_p 1$ and $C, D \to_p 0$.

Recalling (from the proof of Theorem 1) that $\sum_1^n r_i x_{i,1}/(n\mathbf{E}r_n^2) \to_p 1$, the term A can be rewritten as

$$\begin{aligned}
A &= \frac{1}{n\mathbf{E}r_n^2}\sum_{i=1}^n r_i x_{ik} + \frac{1}{n\mathbf{E}r_n^2}\sum_{i=1}^n (\hat{r}_i - r_i) x_{i,1} \\
&= 1 + o_p(1) + \langle\gamma - \hat{\gamma}, \frac{1}{n\mathbf{E}r_n^2}\sum_{i=1}^n x_{i,1} x_{i,-1}\rangle.
\end{aligned}$$

Thus using Hölder's inequality,

$$|A-1| \le |\langle \gamma - \hat{\gamma}, \frac{1}{n\mathbf{E}r_n^2}\sum_{i=1}^n z_i x_{i,-1}\rangle| + o_p(1) \tag{18}$$
$$\le \|\gamma-\hat{\gamma}\|_1 \cdot \max_{j\neq 1}\frac{1}{n\mathbf{E}r_n^2}\left|\sum_{i=1}^n x_{i,1}x_{i,j}\right| \tag{19}$$

It follows form the subgaussianity of x_i that $x_{i,1}x_{i,j}$ is subexponential with

$$\|x_{i,1}x_{i,j} - \Sigma_{1,j}\|_{\psi_1} \le c\sigma_x^2$$

Using the triangle inequality and Lemmas 23 and 26,

$$\begin{aligned}\max_{j\neq 1}\frac{1}{n\mathbf{E}r_n^2}\left|\sum_{i=1}^n x_{i,1}x_{i,j}\right| &\le \max_{j\neq 1}\frac{1}{n\mathbf{E}r_n^2}\left|\sum_{i=1}^n x_{i,1}x_{i,j} - \Sigma_{1,j}\right| + \max_j \frac{|\Sigma_{1,j}|}{\mathbf{E}r_n^2}\\ &\le \mathcal{O}_p\left(\frac{\sigma_x^2}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right) + \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)},\end{aligned}$$

which is bounded above (uniformly in n) since $\log(p)/n \to 0$. Consequently, the ℓ^1 consistency of $\hat{\gamma}$ and the bound in inequality (19) implies $A \to_p 1$.

Next, consider C. Use the definition of $r_i, \hat{r}_i$ to rewrite C as

$$C = \frac{\sum_{i=1}^n \langle \hat{\gamma}-\gamma, x_{i,-1}\rangle z_i}{\sqrt{n}\mathbf{E}r_n^2}.$$

Note that $\hat{\gamma}$ is by construction independent of $(x_i, z_i)_{i=1}^n$. Thus, given $\hat{\gamma}$, the latter is a sum of conditionally i.i.d. and mean-zero random variables, since x_i and z_i are uncorrelated. Moreover, since $U := \mathbf{1}(\|\hat{\gamma}-\gamma\|_1 \le c_\gamma s\sqrt{\log(p)/n}) \to_p 1$ by assumption, it suffices to prove that $\tilde{C} := U \cdot C = o_p(1)$. Using the tower property of expectations,

$$\begin{aligned}\mathbf{E}\tilde{C}^2 &= \mathbf{E}[\mathbf{E}[U^2\tilde{C}^2 \mid \hat{\gamma}]]\\ &= \frac{1}{(\mathbf{E}r_n^2)^2}\mathbf{E}[\mathbf{E}[U^2\langle\hat{\gamma}-\gamma, x_{n,-1}\rangle^2 z_n^2 \mid \hat{\gamma}]].\end{aligned}$$

Let $\tilde{q} = q/2 = 1+\alpha/2$ and recall (from the proof of Theorem 1, inequality(17)) that $\|z_n\|_{L^q} \le c'$ for some $c' < \infty$ independent of n. Also, Let us write $\|\cdot\|_{L^q_{\text{con}}} = (\mathbf{E}[|\cdot|^q \mid \hat{\gamma}])^{1/q}$ for the L^q norm with respect to the conditional expectations given $\hat{\gamma}$. By Hölder's inequality, for $\tilde{q}' = (1-1/\tilde{q})^{-1}$ we obtain

$$\begin{aligned}\mathbf{E}[U^2\langle\hat{\gamma}-\gamma, x_{n,-1}\rangle^2 z_n^2 \mid \hat{\gamma}] &\le \|U^2\langle\hat{\gamma}-\gamma, x_{n,-1}\rangle^2\|_{L^{\tilde{q}'}_{\text{con}}} \cdot \|z_n^2\|_{L^{\tilde{q}}_{\text{con}}}\\ &= \|U\langle\hat{\gamma}-\gamma, x_{n,-1}\rangle\|^2_{L^{2q'}_{\text{con}}} \cdot \|z_n\|^2_{L^q}\\ &\le 2c^2 U \cdot \|\hat{\gamma}-\gamma\|_2^2 \sigma_x^2 q' c'^2,\end{aligned}$$

where in the last inequality the subgaussianity of x_n was used to bound the $L^{2q'}$ norm of the (conditionally) subgaussian random variable $U\langle\hat{\gamma}-\gamma, x_{n,-1}\rangle$. Now using the inequality $\|\hat{\gamma}-\gamma\|_2^2 \leq \|\hat{\gamma}-\gamma\|_1^2$, it follows that

$$\begin{aligned}\mathbf{E}\tilde{C}^2 &\leq \frac{1}{\mathbf{E}r_n^2}\mathbf{E}[2c^2c'^2U \cdot \|\hat{\gamma}-\gamma\|_1^2\sigma_x^2 q'] \\ &= \mathcal{O}(s^2\log(p)/n) \\ &= o(1).\end{aligned}$$

This implies $\tilde{C} \to_p 0$, and therefore, $C \to_p 0$.

The proof will be complete if we show $D \to_p 0$. Use the definition of $r_i, \hat{r}_i$ to rewrite D as

$$\begin{aligned}D &= \frac{1}{\sqrt{n}\mathbf{E}r_n^2}\sum_{i=1}^n \sum_{2\leq j,k\leq p} x_{ij}x_{ik}(\hat{\gamma}_j-\gamma_j)(\hat{\beta}_j-\beta_j) \\ &= \frac{1}{\mathbf{E}r_n^2}\sum_{2\leq j,k\leq p}\frac{1}{\sqrt{n}}\sum_{i=1}^n x_{ij}x_{ik}(\hat{\gamma}_j-\gamma_j)(\hat{\beta}_j-\beta_j).\end{aligned}$$

We can therefore upper bound $|D|$ by

$$|D| \leq \frac{1}{\mathbf{E}r_n^2}\max_{1\leq j,k\leq p}\frac{1}{\sqrt{n}}|\sum_{i=1}^n x_{i,j}x_{i,k}| \cdot \|\hat{\beta}-\beta\|_1 \cdot \|\hat{\gamma}-\gamma\|_1.$$

Using the triangle inequality,

$$\max_{1\leq j,k\leq p}\frac{1}{\sqrt{n}}|\sum_{i=1}^n x_{i,j}x_{i,k}| \leq \max_{1\leq j,k\leq p}\frac{1}{\sqrt{n}}|\sum_{i=1}^n x_{i,j}x_{i,k}-\Sigma_{jk}| + \sqrt{n}\max_{j,k}\Sigma_{jk}.$$

An application of Bernstein's inequality followed by a union bound (as in Lemma 23) shows that for absolute constants c_1, c_2,

$$\mathbf{P}\left(\max_{1\leq j,k\leq p}\frac{1}{\sqrt{n}}|\sum_{i=1}^n x_{i,j}x_{i,k}-\Sigma_{jk}| \geq c_1\sigma_x^2\sqrt{\log p}\right) \to 0, \quad \forall n \geq c_2.$$

Since $\Sigma_{jk} \leq \Sigma_{jj} \leq \lambda_{\max}(\Sigma)$, for large enough n we obtain the upper bound

$$\begin{aligned}|D| &\leq \frac{1}{\mathbf{E}r_n^2}(c_1\lambda_{\max}(\Sigma)\sigma_x^2\sqrt{\log p} + \lambda_{\max}(\Sigma)\sqrt{n})\|\hat{\beta}-\beta\|_1 \cdot \|\hat{\gamma}-\gamma\|_1 \\ &\leq c_3 s\sqrt{\log p}\|\hat{\beta}-\beta\|_1 \to_p 0,\end{aligned}$$

by assumptions.

■

B.7 Proof of Theorem (18)

Some well-known properties of Hermite polynomials are collected in the following proposition. Definitions and proofs of the first two statements can be found in (O'Donnell, 2014, §11.2). Statements 2 and 3 are easy to verify from the definition. The last statement is proved in (Larsson-Cohn, 2002, Theorem 2.1).

Proposition 27 *For Hermite polynomials $\{h_j\}_{j=0}^{\infty}$ defined by (12) the following are true:*

1. *$\{h_j\}_{j=0}^{\infty}$ forms an orthonormal basis of $L^2(\mathbf{R}, N(0,1))$.*
2. *For (deterministic) unit vectors $\tau, \hat{\tau} \in \mathbf{R}^p$ and $x \sim N(0, I_p)$ we have*
$$\mathbf{E}[h_j(\langle \tau, x\rangle) h_k(\langle \hat{\tau}, x\rangle)] = \langle \tau, \hat{\tau}\rangle^j \cdot \mathbf{1}[j = k]. \tag{20}$$
3. *For all $j \geq 1$, $h_j' = \sqrt{j} h_{j-1}$.*
4. *For all $j \geq 1$, $\xi h_j(\xi) = \sqrt{j+1} h_{j+1} + \sqrt{j} h_{j-1}(\xi)$*
5. *For $q > 2$ the L^q norms (w.r.t. the Gaussian measure) of Hermite polynomials satisfy*
$$\|h_j\|_{L_q} = \frac{c(q)}{j^{1/4}}(q-1)^{j/2}\left(1 + \mathcal{O}(\frac{1}{j})\right),$$
as $j \to \infty$, where $c(q) = (1/\pi)^{1/4}((q-1)/(2q-4))^{(q-1)/(2q)}$.

The following lemma relates the smoothness of a function $g = \sum_{j=0}^{\infty} \mu_j h_j$ to the decay of the sequence $\{\mu_j\}_j$. The result and its proof are direct analogues of (Tsybakov, 2008, Lemma A.3) which concerns the trigonometric basis.

Lemma 28 *Suppose that $g(\xi) = \sum_{j=0}^{\infty} \mu_j h_j(\xi)$. Assume also that g is k-times continuously differentiable and that*
$$\mathop{\mathbf{E}}_{\xi \sim N(0,1)} |g^{(k)}(\xi)|^2 \leq L^2.$$
Then we have
$$\sum_{j=0}^{\infty} \frac{(j+k)!}{j!} \mu_{j+k}^2 \leq L^2.$$

Proof Let $\mu_j(k) := \mathbf{E} g^{(k)}(\xi) h_j(\xi)$ for $k \geq 1$ and $\mu_j(0) = \mu_j$. Using integration by parts, for $k \geq 1$ we have
$$\begin{aligned}
\mu_j(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} g^{(k)}(\xi) h_j(\xi) e^{-\xi^2/2} d\xi \\
&= \frac{1}{\sqrt{2\pi}} \left[g^{(k-1)}(\xi) h_j(\xi) e^{-\xi^2/2}\right]_{-\infty}^{+\infty} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g^{(k-1)}(\xi)(\sqrt{j} h_{j-1}(\xi) - \xi h_j(\xi)) e^{-\xi^2/2} d\xi \\
&= 0 - \sqrt{j+1} \mathbf{E} g^{(k-1)}(\xi) h_{j+1}(\xi) \\
&= \sqrt{j+1} \mu_{j+1}(k-1)
\end{aligned}$$

where in the second and third equalities we used the parts (3) and (4) of proposition 27. From the recursion $\mu_j(k) = \sqrt{j+1}\mu_{j+1}(k-1)$ it follows that $\mu_j^2(k) = \frac{(j+k)!}{j!}\mu_{j+k}^2$. Using the latter and Parseval's identity.

$$\sum_{j=0}^{\infty} \frac{(j+k)!}{j!}\mu_{j+k}^2 = \sum_{j=0}^{\infty} \mu_j(k)^2 = \mathop{\mathbf{E}}_{\xi\sim N(0,1)} |g^{(k)}(\xi)|^2 \leq L^2.$$

■

Proof of Theorem 18. We can assume without loss of generality that $\mu > 0$, since the parameters $(\tau, g(\cdot))$ and $(-\tau, g(-\cdot))$ give rise to the same data distribution on (x, y), and exactly one of these choices leads to $\mu > 0$ and the other gives $\mu < 0$, assuming $\mu \neq 0$.

First we show that $\hat{\tau}$ has the same rate of convergence as $\hat{\beta}$. To see this, use the triangle inequality to write

$$\begin{aligned}
\|\hat{\tau} - \tau\|_2 &= \left\| \frac{\hat{\beta}}{\hat{\mu}_1} - \frac{\beta}{\mu_1} \right\|_2 \\
&\leq \left\| \frac{\hat{\beta}}{\hat{\mu}_1} - \frac{\hat{\beta}}{\mu_1} \right\|_2 + \left\| \frac{\hat{\beta}}{\mu_1} - \frac{\beta}{\mu_1} \right\|_2 \\
&= \frac{1}{\mu_1}\left(|\hat{\mu}_1 - \mu_1| \cdot \frac{\|\hat{\beta}\|_2}{\hat{\mu}_1} + \|\hat{\beta} - \beta\|_2 \right).
\end{aligned}$$

Since $|\mu| = \mu > 0$, we can now write

$$\begin{aligned}
|\hat{\mu}_1 - \mu| &= \left| \|\Sigma^{\frac{1}{2}}\hat{\beta}\|_2 - \|\Sigma^{\frac{1}{2}}\beta\|_2 \right| \\
&\leq \|\Sigma^{\frac{1}{2}}(\hat{\beta} - \beta)\|_2 \\
&\leq \lambda_{\max}(\Sigma^{\frac{1}{2}}) \cdot \|\hat{\beta} - \beta\|_2.
\end{aligned}$$

On the other hand,

$$\frac{\|\hat{\beta}\|_2}{\hat{\mu}_1} = \frac{\|\hat{\beta}\|_2}{\|\Sigma^{\frac{1}{2}}\hat{\beta}\|_2} \leq \max_{\theta\in\mathbf{R}^p} \frac{\|\theta\|_2}{\|\Sigma^{\frac{1}{2}}\theta\|_2} = \left(\min_{\theta\in\mathbf{R}^p} \frac{\|\Sigma^{\frac{1}{2}}\theta\|_2}{\|\theta\|_2} \right)^{-1} = \lambda_{\min}^{-1}(\Sigma^{\frac{1}{2}}).$$

The last three bounds put together yield

$$\|\hat{\tau} - \tau\|_2 \leq \frac{1 + \kappa(\Sigma)^{\frac{1}{2}}}{\mu_1} \|\hat{\beta} - \beta\|_2,$$

where $\kappa(\Sigma) = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ is the condition number of Σ. By assumptions (2) and (1), the parameters μ_1 and $\kappa(\Sigma)$ are bounded away from zero and infinity, respectively, showing that $\hat{\tau}$ has the same rate of convergence as $\hat{\beta}$.

We can now write

$$
\begin{aligned}
\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} r_i x_{i1})\frac{(\tilde{\beta}_1-\beta_1)}{\sqrt{\mathbf{E}r_n^2e_n^2}} &= \frac{1}{\sqrt{n\mathbf{E}r_n^2e_n^2}}\sum_{i=1}^{n} r_i e_i && \text{(A)}\\
&+ \frac{1}{\sqrt{n\mathbf{E}r_n^2e_n^2}}\sum_{i=1}^{n} r_i(\mu_0-\hat{\mu}_0) && \text{(B)}\\
&+ \frac{1}{\sqrt{n\mathbf{E}r_n^2e_n^2}}\sum_{i=1}^{n} r_i\langle \mu_1\tau_{-1}-\hat{\mu}_1\hat{\tau}_{-1}, x_{i,-1}\rangle && \text{(C)}\\
&+ \frac{1}{\sqrt{n\mathbf{E}r_n^2e_n^2}}\sum_{i=1}^{n} r_i\sum_{j=2}^{m}\mu_j[h_j(\langle\tau,x_i\rangle)-h_j(\langle\hat{\tau},x_i\rangle)] && \text{(D)}\\
&+ \frac{1}{\sqrt{n\mathbf{E}r_n^2e_n^2}}\sum_{i=1}^{n}\sum_{j=2}^{m} r_i(\mu_j-\hat{\mu}_j)h_j(\langle\hat{\tau},x_i\rangle) && \text{(E)}\\
&+ \frac{1}{\sqrt{n\mathbf{E}r_n^2e_n^2}}\sum_{i=1}^{n} r_i\sum_{j=m+1}^{\infty}\mu_j h_j(\langle\tau,x_i\rangle). && \text{(F)}
\end{aligned}
$$

We will show that the first term (A) converges in law to a normal distribution and the other terms (B-F) converge to zero in probability. In order to prove the negligibility in probability of the terms B - F we show that the conditional second moments of these terms are $o_p(1)$, which proves that these terms are themselves $o_p(1)$.

(A). As argued in the proof of Theorem 1, an application of Hölder's inequality shows that $\mathbf{E}|e_n|^{2+\alpha} < M$ implies $\mathbf{E}|r_ne_n|^{2+\alpha}$ is uniformly bounded above. Therefore the Lyapunov condition for the central limit theorem is satisfied and the first term (A) converges to $N(0,1)$ in distribution.

(B). Since $\mathbf{E}[B|\mathcal{S}_2]=0$, and $\mathbf{E}\hat{\mu}_0=\mu_0$, the variance of (B) evaluates to

$$
\begin{aligned}
\mathbf{Var}(B) &= \mathbf{Var}[\mathbf{E}[B|\mathcal{S}_2]] + \mathbf{E}[\mathbf{Var}[B|\mathcal{S}_2]]\\
&= \frac{\mathbf{E}(\mu_0-\hat{\mu}_0)^2\mathbf{E}r_n^2}{\mathbf{E}r_n^2e_n^2}\\
&\le \frac{\rho\mathbf{E}y_n^2}{\lceil n/2\rceil\mathbf{E}r_n^2e_n^2} \to 0.
\end{aligned}
$$

(C). The "linear" term (C) has been handled in the proof of Theorem (1), as by definition $\hat{\mu}_1\hat{\tau}=\hat{\beta}$ and $\mu_1\tau=\beta$.

(D). To simplify notation, for each $1\le i\le n$ let us write $\xi_i=\langle x_i,\tau\rangle$ and $\hat{\xi}_i=\langle x_i,\hat{\tau}\rangle$, and use the (conditional on $\mathcal{S}_2$) Gaussian decomposition $r_i=\alpha\xi_i+\hat{\alpha}\hat{\xi}_i+u_i$, where, given $\mathcal{S}_2$, u_i is independent of both ξ_i and $\hat{\xi}_i$ (such a decomposition is possible because, given $\mathcal{S}_2$, . Then (D) can be written as

$$D = \frac{1}{\sqrt{n\mathbf{E}r_n^2e_n^2}} \sum_{i=1}^{n} \alpha\xi_i \sum_{j=2}^{m} \mu_j[h_j(\xi_i) - h_j(\hat{\xi}_i)] \tag{D_1}$$

$$+ \frac{1}{\sqrt{n\mathbf{E}r_n^2e_n^2}} \sum_{i=1}^{n} \hat{\alpha}\hat{\xi}_i \sum_{j=2}^{m} \mu_j[h_j(\xi_i) - h_j(\hat{\xi}_i)] \tag{D_2}$$

$$+ \frac{1}{\sqrt{n\mathbf{E}r_n^2e_n^2}} \sum_{i=1}^{n} u_i \sum_{j=2}^{m} \mu_j[h_j(\xi_i) - h_j(\hat{\xi}_i)]. \tag{D_3}$$

To show that (D_3) is negligible in probability, note that after conditioning on $\mathcal{S}_2$, the variable u_i is independent of $\xi_i, \hat{\xi}_i$, leading to

$$\mathbf{E}[D_3^2|\mathcal{S}_2] = \frac{\mathbf{E}[u_n^2|\mathcal{S}_2]}{\mathbf{E}r_n^2e_n^2}\mathbf{E}[(\sum_{j=2}^{m} \mu_j[h_j(\xi_n) - h_j(\hat{\xi}_n)])^2|\mathcal{S}_2].$$

We can write $\langle \tau, x_i \rangle = \langle \Sigma^{\frac{1}{2}}\tau, \Sigma^{-\frac{1}{2}}x_i \rangle$, where by assumption, $\|\Sigma^{\frac{1}{2}}\tau\|_2 = 1$ and $\Sigma^{-\frac{1}{2}}x_i \sim N(0, I)$. Also by construction we have $\|\Sigma^{\frac{1}{2}}\hat{\tau}\|_2 = 1$. Thus the second part of proposition (27) applies and for $2 \le j \ne k \le n$ we have

$$\mathbf{E}[(h_j(\xi_n) - h_j(\hat{\xi}_n))(h_k(\xi_n) - h_k(\hat{\xi}_n))|\mathcal{S}_2] = 0.$$

The last two equations imply

$$\mathbf{E}[D_3^2|\mathcal{S}_2] = \frac{\mathbf{E}[u_n^2|\mathcal{S}_2]}{\mathbf{E}r_n^2e_n^2} \sum_{j=2}^{m} \mu_j^2\mathbf{E}[(h_j(\xi_n) - h_j(\hat{\xi}_n))^2|\mathcal{S}_2].$$

Using once again the second part of (27), we get

$$\begin{aligned}
\mathbf{E}[(h_j(\xi_n) - h_j(\hat{\xi}_n))^2|\mathcal{S}_2] &= \mathbf{E}h_j(\xi_n)^2 + \mathbf{E}[h_j(\hat{\xi}_n)^2|\mathcal{S}_2] - 2\mathbf{E}[h_j(\xi_n)h_j(\hat{\xi}_n)|\mathcal{S}_2] \\
&= 2(1 - \langle \Sigma^{\frac{1}{2}}\tau, \Sigma^{\frac{1}{2}}\hat{\tau}\rangle^j) \\
&= 2[1 - (1 - \|\Sigma^{\frac{1}{2}}(\tau - \hat{\tau})\|_2^2/2)^j] \\
&\le j\|\Sigma^{\frac{1}{2}}(\tau - \hat{\tau})\|_2^2,
\end{aligned}$$

where to get the last line we used the inequality $1 - (1 - \theta)^j \le j\theta$ which is valid for $j \ge 1$ and $0 \le \theta \le 2$, with $\theta = \|\Sigma^{\frac{1}{2}}(\tau - \hat{\tau})\|_2^2/2$.

Thus the conditional variance of D_3 is bounded by

$$\mathbf{E}[D_3^2|\mathcal{S}_2] \le \frac{\mathbf{E}[u_n^2|\mathcal{S}_2]\lambda_{\max}(\Sigma)}{\mathbf{E}r_n^2e_n^2}(\sum_{j=2}^{m} j\mu_j^2)\|\tau - \hat{\tau}\|_2^2.$$

By symmetry (of ξ_i and $\hat{\xi}_i$, after conditioning on $\mathcal{S}_2$), the two terms D_1 and D_2 have similar conditional variance, so we only consider D_1. Note that we can not apply the same reasoning as we did in the case of D_3, since ξ_i is no longer independent of $h_j(\xi_i) - h_j(\hat{\xi}_i)$. The conditional variance is

$$\mathbf{E}[D_1^2|\mathcal{S}_2] = \frac{\mathbf{E}[\alpha^2|\mathcal{S}_2]}{\mathbf{E}r_n^2 e_n^2}\mathbf{E}[\xi_n^2(\sum_{j=2}^m \mu_j[h_j(\xi_n) - h_j(\hat{\xi}_n)])^2|\mathcal{S}_2]$$

Let us use the Gaussian decomposition $\hat{\xi}_n = \langle \Sigma\tau, \hat{\tau}\rangle \cdot \xi_n + w_n$ where w_n is independent of ξ_n, given $\mathcal{S}_2$. To evaluate the conditional variance, apply Stein's lemma twice to the function $\psi_w(\xi_n) = \sum_{j=2}^m \mu_j[h_j(\xi_n) - h_j(\langle\tau, \hat{\tau}\rangle\xi_n + w)]$ to get

$$\mathbf{E}[\xi_n^2 \psi_{w_n}(\xi_n)^2|\mathcal{S}_2, w_n] = \mathbf{E}[\psi_{w_n}(\xi_n)^2|\mathcal{S}_2, w_n] + 2\mathbf{E}[\psi'_{w_n}(\xi_n)^2|\mathcal{S}_2, w_n] \tag{21}$$
$$+2\mathbf{E}[\psi_{w_n}(\xi_n)\psi''_{w_n}(\xi_n)|\mathcal{S}_2, w_n]. \tag{22}$$

Using the tower property of conditional expectations and the last identity, we obtain

$$\begin{aligned}
\mathbf{E}[D_1^2|\mathcal{S}_2] &= \mathbf{E}[\mathbf{E}[D_1^2|\mathcal{S}_2, w_n]|\mathcal{S}_2] \\
&= \frac{\mathbf{E}[\alpha^2|\mathcal{S}_2]}{\mathbf{E}r_n^2 e_n^2}(\mathbf{E}[(\sum_{j=2}^m \mu_j[h_j(\xi_n) - h_j(\hat{\xi}_n)])^2|\mathcal{S}_2] \\
&+ 2\mathbf{E}[(\sum_{j=2}^m \mu_j[\sqrt{j}h_{j-1}(\xi_n) - \langle\Sigma\tau, \hat{\tau}\rangle\sqrt{j}h_{j-1}(\hat{\xi}_n)])^2|\mathcal{S}_2] \\
+2\mathbf{E}[(\sum_{j=2}^m \mu_j[h_j(\xi_n) &- h_j(\hat{\xi}_n)])(\sum_{j'=2}^m \mu_{j'}\sqrt{j'(j'-1)}[h_{j'-2}(\xi_n) - \langle\Sigma\tau, \hat{\tau}\rangle^2 h_{j'-2}(\hat{\xi}_n)])|\mathcal{S}_2])
\end{aligned}$$

Using the first and second part of proposition (27), this can be simplified and bounded by

$$\begin{aligned}
\mathbf{E}[D_1^2|\mathcal{S}_2] \leq \frac{\mathbf{E}[\alpha^2|\mathcal{S}_2]}{\mathbf{E}r_n^2 e_n^2}(&\sum_{j=2}^m 2\mu_j^2(1 - \langle\Sigma\tau, \hat{\tau}\rangle^j) + 4j\mu_j^2(1 - \langle\Sigma\tau, \hat{\tau}\rangle^j)) \\
&+ 2\sum_{j=2}^{m-2} \mu_j\mu_{j+2}\sqrt{(j+1)(j+2)}(1 + \langle\Sigma\tau, \hat{\tau}\rangle^2)(1 - \langle\Sigma\tau, \hat{\tau}\rangle^j).
\end{aligned}$$

Using once again the inequality $2(1 - \langle\Sigma\tau, \hat{\tau}\rangle^j) \leq j\|\Sigma^{\frac{1}{2}}(\tau - \hat{\tau})\|_2^2$, we obtain the bound

$$\begin{aligned}
\mathbf{E}[D_1^2|\mathcal{S}_2] &\leq \frac{\mathbf{E}[\alpha^2|\mathcal{S}_2] \cdot \|\Sigma^{\frac{1}{2}}(\tau - \hat{\tau})\|_2^2}{\mathbf{E}r_n^2 e_n^2}(\sum_{j=2}^m j\mu_j^2 + 2j^2\mu_j^2 + 2\sum_{j=2}^{m-2} j\mu_j\mu_{j+2}\sqrt{(j+1)(j+2)}) \\
&\leq \frac{\mathbf{E}[\alpha^2|\mathcal{S}_2] \cdot \lambda_{\max}(\Sigma)\|\tau - \hat{\tau}\|_2^2}{\mathbf{E}r_n^2 e_n^2}(7\sum_{j=2}^m j^2\mu_j^2).
\end{aligned}$$

(E). The same technique can be applied to compute the conditional variance of (E). Use a Gaussian decomposition $r_i = \alpha'\hat{\xi}_i + v_i$ where after conditioning on $\mathcal{S}_2$, the random variable v_i is independent of $\hat{\xi}_i$, to rewrite (E) as

$$E = \frac{1}{\sqrt{n\mathbf{E}r_n^2 e_n^2}} \sum_{i=1}^{n} \sum_{j=2}^{m} \alpha' \hat{\xi}_i (\mu_j - \hat{\mu}_j) h_j(\langle \hat{\tau}, x_i \rangle) \tag{E_1}$$

$$+ \frac{1}{\sqrt{n\mathbf{E}r_n^2 e_n^2}} \sum_{i=1}^{n} \sum_{j=2}^{m} v_i (\mu_j - \hat{\mu}_j) h_j(\langle \hat{\tau}, x_i \rangle). \tag{E_2}$$

The conditional variance of E_2 is

$$\mathbf{E}[E_2^2 | \mathcal{S}_2] = \frac{\mathbf{E}[v_n^2 | \mathcal{S}_2]}{\mathbf{E}r_n^2 e_n^2} \sum_{j=2}^{m} (\mu_j - \hat{\mu}_j)^2.$$

From the definition of v_n, it can be seen that v_n depends on $\mathcal{S}_2$ only through $\hat{\tau}$, so that we have $\mathbf{E}[v_n^2 | \mathcal{S}_2] = \mathbf{E}[v_n^2 | \hat{\tau}]$. Also note that $\hat{\tau}$ is computed on $\mathcal{S}_{21} \subset \mathcal{S}_2$, so that $\mathcal{S}_2$ contains all the information about $\hat{\tau}$ (In terms of σ-algebras we have $\sigma(\hat{\tau}) \subset \sigma(\mathcal{S}_2)$). This allows us to use the tower property of conditional expectations to write

$$\begin{aligned}
\mathbf{E}[E_2^2 | \hat{\tau}] &= \mathbf{E}[\mathbf{E}[E_2^2 | \mathcal{S}_2] | \hat{\tau}] \\
&= \frac{\mathbf{E}[v_n^2 | \hat{\tau}]}{\mathbf{E}r_n^2 e_n^2} \sum_{j=2}^{m} \mathbf{E}[(\mu_j - \hat{\mu}_j)^2 | \hat{\tau}] \\
&= \frac{\mathbf{E}[v_n^2 | \hat{\tau}]}{\mathbf{E}r_n^2 e_n^2} \sum_{j=2}^{m} \mathbf{E}[(\hat{\mu}_j - \mathbf{E}[\hat{\mu}_j | \hat{\tau}])^2 | \hat{\tau}] + (\mathbf{E}[\hat{\mu}_j | \hat{\tau}] - \mu_j)^2
\end{aligned}$$

The conditional mean and variance of $\hat{\mu}_j$ are

$$\begin{aligned}
\mathbf{E}[\hat{\mu}_j \mid \hat{\tau}] &= \mathbf{E}[y_{2n} h_j(\langle x_{2n}, \hat{\tau} \rangle)] = \mu_j \langle \Sigma\tau, \hat{\tau} \rangle^j, \\
\mathbf{Var}[\hat{\mu}_j | \hat{\tau}] &= \frac{1}{\lceil n/2 \rceil} (\mathbf{E}y_{2n}^2 h_j^2(\langle x_{2n}, \hat{\tau} \rangle) - \mu_j^2 \langle \Sigma\tau, \hat{\tau} \rangle^{2j}).
\end{aligned}$$

To bound the conditional variance, we use the Cauchy-Schwarz inequality and the last part of proposition 27 to obtain

$$\begin{aligned}
\mathbf{E}y_{2n}^2 h_j^2(\langle x_{2n}, \hat{\tau} \rangle) &\leq \|y_{2n}\|_4^2 \cdot \|h_j\|_4^2 \\
&\leq C_y^2 \cdot C_h^2 \cdot \frac{3^j}{j^{1/2}},
\end{aligned}$$

for some absolute constant C_h. Using the latter to bound the conditional variance of E_2 we obtain

$$\begin{aligned}
\mathbf{E}[E_2^2 \mid \hat{\tau}] &\leq \frac{\mathbf{E}[v_n^2 | \hat{\tau}]}{\mathbf{E}r_n^2 e_n^2} \left(C_y^2 C_h^2 \sum_{j=2}^{m} \frac{3^j}{j^{1/2} \lceil n/2 \rceil} + \mu_j^2 (1 - \langle \Sigma\tau, \hat{\tau} \rangle^j)^2 \right) \\
&\leq \frac{\mathbf{E}[v_n^2 | \hat{\tau}]}{\mathbf{E}r_n^2 e_n^2} \left(C_y^2 C_h^2 \frac{3^{m+1}}{\lceil n/2 \rceil} + \frac{\lambda_{\max}^2(\Sigma) \|\tau - \hat{\tau}\|_2^4}{4} \sum_{j=2}^{m} j^2 \mu_j^2 \right).
\end{aligned}$$

Let us now consider the term (E_1). The conditional variance is

$$\mathbf{E}[E_1^2|\mathcal{S}_2] = \frac{(\alpha')^2}{\mathbf{E}r_n^2e_n^2}(\mathbf{E}[\hat{\xi}_n^2(\sum_{j=2}^m(\mu_j - \hat{\mu}_j)h_j(\hat{\xi}_n))^2|\mathcal{S}_2]).$$

Applying identity (21) to $\psi(\hat{\xi}_n) = \sum_{j=2}^m(\mu_j - \hat{\mu}_j)h_j(\hat{\xi}_n)$, we obtain

$$\begin{aligned}\mathbf{E}[E_1^2|\mathcal{S}_2] = \frac{(\alpha')^2}{\mathbf{E}r_n^2e_n^2}(\mathbf{E}[(&\sum_{j=2}^m(\mu_j - \hat{\mu}_j)h_j(\hat{\xi}_n))^2|\mathcal{S}_2] \\ &+ 2\mathbf{E}[(\sum_{j=2}^m\sqrt{j}(\mu_j - \hat{\mu}_j)h_{j-1}(\hat{\xi}_n))^2|\mathcal{S}_2] \\ &+ 2\mathbf{E}[(\sum_{j=2}^m(\mu_j - \hat{\mu}_j)h_j(\hat{\xi}_n))(\sum_{j=2}^m\sqrt{j(j-1)}(\mu_j - \hat{\mu}_j)h_{j-2}(\hat{\xi}_n))])\end{aligned}$$

By the orthonormality of Hermite polynomials, this simplifies to

$$\begin{aligned}\mathbf{E}[E_1^2|\mathcal{S}_2] = \frac{(\alpha')^2}{\mathbf{E}r_n^2e_n^2}(&\sum_{j=2}^m(\mu_j - \hat{\mu}_j)^2 + 2j(\mu_j - \hat{\mu}_j)^2 \\ &+ 2\sum_{j=2}^{m-2}\sqrt{(j+1)(j+2)}(\mu_j - \hat{\mu}_j)(\mu_{j+2} - \hat{\mu}_{j+2})) \\ &\leq \frac{5(\alpha')^2}{\mathbf{E}r_n^2e_n^2}\sum_{j=2}^m j(\mu_j - \hat{\mu}_j)^2.\end{aligned}$$

Using our previous calculations for the conditional mean and variance of $\hat{\mu}_j$, we obtain

$$\begin{aligned}\mathbf{E}[E_1^2|\hat{\tau}] &\leq \frac{3(\alpha')^2}{\mathbf{E}r_n^2e_n^2}\sum_{j=2}^m j(\mathbf{E}[(\hat{\mu}_j - \mathbf{E}[\hat{\mu}_j|\hat{\tau}])^2 \mid \hat{\tau}]) + j(\mathbf{E}[\hat{\mu}_j|\hat{\tau}] - \mu_j) \\ &\leq \frac{5(\alpha')^2}{\mathbf{E}r_n^2e_n^2}(C_y^2C_h^2\sum_{j=2}^m\frac{j3^j}{j^{1/2}\lceil n/2\rceil}) + \frac{5(\alpha')^2}{\mathbf{E}r_n^2e_n^2}(\sum_{j=2}^m j\mu_j^2(1 - \langle\Sigma\tau, \hat{\tau}\rangle^j)^2) \\ &\leq \frac{5(\alpha')^2C_y^2C_h^2}{\mathbf{E}r_n^2e_n^2}\cdot\frac{m3^{m+1}}{\lceil n/2\rceil} + \frac{5(\alpha')^2}{\mathbf{E}r_n^2e_n^2}(\lambda_{\max}^2(\Sigma)\frac{\|\tau - \hat{\tau}\|_2^4}{4}\sum_{j=2}^m j^3\mu_j^2).\end{aligned}$$

(F). Finally, we consider the term (F) by using once again the Gaussian decomposition $r_i = \alpha''\xi_i + v_i'$ with v_i' independent of ξ_i to rewrite (F) as

$$F = \frac{1}{\sqrt{n\mathbf{E}r_n^2e_n^2}}\sum_{i=1}^n\alpha''\xi_i\sum_{j=m+1}^\infty\mu_jh_j(\xi_i) \qquad (F_1)$$

$$+ \frac{1}{\sqrt{n\mathbf{E}r_n^2e_n^2}}\sum_{i=1}^n v_i'\sum_{j=m+1}^\infty\mu_jh_j(\xi_i). \qquad (F_2)$$

The second term (F_2) has variance

$$\mathbf{E}[F_2^2] = \frac{\mathbf{E}(v_n')^2}{\mathbf{E}r_n^2 e_n^2} \sum_{j=m+1}^{\infty} \mu_j^2.$$

The first term has variance

$$\mathbf{E}[F_1^2] \leq \frac{(\alpha'')^2}{\mathbf{E}r_n^2 e_n^2} \mathbf{E}\xi_n^2 (\sum_{j=m+1}^{\infty} \mu_j h_j(\xi_n))^2.$$

Using identity (21) on $\psi(\xi_n) = \sum_{j=m+1}^{\infty} \mu_j h_j(\xi_n)$, we obtain

$$\begin{aligned}
\mathbf{E}\xi_n^2 (\sum_{j=m+1}^{\infty} \mu_j h_j(\xi_n))^2 &= \mathbf{E}(\sum_{j=m+1}^{\infty} \mu_j h_j(\xi_n))^2 \\
&\quad + 2\mathbf{E}(\sum_{j=m+1}^{\infty} \sqrt{j}\mu_j h_{j-1}(\xi_n))^2 \\
&\quad + 2\mathbf{E}(\sum_{j=m+1}^{\infty} \mu_j h_j(\xi_n))(\sum_{j=m+1}^{\infty} \sqrt{j(j-1)}\mu_j h_{j-2}(\xi_n)) \\
&= \sum_{j=m+1}^{\infty} (2j+1)\mu_j^2 + 2\sqrt{(j+1)(j+2)}\mu_j \mu_{j+2} \\
&\leq \sum_{j=m+1}^{\infty} (4j+1)\mu_j^2 \\
&\leq 5 \sum_{j=m+1}^{\infty} j\mu_j^2.
\end{aligned}$$

Let us consider the random variables such as α, α', u_n etc. that result from the Gaussian decomposition of r_i. All these term can be shown to have bounded variance because they result from orthogonal decompositions and r_i itself has uniformly bounded second moment. For example, let us take a closer look at $\alpha, \hat{\alpha}, u_i$ appearing in $r_n = \alpha\xi_n + \hat{\alpha}\hat{\xi}_n + u_n$. Note that by construction, u_n is independent of ξ_n and $\hat{\xi}_n$ given $\mathcal{S}_2$, so that

$$\begin{aligned}
\mathbf{E}[r_n^2] &= \mathbf{E}[\mathbf{E}[(\alpha\xi_n + \hat{\alpha}\hat{\xi}_n + u_n)^2 | \mathcal{S}_2]] \\
&= \mathbf{E}[\alpha^2 + \hat{\alpha}^2 + u_n^2 + 2\langle \Sigma\tau, \hat{\tau}\rangle].
\end{aligned}$$

Since $\mathbf{E}r_n^2 \leq \mathbf{E}x_{n,1}^2$ and $|\langle \Sigma\tau, \hat{\tau}\rangle| \leq 1$, we have

$$\mathbf{E}\alpha^2 + \mathbf{E}\hat{\alpha}^2 + \mathbf{E}u_n^2 \leq \Sigma_{11} + 2 \leq \lambda_{\max}(\Sigma) + 2.$$

Thus the variances of all these terms is bounded above by $C + 2$ where C is by assumption (1) a constant independent of n, implying that these variables are all $\mathcal{O}_p(1)$.

Ignoring the constants and $\mathcal{O}_p(1)$ terms, it suffices to show that the following dominant terms converge to zero:

$$\|\tau-\hat{\tau}\|_2^2\sum_{j=2}^{m} j^2\mu_j^2, \quad \|\tau-\hat{\tau}\|_2^4\sum_{j=2}^{m} j^3\mu_j^2, \quad \sum_{j=m+1}^{\infty} j\mu_j^2, \quad \frac{m3^{m+1}}{\lceil n/2\rceil}. \tag{23}$$

The last term converges to zero by the choice of $m=\lfloor \log^{\frac{2}{3}}(n)\rfloor$. For the first three terms we need the smoothness of g. Since by assumption $\|g'\|_{L^2}^2 \leq L^2$, Lemma (28) implies that $\sum_{j=1}^{\infty} j\mu_j^2 < L^2$, which immediately proves the term $\sum_{j=m+1}^{\infty} j\mu_j^2$ converges to zero. Using the Cauchy-Schwarz inequality, we have

$$\sum_{j=2}^{m} j^2\mu_j^2 \leq (\sum_{j=2}^{m}(j\mu_j^2)^2)^{\frac{1}{2}}(\sum_{j=1}^{m} j^2)^{\frac{1}{2}} = \mathcal{O}(L^2 m^{\frac{3}{2}}), \quad \text{and,}$$
$$\sum_{j=2}^{m} j^3\mu_j^2 \leq (\sum_{j=2}^{m}(j\mu_j^2)^2)^{\frac{1}{2}}(\sum_{j=1}^{m} j^4)^{\frac{1}{2}} = \mathcal{O}(L^2 m^{\frac{5}{2}}),$$

Now using $p\geq n$ and $\|\hat{\tau}-\tau\|_2^2 \leq \|\hat{\tau}-\tau\|_1^2$, we have

$$m^{\frac{3}{2}}\cdot\|\hat{\tau}-\tau\|_2^2 \leq \log(p)\cdot\|\hat{\tau}-\tau\|_1^2 \to_p 0.$$

Finally, we can write

$$m^{\frac{5}{2}}\|\hat{\tau}-\tau\|_2^4 \leq \left(m^{\frac{3}{2}}\|\hat{\tau}-\tau\|_2^2\right)^2 \to_p 0.$$

Appendix C. Simulations

In this subsection we present the coverage rates of confidence intervals based on the debiased estimator $\tilde{\beta}$ defined by (7) and (9). We consider the combinations $n\in\{200,500\}$, $s\in\{5,10\}$ and design covariance matrices Σ_κ with $(\Sigma_\kappa)_{ij}=\kappa^{|i-j|}$ for $\kappa\in\{0,0.5\}$ and the number of covariates $p=2n$. In each case τ is defined by $\tau=\tilde{\tau}/\|\Sigma_\kappa^{\frac{1}{2}}\tilde{\tau}\|_2$, where

$$\tilde{\tau}_j = \begin{cases} s-j+1 & : 1\leq j\leq s, \\ 0 & : s<j\leq p. \end{cases}$$

We also consider two different link functions:

$$\textbf{Model 1: } y_i=\operatorname{sign}(\langle x_i,\tau\rangle)+\epsilon_i,\ \epsilon_i\sim N(0,1)$$
$$\textbf{Model 2: } y_i=U_i\cdot e^{\langle x_i,\tau\rangle},$$

where $U_1,\dots,U_n$ are iid draws from the exponential distribution with rate one, independent of x_i.

Construction of Confidence Intervals.[3] The approximate variance of the debiased estimators in Theorems 1 and 5 is equal to $\mathbf{E}r_n^2z_n^2/(\mathbf{E}r_nx_{n,k})^2$. We estimate this variance by

3. The R code for simulations is available at `https://github.com/ehamid/sim_debiasing`.

replacing the expectation with empirical averages of natural estimates of r_n, z_n. The 95% confidence intervals for β_k are then constructed using

$$\tilde{\beta}_k \pm q_{0.025} \cdot \left(\frac{\sqrt{\sum_{i=1}^n (y_i - x_i^T \hat{\beta})^2 \hat{r}_i^2}}{\sum_{i=1}^n \hat{r}_i x_{ik}} \right),$$

where $q_{0.025}$ is the 0.025 quantile of the standard normal distribution, and $\hat{r}_i$ is computed according to equation (6) or (8) depending on whether or not Σ is assumed known. The pilot estimate $\hat{\beta}$ was computed using the lasso, with the tuning parameter found using ten-fold cross-validation.[4] In the case of unknown Σ, the tuning parameter of the node-wise lasso (10) was chosen by

$$\lambda_k = \max \left\{ \lambda > 0 : \max_{j \neq k} \left| \frac{\sum_{i=1}^n \hat{r}_i x_{ij}}{\sqrt{\sum_{i=1}^n \hat{r}_i^2}} \right| < \sqrt{\log(p)} \right\}.$$

This choice of λ_k is motivated by the fact that $\max_{j \neq k} |\sum_1^n \hat{r}_i x_{ij}| / \sqrt{\sum_{i=1}^n \hat{r}_i^2} = \mathcal{O}_p(\sqrt{\log p})$, and the value of λ_k maintains a trade-off between the bias and variance of the debiased estimator, see (Zhang and Zhang, 2014, Table 2) for a similar tuning method and a detailed explanation of the trade-off. The following measures were computed:

- $\overline{cov}(S)$: computed by averaging the coverage rates of confidence intervals for non-zero coefficients.
- $\overline{cov}(S^c)$: computed by averaging the coverage rates of confidence intervals for 10 randomly chosen (at each of 200 replicates) coefficients in S^c.
- $\bar{l}(S)$ average length of confidence intervals for coefficients in S.
- $\bar{l}(S^c)$ average length of confidence intervals for coefficients in S^c, computed by averaging the lengths of confidence intervals for 10 randomly chosen (at each of 200 replicates) coefficients in S^c.
- FPR: The average False Positive Rate corresponding to 10 randomly chosen coefficients in S^c. (Proportion of confidence intervals corresponding to S^c that did not include zero.)
- TPR: The average True Positive Rate for coefficients in S. (Proportion of confidence intervals over S that did not include zero.)
- TPR(j): The True Positive Rate for confidence intervals corresponding to β_j for $1 \leq j \leq 5$.

Finally, the last four tables report simulation results for Model 1 when x has non-zero mean, $\mathbf{E}x = \mathbf{1}$. In this case the sample column means were used to center X before computing the estimators, i.e. $\tilde{X} := (I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)X$ was used in the procedure.

Some general observations are as follows:

4. The function `cv.glmnet` in the R package `glmnet` (Friedman et al., 2010) was used.

1. In general, coverage rates are close to the nominal level (95%) and coverage is improved as the sample size increases from 200 to 500.

2. Coverage rates for the null coefficients almost always dominate the coverage rate over the support set S.

3. Introducing correlation among covariates (increasing κ from 0 to 0.5) leads to longer confidence intervals and decreases the power of the tests (the TPRs). The coverage rates do not necessarily suffer from this correlation.

4. Similarly, increasing the number of non-null coefficients from 5 to 10 decreases the power.

Measure / (n, κ, s)		$\overline{cov}(S)$	$\overline{cov}(S^c)$	$\bar{l}(S)$	$\bar{l}(S^c)$
1	(200, 0, 5)	0.90	0.94	0.29	0.30
2	(200, 0, 10)	0.91	0.94	0.30	0.30
3	(200, 0.5, 5)	0.92	0.95	0.39	0.40
4	(200, 0.5, 10)	0.92	0.95	0.39	0.39
5	(500, 0, 5)	0.93	0.94	0.19	0.20
6	(500, 0, 10)	0.93	0.95	0.19	0.20
7	(500, 0.5, 5)	0.93	0.94	0.25	0.26
8	(500, 0.5, 10)	0.94	0.95	0.25	0.26

Table 1: Coverage results for model 1 when Σ is known

Measure / (n, κ, s)		FPR	TPR	TPR(1)	TPR(2)	TPR(3)	TPR(4)	TPR(5)
1	(200, 0, 5)	0.06	0.78	1.00	1.00	0.96	0.72	0.22
2	(200, 0, 10)	0.06	0.61	0.99	0.98	0.96	0.88	0.79
3	(200, 0.5, 5)	0.05	0.57	0.98	0.79	0.57	0.40	0.11
4	(200, 0.5, 10)	0.05	0.34	0.78	0.61	0.56	0.41	0.34
5	(500, 0, 5)	0.06	0.90	1.00	1.00	1.00	0.99	0.52
6	(500, 0, 10)	0.05	0.78	1.00	1.00	1.00	0.99	0.98
7	(500, 0.5, 5)	0.06	0.73	1.00	1.00	0.91	0.59	0.17
8	(500, 0.5, 10)	0.05	0.55	0.99	0.93	0.85	0.78	0.65

Table 2: Average True/False positive rates for model 1 when Σ is known

Measure / (n, κ, s)		$\overline{cov}(S)$	$\overline{cov}(S^c)$	$\bar{l}(S)$	$\bar{l}(S^c)$
1	(200, 0, 5)	0.92	0.94	0.38	0.38
2	(200, 0, 10)	0.92	0.94	0.39	0.39
3	(200, 0.5, 5)	0.91	0.95	0.46	0.47
4	(200, 0.5, 10)	0.92	0.95	0.46	0.46
5	(500, 0, 5)	0.95	0.94	0.25	0.26
6	(500, 0, 10)	0.93	0.95	0.25	0.25
7	(500, 0.5, 5)	0.93	0.94	0.31	0.32
8	(500, 0.5, 10)	0.94	0.95	0.31	0.31

Table 3: Coverage results for model 1 when Σ is unknown

Measure / (n, κ, s)		FPR	TPR	TPR(1)	TPR(2)	TPR(3)	TPR(4)	TPR(5)
1	(200, 0, 5)	0.06	0.71	1.00	0.97	0.90	0.51	0.19
2	(200, 0, 10)	0.06	0.54	0.99	0.97	0.85	0.76	0.58
3	(200, 0.5, 5)	0.05	0.54	0.91	0.76	0.55	0.34	0.14
4	(200, 0.5, 10)	0.05	0.34	0.67	0.57	0.59	0.45	0.35
5	(500, 0, 5)	0.06	0.84	1.00	1.00	1.00	0.90	0.31
6	(500, 0, 10)	0.05	0.71	1.00	1.00	1.00	0.97	0.93
7	(500, 0.5, 5)	0.06	0.69	1.00	0.97	0.83	0.49	0.15
8	(500, 0.5, 10)	0.05	0.50	0.94	0.85	0.82	0.67	0.56

Table 4: Average True/False positive rates for model 1 when Σ is unknown

Measure / (n, κ, s)		$\overline{cov}(S)$	$\overline{cov}(S^c)$	$\bar{l}(S)$	$\bar{l}(S^c)$
1	(200, 0, 5)	0.88	0.96	1.09	0.87
2	(200, 0, 10)	0.89	0.95	0.98	0.87
3	(200, 0.5, 5)	0.92	0.96	1.18	1.14
4	(200, 0.5, 10)	0.93	0.95	1.12	1.10
5	(500, 0, 5)	0.90	0.96	0.71	0.56
6	(500, 0, 10)	0.91	0.96	0.63	0.55
7	(500, 0.5, 5)	0.93	0.95	0.80	0.75
8	(500, 0.5, 10)	0.93	0.96	0.74	0.71

Table 5: Coverage results for model 2 when Σ is known

	(n, κ, s) \ Measure	FPR	TPR	TPR(1)	TPR(2)	TPR(3)	TPR(4)	TPR(5)
1	(200, 0, 5)	0.04	0.63	0.98	0.93	0.70	0.38	0.15
2	(200, 0, 10)	0.04	0.46	0.89	0.82	0.73	0.59	0.56
3	(200, 0.5, 5)	0.04	0.40	0.74	0.52	0.43	0.23	0.07
4	(200, 0.5, 10)	0.05	0.20	0.48	0.41	0.29	0.23	0.15
5	(500, 0, 5)	0.04	0.81	1.00	0.99	0.99	0.79	0.27
6	(500, 0, 10)	0.04	0.67	1.00	0.99	0.98	0.94	0.90
7	(500, 0.5, 5)	0.05	0.61	1.00	0.90	0.69	0.35	0.10
8	(500, 0.5, 10)	0.04	0.38	0.81	0.74	0.57	0.51	0.41

Table 6: Average True/False positive rates for model 2 when Σ is known

	(n, κ, s) \ Measure	$\overline{cov}(S)$	$\overline{cov}(S^c)$	$\bar{l}(S)$	$\bar{l}(S^c)$
1	(200, 0, 5)	0.88	0.96	1.27	1.08
2	(200, 0, 10)	0.91	0.95	1.20	1.09
3	(200, 0.5, 5)	0.94	0.95	1.34	1.22
4	(200, 0.5, 10)	0.94	0.95	1.28	1.25
5	(500, 0, 5)	0.92	0.96	0.85	0.72
6	(500, 0, 10)	0.92	0.95	0.77	0.71
7	(500, 0.5, 5)	0.93	0.95	0.94	0.87
8	(500, 0.5, 10)	0.94	0.96	0.89	0.86

Table 7: Coverage results for model 2 when Σ is unknown

	(n, κ, s) \ Measure	FPR	TPR	TPR(1)	TPR(2)	TPR(3)	TPR(4)	TPR(5)
1	(200, 0, 5)	0.04	0.53	0.94	0.78	0.54	0.34	0.08
2	(200, 0, 10)	0.05	0.35	0.75	0.66	0.51	0.41	0.39
3	(200, 0.5, 5)	0.05	0.34	0.67	0.54	0.30	0.15	0.07
4	(200, 0.5, 10)	0.05	0.21	0.37	0.40	0.39	0.24	0.28
5	(500, 0, 5)	0.04	0.76	1.00	1.00	0.96	0.67	0.15
6	(500, 0, 10)	0.05	0.59	0.99	0.94	0.91	0.89	0.71
7	(500, 0.5, 5)	0.05	0.56	0.94	0.81	0.63	0.30	0.10
8	(500, 0.5, 10)	0.04	0.32	0.71	0.62	0.44	0.42	0.33

Table 8: Average True/False positive rates for model 2 when Σ is unknown

(n,κ,s) \ Measure		$\overline{cov}(S)$	$\overline{cov}(S^c)$	$\bar{l}(S)$	$\bar{l}(S^c)$
1	(200, 0, 5)	0.90	0.94	0.29	0.30
2	(200, 0, 10)	0.91	0.94	0.30	0.31
3	(200, 0.5, 5)	0.92	0.95	0.39	0.40
4	(200, 0.5, 10)	0.93	0.95	0.39	0.39
5	(500, 0, 5)	0.93	0.94	0.19	0.20
6	(500, 0, 10)	0.93	0.95	0.19	0.20
7	(500, 0.5, 5)	0.93	0.94	0.25	0.26
8	(500, 0.5, 10)	0.94	0.95	0.25	0.26

Table 9: Coverage results for model 1 when Σ is known and $\mathbf{E}x_i = \mathbf{1}$

(n,κ,s) \ Measure		FPR	TPR	TPR(1)	TPR(2)	TPR(3)	TPR(4)	TPR(5)
1	(200, 0, 5)	0.06	0.78	1.00	1.00	0.97	0.68	0.24
2	(200, 0, 10)	0.06	0.61	0.99	0.99	0.95	0.88	0.75
3	(200, 0.5, 5)	0.05	0.56	0.96	0.78	0.57	0.37	0.12
4	(200, 0.5, 10)	0.05	0.34	0.77	0.61	0.53	0.42	0.34
5	(500, 0, 5)	0.06	0.90	1.00	1.00	1.00	0.98	0.51
6	(500, 0, 10)	0.05	0.78	1.00	1.00	1.00	1.00	0.99
7	(500, 0.5, 5)	0.06	0.73	1.00	0.99	0.91	0.58	0.17
8	(500, 0.5, 10)	0.05	0.55	0.99	0.93	0.86	0.78	0.65

Table 10: Average True/False positive rates for model 1 when Σ is known and $\mathbf{E}x_i = \mathbf{1}$

(n,κ,s) \ Measure		$\overline{cov}(S)$	$\overline{cov}(S^c)$	$\bar{l}(S)$	$\bar{l}(S^c)$
1	(200, 0, 5)	0.91	0.94	0.38	0.39
2	(200, 0, 10)	0.93	0.94	0.39	0.40
3	(200, 0.5, 5)	0.91	0.95	0.47	0.48
4	(200, 0.5, 10)	0.92	0.95	0.46	0.46
5	(500, 0, 5)	0.95	0.95	0.25	0.26
6	(500, 0, 10)	0.94	0.94	0.25	0.26
7	(500, 0.5, 5)	0.93	0.95	0.31	0.32
8	(500, 0.5, 10)	0.93	0.95	0.31	0.31

Table 11: Coverage results for model 1 when Σ is unknown and $\mathbf{E}x_i = \mathbf{1}$

	(n, κ, s) \ Measure	FPR	TPR	TPR(1)	TPR(2)	TPR(3)	TPR(4)	TPR(5)
1	(200, 0, 5)	0.06	0.71	1.00	0.97	0.88	0.54	0.18
2	(200, 0, 10)	0.06	0.53	0.95	0.95	0.85	0.72	0.56
3	(200, 0.5, 5)	0.05	0.54	0.90	0.76	0.56	0.35	0.15
4	(200, 0.5, 10)	0.05	0.34	0.66	0.56	0.56	0.47	0.36
5	(500, 0, 5)	0.05	0.84	1.00	1.00	1.00	0.89	0.32
6	(500, 0, 10)	0.06	0.71	1.00	1.00	1.00	0.96	0.94
7	(500, 0.5, 5)	0.05	0.68	0.99	0.96	0.83	0.47	0.15
8	(500, 0.5, 10)	0.05	0.50	0.93	0.83	0.80	0.67	0.57

Table 12: Average True/False positive rates for model 1 when Σ is unknown and $\mathbf{E}x_i = \mathbf{1}$

References

Alin Achim, Adrian Basarab, George Tzagkarakis, Panagiotis Tsakalides, and Denis Kouamé. Reconstruction of compressively sampled ultrasound images using dual prior information. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1283–1286. IEEE, 2014.

Richard Baraniuk and Philippe Steeghs. Compressive radar imaging. In *2007 IEEE radar conference*, pages 128–133. IEEE, 2007.

Peter J Bickel, Yaacov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

DR Brillinger. A generalized linear model with "Gaussian" regressor variables, a Festschrift for Erich L. Lehmann, 1982.

Raymond J Carroll, Jianqing Fan, Irene Gijbels, and Matt P Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

Longjie Cheng, Peng Zeng, Yu Zhu, et al. BS-SIM: An effective variable selection method for high-dimensional single index model. *Electronic Journal of Statistics*, 11(2):3522–3548, 2017.

Jared C Foster, Jeremy MG Taylor, and Bin Nan. Variable selection in monotone single-index models via the adaptive lasso. *Statistics in medicine*, 32(22):3944–3954, 2013.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

Ravi Ganti, Nikhil Rao, Rebecca M Willett, and Robert Nowak. Learning single index models in high dimensions. *arXiv preprint arXiv:1506.08910*, 2015.

Larry Goldstein, Stanislav Minsker, and Xiaohan Wei. Structured signal recovery from non-linear and heavy-tailed measurements. *IEEE Transactions on Information Theory*, 64(8):5513–5530, 2018.

David A Hirshberg and Stefan Wager. Debiased inference of average partial effects in single-index models. *arXiv preprint arXiv:1811.02547*, 2018.

Joel L Horowitz. *Semiparametric and nonparametric methods in econometrics*, volume 12. Springer, 2009.

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

Jeankyung Kim, David Pollard, et al. Cube root asymptotics. *The Annals of Statistics*, 18 (1):191–219, 1990.

Lars Larsson-Cohn. Lp -norms of hermite polynomials and an extremal problem on wiener chaos. *Ark. Mat.*, 40(1):133–144, 04 2002. doi: 10.1007/BF02384506. URL `https://doi.org/10.1007/BF02384506`.

Ker-Chau Li, Naihua Duan, et al. Regression analysis under link violation. *The Annals of Statistics*, 17(3):1009–1052, 1989.

Shikai Luo and Subhashis Ghosal. Forward selection and estimation in high dimensional single index models. *Statistical Methodology*, 33:172–179, 2016.

Charles F Manski. Maximum score estimation of the stochastic utility model of choice. *Journal of econometrics*, 3(3):205–228, 1975.

Charles F Manski. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of econometrics*, 27(3):313–333, 1985.

Debarghya Mukherjee, Moulinath Banerjee, and Ya'acov Ritov. Non-standard asymptotics in high dimensions: Manski's maximum score estimator revisited. *arXiv preprint arXiv:1903.10063*, 2019.

Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016.

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

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pages 10–1, 2012.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981. ISSN 00905364. URL `http://www.jstor.org/stable/2240405`.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*, pages 3420–3428, 2015.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

Sara van de Geer, Peter Bhlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 06 2014. doi: 10.1214/14-AOS1221. URL `https://doi.org/10.1214/14-AOS1221`.

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

Susan Wei and Michael R Kosorok. The change-plane Cox model. *Biometrika*, 105(4): 891–903, 2018.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.