Econometric Theory, **0**, 2021, 1–47.
doi:10.1017/S0266466621000219

UNIFORM-IN-SUBMODEL BOUNDS FOR LINEAR REGRESSION IN A MODEL-FREE FRAMEWORK

ARUN K. KUCHIBHOTLA
Carnegie Mellon University

LAWRENCE D. BROWN
University of Pennsylvania

ANDREAS BUJA
University of Pennsylvania

EDWARD I. GEORGE
University of Pennsylvania

LINDA ZHAO
University of Pennsylvania

For the last two decades, high-dimensional data and methods have proliferated throughout the literature. Yet, the classical technique of linear regression has not lost its usefulness in applications. In fact, many high-dimensional estimation techniques can be seen as variable selection that leads to a smaller set of variables (a "submodel") where classical linear regression applies. We analyze linear regression estimators resulting from model selection by proving estimation error and linear representation bounds uniformly over sets of submodels. Based on deterministic inequalities, our results provide "good" rates when applied to both independent and dependent data. These results are useful in meaningfully interpreting the linear regression estimator obtained after exploring and reducing the variables and also in justifying post-model-selection inference. All results are derived under no model assumptions and are nonasymptotic in nature.

1. INTRODUCTION AND MOTIVATION

Least-squares linear regression is one of the most widely used prediction tools in practical data analysis. With its simple form, linear regression leads to interpretable results and in many cases has predictive performance on par with sophisticated/complex models. It is, however, an open secret that in most cases the set of

We would like to thank Abhishek Chakrabortty for discussions that led to Remark 4.5. We would also like to thank the reviewers and the Editor for their constructive comments which have led to a better presentation. Address correspondence to Arun K. Kuchibhotla, Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, USA; e-mail: arunku@cmu.edu.

covariates used in the final linear regression model is rarely the same as the set of covariates initially considered by the data analyst. This is typically a consequence of the selection of a good predictive submodel based on an estimate of the out-of-sample prediction risk. We use "submodel" here to denote a subset of the full set of covariates.

Traditional analysis of the least-squares linear regression estimator restricts attention to a single set of covariates to prove consistency as well as asymptotic normality; see White (1980a, 1980b) and Buja et al. (2019). In this case, it was proved that the least-squares estimator is weakly and strongly consistent to the population least-squares functional; see (10) below. Also, a properly normalized estimator has an asymptotic normal distribution. However, the theoretical understanding and practical usefulness of submodel least-squares estimators resulting from a covariate selection procedure requires simultaneous consistency and (asymptotic) normality of all the estimators under consideration. Such simultaneous consistency and normality properties are the major focus of the current article. These are what we call *uniform-in-submodel* results. To be more concrete, suppose $\mathcal{M} = \{M_1, M_2, \ldots, M_L\}$ denotes a collection of submodels, where M_j represents a subset of covariates for $1 \le j \le L$. Also, let $\hat{\beta}_{M_j}$ represent the least-squares estimator for the linear regression of the response on the covariates in M_j. By simultaneous consistency, we mean the existence of target vectors $\{\beta_{M_j} : 1 \le j \le L\}$ such that

$$\sup_{M \in \mathcal{M}} \|\hat{\beta}_M - \beta_M\| = o_p(1), \quad \text{as} \quad n \to \infty, \tag{1}$$

for some norm $\|\cdot\|$. To claim simultaneous asymptotic normality, we prove the existence of functions $\{\psi_{M_j}(\cdot) : 1 \le j \le L\}$ such that

$$\sup_{M \in \mathcal{M}} \left\| \sqrt{n}\left(\hat{\beta}_M - \beta_M\right) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_M(Z_i) \right\| = o_p(1), \quad \text{as} \quad n \to \infty. \tag{2}$$

Here, n represents the sample size and $Z_i = (X_i, Y_i), 1 \le i \le n$, represent the regression data, with detailed notation given in Section 2. Equation (2) provides the well-known "asymptotic uniform linear representation" in the special case of the least-squares linear regression estimator. This uniform linear representation is very crucial in providing inference after variable selection via simultaneous inference (Bachoc, Preinerstorfer, and Steinberger, 2019b). If $\widehat{M}$ is a selected model, then one can perform inference on $\beta_{\widehat{M}}$ by estimating the distribution of $\widehat{\beta}_{\widehat{M}}$. This can be a tricky problem to deal with as shown in the works of Leeb and Pötscher (2005, 2006a, 2006b, 2008).

Although various model-selection criteria like C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), and lasso have been recommended for covariate selection in linear regression, results of the type (1) and (2) have not been established in the literature (at least not in the full generality considered here). Our method of attack is quite nonstandard. Instead of assuming that the observations are independent and identically distributed (i.i.d.), we prove a purely

deterministic inequality to bound the left-hand sides of (1) and (2) using maxima of several averages. We then control these averages under both independence and functional dependence to obtain explicit rates of convergence; cf. White (2001) where a detailed classical analysis of the least-squares regression estimator is provided. The functional dependence structure of data, introduced in Wu (2005), is based on the idea of coupling and covers the setting of many linear and nonlinear time series. This dependence concept is very closely related to the L_p-approximability concept introduced in Pötscher and Prucha (1997).

Some noteworthy aspects of our results are as follows.

1. We provide a purely deterministic inequality for the least-squares linear regression estimator which does not require any stochasticity of the regression data and holds for any sample size n. These deterministic results are sharp and by nature more widely applicable than any asymptotic results. Some deterministic inequalities for linear regression appeared in Kuchibhotla et al. (2019). Although these inequalities led to suboptimal rates, the structures of those deterministic inequalities were useful for the context in that paper.
2. All our results allow misspecification of the linear model. This means that the classical Gauss–Markov linear model need not hold true for any of the submodels under consideration; see Chapter 4 of Monahan (2008). Two important objections (for us) to the classical model are the impositions of fixed design and linearity structure on the data generating distribution. Since our setting allows for misspecification, we call our framework "model-free." We note here that our results do apply to the setting of fixed covariates.
3. When studied assuming a suitable randomness structure (such as independence or functional dependence), our results are precise concentration inequalities applicable in finite samples and apply to high-dimensional observations. Another interesting facet of our results is that we do not assume the observations are identically distributed. This is an important generalization needed to include the case of fixed covariates.
4. For concreteness, we take the set of submodels $\mathcal{M}$ to be the set of all submodels of size bounded by k (for some $1 \le k \le p$). Here, p represents the total number of available covariates. Under certain regularity conditions, the rates of convergence we obtain in this case for simultaneous consistency (1) and normality (2) with euclidean norm are $\sqrt{k \log(ep/k)/n}$ and $k \log(ep/k)/\sqrt{n}$, respectively (up to a lower-order additive term). Interestingly, the simultaneous consistency rate matches the minimax optimal rate of a well-specified high-dimensional sparse linear regression; see Raskutti, Wainwright, and Yu (2011). It should be noted that even though the rates match with the setting of *well-specified high-dimensional linear regression*, we do NOT require a well-specified model in this article.
5. In the process of applying our results to functionally dependent observations, we prove a tail bound for zero-mean-dependent sums, thereby extending the results of Wu and Wu (2016). For independent observations, we use the precise concentration inequality results of Kuchibhotla and Chakrabortty (2020).

In addition to the important general model-selection problem above where the results of the type (1) and (2) are required, our simultaneity results can be seen to provide essential inferential validity guarantees for the following setting of growing importance. In the vast literature on high-dimensional linear regression, it has become customary to assume an underlying linear model along with a sparsity constraint on the true regression parameter. But suppose statisticians are not willing to assume sparsity of the parameter, and neither are they willing to assume a linear model. Such unwillingness is not unreasonable in light of the fact that any model is just an approximation, and sparsity is just an assumption of convenience. Now, consider the following stylized description of approaches to high-dimensional data as widely practiced in applied statistics and data science: High-dimensional data are first explored either in a formal algorithmic way (e.g., using lasso or best subset selection) or in an informal exploratory way (e.g., using residual and leverage plots) to select a manageable small set of variables. Subsequently, the reduced data are subjected to linear regression. The combination of variable selection and linear regression is thought of as one procedure, a "high-dimensional linear regression." Even though the procedure uses only a reduced set of variables in the final regression, it uses all the variables in the preceding selection phase. Suppose $\hat{M} \in \mathcal{M}$ is the final selected submodel (from some collection of models $\mathcal{M}$) and $\hat{\beta}_{\hat{M}}$ is the least-squares linear regression estimator thus obtained. The estimator $\hat{\beta}_{\hat{M}}$ is known as the postregularization estimator in the high-dimensional statistics literature if $\hat{M}$ is obtained from some regularized least-squares procedure. An important question now is "what does $\hat{\beta}_{\hat{M}}$ estimate (consistently)?" A simultaneous result answers this question through the *trivial* bound

$$\|\hat{\beta}_{\hat{M}} - \beta_{\hat{M}}\| \leq \sup_{M \in \mathcal{M}} \|\hat{\beta}_M - \beta_M\| = o_p(1).$$

Therefore, $\hat{\beta}_{\hat{M}}$ is estimating the quantity $\beta_{\hat{M}}$ which is random through $\hat{M}$. If the model-selection procedure is such that $\hat{M}$ does not stabilize as $n \to \infty$, then $\hat{\beta}_{\hat{M}}$ is only consistent for the random quantity $\beta_{\hat{M}}$ and may not be consistent for any nonrandom quantity. By comparison, if $\mathbb{P}(\hat{M} = M_0) \to 1$ as $n \to \infty$ for some submodel M_0, then with probability converging to one, $\beta_{\hat{M}} = \beta_{M_0}$ and hence $\hat{\beta}_{\hat{M}}$ is consistent for the nonrandom quantity β_{M_0}.

1.1. Literature Review

Results of the simultaneous type described in (1) and (2) are not readily available in the literature. Some works that are closely related to ours are Belloni and Chernozhukov (2013), Bachoc et al. (2018), and Chakrabortty, Nandy, and Li (2021). Although some of these works consider a simultaneous problem, their results are only restricted to certain special cases (e.g., independent observations and/or fixed design) of our framework. Belloni and Chernozhukov (2013) prove the rate of convergence of the least-squares linear regression estimator obtained

after covariate selection using lasso. Bachoc et al. (2018) prove the rate of convergence of

$$\sup_{M\in\mathcal{M}} \left\| \hat{\beta}_M - \beta_M \right\|_\infty$$

under the restricted isometry property (RIP). (Here, $\|v\|_\infty$ for a vector v denotes the maximum absolute entry in the vector.) Also, they only consider fixed covariates. We do not assume RIP, because it is not a practical assumption, and also we prove the simultaneous convergence guarantee with the euclidean norm rather than $\|\cdot\|_\infty$. It should also be mentioned that Bachoc et al. (2018) appeared after the initial version of the current work Kuchibhotla et al. (2018). Chakrabortty et al. (2021) independently prove results very similar to ours in the case of independent observations with sub-Gaussian tails. They consider a more general collection of submodels $\mathcal{M}$ than the set of k-sparse submodels; see Section 5 of Chakrabortty et al. (2021) for more details. Because our results are deterministic in nature, they do apply for a general collection of submodels, but for concreteness, we fix the choice of the collection. Under the assumptions of Chakrabortty et al. (2021, Sect. 5), their results match ours exactly. We note, however, that their results are only proved for i.i.d. observations, which is why they do not apply to the case of fixed covariates. Furthermore, our results, including the case of independent observations, are proved for a large class of tail assumptions that subsume their assumptions.

Finally, we mention two recent works that discuss uniform-in-submodel-type results. Rinaldo et al. (2018) in their Theorem 1, as well as Remark 4 that follows, discuss uniform-in-submodel consistency for i.i.d. observations that are bounded. Their rates, however, are suboptimal; for instance, their Theorem 1 only proves a rate $k\sqrt{\log(k)/n}$, while our results imply the optimal rate of $\sqrt{k/n}$. Giessing (2018, Chap. 2), following the initial version of our work, proves uniform-in-submodel consistency as well as linear representation results for quantile regression when the observations are independent. The tail assumptions on the observations there are weaker than ours, but this is expected, at least for the response, because the loss is Lipschitz in the response.

1.2. Organization

The remainder of our paper is organized as follows. In Section 2, we introduce our notation and general framework. In Section 3, we derive various deterministic inequalities for linear regression that form the core of the paper. The application of these results to the case of independent observations is considered in Section 4. The application of the deterministic inequalities to the case of (functionally) dependent observations is considered in Section 5. A discussion of our results along with their implications for postselection inference is given in Section 6. Some auxiliary probability results for sums of independent and functionally dependent random variables are given in Appendixes A and B, respectively.

2. NOTATION

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are n random vectors in $\mathbb{R}^p \times \mathbb{R}$. Throughout the paper, we implicitly think of p as a function of n, and so the sequence of random vectors should be thought of as a triangular array. The term "submodel" is used to specify the subset of covariates used in the regression and does not refer to any probability model. We do *not* assume a linear model (in any sense) to be true anywhere for any choice of covariates in any section of the paper. In this sense, all our results are applicable in the case of misspecified linear regression models.

For any vector $v \in \mathbb{R}^q$, for $q \geq 1$ and $1 \leq j \leq q$, let $v(j)$ denote the jth coordinate of v. For any nonempty submodel M given by a subset of $\{1, 2, \ldots, q\}$, let $v(M)$ denote a subvector of v with indices in M. For instance, if $M = \{2, 4\}$ and $q \geq 4$, then $v(M) = (v(2), v(4))$. The notation $|M|$ is used to denote the cardinality of M. For any nonempty submodel $M \subseteq \{1, 2, \ldots, q\}$ and any symmetric matrix $A \in \mathbb{R}^{q \times q}$, let $A(M)$ denote the submatrix of A with indices in $M \times M$. For $1 \leq j, k \leq q$, let $A(j, k)$ denote the value at the jth row and the kth column of A. Define the r-norm of a vector $v \in \mathbb{R}^q$, for $1 \leq r \leq \infty$, as

$$\|v\|_r^r := \sum_{j=1}^{q} |v(j)|^r, \quad \text{for} \quad 1 \leq r < \infty, \quad \text{and} \quad \|v\|_\infty := \max_{1 \leq j \leq q} |v(j)|.$$

Let $\|v\|_0$ denote the number of nonzero entries in v (note this is not a norm). For any square matrix A, let $\lambda_{\min}(A)$ denote the minimum eigenvalue of A. Also, let the elementwise maximum and the operator norm be defined, respectively, as

$$|||A|||_\infty := \max_{1 \leq j, k \leq q} |A(j, k)|, \quad \text{and} \quad \|A\|_{op} := \sup_{\|\delta\|_2 \leq 1} \|A\delta\|_2 .$$

The following simple inequalities are useful. For any matrix $A \in \mathbb{R}^{q \times q}$ and $v \in \mathbb{R}^q$,

$$\|v\|_1 \leq \|v\|_0^{1/2} \|v\|_2, \quad \|Av\|_\infty \leq |||A|||_\infty \|v\|_1, \quad \text{and} \quad |v^\top A v| \leq |||A|||_\infty \|v\|_1^2. \quad \textbf{(3)}$$

For any $1 \leq k \leq p$, define the set of k-sparse submodels

$$\mathcal{M}(k) := \{M : M \subseteq \{1, 2, \ldots, p\},\ 1 \leq |M| \leq k\},$$

so that $\mathcal{M}(p)$ is the power set of $\{1, 2, \ldots, p\}$ with the deletion of the empty set. Thus, the set $\mathcal{M}(k)$ denotes the set of all nonempty submodels of size bounded by k. The most important aspect of our results is the "uniform-in-submodel" feature. These results are proved uniform over $M \in \mathcal{M}(k)$, for some k, that is allowed to diverge with n.

When fitting a linear regression, it is common to include an intercept term. To avoid extra notation, we assume that all covariates under consideration are included in the vectors X_i. So, take the first coordinate of all X_i's to be 1, that is, $X_i(1) = 1$,

for all $1 \le i \le n$, if an intercept is required. For any $M \subseteq \{1, 2, \ldots, p\}$, define the ordinary least-squares (OLS) empirical risk (or objective) function as

$$\hat{R}_n(\theta; M) := \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - X_i^\top(M)\theta \right\}^2, \quad \text{for} \quad \theta \in \mathbb{R}^{|M|}.$$

Expanding the square function, it is clear that

$$\hat{R}_n(\theta; M) = \frac{1}{n} \sum_{i=1}^{n} Y_i^2 - \frac{2}{n} \sum_{i=1}^{n} Y_i X_i^\top(M)\theta + \theta^\top \left(\frac{1}{n} \sum_{i=1}^{n} X_i(M) X_i^\top(M) \right) \theta. \tag{4}$$

Only the second and third terms depend on θ. Because the quantities in these terms play a significant role in our analysis, define

$$\hat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top \in \mathbb{R}^{p \times p}, \quad \text{and} \quad \hat{\Gamma}_n := \frac{1}{n} \sum_{i=1}^{n} X_i Y_i \in \mathbb{R}^p. \tag{5}$$

The least-squares linear regression estimator $\hat{\beta}_{n,M}$ is defined as

$$\hat{\beta}_{n,M} := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{|M|}} \hat{R}_n(\theta; M) = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{|M|}} \{\theta^\top \hat{\Sigma}_n(M)\theta - 2\theta^\top \hat{\Gamma}_n(M)\}. \tag{6}$$

The notation $\arg\min_\theta f(\theta)$ denotes the minimizer of $f(\theta)$. Based on the quadratic expansion (4) of the empirical objective $\hat{R}_n(\theta; M)$, the estimator $\hat{\beta}_{n,M}$ is given by the closed form expression

$$\hat{\beta}_{n,M} = [\hat{\Sigma}_n(M)]^{-1} \hat{\Gamma}_n(M), \tag{7}$$

assuming nonsingularity of $\hat{\Sigma}_n(M)$. Note that $[\hat{\Sigma}_n(M)]^{-1}$ is *not* equal to $\hat{\Sigma}_n^{-1}(M)$. The matrix $\hat{\Sigma}_n(M)$ being the average of n rank-one matrices in $\mathbb{R}^{|M| \times |M|}$, its rank is at most $\min\{|M|, n\}$. This implies that the least-squares estimator $\hat{\beta}_{n,M}$ is not uniquely defined unless $|M| \le n$.

It is clear from (7) that $\hat{\beta}_{n,M}$ is a smooth (nonlinear) function of two averages $\hat{\Sigma}_n(M)$ and $\hat{\Gamma}_n(M)$. Assuming for a moment that the random vectors (X_i, Y_i) are i.i.d. with finite fourth moments, it follows that $\hat{\Sigma}_n(M)$ and $\hat{\Gamma}_n(M)$ converge in probability to their expectations. The i.i.d. assumption here can be relaxed to weak dependence and nonidentically distributed random vectors; see White (2001) for more details.

Getting back to the general context, define the "expected" matrix and vector as

$$\Sigma_n := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[X_i X_i^\top\right] \in \mathbb{R}^{p \times p}, \quad \text{and} \quad \Gamma_n := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[X_i Y_i\right] \in \mathbb{R}^p. \tag{8}$$

Note that we write Σ_n or Γ_n (indexing by the sample size n) for two reasons. First, we do not assume the random vectors are identically distributed, and hence, the expected matrix changes with n even if the dimension is fixed. Second, the dimension in our setting is allowed to change with n, and hence, even if the

observations are identically distributed, the expectation matrix changes with the sample size.

To define a target vector that is being consistently estimated by $\hat{\beta}_{n,M}$, consider the following simple calculation in a simpler setting where $|M|$ does not change with n. As noted above $\hat{\beta}_{n,M} = [\hat{\Sigma}_n(M)]^{-1}\hat{\Gamma}_n(M)$, and if

$$(\hat{\Sigma}_n - \Sigma_n, \hat{\Gamma}_n - \Gamma_n) \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty,$$

then by a Slutsky-type argument, it follows that

$$\hat{\beta}_{n,M} - \beta_{n,M} \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty, \tag{9}$$

where

$$\beta_{n,M} := [\Sigma_n(M)]^{-1}\Gamma_n(M) = \underset{\theta \in \mathbb{R}^{|M|}}{\arg\min}\{\theta^\top \Sigma_n(M)\theta - 2\theta^\top \Gamma_n(M)\}. \tag{10}$$

The convergence statement (9) only concerns a single submodel M and is not uniform over M. By uniform-in-submodel $\|\cdot\|_2$-norm consistency of $\hat{\beta}_{n,M}$ to $\beta_{n,M}$, for $M \in \mathcal{M}(k)$, we mean that

$$\sup_{M \in \mathcal{M}(k)} \left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_2 = o_p(1) \quad \text{as} \quad n \to \infty.$$

As shown above, convergence of $\hat{\beta}_{n,M}$ to $\beta_{n,M}$ only requires convergence of $\hat{\Sigma}_n(M)$ to $\Sigma_n(M)$ and $\hat{\Gamma}_n(M)$ to $\Gamma_n(M)$. It is not required that these matrices and vectors are averages of random matrices and random vectors.

In the following section, in proving deterministic inequalities, we generalize the linear regression estimator by the function $\beta_M : \mathbb{R}^{p\times p} \times \mathbb{R}^p \to \mathbb{R}^{|M|}$ as

$$\beta_M(\Sigma, \Gamma) = [\Sigma(M)]^{-1}\Gamma(M), \tag{11}$$

assuming the existence of the inverse of $\Sigma(M)$. We call this $\beta_M(\cdot,\cdot)$ the *linear regression map*. It is evident that

$$\hat{\beta}_{n,M} = \beta_M(\hat{\Sigma}_n, \hat{\Gamma}_n) \quad \text{and} \quad \beta_{n,M} = \beta_M(\Sigma_n, \Gamma_n).$$

There are many potential applications that require replacing the sample average matrices in the linear regression estimator by a suitable nonaverage version, e.g., shrinkage or robust estimators. Three of these applications are listed in Section 3.3. To distinguish the estimator $\hat{\beta}_{n,M}$ with sample averages from the linear regression map, we call $\hat{\beta}_{n,M}$ as the OLS estimator.

In the next section, we shall prove a bound of the type

$$\|\beta_M(\Sigma_1, \Gamma_1) - \beta_M(\Sigma_2, \Gamma_2)\|_2 \le F_M(\Sigma_1 - \Sigma_2, \Gamma_1 - \Gamma_2) \quad \text{for all} \quad M \in \mathcal{M}(k) \tag{12}$$

and for some function $F_M(\cdot,\cdot)$. Taking $(\Sigma_1, \Gamma_1) = (\hat{\Sigma}_n, \hat{\Gamma}_n)$ and $(\Sigma_2, \Gamma_2) = (\Sigma_n, \Gamma_n)$, inequality (12) is useful for the purpose of proving (1). In regard to (12), thinking of β_M as a function of (Σ, Γ), our results are essentially about studying Lipschitz continuity properties and understanding what kind of norms are best suited for

this purpose. Using the smoothness of the linear regression map, we also obtain a bound on

$$\sup_{M\in\mathcal{M}(k)} \|\beta_M(\Sigma_1,\Gamma_1)-\beta_M(\Sigma_2,\Gamma_2)-\nabla\beta_M(\Sigma_2,\Gamma_2)(\Sigma_1-\Sigma_2,\Gamma_1-\Gamma_2)\|_2,$$

where $\nabla\beta_M(\cdot,\cdot)$ represents the gradient of the linear regression map. The following error norms will be very useful for these results:

$$\begin{aligned}\mathrm{RIP}(k,\Sigma_1-\Sigma_2) &:= \sup_{M\in\mathcal{M}(k)} \|\Sigma_1(M)-\Sigma_2(M)\|_{op},\\ \mathcal{D}(k,\Gamma_1-\Gamma_2) &= \sup_{M\in\mathcal{M}(k)} \|\Gamma_1(M)-\Gamma_2(M)\|_2. \end{aligned} \tag{13}$$

The quantity RIP is a norm for any $k \geq 2$ and is not a norm for $k = 1$. This error norm is very closely related to the RIP used in the compressed sensing and high-dimensional linear regression literature where Σ_2 is the identity matrix. Also, define the k-sparse minimum singular value of a matrix $A \in \mathbb{R}^{p\times p}$ as

$$\Lambda(k;A) = \inf_{\theta\in\mathbb{R}^p,\,\|\theta\|_0\leq k} \frac{\|A\theta\|_2}{\|\theta\|_2}. \tag{14}$$

Even though all the results in the next section are written in terms of the linear regression map (11), our main focus will still be the matrices and vectors defined in (5) and (8).

3. DETERMINISTIC RESULTS FOR LINEAR REGRESSION

3.1. Can We Expect Deterministic Inequalities?

Classical asymptotic theory for linear regression or for that matter any estimation problem usually starts with an assumption that the observations are independent or otherwise follow a specific stochastic dependence. What we are aiming for is a purely deterministic inequality that does not even assume randomness of the observations.

To see whether we can at all expect a deterministic inequality, let us consider a simple example with only one submodel $M = \{1\}$, that is, a simple regression through the origin based on one regressor. For this case, let us write

$$\hat{\sigma}_n^2 := \hat{\Sigma}_n(M), \quad \hat{\gamma}_n := \hat{\Gamma}_n(M), \quad \sigma_n^2 := \Sigma_n(M), \quad \text{and} \quad \gamma_n := \Gamma_n(M).$$

Note that these are all scalar quantities. Now, the regression estimator and targets become

$$\hat{\beta}_{n,M} = \frac{\hat{\gamma}_n}{\hat{\sigma}_n^2} \quad \text{and} \quad \beta_{n,M} = \frac{\gamma_n}{\sigma_n^2}.$$

Observe that

$$\begin{aligned}\left|\hat{\beta}_{n,M}-\beta_{n,M}\right| &= \left|\frac{\hat{\gamma}_n}{\hat{\sigma}_n^2}-\frac{\gamma_n}{\sigma_n^2}\right| \\ &\le \left|\frac{1}{\hat{\sigma}_n^2}-\frac{1}{\sigma_n^2}\right|\hat{\gamma}_n+\frac{1}{\sigma_n^2}\left|\hat{\gamma}_n-\gamma_n\right| \\ &\le \sigma_n^{-2}\left|\hat{\sigma}_n^2-\sigma_n^2\right|\times\left|\hat{\beta}_{n,M}\right|+\sigma_n^{-2}\left|\hat{\gamma}_n-\gamma_n\right| \\ &\le \sigma_n^{-2}\left|\hat{\sigma}_n^2-\sigma_n^2\right|\times\left|\hat{\beta}_{n,M}-\beta_{n,M}\right|+\sigma_n^{-2}\left|\hat{\sigma}_n^2-\sigma_n^2\right| \\ &\quad\times\left|\beta_{n,M}\right|+\sigma_n^{-2}\left|\hat{\gamma}_n-\gamma_n\right|.\end{aligned}$$

Solving this inequality for $|\hat{\beta}_{n,M}-\beta_{n,M}|$, we get

$$\left|\hat{\beta}_{n,M}-\beta_{n,M}\right| \le \frac{\left|\hat{\sigma}_n^2-\sigma_n^2\right|\times\left|\beta_{n,M}\right|+\left|\hat{\gamma}_n-\gamma_n\right|}{\sigma_n^2-\left|\hat{\sigma}_n^2-\sigma_n^2\right|}.$$

This is a deterministic inequality that does not require any probabilistic structure on the data, and more importantly, the right-hand side tends to zero if $\hat{\sigma}_n^2-\sigma_n^2=o(\sigma_n^2)$ and $\hat{\gamma}_n-\gamma_n=o(\sigma_n^2)$. Because this bound is a deterministic inequality, taking a supremum over a collection of submodels does not invalidate the inequality. This is *not* the case if we only have an asymptotic result. All our deterministic inequalities to be stated/proved in the forthcoming sections are variations of the calculation above. One might suspect that the closed form expression of the linear regression map made a deterministic inequality possible, but as shown in Kuchibhotla (2018), most "smooth" M-estimators satisfy this type of result.

3.2. Main Results

All our results in this section depend on the error norms $\mathrm{RIP}(k,\Sigma_1-\Sigma_2)$ and $\mathcal{D}(k,\Gamma_1-\Gamma_2)$ in (13). These are, respectively, the maximal k-sparse eigenvalue of $\Sigma_1-\Sigma_2$ and the maximal k-sparse euclidean norm of $\Gamma_1-\Gamma_2$. At first glance, it may not be clear how these quantities behave. We first present a simple inequality for RIP and $\mathcal{D}$ in terms of $|||\cdot|||_\infty$ and $\|\cdot\|_\infty$.

PROPOSITION 3.1. *For any* $k \ge 1$,

$$\sup_{M\in\mathcal{M}(k)}\|\Sigma_1(M)-\Sigma_2(M)\|_{op} \le k|||\Sigma_1-\Sigma_2|||_\infty,$$

$$\sup_{M\in\mathcal{M}(k)}\|\Gamma_1(M)-\Gamma_2(M)\|_2 \le k^{1/2}\|\Gamma_1-\Gamma_2\|_\infty.$$

Proof. See Appendix C for a proof. ∎

In many cases, it is much easier to control the maximum elementwise norm rather than the RIP error norm. However, the factor k on the right-hand side often leads to suboptimal dependence in the dimension. For the special cases of

independent and dependent random vectors (to be discussed in Sections 4 and 5), we directly control RIP and $\mathcal{D}$.

The sequence of results to follow are related to uniform consistency in $\|\cdot\|_2$- and $\|\cdot\|_1$-norms. To state these results, we require the following quantities representing the strength of regression (or linear association). For $r, k \geq 1$,

$$S_{r,k}(\Sigma, \Gamma) := \sup_{M \in \mathcal{M}(k)} \|\beta_M(\Sigma, \Gamma)\|_r = \sup_{M \in \mathcal{M}(k)} \left\| [\Sigma(M)]^{-1} \Gamma(M) \right\|_r. \tag{15}$$

For the following theorem, recall the k-sparse minimum singular value $\Lambda(\cdot;\cdot)$ defined in (14) and the error metrics defined in (13).

THEOREM 3.1 (Uniform L_2-consistency). *Let $k \geq 1$ be any integer such that*

$$RIP(k, \Sigma_1 - \Sigma_2) \leq \Lambda(k; \Sigma_2). \tag{16}$$

Then, simultaneously, for all $M \in \mathcal{M}(k)$,

$$\|\beta_M(\Sigma_1, \Gamma_1) - \beta_M(\Sigma_2, \Gamma_2)\|_2 \leq \frac{\mathcal{D}(k, \Gamma_1 - \Gamma_2) + RIP(k, \Sigma_1 - \Sigma_2)\, \|\beta_M(\Sigma_2, \Gamma_2)\|_2}{\Lambda(k; \Sigma_2) - RIP(k, \Sigma_1 - \Sigma_2)}.$$

Proof. Recall from the linear regression map (11) that

$$\beta_M(\Sigma_1, \Gamma_1) = [\Sigma_1(M)]^{-1} \Gamma_1(M) \quad \text{and} \quad \beta_M(\Sigma_2, \Gamma_2) = [\Sigma_2(M)]^{-1} \Gamma_2(M).$$

Fix $M \in \mathcal{M}(k)$. Then,

$$\begin{aligned} \|\beta_M(\Sigma_1, \Gamma_1) - \beta_M(\Sigma_2, \Gamma_2)\|_2 &= \left\| [\Sigma_1(M)]^{-1} \Gamma_1(M) - [\Sigma_2(M)]^{-1} \Gamma_2(M) \right\|_2 \\ &\leq \left\| \left([\Sigma_1(M)]^{-1} - [\Sigma_2(M)]^{-1} \right) \Gamma_1(M) \right\|_2 \\ &\quad + \left\| [\Sigma_2(M)]^{-1} (\Gamma_1(M) - \Gamma_2(M)) \right\|_2 \\ &=: \Delta_1 + \Delta_2. \end{aligned}$$

By definition of the operator norm,

$$\Delta_2 \leq [\Lambda(k; \Sigma_2)]^{-1} \|\Gamma_1(M) - \Gamma_2(M)\|_2 \leq [\Lambda(k; \Sigma_2)]^{-1} \mathcal{D}(k, \Gamma_1 - \Gamma_2).$$

To control Δ_1, note that

$$\begin{aligned} \Delta_1 &\leq \left\| \left(I_M - [\Sigma_2(M)]^{-1} \Sigma_1(M) \right) [\Sigma_1(M)]^{-1} \Gamma_1(M) \right\|_2 \\ &\leq \left\| \left(I_M - [\Sigma_2(M)]^{-1} \Sigma_1(M) \right) \right\|_{op} \|\beta_M(\Sigma_1, \Gamma_1)\|_2 \\ &\leq [\Lambda(k; \Sigma_2)]^{-1} \|\Sigma_1(M) - \Sigma_2(M)\|_{op} \|\beta_M(\Sigma_1, \Gamma_1)\|_2 \\ &\leq [\Lambda(k; \Sigma_2)]^{-1} \mathrm{RIP}(k, \Sigma_1 - \Sigma_2) \|\beta_M(\Sigma_1, \Gamma_1)\|_2, \end{aligned}$$

where I_M represents the identity matrix of dimension $|M| \times |M|$. Now, combining bounds on Δ_1, Δ_2, we get

$$\|\beta_M(\Sigma_1, \Gamma_1) - \beta_M(\Sigma_2, \Gamma_2)\|_2 \leq \frac{\mathcal{D}(k, \Gamma_1 - \Gamma_2) + \mathrm{RIP}(k, \Sigma_1 - \Sigma_2) \|\beta_M(\Sigma_1, \Gamma_1)\|_2}{\Lambda(k; \Sigma_2)}.$$

Subtracting and adding $\beta_M(\Sigma_2, \Gamma_2)$ from $\beta_M(\Sigma_1, \Gamma_1)$, we get

$$\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\|_2 \le \frac{\mathcal{D}(k,\Gamma_1-\Gamma_2) + \mathrm{RIP}(k,\Sigma_1-\Sigma_2)\,\|\beta_M(\Sigma_2,\Gamma_2)\|_2}{\Lambda(k;\Sigma_2)} + \frac{\mathrm{RIP}(k,\Sigma_1-\Sigma_2)}{\Lambda(k,\Sigma_2)}\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\|_2.$$

Solving this inequality under assumption (16), it follows, for all $M \in \mathcal{M}(k)$, that

$$\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\|_2 \le \frac{\mathcal{D}(k,\Gamma_1-\Gamma_2) + \mathrm{RIP}(k,\Sigma_1-\Sigma_2)\,\|\beta_M(\Sigma_2,\Gamma_2)\|_2}{\Lambda(k;\Sigma_2) - \mathrm{RIP}(k;\Sigma_2)}.$$

This proves the result. ∎

As will be seen in the application of Theorem 3.1, the complicated looking bound provided above gives the "optimal" bound. Combining Proposition 3.1 and Theorem 3.1, we get the following simple corollary that gives suboptimal rates.

COROLLARY 3.1. *Let $k \ge 1$ be any integer such that*

$$k|||\Sigma_1 - \Sigma_2|||_\infty \le \Lambda(k;\Sigma_2).$$

Then,

$$\sup_{M\in\mathcal{M}(k)} \|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\|_2 \le \frac{k^{1/2}\,\|\Gamma_1-\Gamma_2\|_\infty + k|||\Sigma_1-\Sigma_2|||_\infty S_{2,k}(\Sigma_2,\Gamma_2)}{\Lambda(k;\Sigma_2) - k|||\Sigma_1-\Sigma_2|||_\infty}.$$

Remark 3.1 (Bounding $S_{2,k}$ in (15)). The bound for uniform L_2-consistency requires a bound on $\|\beta_M(\Sigma_2,\Gamma_2)\|_2$ in addition to bounds on the error norms related to Σ-matrices and Γ-vectors. It is a priori not clear how this quantity might vary as the dimension of the submodel M changes. In the classical analysis of linear regression where a true linear model is assumed, the true parameter vector β is seen as something chosen by nature, and hence, its norm is not under control of the statistician. Hence, in the classical analysis, a growth rate on $\|\beta\|_2$ is imposed as an assumption.

From the viewpoint taken in this paper, under misspecification nature picks the whole distribution sequence of random vectors and hence the quantity $\beta_M(\cdot,\cdot)$ that came up in the analysis. In the full generality of linear regression maps considered here, we do not know of any techniques to bound the norm of this vector. It is, however, possible to bound it if $\beta_M(\cdot,\cdot)$ is defined by a least-squares linear regression problem. Recall the definition of Σ_n, Γ_n from (8) and $\beta_{n,M}$ from (10). Observe that by definition of $\beta_{n,M}$,

$$0 \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\{Y_i - X_i^\top(M)\beta_{n,M}\right\}^2\right] \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[Y_i^2\right] - \beta_{n,M}^\top \Sigma_n(M)\beta_{n,M}.$$

This holds because $\beta_{n,M}$ satisfies $n^{-1}\sum_{i=1}^n \mathbb{E}[X_i(M)Y_i] = n^{-1}\sum_{i=1}^n \mathbb{E}[X_i(M)X_i^\top(M)\beta_{n,M}] = \Sigma_n(M)\beta_{n,M}$. Hence, for every $M \in \mathcal{M}(p)$,

$$\|\beta_{n,M}\|_2^2 \lambda_{\min}(\Sigma_n(M)) \le \beta_{n,M}\Sigma_n(M)\beta_{n,M} \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[Y_i^2\right].$$

Therefore, using the definitions of $\Lambda(k;\Sigma_n)$ and $S_{r,k}$ in (14) and (15),

$$S_{2,k}(\Sigma_n,\Gamma_n) \le \left(\frac{1}{n\Lambda(k;\Sigma_n)}\sum_{i=1}^n \mathbb{E}\left[Y_i^2\right]\right)^{1/2},$$

$$S_{1,k}(\Sigma_n,\Gamma_n) \le \left(\frac{k}{n\Lambda(k;\Sigma_n)}\sum_{i=1}^n \mathbb{E}\left[Y_i^2\right]\right)^{1/2}.$$

It is immediate from these results that if the second moment of the response is uniformly bounded, then $S_{2,k}$ behaves like a constant when Σ_n is well-conditioned. See Foygel and Srebro (2011) for a similar calculation. ◊

Based on the uniform-in-submodel $\|\cdot\|_2$-bound, the following result is trivially proved.

THEOREM 3.2 (Uniform L_1-consistency). *Let $k \ge 1$ be such that*

$$RIP(k,\Sigma_1-\Sigma_2) \le \Lambda(k;\Sigma_2).$$

Then, simultaneously, for all $M \in \mathcal{M}(k)$,

$$\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\|_1 \le |M|^{1/2}\frac{\mathcal{D}(k,\Gamma_1-\Gamma_2) + RIP(k,\Sigma_1-\Sigma_2)\,\|\beta_M(\Sigma_2,\Gamma_2)\|_2}{\Lambda(k;\Sigma_2) - RIP(k,\Sigma_1-\Sigma_2)}.$$

Proof. The proof follows by using the first inequality in (3). ■

The results above only prove a rate of convergence that gives uniform consistency. They are therefore not readily applicable for (asymptotic) inference. For inference about a parameter, an asymptotic distribution result is required, usually asymptotic normality, which is typically proved by way of an asymptotic linear representation. In what follows, we derive a uniform-in-submodel linear representation for the linear regression map. The result in terms of the regression map itself is somewhat abstract; hence, it might be helpful to revisit the usual estimators $\hat\beta_{n,M}$ and $\beta_{n,M}$ from (6) and (10) to understand what kind of representation is possible. From the definition of $\hat\beta_{n,M}$, we have

$$\hat\Sigma_n(M)\hat\beta_{n,M} = \hat\Gamma_n(M) \quad \Rightarrow \quad \hat\Sigma_n(M)\left(\hat\beta_{n,M} - \beta_{n,M}\right) = \hat\Gamma_n(M) - \hat\Sigma_n(M)\beta_{n,M}.$$

Assuming $\hat\Sigma_n(M)$ and $\Sigma_n(M)$ are close, one would expect

$$\left\|\hat\beta_{n,M} - \beta_{n,M} - [\Sigma_n(M)]^{-1}\left(\hat\Gamma_n(M) - \hat\Sigma_n(M)\beta_{n,M}\right)\right\|_2 \approx 0. \tag{17}$$

Note, by substituting all the definitions, that

$$[\Sigma_n(M)]^{-1}\left(\hat{\Gamma}_n(M) - \hat{\Sigma}_n(M)\beta_{n,M}\right) = \frac{1}{n}\sum_{i=1}^{n}[\Sigma_n(M)]^{-1}X_i(M)(Y_i - X_i^\top(M)\beta_{n,M}).$$

This being an average (a linear functional), the left-hand side quantity in (17) is called the linear representation error. Now, using the same argument and substituting Σ_1 and Σ_2 for $\hat{\Sigma}_n$ and Σ_n, respectively, we get the following result. Recall the notations $S_{2,k}(\cdot,\cdot)$ and $\Lambda(\cdot,\cdot)$ from equations (14) and (15).

THEOREM 3.3 (Uniform linear representation). *Let $k \geq 1$ be any integer such that*

$$RIP(k, \Sigma_1 - \Sigma_2) \leq \Lambda(k; \Sigma_2).$$

Then, for all submodels $M \in \mathcal{M}(k)$,

$$\begin{aligned}\left\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2) - [\Sigma_2(M)]^{-1}(\Gamma_1(M) - \Sigma_1(M)\beta_M(\Sigma_2,\Gamma_2))\right\|_2 \\ \leq \frac{RIP(k,\Sigma_1-\Sigma_2)}{\Lambda(k;\Sigma_2)}\left\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\right\|_2. \end{aligned} \tag{18}$$

Furthermore, using Theorem 3.1, we get

$$\begin{aligned}\sup_{M\in\mathcal{M}(k)}\left\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2) - [\Sigma_2(M)]^{-1}(\Gamma_1(M) - \Sigma_1(M)\beta_M(\Sigma_2,\Gamma_2))\right\|_2 \\ \leq \frac{RIP(k,\Sigma_1-\Sigma_2)}{\Lambda(k;\Sigma_2)}\frac{\mathcal{D}(k,\Gamma_1-\Gamma_2) + RIP(k,\Sigma_1-\Sigma_2)S_{2,k}(\Sigma_2,\Gamma_2)}{\Lambda(k;\Sigma_2) - RIP(k,\Sigma_1-\Sigma_2)}. \end{aligned} \tag{19}$$

Proof. From the definition (11) of $\beta_M(\Sigma,\Gamma)$, we have

$$\begin{aligned}\Sigma_1(M)\beta_M(\Sigma_1,\Gamma_1) - \Gamma_1(M) = 0, \\ \Sigma_2(M)\beta_M(\Sigma_2,\Gamma_2) - \Gamma_2(M) = 0. \end{aligned} \tag{20}$$

Adding and subtracting $\beta_M(\Sigma_2,\Gamma_2)$ from $\beta_M(\Sigma_1,\Gamma_1)$ in (20), it follows that

$$\Sigma_1(M)\left(\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\right) = \Gamma_1(M) - \Sigma_1(M)\beta_M(\Sigma_2,\Gamma_2).$$

Now, adding and subtracting $\Sigma_2(M)$ from $\Sigma_1(M)$ in this equation, we get

$$\begin{aligned}(\Sigma_2(M) - \Sigma_1(M))\left(\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\right) \\ = \Sigma_2(M)\left(\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\right) - [\Gamma_1(M) - \Sigma_1(M)\beta_M(\Sigma_2,\Gamma_2)]. \end{aligned} \tag{21}$$

The right-hand side is almost the quantity we need to bound to establish the result. Multiplying both sides of the equation by $[\Sigma_2(M)]^{-1}$ and then applying the euclidean norm implies that, for $M \in \mathcal{M}(k)$,

$$\begin{aligned}\left\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2) - [\Sigma_2(M)]^{-1}\{\Gamma_1(M) - \Sigma_1(M)\beta_M(\Sigma_2,\Gamma_2)\}\right\|_2 \\ \leq \frac{\|\Sigma_1(M) - \Sigma_2(M)\|_{op}}{\Lambda(k;\Sigma_2)} \|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\|_2 .\end{aligned}$$

This proves the first part of the result. The second part of the result follows by the application of Theorem 3.1. ■

Remark 3.2 (Matching lower bounds). The bound (18) only proves an upper bound. It can, however, be seen from equation (21) that, for any $M \in \mathcal{M}(k)$,

$$\begin{aligned}\left\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2) - [\Sigma_2(M)]^{-1}(\Gamma_1(M) - \Sigma_1(M)\beta_M(\Sigma_2,\Gamma_2))\right\|_2 \\ = \left\|[\Sigma_2(M)]^{-1}(\Sigma_1(M) - \Sigma_2(M))(\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2))\right\|_2 \\ \geq C_*(k,\Sigma_2)\Lambda(k,\Sigma_1-\Sigma_2)\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\|_2 ,\end{aligned}$$

where

$$C_*(k,\Sigma_2) := \min_{M\in\mathcal{M}(k)} \lambda_{\min}\left([\Sigma_2(M)]^{-1}\right) = [\mathrm{RIP}(k,\Sigma_2)]^{-1} .$$

Recall, from equations (13) and (14), that

$$\mathrm{RIP}(k,\Sigma_2) = \sup_{M\in\mathcal{M}(k)} \|\Sigma_2(M)\|_{op} \quad \text{and}$$

$$\Lambda(k,\Sigma_1-\Sigma_2) = \inf_{\theta\in\mathbb{R}^p,\,\|\theta\|_0\leq k} \frac{\|(\Sigma_1-\Sigma_2)\theta\|_2}{\|\theta\|_2}.$$

If the minimal and maximal k-sparse singular values of $\Sigma_1 - \Sigma_2$ are of the same order, then the upper and lower bounds for the linear representation error match up to the order under the additional assumption that the minimal and maximal sparse eigenvalues of Σ_2 are of the same order. ◊

Remark 3.3 (Improved $\|\cdot\|_2$-error bounds). Uniform linear representation error bounds (18) and (19) prove more than just a linear representation. These bounds allow us to improve the bounds provided for uniform L_2-consistency. Bound (18) is of the form

$$\|u-v\|_2 \leq \delta\|u\|_2 \quad \Rightarrow \quad (1-\delta)\|u\|_2 \leq \|v\|_2 \leq (1+\delta)\|u\|_2 .$$

Therefore, assuming $\mathrm{RIP}(k,\Sigma_1-\Sigma_2) \leq \Lambda(k;\Sigma_2)/2$, it follows that, for all $M \in \mathcal{M}(k)$,

$$\begin{aligned}\frac{1}{2}\left\|[\Sigma_2(M)]^{-1}(\Gamma_1(M) - \Sigma_1(M)\beta_M(\Sigma_2,\Gamma_2))\right\|_2 \\ \leq \|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\|_2 \\ \leq 2\left\|[\Sigma_2(M)]^{-1}(\Gamma_1(M) - \Sigma_1(M)\beta_M(\Sigma_2,\Gamma_2))\right\|_2 . \qquad \textbf{(22)}\end{aligned}$$

This is a more precise result than informed by Theorem 3.1, because here we characterize the estimation error exactly up to a factor of 2. Also, note that in case of the least-squares estimator and target, $\hat{\beta}_{n,M}$ and $\beta_{n,M}$, the upper and lower bounds here are euclidean norms of averages of random vectors. Dealing with linear functionals like averages is much simpler than dealing with nonlinear functionals such as $\hat{\beta}_{n,M}$.

If $\mathrm{RIP}(k, \Sigma_1 - \Sigma_2)$ converges to zero, then the right-hand side of bound (18) is of smaller order than both the terms appearing on the left-hand side (which are the same as those appearing in (22)). This means that the linear representation error is of strictly smaller order than the estimator error simultaneously over all $M \in \mathcal{M}(k)$. ◊

Remark 3.4 (Alternative to RIP). A careful inspection of the proof of Theorems 3.1 and 3.3 reveals that the bounds can be written in terms of

$$\sup_{M \in \mathcal{M}(k)} \left\| [\Sigma_2(M)]^{-1/2} \Sigma_1(M) [\Sigma_2(M)]^{-1/2} - I_{|M|} \right\|_{op},$$

instead of $\mathrm{RIP}(k, \Sigma_1 - \Sigma_2)$. Here, $I_{|M|}$ is the identity matrix in $\mathbb{R}^{|M| \times |M|}$. Bounding this quantity might not require a bounded condition number of Σ_2; however, we will only deal with $\mathrm{RIP}(k, \Sigma_1 - \Sigma_2)$ in the following sections for convenience. ◊

Summarizing all the results in this section, it is sufficient to control

$$\mathrm{RIP}(k, \Sigma_1 - \Sigma_2) \quad \text{and} \quad \mathcal{D}(k, \Gamma_1 - \Gamma_2)$$

to derive uniform-in-submodel results in any linear-regression-type problem. In this respect, these are the norms in which one should measure the accuracy of the Gram matrix and the inner product of covariates and response. Hence, if one wishes to use shrinkage estimators, for example, because Σ and Γ are high-dimensional "objects," then the estimation accuracy should be measured with respect to RIP and $\mathcal{D}$ for uniform-in-submodel-type results.

3.3. Applications of the Linear Regression Map

Before proceeding to the rates of convergence of these error norms for independent and dependent data, we describe the importance of defining the linear regression map with general matrices instead of just Gram matrices. The generality achieved so far would be worthless if no interesting applications existed. The goal now is to provide a few such interesting examples.

1. *Heavy-Tailed Observations:* The $\mathrm{RIP}(\cdot, \cdot)$-norm is a supremum over all submodels of size k or less; hence, the supremum is over

$$\sum_{s=1}^{k} \binom{p}{s} \leq \sum_{s=1}^{k} \frac{p^s}{s!} = \sum_{s=1}^{k} \frac{k^s}{s!} \left(\frac{p}{k}\right)^s \leq \left(\frac{ep}{k}\right)^k$$

number of submodels. This bound is polynomial in the total number of covariates but is exponential in the size of the largest submodel under consideration. Therefore, if the total number of covariates p is allowed to diverge, then the question we are interested in is inherently high-dimensional. If the usual Gram matrices are used, then

$$\mathrm{RIP}(k, \hat{\Sigma}_n - \Sigma_n) = \sup_{|M| \le k} \left\| \hat{\Sigma}_n(M) - \Sigma_n(M) \right\|_{op};$$

hence, RIP in this case is the supremum in the order of $(ep/k)^k$ many averages. As is well-understood from the literature on concentration of measure or even the union bound, one would require exponential tails on the initial random vectors to allow a good control on $\mathrm{RIP}(\cdot, \cdot)$ if the usual Gram matrix is used. Does this mean that the situation is hopeless if the initial random vectors do not have exponential tails? The short answer is "not necessarily." Viewing the matrix Σ_n (the "population" Gram matrix) as a target, there have been many variations of sample mean Gram matrix estimators that are shown to provide exponential tails even though the initial observations are heavy tailed. See, for example, Catoni (2012), Wei and Minsker (2017), and Catoni and Giulini (2017), along with the references therein, for more details on a specific estimator and its properties. It should be noted that these authors do not study the estimator accuracy with respect to the RIP-norm.

2. *Outlier Contamination:* Real data, more often than not, are contaminated with outliers, and it is a difficult problem to remove or downweight observations when contamination is present. Robust statistics provide estimators that can ignore or downweight the observations suspected to be outliers and yet perform comparably when there is no contamination present in the data. Some simple examples include entrywise medians or trimmed means. See Minsker (2015) and the references therein for some more examples. Almost none of these estimators are simple averages but behave regularly in the sense that they can be expressed as averages up to a negligible asymptotic remainder term. Chen, Caramanis, and Mannor (2013) provide a simple estimator of the Gram matrix under adversarial corruption and casewise contamination.
3. *Indirect Observations:* This example is taken from Loh and Wainwright (2012). The setting is as follows. Instead of observing the real random vectors (X_1, Y_1), $\ldots$, (X_n, Y_n), we observe a sequence $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ with Z_i linked with X_i via some conditional distribution, that is, for $1 \le i \le n$,

   $$Z_i \sim Q(\cdot | X_i).$$

   As discussed on page 4 of Loh and Wainwright (2012), this setting includes some interesting cases such as missing data and noisy covariates. A brief hint of the settings is given below:
   - If $Z_i = X_i + W_i$ where W_i is independent of X_i and has mean zero with a known covariance matrix.

– For some fraction $\rho \in [0, 1)$, we observe a random vector $Z_i \in \mathbb{R}^p$ such that for each component j, we independently observe $Z_i(j) = X_i(j)$ with probability $1 - \rho$ and $Z_i(j) = *$ with probability ρ. (Here, $*$ means a missing value.)
– If $Z_i = X_i \odot u_i$, where $u_i \in \mathbb{R}^p$ is again a random vector independent of X_i and $\odot$ is the Hadamard (coordinatewise) product. The problem of missing data is a special case.

On page 6, Loh and Wainwright (2012) provide various estimators in place of $\hat{\Sigma}_n$ in (5). The assumption in Lemma 12 of Loh and Wainwright (2012) is essentially a bound on the RIP-norm in our notation, and they verify this assumption in all the examples above. Hence, all our results in this section apply to these settings.

3.4. Application of Deterministic Inequalities to OLS

In the following two sections, we prove finite sample nonasymptotic bounds for $\mathrm{RIP}(k, \Sigma_1 - \Sigma_2)$ and $\mathcal{D}(k, \Gamma_1 - \Gamma_2)$ when

$$\Sigma_1 = \hat{\Sigma}_n, \quad \Sigma_2 = \Sigma_n \quad \text{and} \quad \Gamma_1 = \hat{\Gamma}_n, \quad \Gamma_2 = \Gamma_n.$$

See equations (5) and (8). For convenience, we rewrite Theorem 3.3 for this setting. Also, for notational simplicity, let

$$\Lambda_n(k) := \Lambda(k, \Sigma_n), \ \mathrm{RIP}_n(k) := \mathrm{RIP}(k, \hat{\Sigma}_n - \Sigma_n) \quad \text{and} \quad \mathcal{D}_n(k) := \mathcal{D}(k, \hat{\Gamma}_n - \Gamma_n). \tag{23}$$

Recall the definition of $\hat{\beta}_{n,M}$, $\beta_{n,M}$, and $S_{2,k}$ from (7), (10), and (15).

THEOREM 3.4. *Let $k \geq 1$ be any integer such that $RIP_n(k) \leq \Lambda_n(k)$. Then, for all submodels $M \in \mathcal{M}(k)$,*

$$\sup_{M \in \mathcal{M}(k)} \left\| \hat{\beta}_{n,M} - \beta_{n,M} - \frac{1}{n} \sum_{i=1}^{n} [\Sigma_n(M)]^{-1} X_i(M) \left(Y_i - X_i^\top(M) \beta_{n,M} \right) \right\|_2 \leq \frac{RIP_n(k)}{\Lambda_n(k)} \left(\frac{\mathcal{D}_n(k) + RIP_n(k) S_{2,k}(\Sigma_n, \Gamma_n)}{\Lambda_n(k) - RIP_n(k)} \right).$$

Recall here that Γ_n and Σ_n are *nonrandom* vectors/matrices given in (8). So, Theorem 3.4 (which is still a deterministic inequality) can be used to prove an asymptotic uniform linear representation.

Remark 3.5 (Nonuniform bounds). The bound above applies for any k satisfying the assumption $\mathrm{RIP}_n(k) \leq \Lambda_n(k)$. Noting that, for $M \in \mathcal{M}(k)$, $\mathrm{RIP}_n(|M|) \leq \mathrm{RIP}_n(k)$ as well as $\Lambda_n(|M|) \geq \Lambda_n(k)$, Theorem 3.4 implies that

$$\left\| \hat{\beta}_{n,M} - \beta_{n,M} - \frac{1}{n} \sum_{i=1}^{n} [\Sigma_n(M)]^{-1} X_i(M) \left(Y_i - X_i^\top(M) \beta_{n,M} \right) \right\|_2 \leq \frac{\mathrm{RIP}_n(|M|)}{\Lambda_n(|M|)} \left(\frac{\mathcal{D}_n(|M|) + \mathrm{RIP}_n(|M|) S_{2,|M|}(\Sigma_n, \Gamma_n)}{\Lambda_n(|M|) - \mathrm{RIP}_n(|M|)} \right).$$

The point made here is that even though the bound in Theorem 3.4 only uses the maximal submodel size, it can recover submodel size-dependent bounds, because the result is proved for every k. ◊

Remark 3.6 (Postselection consistency). One of the main aspects of our results is in proving consistency of the least-squares linear regression estimator after data exploration. Suppose a random submodel $\hat{M}$ chosen based on data satisfies $|\hat{M}| \le k$ with probability converging to one, that is, $\mathbb{P}(\hat{M} \in \mathcal{M}(k)) \to 1$. Then, with probability converging to one,

$$\left\| \hat{\beta}_{n,\hat{M}} - \beta_{n,\hat{M}} \right\|_2 \le \sup_{M \in \mathcal{M}(k)} \left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_2.$$

A similar bound also holds for the linear representation error. Therefore, the uniform-in-submodel results above allow us to prove consistency and asymptotic normality of the least-squares linear regression estimator after data exploration. See Belloni and Chernozhukov (2013) for related applications and methods of choosing the random submodel $\hat{M}$. ◊

Remark 3.7 (Bounding $S_{2,k}$). As shown in Remark 3.1, for the setting of averages,

$$S_{2,k}(\Sigma_n, \Gamma_n) \le \left(\frac{1}{n\Lambda_n(k)} \sum_{i=1}^{n} \mathbb{E}\left[Y_i^2 \right] \right)^{1/2}. \tag{24}$$

The quantity on the right-hand side of (24) is of the order $\Lambda_n^{-1/2}(k)$ under the assumption of bounded second moments of the Y_i's. Therefore, we will not further write $S_{2,k}$ explicitly and just use $\Lambda_n^{-1/2}(k)$ instead. ◊

4. RATES FOR INDEPENDENT OBSERVATIONS

In this section, we derive bounds for $\mathrm{RIP}_n(k)$ and $\mathcal{D}_n(k)$ defined in (23) under the assumption of independence and weak exponential tails. The setting is as follows. Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are a sequence of independent random vectors in $\mathbb{R}^p \times \mathbb{R}$. Consider the following assumptions:

(MExp) Assume that there exist positive numbers $\alpha > 0$, and $K_{n,p} > 0$, such that

$$\max_{1 \le j \le p} \max \left\{ \|X_i(j)\|_{\psi_\alpha}, \|Y_i\|_{\psi_\alpha} \right\} \le K_{n,p} \quad \text{for all} \quad 1 \le i \le n.$$

(JExp) Assume that there exist positive numbers $\alpha > 0$, and $K_{n,p} > 0$, such that

$$\max \left\{ \left\| X_i^\top \theta \right\|_{\psi_\alpha}, \|Y_i\|_{\psi_\alpha} \right\} \le K_{n,p} \quad \text{for all} \quad \theta \in \mathbb{R}^p,\ \|\theta\|_2 \le 1,\ 1 \le i \le n.$$

Recall that $X_i(j)$ means the jth coordinate of X_i. The notation $\|\cdot\|_{\psi_\alpha}$ refers to a quasinorm defined by

$$\|W\|_{\psi_\alpha} := \inf\left\{C > 0 : \mathbb{E}\left[\exp\left(\frac{|W|^\alpha}{C^\alpha}\right)\right] \le 2\right\},$$

for any random variable W. Random variables W satisfying $\|W\|_{\psi_\alpha} < \infty$ are referred to as sub-Weibull of order α, because $\|W\|_{\psi_\alpha} < \infty$ implies that, for all $t \ge 0$,

$$\mathbb{P}(|W| \ge t) \le 2\exp\left(-\frac{t^\alpha}{\|W\|_{\psi_\alpha}^\alpha}\right),$$

where the right-hand side resembles the survival function of a Weibull random variable of order $\alpha > 0$ (see Kuchibhotla and Chakrabortty (2020) for more details). The special cases $\alpha = 1, 2$ are very much used in the high-dimensional literature as assumed tail behaviors. A random variable W satisfying $\|W\|_{\psi_\alpha} < \infty$ with $\alpha = 2$ is called sub-Gaussian, and with $\alpha = 1$ it is called subexponential (see van der Vaart and Wellner (1996) for more details).

It is easy to see that Assumption (JExp) implies Assumption (MExp). We refer to Assumption (MExp) as a marginal assumption and Assumption (JExp) as a joint assumption. It should be noted that Assumption (JExp) is much stronger than Assumption (MExp), because Assumption (JExp) implies that the coordinates of X_i should be "almost" independent (see Chapter 3 of Vershynin (2018) and Kuchibhotla and Chakrabortty (2020) for further discussion).

The following results bound $\mathcal{D}_n(k)$ and $\mathrm{RIP}_n(k)$ based on Theorem A.1 in Appendix A. Because $\mathrm{RIP}_n(k)$ involves operator norms over k-sparse unit balls, we will bound it using ε-nets for the union of these unit balls. This will also be useful for bounding $\mathcal{D}_n(k)$. Before stating the results, we need the following preliminary calculations and notations. For any set K with metric $d(\cdot,\cdot)$, a set $\mathcal{N}$ is called a γ-net of K with respect to d if $\mathcal{N} \subset K$, and for any $z \in K$, there exists an $x \in \mathcal{N}$ such that $d(x,z) \le \gamma$. Let $\|\cdot\|_2$ denote the euclidean norm and define the d-dimensional unit ball by

$$\mathcal{B}_{2,d} := \left\{x \in \mathbb{R}^d : \|x\|_2 \le 1\right\}.$$

Let $\mathcal{N}_d(\varepsilon)$ represent an ε-net of $\mathcal{B}_{2,d}$ with respect to the euclidean norm. Define the k-sparse subset of the unit ball in $\mathbb{R}^p$ as

$$\Theta_k := \{\theta \in \mathbb{R}^p : \|\theta\|_0 \le k,\ \|\theta\|_2 \le 1\}. \tag{25}$$

With some abuse of notation, a disjoint decomposition of Θ_k can be written as

$$\Theta_k = \bigcup_{s=1}^{k} \bigcup_{|M|=s} \mathcal{B}_{2,s}.$$

The last union includes repetition of $\mathcal{B}_{2,s}$ as subsets of $\mathbb{R}^p$ with unequal supports. Using this decomposition, it follows that a $\frac{1}{4}$-net $\mathcal{N}(\varepsilon, \Theta_k)$ of Θ_k with respect to the euclidean norm on $\mathbb{R}^p$ can be chosen to satisfy

$$\mathcal{N}(\varepsilon, \Theta_k) \subseteq \bigcup_{s=1}^{k} \bigcup_{|M|=s} \mathcal{N}_s(\varepsilon),$$

and, hence, can be bounded in cardinality by

$$|\mathcal{N}(\varepsilon, \Theta_k)| \leq \sum_{s=1}^{k} \binom{p}{s} |\mathcal{N}_s(\varepsilon)|.$$

Applying Lemma 4.1 of Pollard (1990), it follows that

$$|\mathcal{N}_s(\varepsilon)| \leq (1+\varepsilon^{-1})^s \quad \Rightarrow \quad |\mathcal{N}(\varepsilon, \Theta_k)| \leq \sum_{s=1}^{k} \binom{p}{s} (1+\varepsilon^{-1})^s \leq \left(\frac{(1+\varepsilon^{-1})ep}{k} \right)^k.$$

(Lemma 4.1 of Pollard (1990) provides the bound on the covering number to be $(\frac{3}{\varepsilon})^d$, but it can be improved from the proof to $(1+\frac{1}{\varepsilon})^d$.) Here, one can choose the elements of the covering set $\mathcal{N}_s(\varepsilon)$ to be s-sparse in $\mathbb{R}^p$. See Lemma 3.3 of Plan and Vershynin (2013) for a similar result. Based on these calculations and the covering set $\mathcal{N}(\varepsilon, \Theta_k)$, we bound $\mathcal{D}_n(k)$ and $\mathrm{RIP}_n(k)$ by a finite maximum of mean-zero averages.

Observe that

$$\begin{aligned}
\mathcal{D}_n(k) &= \sup_{\theta \in \Theta_k} \theta^\top \left(\hat{\Gamma}_n - \Gamma_n \right) \\
&\leq \sup_{\alpha \in \mathcal{N}(1/2, \Theta_k)} \alpha^\top \left(\hat{\Gamma}_n - \Gamma_n \right) + \sup_{\beta \in \Theta_k/2} \beta^\top \left(\hat{\Gamma}_n - \Gamma_n \right) \\
&= \sup_{\alpha \in \mathcal{N}(1/2, \Theta_k)} \alpha^\top \left(\hat{\Gamma}_n - \Gamma_n \right) + \frac{1}{2} \sup_{\beta \in \Theta_k} \beta^\top \left(\hat{\Gamma}_n - \Gamma_n \right).
\end{aligned}$$

Therefore,

$$\mathcal{D}_n(k) \leq 2 \sup_{\theta \in \mathcal{N}(1/2, \Theta_k)} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \theta^\top X_i Y_i - \mathbb{E}\left[\theta^\top X_i Y_i \right] \right\} \right|. \tag{26}$$

It is clear that the bound is sharp up to a constant factor. By a similar calculation, it can be shown that

$$\mathrm{RIP}_n(k) \leq 2 \sup_{\theta \in \mathcal{N}(1/4, \Theta_k)} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \left(X_i^\top \theta \right)^2 - \mathbb{E}\left[\left(X_i^\top \theta \right)^2 \right] \right\} \right|. \tag{27}$$

See Lemma 2.2 of Vershynin (2012) for a derivation. Importantly, independence of the random vectors is not used in any of these calculations. Replacing the continuous supremum by a finite maximum works irrespective of how the random vectors are distributed.

As an immediate corollary, we get the following rate of convergence results.

THEOREM 4.1. *Define, for* $k \geq 1$,

$$\Upsilon^{\Gamma}_{n,k} := \sup_{\theta \in \Theta_k} \frac{1}{n} \sum_{i=1}^{n} Var\left(\theta^\top X_i Y_i\right), \quad \textit{and} \quad \Upsilon^{\Sigma}_{n,k} := \sup_{\theta \in \Theta_k} \frac{1}{n} \sum_{i=1}^{n} Var\left(\left(\theta^\top X_i\right)^2\right).$$

Then, the following rates of convergence hold if $K_{n,p} = O(1)$*:*

(a) *Under Assumption (MExp),*

$$\mathcal{D}_n(k) = O_p\left(\sqrt{\frac{\Upsilon^{\Gamma}_{n,k} k \log(ep/k)}{n}} + \frac{k^{1/2} (\log n)^{2/\alpha} (k \log(ep/k))^{1/T_1(\alpha/2)}}{n}\right),$$

$$RIP_n(k) = O_p\left(\sqrt{\frac{\Upsilon^{\Sigma}_{n,k} k \log(ep/k)}{n}} + \frac{k (\log n)^{2/\alpha} (k \log(ep/k))^{1/T_1(\alpha/2)}}{n}\right).$$

Here, $T_1(\alpha) = \min\{\alpha, 1\}$.

(b) *Under Assumption (JExp),*

$$\mathcal{D}_n(k) = O_p\left(\sqrt{\frac{\Upsilon^{\Gamma}_{n,k} k \log(ep/k)}{n}} + \frac{(\log n)^{2/\alpha} (k \log(ep/k))^{1/T_1(\alpha/2)}}{n}\right),$$

$$RIP_n(k) = O_p\left(\sqrt{\frac{\Upsilon^{\Sigma}_{n,k} k \log(ep/k)}{n}} + \frac{(\log n)^{2/\alpha} (k \log(ep/k))^{1/T_1(\alpha/2)}}{n}\right).$$

For simplicity, we provide here only rates of convergence. A more precise tail bound is given in Theorem A.2 of Appendix A.

Remark 4.1 (Simplified rates of convergence). In most cases, the second term in the rate of convergence is of lower order than the first term. Hence, under both the assumptions (MExp) and (JExp), we get

$$\mathcal{D}_n(k) = O_p\left(\sqrt{\frac{\Upsilon^{\Gamma}_{n,k} k \log(ep/k)}{n}}\right) \quad \text{and} \quad \text{RIP}_n(k) = O_p\left(\sqrt{\frac{\Upsilon^{\Sigma}_{n,k} k \log(ep/k)}{n}}\right).$$

We believe these to be optimal, because if X and Y are independent and jointly Gaussian, then the rates would be $\sqrt{k \log(ep/k)/n}$; see Theorem 3.3 of Cai and Yuan (2012) and Lemma 15 of Loh and Wainwright (2012) for related results. $\diamondsuit$

A direct application of Theorem 4.1 to Theorem 3.4 implies the following uniform linear representation result for linear regression under independence. Recall the notation $\Lambda_n(k)$ from (23) and also $\hat{\beta}_{n,M}$ and $\beta_{n,M}$ from (7) and (10).

THEOREM 4.2. *If* $(\Lambda_n(k))^{-1} = O(1)$ *as* $n,p \to \infty$, *then the following rates of convergence hold as* $n \to \infty$:

(a) *Under Assumption (MExp),*

$$\sup_{M\in\mathcal{M}(k)} \left\|\hat{\beta}_{n,M} - \beta_{n,M}\right\|_2 = O_p\left(\sqrt{\frac{\Upsilon^{\Gamma}_{n,k} k\log(ep/k)}{n}} + K^2_{n,p}\frac{k(\log n)^{2/\alpha}(k\log(ep/k))^{1/T_1(\alpha/2)}}{n}\right),$$

and

$$\sup_{M\in\mathcal{M}(k)} \left\|\hat{\beta}_{n,M} - \beta_{n,M} - \frac{1}{n}\sum_{i=1}^{n}[\Sigma_n(M)]^{-1}X_i(M)\left(Y_i - X_i^{\top}(M)\beta_{n,M}\right)\right\|_2 = O_p\left(\frac{\max\{\Upsilon^{\Gamma}_{n,k},\Upsilon^{\Sigma}_{n,k}\}k\log(ep/k)}{n} + K^4_{n,p}\frac{k^2(\log n)^{4/\alpha}(k\log(ep/k))^{2/T_1(\alpha/2)}}{n^2}\right).$$

(b) *Under Assumption (JExp),*

$$\sup_{M\in\mathcal{M}(k)} \left\|\hat{\beta}_{n,M} - \beta_{n,M}\right\|_2 = O_p\left(\sqrt{\frac{\Upsilon^{\Gamma}_{n,k} k\log(ep/k)}{n}} + K^2_{n,p}\frac{(\log n)^{2/\alpha}(k\log(ep/k))^{1/T_1(\alpha/2)}}{n}\right),$$

and

$$\sup_{M\in\mathcal{M}(k)} \left\|\hat{\beta}_{n,M} - \beta_{n,M} - \frac{1}{n}\sum_{i=1}^{n}[\Sigma_n(M)]^{-1}X_i(M)\left(Y_i - X_i^{\top}(M)\beta_{n,M}\right)\right\|_2 = O_p\left(\frac{\max\{\Upsilon^{\Gamma}_{n,k},\Upsilon^{\Sigma}_{n,k}\}k\log(ep/k)}{n} + K^4_{n,p}\frac{(\log n)^{4/\alpha}(k\log(ep/k))^{2/T_1(\alpha/2)}}{n^2}\right).$$

Remark 4.2 (Simplified rates of convergence). The result can be made much more precise by giving the exact tail bound for all the quantities using the exact result of Theorem A.2. We leave the details to the reader. From Theorem 4.2, it is clear that if $k\log(ep/k)^{2/T_1(\alpha)} = o(n)$, then the least-squares linear regression estimator is uniformly consistent at the rate of $\sqrt{k\log(ep/k)/n}$, which is well known to be the minimax optimal rate of convergence for high-dimensional linear regression estimators under a true linear model with a sparse parameter vector. We conjecture these rates to be optimal. However, we have not derived minimax rates for this problem. Also, our results are uniform over all probability distributions of the random vectors (X_i, Y_i) satisfying either of the Assumptions (MExp) or (JExp) with $K_{n,p} \le K$ for some fixed constant $K < \infty$. ◊

Remark 4.3 (Fixed covariates). The results in this section do not require any special properties of the data generating distribution such as linearity and Gaussianity. The results only require independence of random vectors with weak exponential tails, but it is *not* assumed that (X_i, Y_i) have identical distributions for $1 \le i \le n$.

It is worth mentioning a special case of our setting that is popular in the classical as well as modern linear regression literature: the setting of fixed covariates. As explained in Buja et al. (2019), this assumption has its roots in the ancillarity theory assuming the truth of a linear model. If the covariates are nonstochastic, then

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i X_i^\top\right] = \Sigma_n,$$

so that $\mathrm{RIP}_n(k) = 0$, for all n and k. Therefore, the bounds in Theorem 3.4 become trivial in the sense that the uniform linear representation error becomes zero. The result applies because assumption (MExp) holds with

$$K_{n,p} = \max\{\max_{1\le i\le n} \|X_i\|_\infty, \max_{1\le i\le n} \|Y_i\|_{\psi_\alpha}\}.$$

Also, note from Theorem 3.1 that

$$\sup_{M\in\mathcal{M}(k)} \left\|\hat{\beta}_{n,M} - \beta_{n,M}\right\|_2 \le \frac{\mathcal{D}_n(k)}{\Lambda_n(k)},$$

which again leads to the same rate of convergence $\sqrt{k\log(ep/k)/n}$. An interesting observation here is that there is no dependence on the strength of linear association $S_{2,k}(\Sigma_n, \Gamma_n)$ defined in equation (15) in the case of fixed covariates. ◊

Remark 4.4 (Are the rates optimal?). We believe the rates for the uniform linear representation error to be optimal; cf. Theorem 5.1 of Javanmard et al. (2018). An intuitive reason is as follows. Any symmetric function of independent random variables can be expanded as a sum of degenerate U-statistics of increasing order according to the Hoeffding decomposition; see van Zwet (1984). That is,

$$f(W_1, \ldots, W_n) = \mathcal{U}_{1n} + \mathcal{U}_{2n} + \cdots + \mathcal{U}_{nn},$$

for any symmetric function f of independent random variables $W_1, \ldots, W_n$. Here, $\mathcal{U}_{in}$ represents an ith order degenerate U-statistics.

For the statistic $\hat{\beta}_{n,M} - \beta_{n,M}$, the first-order term $\mathcal{U}_{1n}$ in the decomposition is given by

$$\mathcal{U}_{1n}^{(M)} = \frac{1}{n}\sum_{i=1}^{n} [\Sigma_n(M)]^{-1} X_i(M)\left(Y_i - X_i^\top(M)\beta_{n,M}\right).$$

Hence, the difference $\hat{\beta}_{n,M} - \beta_{n,M} - \mathcal{U}_{1n}^{(M)}$ is of the same order as the second-order U-statistics $\mathcal{U}_{2n}^{(M)}$ next in the decomposition. It is well known that under mild conditions, a second-order degenerate U-statistics is of order $\frac{1}{n}$; see Serfling (1980,

Chap 5) for precise results. Therefore, bounding the supremum of the $\|\cdot\|_2$-norm in the uniform linear representation by

$$2\max_{|M|\le k}\max_{\theta\in\mathbb{R}^{|M|},\,\|\theta\|_2\le 1}\theta^\top\left(\hat\beta_{n,M}-\beta_{n,M}-\mathcal{U}_{1n}^{(M)}\right)\quad\approx\quad 2\max_{|M|\le k}\max_{\theta\in\mathbb{R}^{|M|},\,\|\theta\|_2\le 1}\theta^\top\mathcal{U}_{2n}^{(M)},$$

we see that this is a maximum of at most $(\frac{5ep}{k})^k$ many degenerate U-statistics of order 2, which is expected to be of order $(\log(5ep/k)^k)/n = (k\log(5ep/k))/n$. See de la Peña and Giné (1999) for results about suprema of degenerate U-statistics. ◊

Remark 4.5 (Using covariance matrices instead of Gram matrices). The quantities $\Upsilon_{n,k}^{\Gamma}$ and $\Upsilon_{n,k}^{\Sigma}$ play an important role in determining the exact rates of convergence in Theorem 4.2. Under Assumption (JExp), it can be easily shown that these quantities are of the same order as $K_{n,p}$. In cases where the dimension grows, Assumption (JExp) cannot be justified with nonzero mean of X_i's unless $\|\mathbb{E}[X_i]\|_2 = O(1)$. Under Assumption (MExp), $\Upsilon_{n,k}^{\Gamma}$ and $\Upsilon_{n,k}^{\Sigma}$ can grow with k, and it is hard to pinpoint their growth rate. In many cases, it is reasonable to assume a bounded operator norm of the covariance matrix instead of the second moment (or Gram) matrix. For this reason, it is of interest to analyze the least-squares estimators with centered random vectors. In this case, $\hat\Sigma_n$ and $\hat\Gamma_n$ should be replaced by

$$\hat\Sigma_n^* := \frac{1}{n}\sum_{i=1}^n\left(X_i-\bar X\right)\left(X_i-\bar X\right)^\top\quad\text{and}\quad\hat\Gamma_n^* := \frac{1}{n}\sum_{i=1}^n\left(X_i-\bar X\right)\left(Y_i-\bar Y\right).$$

Here, $\bar X$ and $\bar Y$ represent the sample means of the covariates and the response, respectively. Without the assumption of equality of $\mathbb{E}[X_i]$, for $1\le i\le n$, $\hat\Sigma_n^*$ is not consistent for the covariance matrix of $\bar X$. Define

$$\bar\mu_n^X := \frac{1}{n}\sum_{i=1}^n\mathbb{E}[X_i]\quad\text{and}\quad\bar\mu_n^Y := \frac{1}{n}\sum_{i=1}^n\mathbb{E}[Y_i].$$

It is easy to prove that

$$\begin{aligned}\hat\Sigma_n^* &= \frac{1}{n}\sum_{i=1}^n\left(X_i-\bar\mu_n^X\right)\left(X_i-\bar\mu_n^X\right)^\top-\left(\bar X_n-\bar\mu_n^X\right)\left(\bar X_n-\bar\mu_n^X\right)^\top\\ &= \tilde\Sigma_n-\left(\bar X_n-\bar\mu_n^X\right)\left(\bar X_n-\bar\mu_n^X\right)^\top,\end{aligned}$$

where

$$\tilde\Sigma_n := \frac{1}{n}\sum_{i=1}^n\left(X_i-\bar\mu_n^X\right)\left(X_i-\bar\mu_n^X\right)^\top.$$

Similarly, we get

$$\hat\Gamma_n^* = \tilde\Gamma_n-\left(\bar X-\bar\mu_n^X\right)\left(\bar Y-\bar\mu_n^Y\right),\quad\text{where}\quad\tilde\Gamma_n := \frac{1}{n}\sum_{i=1}^n\left(X_i-\bar\mu_n^X\right)\left(Y_i-\bar\mu_n^Y\right).$$

Note that $\tilde{\Gamma}_n$ and $\tilde{\Sigma}_n$ are averages of independent random vectors and random matrices, and so the theory before applies with the target vector and matrix given by

$$\Gamma_n^* = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left(X_i - \bar{\mu}_n^X\right)\left(Y_i - \bar{\mu}_n^Y\right)\right] \text{ and } \Sigma_n^* = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left(X_i - \bar{\mu}_n^X\right)\left(X_i - \bar{\mu}_n^X\right)^\top\right].$$

It is important to recognize that Theorem 3.4 is not directly applicable, since the forms of $\hat{\Sigma}_n^*$ and $\hat{\Gamma}_n^*$ do not match the structure required. One has to apply Theorem 3.3 to obtain

$$\begin{aligned}\sup_{M\in\mathcal{M}(k)} \Bigg\| \hat{\beta}_M^* - \beta_M^* - \frac{1}{n}\sum_{i=1}^n \left[\Sigma_n^*(M)\right]^{-1}\left(X_i - \bar{\mu}_n^X\right)(M) \\ \left\{Y_i - \bar{\mu}_n^Y - (X_i(M) - \bar{\mu}_n^X(M))^\top \beta_M^*\right\}\Bigg\|_2 \\ \leq \frac{\mathcal{D}(k, \bar{X} - \bar{\mu}_n^X)\left[|\bar{Y} - \bar{\mu}_n^Y| + \mathcal{D}(k, \bar{X} - \bar{\mu}_n^X) S_{2,k}^*\right]}{\Lambda_n^*(k)} \\ + \frac{\mathrm{RIP}_n^*(k)}{\Lambda_n^*(k)} \times \frac{\mathcal{D}_n^*(k) + \mathrm{RIP}_n^*(k) S_{2,k}^*}{\Lambda_n^*(k) - \mathrm{RIP}_n^*(k)},\end{aligned}$$

where

$$\hat{\beta}_M^* := \beta_M(\hat{\Sigma}_n^*, \hat{\Gamma}_n^*), \quad \beta_M^* := \beta_M(\Sigma_n^*, \Gamma_n^*), \quad S_{2,k}^* := S_{2,k}(\Sigma_n^*, \Gamma_n^*),$$

and

$$\mathrm{RIP}_n^*(k) := \mathrm{RIP}(k, \hat{\Sigma}_n^* - \Sigma_n^*), \quad \mathcal{D}_n^*(k) := \mathcal{D}(k, \hat{\Gamma}_n^* - \Gamma_n^*), \quad \text{and}$$
$$\Lambda_n^*(k) := \Lambda(k; \Sigma_n^*).$$

From the calculations presented above, it follows that

$$\begin{aligned}\mathrm{RIP}_n^*(k) &\leq \mathrm{RIP}(k, \tilde{\Sigma}_n - \Sigma_n^*) + \mathcal{D}^2(k, \bar{X} - \bar{\mu}_n^X), \\ \mathcal{D}_n^*(k) &\leq \mathcal{D}(k, \tilde{\Gamma}_n - \Gamma_n^*) + \mathcal{D}(k, \bar{X} - \bar{\mu}_n^X)\left|\bar{Y} - \mu_n^Y\right|.\end{aligned}$$

The right-hand side terms above can be controlled using Theorem A.1. Thus, the linear representation changes when using the sample covariance matrix. See Section 4.1.1 of Kuchibhotla and Chakrabortty (2020) for more details. ◊

5. RATES FOR FUNCTIONALLY DEPENDENT OBSERVATIONS

In this section, we extend all the results presented in the previous section to dependent data. The dependence structure on the observations we use is based on a notion developed by Wu (2005). It is possible to derive these results also under the classical dependence notions like α-, β-, ρ-mixing; however, verifying the mixing assumptions can often be hard and many well-known processes do not satisfy them. See Wu (2005) for more details. It has also been shown that many econometric time series can be studied under the notion of functional dependence; see Wu and Mielniczuk (2010), Liu, Xiao, and Wu (2013), and Wu and Wu (2016). For a study

of dependent processes under a similar framework called L_p-approximability, see Pötscher and Prucha (1997).

The dependence notion of Wu (2005) is written in terms of an input–output process that is easy to analyze in many settings. The process is defined as follows. Let $\{\varepsilon_i, \varepsilon_i' : i \in \mathbb{Z}\}$ denote a sequence of i.i.d. random variables on some measurable space $(\mathcal{E}, \mathcal{B})$. Define the q-dimensional process W_i with causal representation as

$$W_i = G_i(\ldots, \varepsilon_{i-1}, \varepsilon_i) \in \mathbb{R}^q, \tag{28}$$

for some vector-valued function $G_i(\cdot) = (g_{i1}(\cdot), \ldots, g_{iq}(\cdot))$. By Wold representation theorem for stationary processes, this causal representation holds in many cases. Define the nondecreasing filtration

$$\mathcal{F}_i := \sigma(\ldots, \varepsilon_{i-1}, \varepsilon_i).$$

Using this filtration, we also use the notation $W_i = G_i(\mathcal{F}_i)$. To measure the strength of dependence, define, for $r \geq 1$ and $1 \leq j \leq q$, the *functional dependence measure*

$$\delta_{s,r,j} := \max_{1 \leq i \leq n} \left\| W_i(j) - W_{i,s}(j) \right\|_r, \quad \text{and} \quad \Delta_{m,r,j} := \sum_{s=m}^{\infty} \delta_{s,r,j}, \tag{29}$$

where

$$W_{i,s}(j) := g_{ij}(\mathcal{F}_{i,i-s}) \quad \text{with} \quad \mathcal{F}_{i,i-s} := \sigma\left(\ldots, \varepsilon_{i-s-1}, \varepsilon_{i-s}', \varepsilon_{i-s+1}, \ldots, \varepsilon_{i-1}, \varepsilon_i\right). \tag{30}$$

The σ-field $\mathcal{F}_{i,i-s}$ represents a coupled version of $\mathcal{F}_i$. The quantity $\delta_{s,r,j}$ measures the dependence using the distance in terms of $\|\cdot\|_r$-norm between $g_{ij}(\mathcal{F}_i)$ and $g_{ij}(\mathcal{F}_{i,i-s})$. In other words, it is quantifying the impact of changing ε_{i-s} on $g_{ij}(\mathcal{F}_i)$; see Definition 1 of Wu (2005). The *dependence adjusted norm* for the jth coordinate is given by

$$\|\{W(j)\}\|_{r,\nu} := \sup_{m \geq 0} (m+1)^{\nu} \Delta_{m,r,j}, \quad \nu \geq 0.$$

To summarize these measures for the vector-valued process, define

$$\|\{W\}\|_{r,\nu} := \max_{1 \leq j \leq q} \|\{W(j)\}\|_{r,\nu} \quad \text{and} \quad \|\{W\}\|_{\psi_\alpha,\nu} := \sup_{r \geq 2} r^{-1/\alpha} \|\{W\}\|_{r,\nu}.$$

Remark 5.1 (Independent sequences). Any notion of dependence should at least include independent random variables. It might be helpful to understand how independent random variables fit into this framework of dependence. For independent random vectors W_i, the causal representation reduces to

$$W_i = G_i(\ldots, \varepsilon_{i-1}, \varepsilon_i) = G_i(\varepsilon_i) \in \mathbb{R}^q.$$

It is not a function of any of the previous $\varepsilon_j, j < i$. This implies by the definition (30) that

$$W_{i,s} = \begin{cases} G_i(\varepsilon_i) = W_i, & \text{if } s \geq 1, \\ G_i(\varepsilon_i') =: W_i', & \text{if } s = 0. \end{cases}$$

Here, W_i' represents an i.i.d. copy of W_i. Hence,

$$\delta_{s,r,j} = \begin{cases} 0, & \text{if } s \geq 1, \\ \left\| W_i(j) - W_i'(j) \right\|_r \leq 2 \left\| W_i(j) \right\|_r, & \text{if } s = 0. \end{cases}$$

It is now clear that, for any $\nu > 0$,

$$\|\{W\}\|_{r,\nu} = \sup_{m \geq 0} (m+1)^{\nu} \Delta_{m,r} = \Delta_{0,r} \leq 2 \max_{1 \leq j \leq q} \|W_i(j)\|_r .$$

Hence, if the independent sequence W_i satisfies assumption (MExp), then $\|\{W\}\|_{\psi_\alpha,\nu} < \infty$, for all $\nu > 0$, in particular for $\nu = \infty$. Therefore, independence corresponds to $\nu = \infty$. As ν decreases to zero, the random vectors become more and more dependent. ◊

All our results in this section are based on the following tail bound for the maximum of averages of functionally dependent variables which is an extension of Theorem 2 of Wu and Wu (2016). This result is similar to Theorem A.1. For this result, define

$$s(\lambda) := (1/2 + 1/\lambda)^{-1}, \quad \text{and} \quad T_1(\lambda) := \min\{\lambda, 1\} \quad \text{for all} \quad \lambda > 0. \tag{31}$$

THEOREM 5.1. *Suppose $Z_1, \ldots, Z_n$ are random vectors in $\mathbb{R}^q$ with a causal representation such as (28) with mean zero. Assume that, for some $\alpha > 0$ and $\nu > 0$,*

$$\|\{Z\}\|_{\psi_\alpha,\nu} = \sup_{r \geq 2} \sup_{m \geq 0} r^{-1/\alpha} (m+1)^{\nu} \Delta_{m,r} \leq K_{n,q}.$$

Define

$$\Omega_n(\nu) := 2^{\nu} \times \begin{cases} 5/(\nu - 1/2)^3, & \text{if } \nu > 1/2, \\ 2(\log_2 n)^{5/2}, & \text{if } \nu = 1/2, \\ 5(2n)^{(1/2-\nu)}/(1/2-\nu)^3, & \text{if } \nu < 1/2. \end{cases}$$

Then, for all $t \geq 0$, with probability at least $1 - 8e^{-t}$,

$$\max_{1 \leq j \leq q} \left| \sum_{i=1}^{n} Z_i(j) \right| \leq e\sqrt{n} \|\{Z\}\|_{2,\nu} B_\nu \sqrt{t + \log(q+1)} + C_\alpha K_{n,q} (\log n)^{1/s(\alpha)} \Omega_n(\nu) (t + \log(q+1))^{1/T_1(s(\alpha))}.$$

Here, B_ν and C_α are constants depending only on ν and α, respectively.

Proof. The proof follows from Theorem B.1 proved in Appendix B and a union bound. ■

Getting back to the application of uniform-in-submodel results for linear regression, we assume that the random vectors are elements of a causal process with exponential tails. Formally, suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are random vectors in $\mathbb{R}^p \times \mathbb{R}$ satisfying the following assumption:

(DEP) Assume that there exist n vector-valued functions G_i and an i.i.d. sequence $\{\varepsilon_i : i \in \mathbb{Z}\}$ such that

$$W_i := (X_i, Y_i) = G_i(\ldots, \varepsilon_{i-1}, \varepsilon_i) \in \mathbb{R}^{p+1}.$$

Also, for some $\nu, \alpha > 0$,

$$\|\{W\}\|_{\psi_\alpha, \nu} \le K_{n,p} \quad \text{and} \quad \max_{1 \le i \le n} \max_{1 \le j \le p+1} |\mathbb{E}[W_i(j)]| \le K_{n,p}.$$

Based on Remark 5.1, Assumption (DEP) is equivalent to Assumption (MExp) for independent data. For independent random variables, the second part of Assumption (DEP) about the expectations follows from the ψ_α-bound assumption. The reason for this expectation bound in the assumption here is that the functional dependence measure $\delta_{s,r}$ does not have any information about the expectation, since

$$\left\| W_i(j) - W_{i,s}(j) \right\|_r = \left\| (W_i(j) - \mathbb{E}[W_i(j)]) - \left(W_{i,s}(j) - \mathbb{E}\left[W_{i,s}(j)\right]\right) \right\|_r.$$

The coupled random variable $W_{i,s}$ has the same expectation as W_i. Since the quantities we need to bound involve products of random variables, such a bound on the expectations is needed for our analysis.

We are now ready to state the final results of this section. Only results similar to Theorems A.2 and 4.2 are stated. Also, we only state the results under marginal moment assumption, and the version with joint moment assumption can easily be derived based on the proof. These results are based on Theorem 5.1. Recall from inequalities (26) and (27) that

$$\mathcal{D}_n(k) \le 2 \sup_{\theta \in \mathcal{N}(1/2, \Theta_k)} \left| \frac{1}{n} \sum_{i=1}^n \left\{ \theta^\top X_i Y_i - \mathbb{E}\left[\theta^\top X_i Y_i\right] \right\} \right|,$$

$$\mathrm{RIP}_n(k) \le 2 \sup_{\theta \in \mathcal{N}(1/4, \Theta_k)} \left| \frac{1}{n} \sum_{i=1}^n \left(X_i^\top \theta\right)^2 - \mathbb{E}\left[\left(X_i^\top \theta\right)^2 \right] \right|.$$

Note that these quantities involve linear combinations $(\theta^\top X_i)$ and products $(\theta^\top X_i Y_i)$ of functionally dependent random variables. It is clear that all linear combinations and products of functionally dependent random variables have a causal representation, since if $W_i^{(1)} := h_i^{(1)}(\mathcal{F}_i)$ and $W_i^{(2)} := h_i^{(2)}(\mathcal{F}_i)$, then

$$\alpha W_i^{(1)} + \beta W_i^{(2)} = \alpha h_i^{(1)}(\mathcal{F}_i) + \beta h_i^{(2)}(\mathcal{F}_i) \quad \text{and} \quad W_i^{(1)} W_i^{(2)} = h_i^{(1)}(\mathcal{F}_i) h_i^{(2)}(\mathcal{F}_i).$$

Thus, they can be studied under the same framework of dependence. In Lemma B.4, we bound the functional dependence measure of such linear combination and product processes.

For the main results of this section, define, for $\theta \in \Theta_k$ (see (25)),

$$\vartheta_4^{(\Gamma)}(\theta) := \left(\left\| \{\theta^\top X\} \right\|_{4,0} + \max_{1 \le i \le n} \left| \mathbb{E}\left[\theta^\top X_i \right] \right| \right) \|\{Y\}\|_{4,\nu} + \left(\|\{Y\}\|_{4,0} + \max_{1 \le i \le n} |\mathbb{E}[Y_i]| \right) \left\| \{\theta^\top X\} \right\|_{4,\nu},$$

$$\vartheta_4^{(\Sigma)}(\theta) := 2 \left(\left\| \{\theta^\top X\} \right\|_{4,0} + \max_{1 \le i \le n} \left| \mathbb{E}\left[\theta^\top X_i \right] \right| \right) \left\| \{\theta^\top X\} \right\|_{4,\nu}.$$

THEOREM 5.2. *Fix $n, k \ge 1$ and let $t \ge 0$ be any real number. Define*

$$\sqrt{\Upsilon^{\Gamma}_{n,k}} := \sup_{\theta \in \Theta_k} \vartheta_4^{(\Gamma)}(\theta), \quad \textit{and} \quad \sqrt{\Upsilon^{\Sigma}_{n,k}} := \sup_{\theta \in \Theta_k} \vartheta_4^{(\Sigma)}(\theta).$$

Then, under Assumption (DEP), with probability at least $1 - 16e^{-t}$, the following inequalities hold simultaneously:

$$\mathcal{D}_n(k) \le 2eB_\nu \sqrt{\frac{\Upsilon^{\Gamma}_{n,k}(t + k\log(3ep/k))}{n}} + C_\alpha K_{n,p}^2 \frac{k^{1/2} (\log n)^{1/s(\alpha/2)} \Omega_n(\nu) (t + k\log(3ep/k))^{1/T_1(s(\alpha/2))}}{n},$$

and

$$RIP_n(k) \le 2eB_\nu \sqrt{\frac{\Upsilon^{\Gamma}_{n,k}(t + k\log(5ep/k))}{n}} + C_\alpha K_{n,p}^2 \frac{k (\log n)^{1/s(\alpha/2)} \Omega_n(\nu) (t + k\log(5ep/k))^{1/T_1(s(\alpha/2))}}{n}.$$

Here, $T_1(\alpha)$ and $s(\alpha)$ are functions given in (31) and B_ν, C_α are constants depending only on ν and α, respectively.

Proof. By Lemma B.4 and Assumption (DEP), it holds that, for all $\theta \in \Theta_k$,

$$\left\| \{\theta^\top XY\} \right\|_{2,\nu} \le \vartheta_4^{\Gamma}(\theta) \quad \text{and} \quad \left\| \{(\theta^\top X)^2\} \right\|_{2,\nu} \le \vartheta_4^{\Sigma}(\theta).$$

Also, using Lemmas B.3 and B.4, it follows that

$$\sup_{r \ge 2} r^{-2/\alpha} \left\| \{\theta^\top XY\} \right\|_{r,\nu} \le 3k^{1/2} K_{n,p}^2 2^{1/\alpha},$$

$$\sup_{r \ge 2} r^{-2/\alpha} \left\| \{(\theta^\top X)^2\} \right\|_{r,\nu} \le 3k K_{n,p}^2 2^{1/\alpha}.$$

Hence, applying Theorem 5.1, the result is proved. ∎

Theorem 5.2 along with Theorem 3.4 implies the following uniform linear representation result for linear regression under functional dependence. Recall the notation $\Lambda_n(k)$ from equation (23) and also $\hat{\beta}_{n,M}$ and $\beta_{n,M}$ from equations (7) and (10).

THEOREM 5.3. *If* $(\Lambda_n(k))^{-1} = O(1)$ *as* $n, p \to \infty$, *then under Assumption (DEP), the following rates of convergence hold as* $n \to \infty$:

$$\sup_{M \in \mathcal{M}(k)} \left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_2 = O_p\left(\sqrt{\frac{\Upsilon^{\Gamma}_{n,k} k \log(ep/k)}{n}} + K^2_{n,p} \frac{k^{1/2} (\log n)^{1/s(\alpha/2)} \Omega_n(\nu) (k \log(ep/k))^{1/T_1(s(\alpha/2))}}{n} \right),$$

and

$$\sup_{M \in \mathcal{M}(k)} \left\| \hat{\beta}_{n,M} - \beta_{n,M} - \frac{1}{n} \sum_{i=1}^{n} [\Sigma_n(M)]^{-1} X_i(M) \left(Y_i - X_i^{\top}(M) \beta_{n,M} \right) \right\|_2 = O_p\left(\frac{\max\{\Upsilon^{\Gamma}_{n,k}, \Upsilon^{\Sigma}_{n,k}\} k \log(ep/k)}{n} \right) + K^4_{n,p} O_p\left(\frac{k^2 (\log n)^{2/s(\alpha/2)} (k \log(ep/k))^{2/T_1(s(\alpha/2))} \Omega^2_n(\nu)}{n^2} \right).$$

In comparison to Theorem 4.2, the rates attained here are very similar except for two changes:

1. The exponent terms $\alpha/2$ and $T_1(\alpha/2)$ are replaced by $s(\alpha/2)$ and $T_1(s(\alpha/2))$, respectively. This is because of the use of a version of Burkholder's inequality from Rio (2009) in the proof of Theorem B.1.
2. The factor $\Omega_n(\nu)$ in the second-order terms above. This factor is due to the dependence of the process. If $\nu > 1/2$ (which corresponds to "weak" dependence), then $\Omega_n(\nu)$ is of order 1, and for the boundary case $\nu = 1/2$, $\Omega_n(\nu)$ is of order $(\log n)^{5/2}$. In both these cases, the rates obtained for functionally dependent ψ_α-random vectors match very closely the rates obtained for independent $\psi_{s(\alpha)}$-random vectors.

Remark 5.2 (Some comments on Assumption (DEP)). Assumption (DEP) is similar to the one used in Theorem 3.3 of Zhang and Wu (2017) for derivation of a high-dimensional central limit theorem with logarithmic dependence on the dimension p. It is worth mentioning that in their notation, α corresponds to the

functional dependence and ν corresponds to the moment assumption. Also, their assumption is written as

$$\sup_{r\geq 2} \frac{\|\{Z\}\|_{r,\nu}}{r^{\alpha}} < \infty \quad \text{(after swapping the dependence and moment parameters).}$$

Our assumption, however, is written as

$$\sup_{r\geq 2} \frac{\|\{Z\}\|_{r,\nu}}{r^{1/\alpha}} < \infty.$$

Hence, our parameters (α, ν) correspond to their parameters $(1/\nu, \alpha)$. Our assumptions are weaker than those used by Zhang and Cheng (2014). From the discussion surrounding equation (28) there, they require geometric decay of $\Delta_{m,r,j}$, while we only require polynomial decay. Zhang and Wu (2017) only deal with stationary sequences and Zhang and Cheng (2014) allows nonstationarity. Some useful examples verifying the bounds on the functional dependence measure are also provided in Zhang and Cheng (2014). ◊

6. DISCUSSION AND CONCLUSIONS

In this paper, we have proved uniform-in-submodel results for the least-squares linear regression estimator under a model-free framework allowing for the total number of covariates to diverge "almost exponentially" in n. Our results are based on deterministic inequalities. The exact rate bounds are provided when the random vectors are independent and functionally dependent. In both cases, the random variables are assumed to have weak exponential tails to provide logarithmic dependence on the dimension p.

In this paper, we have primarily focused on OLS linear regression. The main results, uniform-in-submodel consistency and linear representation, continue to hold for a large class of M-estimators defined by twice differentiable loss function as shown in Kuchibhotla (2018). The implications of these results are that one can use all the information from all the observations to build a submodel (subset of covariates) and apply a general M-estimation technique on the final model selected. These results can be extended to nondifferentiable loss functions using techniques from empirical process theory, in particular, the stochastic uniform equicontinuity assumption. See, for example, Giessing (2018, Chap. 2) for results under independence.

All of our results are free of the assumption of correctly specified models. Therefore, our results provide a "target" $\beta_{n,M}$ for the estimator $\hat{\beta}_{n,M}$ irrespective of whether M is fixed or random as long as $|M| \leq k$. This implication follows from the uniform-in-submodel feature of the results. The conclusion here is that if the statistician has a target in mind, then all they need to check is if $\beta_{n,M}$ is close to the target they are thinking of.

As mentioned in the beginning of the article, one can rethink high-dimensional linear regression as using high-dimensional data for exploration to find a

"significant" set of variables and then applying the "low-dimensional" linear regression technique. If the exploration is not restricted to a very principled method, then inference can be very difficult. This problem is exactly equivalent to the problem of valid postselection inference. Postselection inference has a rich history in both statistics and econometrics. Leeb and Pötscher (2005, 2006a, 2006b, 2008) have provided several impossibility results regarding the estimation of the distribution of $\widehat{\beta}_{\widehat{M}}$, when $\widehat{M}$ represents the data-dependent selected model. One way to avoid this difficulty is by performing inference for all models simultaneously. The results in this paper allow for the construction of a simultaneous inference procedure using a high-dimensional central limit theorem and multiplier bootstrap; see Bachoc et al. (2019a; 2019b), Kuchibhotla et al. (2021), and Belloni et al. (2018, Sect. 2) for more details. A related exploration will be provided in a future manuscript.

APPENDICES

A. Auxiliary Results for Independent Random Vectors

The following result proves a tail bound for a maximum of the average of mean-zero random variables and follows from Theorem 4 of Adamczak (2008). The result there is only stated for $\alpha \in (0, 1]$; however, the proof can be extended to the case $\alpha > 1$. See the forthcoming paper Kuchibhotla and Chakrabortty (2020) for a clear exposition.

THEOREM A.1. *Suppose $W_1, \ldots, W_n$ are mean-zero independent random vectors in $\mathbb{R}^q, q \geq 1$ such that, for some $\alpha > 0$ and $K_{n,q} > 0$,*

$$\max_{1 \leq i \leq n} \max_{1 \leq j \leq q} \|W_i(j)\|_{\psi_\alpha} \leq K_{n,q}.$$

Define

$$\Gamma_{n,q} := \max_{1 \leq j \leq q} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[W_i^2(j)\right].$$

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

$$\max_{1 \leq j \leq q} \left| \frac{1}{n} \sum_{i=1}^{n} W_i(j) \right| \leq 7\sqrt{\frac{\Gamma_{n,q}(t + \log(2q))}{n}} + \frac{C_\alpha K_{n,q} (\log(2n))^{1/\alpha} (t + \log(2q))^{1/T_1(\alpha)}}{n},$$

where $T_1(\alpha) = \min\{\alpha, 1\}$ and C_α is a constant depending only on α.

Proof. Fix $1 \leq j \leq q$ and apply Theorem 4 of Adamczak (2008) with $\mathcal{F} = \{f\}$ where $f(W_i) = W_i(j)$, for $1 \leq i \leq n$. Then, applying the union bound, the result follows. To extend the result to the case $\alpha > 1$, use Theorem 5 of Adamczak (2008) with $\alpha = 1$ to bound the second part of inequality (8) there. ■

Using Theorem A.1, we get the following results for RIP and $\mathcal{D}$ under independence.

THEOREM A.2. *Fix $n,k \geq 1$ and let $t \geq 0$ be any real number. Then, the following probability statements hold true:*

(a) *Under Assumption (MExp), with probability at least $1-6e^{-t}$, the following two inequalities hold simultaneously:*

$$\mathcal{D}_n(k) \leq 14\sqrt{\frac{\Upsilon^{\Gamma}_{n,k}(t+k\log(3ep/k))}{n}} + C_\alpha K^2_{n,p}\frac{k^{1/2}(\log(2n))^{2/\alpha}(t+k\log(3ep/k))^{1/T_1(\alpha/2)}}{n},$$

and

$$RIP_n(k) \leq 14\sqrt{\frac{\Upsilon^{\Sigma}_{n,k}(t+k\log(5ep/k))}{n}} + C_\alpha K^2_{n,p}\frac{k(\log(2n))^{2/\alpha}(t+k\log(5ep/k))^{1/T_1(\alpha/2)}}{n}.$$

(b) *Under Assumption (JExp), with probability at least $1-6e^{-t}$, the following two inequalities hold simultaneously:*

$$\mathcal{D}_n(k) \leq 14\sqrt{\frac{\Upsilon^{\Gamma}_{n,k}(t+k\log(3ep/k))}{n}} + C_\alpha K^2_{n,p}\frac{(\log(2n))^{2/\alpha}(t+k\log(3ep/k))^{1/T_1(\alpha/2)}}{n},$$

and

$$RIP_n(k) \leq 14\sqrt{\frac{\Upsilon^{\Sigma}_{n,k}(t+k\log(5ep/k))}{n}} + C_\alpha K^2_{n,p}\frac{(\log(2n))^{2/\alpha}(t+k\log(5ep/k))^{1/T_1(\alpha/2)}}{n}.$$

Here, $T_1(\alpha) = \min\{\alpha, 1\}$ and C_α is a constant depending only on α.

Proof. These bounds follow from Theorem A.1 and inequalities (26) and (27). To bound $\mathcal{D}_n(k)$, we take

$$W_i := (\theta^\top X_i Y_i)_{\theta \in \mathcal{N}(1/2,\Theta_k)},$$

in Theorem A.1. Because $|\mathcal{N}(1/2,\Theta_k)| \leq (3ep/k)^k$, the result follows. Similarly for $\mathrm{RIP}_n(k)$, we take

$$W_i := ((\theta^\top X_i)^2)_{\theta \in \mathcal{N}(1/4,\Theta_k)},$$

in Theorem A.1. ■

B. Auxiliary Results for Dependent Random Vectors

In this section, we present a moment bound for sum of functionally dependent mean-zero real-valued random variables. The moment bound here is an extension of Theorem 2 of Wu

and Wu (2016) to random variables with exponential tails. The main distinction is that our moment bound exhibits a part Gaussian behavior. For proving these moment bounds, we need a few preliminary results and notation. Suppose $Z_1 \ldots, Z_n$ are mean-zero real-valued random variables with a causal representation

$$Z_i = g_i(\ldots, \varepsilon_{i-1}, \varepsilon_i), \tag{32}$$

for some real-valued function g_i. We write $\delta_{k,r} = \|Z_i - Z_{i,k}\|_r$. The following proposition bounds the rth moment of Z_i in terms of $\|\{Z\}\|_{r,v}$. This is based on the calculation shown after equation (2.8) in Wu and Wu (2016).

PROPOSITION B.1. *Consider the setting above. If* $\mathbb{E}[Z_i] = 0$, *for* $1 \le i \le n$, *then*

$$\|Z_i\|_r \le \|\{Z\}\|_{r,0} \le \|\{Z\}\|_{r,v}, \quad \textit{for any} \quad r \ge 1 \quad \textit{and} \quad v > 0.$$

Proof. Assuming $\mathbb{E}[Z_i] = 0$, for $1 \le i \le n$, it follows that

$$Z_i = \sum_{\ell=-\infty}^{i} \left(\mathbb{E}\left[Z_i \middle| \mathcal{F}_\ell\right] - \mathbb{E}\left[Z_i \middle| \mathcal{F}_{\ell-1}\right]\right),$$

and so,

$$\|Z_i\|_r \le \sum_{\ell=-\infty}^{i} \left\|\mathbb{E}\left[Z_i \middle| \mathcal{F}_\ell\right] - \mathbb{E}\left[Z_i \middle| \mathcal{F}_{\ell-1}\right]\right\|_r = \sum_{\ell=-\infty}^{i} \left\|\mathbb{E}\left[Z_i - Z_{i,i-\ell} \middle| \mathcal{F}_{-\ell}\right]\right\|_r \le \sum_{\ell=0}^{\infty} \delta_{\ell,r}.$$

The last inequality follows from Jensen's inequality and noting that the last bound equals $\Delta_{0,r}$, it follows that $\|Z_i\|_r \le \Delta_{0,r} = \|\{Z\}\|_{r,0}$. ■

The following lemma provides a bound on the moments of a martingale in terms of the moments of the martingale difference sequence. This result is an improvement over the classical Burkholder's inequality.

LEMMA B.1 (Theorem 2.1 of Rio (2009)). *Let* $\{S_n : n \ge 0\}$ *be a martingale sequence with* $S_0 = 0$ *adapted with respect to some nondecreasing filtration* $\mathcal{F}_n, n \ge 0$. *Let* $X_k = S_k - S_{k-1}$ *denote the corresponding martingale difference sequence. Then, for any* $p \ge 2$,

$$\|S_n\|_p \le \sqrt{p-1} \left(\sum_{k=1}^{n} \|X_k\|_p^2\right)^{1/2}.$$

The following simple calculation is also used in Theorem B.1. Define

$$L := \left\lfloor \frac{\log n}{\log 2} \right\rfloor \quad \text{and} \quad \lambda_\ell := \begin{cases} 3\pi^{-2}\ell^{-2}, & \text{if } 1 \le \ell \le L/2, \\ 3\pi^{-2}(L+1-\ell)^{-2}, & \text{if } L/2 < \ell \le L. \end{cases}$$

LEMMA B.2. *The following inequalities hold true:*

(a) *For any $\beta \geq 0$ and $p \geq 2$,*

$$\sum_{\ell=1}^{L} \frac{1}{\lambda_\ell^p 2^{p\ell\beta}} \leq 2 \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{p\ell\beta}} \leq \begin{cases} (5/\beta^3)^p \left(\pi^2/3\right)^{p+1}, & \text{if } \beta > 0, \\ 2(\log_2 n)^{2p+1} \left(\pi^2/3\right)^{p+1}, & \text{if } \beta = 0. \end{cases}$$

(b) *For any $\beta > 0$ and $p \geq 2$,*

$$\sum_{\ell=1}^{L} \frac{2^{p\ell(1/2-\beta)}}{\lambda_\ell^p} \leq \left(\frac{\pi^2}{3}\right)^{p+1} \begin{cases} (5/(\beta - 1/2)^3)^p, & \text{if } \beta > 1/2, \\ 2(\log_2 n)^{2p+1}, & \text{if } \beta = 1/2, \\ (2n)^{(1/2-\beta)p} (5/(1/2-\beta)^3)^p, & \text{if } \beta < 1/2. \end{cases}$$

Proof. (a) Note that, for any $\beta > 0$,

$$\sup_{\ell > 0} \ell^3 2^{-\ell\beta} = \ell^3 \exp(-(\log 2)\ell\beta) \leq \left(\frac{3}{e\beta \log 2}\right)^3 \leq \frac{5}{\beta^3},$$

and so,

$$\begin{aligned} \left(\frac{3}{\pi^2}\right)^p \sum_{\ell=1}^{L} \frac{1}{\lambda_\ell^p 2^{p\ell\beta}} &= \sum_{\ell=1}^{L/2} \left(\frac{\ell^2}{2^{\ell\beta}}\right)^p + \sum_{\ell=L/2+1}^{L} \left(\frac{(L+1-\ell)^2}{2^{\ell\beta}}\right)^p \\ &\leq \sum_{\ell=1}^{L/2} \left(\frac{\ell^2}{2^{\ell\beta}}\right)^p + 2^{-p\beta} \sum_{\ell=1}^{L/2} \left(\frac{\ell^2}{2^{\ell\beta}}\right)^p \\ &\leq 2 \left(\frac{5}{\beta^3}\right)^p \sum_{\ell=1}^{L/2} \frac{1}{\ell^p} \leq \frac{\pi^2}{3} \left(\frac{5}{\beta^3}\right)^p. \end{aligned}$$

Hence, the result (a) follows. The case $\beta = 0$ follows from the calculation in (b).

(b) If $\beta > 1/2$, then

$$\sum_{\ell=1}^{L} \frac{2^{p\ell(1/2-\beta)}}{\lambda_\ell^p} = \sum_{\ell=1}^{L} \frac{1}{\ell^p 2^{p\ell(\beta-1/2)}},$$

and so, the bound for this case follows from (a).
If $\beta = 1/2$, then

$$\sum_{\ell=1}^{L} \frac{2^{p\ell(1/2-\beta)}}{\lambda_\ell^p} = \sum_{\ell=1}^{L} \frac{1}{\lambda_\ell^p} \leq 2 \left(\frac{\pi^2}{3}\right)^p \sum_{\ell=1}^{L/2} \ell^{2p} \leq 2 \left(\frac{\pi^2}{3}\right)^p \left(\frac{\log n}{\log 2}\right)^{2p+1}.$$

If $\beta > 1/2$, then

$$\begin{aligned} &\sum_{\ell=1}^{L} \frac{2^{p\ell(1/2-\beta)}}{\lambda_\ell^p} \\ &\quad = \sum_{\ell=1}^{L/2} \frac{2^{\ell(1/2-\beta)p}}{\lambda_\ell^p} + 2^{(L+1)(1/2-\beta)p} \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{\ell(1/2-\beta)p}} \end{aligned}$$

$$\le \sum_{\ell=1}^{L/2} \frac{2^{\ell(1/2-\beta)p}}{\lambda_\ell^p} + (2n)^{(1/2-\beta)p} \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{\ell(1/2-\beta)p}}$$
$$\le 2^{(L+1)(1/2-\beta)p} \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{(L+1-\ell)(1/2-\beta)p}} + (2n)^{(1/2-\beta)p} \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{\ell(1/2-\beta)p}}$$
$$\le (2n)^{(1/2-\beta)p} \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{\ell(1/2-\beta)p}} + (2n)^{(1/2-\beta)p} \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{\ell(1/2-\beta)p}}$$
$$\le (2n)^{(1/2-\beta)p} \left(\frac{5}{(1/2-\beta)^3}\right)^p \left(\frac{\pi^2}{3}\right)^{p+1}.$$

Hence, the result follows. ■

Define the functions

$$s(\lambda) := (1/2 + 1/\lambda)^{-1}, \quad \text{and} \quad T_1(\lambda) := \min\{\lambda, 1\} \quad \text{for all} \quad \lambda > 0. \tag{33}$$

THEOREM B.1. *Suppose $Z_1, \ldots, Z_n$ are elements of the causal process (32) with mean zero. Assume that, for some $\alpha > 0$ and $\nu > 0$,*

$$\|\{Z\}\|_{\psi_\alpha, \nu} = \sup_{p \ge 2} \sup_{m \ge 0} p^{-1/\alpha} (m+1)^\nu \Delta_{m,p} < \infty. \tag{34}$$

Define

$$\Omega_n(\nu) := 2^\nu \times \begin{cases} 5/(\nu - 1/2)^3, & \text{if } \nu > 1/2, \\ 2(\log_2 n)^{5/2}, & \text{if } \nu = 1/2, \\ 5(2n)^{(1/2-\nu)}/(1/2-\nu)^3, & \text{if } \nu < 1/2. \end{cases}$$

Then, for any $p \ge 2$,

$$\left\| \sum_{i=1}^n Z_i \right\|_p \le \sqrt{pn}\, \|\{Z\}\|_{\psi_\alpha, \nu} B_\nu + C_\alpha \|\{Z\}\|_{\psi_\alpha, \nu} (\log n)^{1/s(\alpha)} p^{1/T_1(s(\alpha))} \Omega_n(\nu), \tag{35}$$

where C_α is a constant depending only on α, and B_ν is a constant depending only on ν given by

$$B_\nu := \sqrt{6}\left[1 + \frac{20\pi^3 2^\nu}{3\sqrt{3}\nu^3}\right], \quad \text{if} \quad \nu > 0.$$

Furthermore, it follows by Markov's inequality that, for all $t \ge 0$,

$$\mathbb{P}\left(\left| \sum_{i=1}^n Z_i \right| \ge e\sqrt{tn}\, \|\{Z\}\|_{2,\nu} B_\nu + C_\alpha \|\{Z\}\|_{\psi_\alpha, \nu} t^{1/T_1(s(\alpha))} (\log n)^{1/s(\alpha)} \Omega_n(\nu) \right) \le 8e^{-t}.$$

Here, C_α is different from the one in the moment bound (35).

Proof. Define

$$S_n := \sum_{i=1}^{n} Z_i, \quad L = \left\lfloor \frac{\log n}{\log 2} \right\rfloor, \quad \text{and} \quad \xi_\ell = \begin{cases} 2^\ell, & \text{if } 0 \le \ell < L, \\ n, & \text{if } \ell = L. \end{cases}$$

Define, for $m \ge 0$,

$$Z_i^{(m)} := \mathbb{E}\left[Z_i \middle| \varepsilon_{i-m}, \ldots, \varepsilon_i\right], \quad \text{and} \quad M_{i,\ell} := \sum_{k=1}^{i} \left(Z_k^{(\xi_\ell)} - Z_k^{(\xi_{\ell-1})} \right).$$

Let

$$S_{n,m} := \sum_{i=1}^{n} Z_i^{(m)},$$

and consider the decomposition

$$S_n = S_{n,0} + \left(S_n - S_{n,n}\right) + \sum_{\ell=1}^{L} \left(S_{n,\xi_\ell} - S_{n,\xi_{\ell-1}}\right) := \mathbf{I} + \mathbf{II} + \mathbf{III}. \tag{36}$$

We prove the moment bound (35) by bounding the moments of each term in the decomposition (36).

Bounding **I**: Regarding the first term **I**, observe that $S_{n,0}$ is a sum of independent random variables $Z_i^{(0)}$ satisfying the tail assumption of Theorem A.1 with $\beta = \alpha$. This verification follows by noting that

$$\left\| Z_i^{(0)} \right\|_p \overset{(a)}{\le} \|Z_i\|_p \overset{(b)}{\le} \|\{Z\}\|_{p,\nu} \overset{(c)}{\le} p^{1/\alpha} \|\{Z\}\|_{\psi_\alpha,\nu}.$$

Inequality (a) follows from Jensen's inequality, (b) follows from Proposition B.1, and (c) follows from assumption (34). Hence, we get that, for any $p \ge 1$,

$$\begin{aligned} \|\mathbf{I}\|_p &= \left\| \sum_{i=1}^{n} \mathbb{E}\left[Z_i \middle| \varepsilon_i\right] \right\|_p \\ &\le \sqrt{6p} \left(\sum_{i=1}^{n} \mathbb{E}\left[Z_i^2\right] \right)^{1/2} + C_\alpha \|\{Z\}\|_{\psi_\alpha,\nu} p^{1/T_1(\alpha)} (\log n)^{1/\alpha}, \end{aligned}$$

for some constant C_α depending only on α. Here, Jensen's inequality is used to bound the variance of $\mathbb{E}\left[Z_i \middle| \varepsilon_i\right]$. By Proposition B.1, $\|Z_i\|_2 \le \|\{Z\}\|_{2,\nu}$, and hence,

$$\left\|S_{n,0}\right\|_p \le \sqrt{6pn} \|\{Z\}\|_{2,\nu} + C_\alpha \|\{Z\}\|_{\psi_\alpha,\nu} p^{1/T_1(\alpha)} (\log n)^{1/\alpha}. \tag{37}$$

Bounding **II**: For the second term, note that

$$S_n = \sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} \mathbb{E}\left[Z_i \middle| \varepsilon_i, \varepsilon_{i-1}, \ldots\right] = S_{n,\infty},$$

and hence,

$$S_n - S_{n,n} = \sum_{m=n}^{\infty} \left(S_{n,m+1} - S_{n,m}\right).$$

Substituting the definition of $S_{n,m}$, we have

$$S_{n,m+1} - S_{n,m} = \sum_{k=1}^{n} \left(\mathbb{E}\left[Z_k \middle| \varepsilon_k, \ldots, \varepsilon_{k-m-1}\right] - \mathbb{E}\left[Z_k \middle| \varepsilon_k, \ldots, \varepsilon_{k-m}\right]\right).$$

We now prove that the summands above form a martingale difference sequence with respect to a filtration. The following construction is taken from the proof of Lemma 1 of Liu and Wu (2010). Define

$$D_{k,m+1} := \mathbb{E}\left[Z_k \middle| \varepsilon_k, \ldots, \varepsilon_{k-m-1}\right] - \mathbb{E}\left[Z_k \middle| \varepsilon_k, \ldots, \varepsilon_{k-m}\right],$$

and the nondecreasing filtration

$$\mathcal{G}_{k,m+1} := \sigma\left(\varepsilon_{k-m-1}, \varepsilon_{k-m-1}, \ldots\right).$$

It is easy to see that

$$\mathbb{E}\left[D_{n-k+1,m+1} \middle| \mathcal{G}_{k-1,m+1}\right] = 0. \tag{38}$$

Therefore, $\{(D_{n-k+1,m+1}, \mathcal{G}_{k,m+1}) : 1 \le k \le n\}$ forms a martingale difference sequence. This implies that $S_{n,m+1} - S_{n,m}$ is a martingale, and hence, by Lemma B.1, we get, for $p \ge 2$,

$$\left\|S_{n,m+1} - S_{n,m}\right\|_p^2 \le p \sum_{k=1}^{n} \left\|D_{k,m+1}\right\|_p^2.$$

To further bound the right-hand side, note that, for $p \ge 2$,

$$\left\|D_{k,m+1}\right\|_p = \left\|\mathbb{E}\left[Z_k - g(\ldots, \varepsilon'_{k-m-1}, \varepsilon_{k-m}, \ldots, \varepsilon_k) \middle| \varepsilon_k, \ldots, \varepsilon_{k-m-1}\right]\right\|_p \le \delta_{m+1,p}. \tag{39}$$

Hence, for $p \ge 2$,

$$\left\|S_{n,m+1} - S_{n,m}\right\|_p \le \sqrt{pn}\delta_{m+1,p},$$

and

$$\left\|S_n - S_{n,n}\right\|_p \le \sum_{m=n}^{\infty} \left\|S_{n,m+1} - S_{n,m}\right\|_p \le \sqrt{pn} \sum_{m=n}^{\infty} \delta_{m+1,p} = \sqrt{pn}\Delta_{n+1,p}.$$

Under assumption (34), we obtain

$$\|\mathbf{II}\|_p = \left\|S_n - S_{n,n}\right\|_p \le \|\{Z\}\|_{\psi_\alpha,\nu} \frac{n^{1/2} p^{1/2+1/\alpha}}{(n+2)^\nu} = \|\{Z\}\|_{\psi_\alpha,\nu}\, n^{1/2-\nu} p^{1/2+1/\alpha}. \tag{40}$$

Bounding **III**: To bound **III**, note by definition of $M_{i,\ell}$ that

$$\mathbf{III} = \sum_{\ell=1}^{L} \sum_{k=1}^{n} \left(Z_k^{(\xi_\ell)} - Z_k^{(\xi_{\ell-1})}\right) = \sum_{\ell=1}^{L} M_{n,\ell}.$$

Now, observe that the summands of $M_{n,\ell}$,

$$\mathcal{D}_{k,\ell} := \left(Z_k^{(\xi_\ell)} - Z_k^{(\xi_{\ell-1})}\right),$$

are ξ_ℓ-dependent in the sense that $\mathcal{D}_{k,\ell}$ and $\mathcal{D}_{s,\ell}$ are independent if $|s-k| > \xi_\ell$. This can be proved as follows. By definition, $\mathcal{D}_{k,\ell}$ is only a function of $(\varepsilon_k, \ldots, \varepsilon_{k-\xi_\ell})$, and by independence of $\varepsilon_k, k \in \mathbb{Z}$, the claim follows. Now, a blocking technique can be used to convert $M_{n,\ell}$ into a sum of independent variables. See Corollary A.1 of Romano and Wolf (2000) for a similar use. Define

$$\mathcal{A}_\ell := \{2\xi_\ell i + j : i \in \mathbb{Z}, 1 \le j \le \xi_\ell\},$$
$$\mathcal{B}_\ell := \{2\xi_\ell i + \xi_\ell + j : i \in \mathbb{Z}, 1 \le j \le \xi_\ell\}.$$

Consider the decomposition of $M_{n,\ell}$ as

$$M_{n,\ell} = \sum_{k=1}^{n} \mathcal{D}_{k,\ell} = A_{n,\ell} + B_{n,\ell},$$

where

$$A_{n,\ell} := \sum_{1 \le k \le n, k \in \mathcal{A}} \mathcal{D}_{k,\ell} \quad \text{and} \quad B_{n,\ell} := \sum_{1 \le k \le n, k \in \mathcal{B}} \mathcal{D}_{k,\ell}.$$

We now provide moment bounds for $M_{n,\ell}$ by giving moment bounds for $A_{n,\ell}$ and $B_{n,\ell}$, which is in turn done by separating the summands of $A_{n,\ell}$ and $B_{n,\ell}$ to form an independent sum. Note that

$$\begin{aligned} A_{n,\ell} &= \sum_{i=1}^{\lfloor \frac{n}{2\xi_\ell} \rfloor} \left(\sum_{j=1}^{\xi_\ell} \mathcal{D}_{2\xi_\ell i + j, \ell} \right) = \sum_{i=1}^{\lfloor \frac{n}{2\xi_\ell} \rfloor} \left(\sum_{k=2\xi_\ell i+1}^{2\xi_\ell i + \xi_\ell} \left(Z_k^{(\xi_\ell)} - Z_k^{(\xi_{\ell-1})} \right) \right) \\ &= \sum_{i=1}^{\lfloor \frac{n}{2\xi_\ell} \rfloor} \left(M_{2\xi_\ell i + \xi_\ell, \ell} - M_{2\xi_\ell i, \ell} \right). \end{aligned} \tag{41}$$

By the ξ_ℓ-independence of the summands of $M_{n,\ell}$, we get that the summands in the final representation of $A_{n,\ell}$ are independent, and so Theorem A.1 applies. In the following, we verify the assumption of Theorem A.1. For $1 \le i < j \le n$, it is clear that

$$\begin{aligned} M_{j,\ell} - M_{i,\ell} &= \sum_{k=i+1}^{j} \left(Z_k^{(\xi_\ell)} - Z_k^{(\xi_{\ell-1})} \right) \\ &= \sum_{k=i+1}^{j} \left(\sum_{t=1+\xi_{\ell-1}}^{\xi_\ell} \left(Z_k^{\xi_\ell} - Z_k^{(\xi_{\ell-1})} \right) \right) \\ &= \sum_{t=1+\xi_{\ell-1}}^{\xi_\ell} \left(\sum_{k=i+1}^{j} \left(Z_k^{(t)} - Z_k^{(t-1)} \right) \right). \end{aligned}$$

By triangle inequality,

$$\left\|M_{j,\ell}-M_{i,\ell}\right\|_p \le \sum_{t=1+\xi_{\ell-1}}^{\xi_\ell} \left\| \sum_{k=i+1}^{j} \left(Z_k^{(t)}-Z_k^{(t-1)}\right)\right\|_p. \tag{42}$$

As proved in (38), the summation for each t represents a martingale, and hence, by Lemma B.1, we get, for $p \ge 2$, that

$$\left\| \sum_{k=i+1}^{j} \left(Z_k^{(t)}-Z_k^{(t-1)}\right)\right\|_p^2 \le p \sum_{k=i+1}^{j} \left\|Z_k^{(t)}-Z_k^{(t-1)}\right\|_p^2 \le p \sum_{k=i+1}^{j} \delta_{t,p}^2 = p(j-i)\delta_{t,p}^2.$$

Here, we used inequality (39). Substituting this in inequality (42) and using $\xi_{\ell-1} \ge \xi_\ell/2$, we get

$$\begin{aligned}\left\|M_{j,\ell}-M_{i,\ell}\right\|_p &\le p^{1/2}(j-i)^{1/2} \sum_{t=1+\xi_{\ell-1}}^{\xi_\ell} \delta_{t,p} \le p^{1/2}(j-i)^{1/2}\Delta_{1+\xi_{\ell-1},p} \\ &\le \|\{Z\}\|_{p,\nu}\, p^{1/2}(j-i)^{1/2}(2+\xi_{\ell-1})^{-\nu} \\ &\le 2^\nu \|\{Z\}\|_{p,\nu}\, p^{1/2}(j-i)^{1/2}\xi_\ell^{-\nu}. \end{aligned} \tag{43}$$

Under assumption (34), we get

$$\begin{aligned}\left\|M_{j,\ell}-M_{i,\ell}\right\|_p &\le 2^\nu \|\{Z\}\|_{\psi_\alpha,\nu}\, p^{1/2+1/\alpha}(j-i)^{1/2}\xi_\ell^{-\nu} \\ &= 2^\nu \|\{Z\}\|_{\psi_\alpha,\nu}\, p^{1/s(\alpha)}(j-i)^{1/2}\xi_\ell^{-\nu}.\end{aligned}$$

See (33) for the definition of $s(\alpha)$. Thus, for all $1 \le i \le \lfloor \frac{n}{2\xi_\ell} \rfloor$,

$$\sup_{p\ge 2} p^{-1/s(\alpha)} \left\|M_{2\xi_\ell i+\xi_\ell,\ell}-M_{2\xi_\ell i,\ell}\right\|_p \le 2^\nu \|\{Z\}\|_{\psi_\alpha,\nu}\, \xi_\ell^{1/2-\nu}.$$

So, the summands of $A_{n,\ell}$ in the final representation in (41) are independent and satisfy the hypothesis of Theorem A.1 with $\beta = s(\alpha)$. Therefore, for $p \ge 2$,

$$\begin{aligned}\left\|A_{n,\ell}\right\|_p &\le \sqrt{6p}\left(\sum_{i=1}^{\lfloor n/(2\xi_\ell)\rfloor} \left\|M_{2\xi_\ell i+\xi_\ell,\ell}-M_{2\xi_\ell i,\ell}\right\|_2^2\right)^{1/2} \\ &\quad + C_\alpha 2^\nu \|\{Z\}\|_{\psi_\alpha,\nu}\,(\log n)^{1/s(\alpha)}\,\xi_\ell^{1/2-\nu} p^{1/T_1(s(\alpha))} \\ &\le \sqrt{12p}\,\|\{Z\}\|_{2,\nu}\,\frac{2^\nu \xi_\ell^{1/2}}{\xi_\ell^\nu}\left(\frac{n}{2\xi_\ell}\right)^{1/2} \\ &\quad + C_\alpha 2^\nu \|\{Z\}\|_{\psi_\alpha,\nu}\,(\log n)^{1/s(\alpha)}\,\xi_\ell^{1/2-\nu} p^{1/T_1(s(\alpha))} \\ &\le \frac{2^\nu}{\xi_\ell^\nu}\left[\|\{Z\}\|_{2,\nu}\sqrt{6pn} + C_\alpha \|\{Z\}\|_{\psi_\alpha,\nu}\, p^{1/T_1(s(\alpha))}\,(\log n)^{1/s(\alpha)}\,\xi_\ell^{1/2}\right].\end{aligned}$$

Here, the second inequality follows from (43).

Similarly, a representation for $B_{n,\ell}$ exists with independent summands satisfying the assumption of Theorem A.1 with $\beta = s(\alpha)$, and so,

$$\left\|B_{n,\ell}\right\|_p \le \frac{2^\nu}{\xi_\ell^\nu}\left[\|\{Z\}\|_{2,\nu}\sqrt{6pn} + C_\alpha \|\{Z\}\|_{\psi_\alpha,\nu} p^{1/T_1(s(\alpha))} (\log n)^{1/s(\alpha)} \xi_\ell^{1/2}\right].$$

Combining the bounds for $A_{n,\ell}$ and $B_{n,\ell}$ implies the bound on $M_{n,\ell}$ as

$$\left\|M_{n,\ell}\right\|_p \le \frac{2^{1+\nu}}{\xi_\ell^\nu}\left[\|\{Z\}\|_{2,\nu}\sqrt{6pn} + C_\alpha \|\{Z\}\|_{\psi_\alpha,\nu} p^{1/T_1(s(\alpha))} (\log n)^{1/s(\alpha)} \xi_\ell^{1/2}\right]. \tag{44}$$

To complete bounding **III**, we need to bound the moments of the sum of $M_{n,\ell}$ over $1 \le \ell \le L$, which are all dependent. For this, define the sequence

$$\lambda_\ell = \begin{cases} 3\pi^{-2}\ell^{-2}, & \text{if } 1 \le \ell \le L/2, \\ 3\pi^{-2}(L+1-\ell)^{-2}, & \text{if } L/2 < \ell \le L. \end{cases}$$

This positive sequence satisfies $\sum_{\ell=1}^L \lambda_\ell < 1$. It is easy to derive from H'´older's inequality that

$$\left|\sum_{\ell=1}^L a_\ell\right|^p \le \sum_{\ell=1}^L \frac{|a_\ell|^p}{\lambda_\ell^p}.$$

Substituting in this inequality $a_\ell = M_{n,\ell}$ and the moment bound (44), we get

$$\mathbb{E}\left[\left|\sum_{\ell=1}^L M_{n,\ell}\right|^p\right] \le 2^{(2+\nu)p}\|\{Z\}\|_{2,\nu}^p (6pn)^{p/2}\sum_{\ell=1}^L \frac{1}{\lambda_\ell^p \xi_\ell^{p\nu}} + C_\alpha^p 2^{(2+\nu)p}\|\{Z\}\|_{\psi_\alpha,\nu}^p p^{p/T_1(s(\alpha))}(\log n)^{p/s(\alpha)}\sum_{\ell=1}^L \frac{\xi_\ell^{p/2}}{\lambda_\ell^p \xi_\ell^{p\nu}}.$$

It follows from Lemma B.2 and the definition of $\Omega_n(\nu)$ that, for $p \ge 2$,

$$\left\|\sum_{\ell=1}^L M_{n,\ell}\right\|_p \le \frac{5\pi^3 2^2}{3\sqrt{3}}\left[\frac{2^\nu \|\{Z\}\|_{2,\nu}\sqrt{6pn}}{\nu^3} + C_\alpha \|\{Z\}\|_{\psi_\alpha,\nu}(\log n)^{1/s(\alpha)}\Omega_n(\nu) p^{1/T_1(s(\alpha))}\right]. \tag{45}$$

Combining the moment bounds (37), (40), and (45), it follows that, for $p \ge 2$,

$$\|S_n\|_p \le \sqrt{6pn}\|\{Z\}\|_{\psi_\alpha,\nu}\left[1 + \frac{20\pi^3 2^\nu}{3\sqrt{3}\nu^3}\right] + \|\{Z\}\|_{\psi_\alpha,\nu} n^{1/2-\nu} p^{1/s(\alpha)} + C_\alpha \|\{Z\}\|_{\psi_\alpha,\nu}(\log n)^{1/s(\alpha)} p^{1/T_1(s(\alpha))}\Omega_n(\nu).$$

Here, the inequalities $s(\alpha) \le \alpha$ and $T_1(s(\alpha)) \le T_1(\alpha)$ are used. Now, noting that $\Omega_n(\nu) \ge n^{1/2-\nu}$, for all $\nu > 0$ and $p^{1/s(\alpha)} \le p^{1/T_1(s(\alpha))}$, the result follows. ∎

In the following two lemmas, we prove that the dependent adjusted norm of linear combinations and products of functionally dependent random variables can be bounded in terms of the individual processes. Recall the definition of Θ_k from (25).

LEMMA B.3. *Suppose Assumption (DEP) holds, then, for any* $\theta \in \Theta_k$,

$$\sup_{\theta \in \Theta_k} \left\| \{\theta^\top X\} \right\|_{r,\nu} \le k^{1/2} K_{n,p}.$$

Proof. Fix $\theta \in \Theta_k$. Set the functional dependence measure (29) for the linear combination $\theta^\top X$ as

$$\delta_{s,r}^{(L)} := \max_{1 \le i \le n} \left\| \theta^\top X_i - \theta^\top X_{i,s} \right\|_r.$$

Note that $\theta \in \Theta_k$ are all k-sparse, and so there are only k nonzero coordinates $\theta(j)$ of θ. Since the functional dependence measure is a norm, it follows that

$$\begin{aligned}\delta_{s,r}^{(L)} &= \max_{1 \le i \le n} \sum_{j=1}^{p} |\theta(j)| \left\| X_i(j) - X_{i,s}(j) \right\|_r \\ &\le \sum_{j=1}^{p} |\theta(j)| \max_{1 \le i \le n} \left\| X_i(j) - X_{i,s}(j) \right\|_r = \sum_{j=1}^{p} |\theta(j)| \delta_{s,r,j}.\end{aligned}$$

Hence, for $m \ge 0$,

$$\Delta_{m,r}^{(L)} := \sum_{s=m}^{\infty} \delta_{s,r}^{(L)} \le \sum_{s=m}^{\infty} \sum_{j=1}^{p} |\theta(j)| \delta_{s,r,j} = \sum_{j=1}^{p} |\theta(j)| \left(\sum_{s=m}^{\infty} \delta_{s,r,j} \right) = \sum_{j=1}^{p} |\theta(j)| \Delta_{m,r,j}.$$

This implies that

$$\Delta_{m,r}^{(L)} \le \|\theta\|_1 \max_{1 \le j \le p} \Delta_{m,r,j} \le k^{1/2} \max_{1 \le j \le p} \Delta_{m,r,j}.$$

Therefore, for $r \ge 1$ and $\nu > 0$,

$$\left\| \{\theta^\top X\} \right\|_{r,\nu} \le k^{1/2} \|\{X\}\|_{r,\nu} \quad \Rightarrow \quad \left\| \{\theta^\top X\} \right\|_{\psi_\alpha,\nu} \le k^{1/2} \|\{X\}\|_{\psi_\alpha,\nu} \le k^{1/2} K_{n,p},$$

proving the result. ∎

LEMMA B.4. *Suppose* $(W_1^{(1)}, W_1^{(2)}), \ldots, (W_n^{(1)}, W_n^{(2)})$ *are* n *functionally dependent real-valued random vectors. Set* $W_i = W_i^{(1)} W_i^{(2)}$, *for* $1 \le i \le n$. *Then, for all* $r \ge 2$ *and* $\nu > 0$,

$$\begin{aligned}\|\{W\}\|_{r/2,\nu} \le{}& \left\| \{W^{(1)}\} \right\|_{r,0} \left\| \{W^{(2)}\} \right\|_{r,\nu} + \max_{1 \le i \le n} \left| \mathbb{E}\left[W_i^{(1)} \right] \right| \left\| \{W^{(2)}\} \right\|_{r,\nu} \\ &+ \left\| \{W^{(2)}\} \right\|_{r,0} \left\| \{W^{(1)}\} \right\|_{r,\nu} + \max_{1 \le i \le n} \left| \mathbb{E}\left[W_i^{(2)} \right] \right| \left\| \{W^{(1)}\} \right\|_{r,\nu}.\end{aligned}$$

Proof. Set, for $j = 1, 2$,

$$\delta_{s,r}^{(j)} := \left\| W_i^{(1)} - W_{i,s}^{(1)} \right\|_r, \quad \text{and} \quad \Delta_{m,r}^{(j)} := \sum_{s=m}^{\infty} \delta_{s,r}^{(j)}.$$

Fix $1 \le i \le n$ and consider

$$\begin{aligned}
\varphi_{s,r/2,i} &:= \left\| W_i^{(1)} W_i^{(2)} - W_{i,s}^{(1)} W_{i,s}^{(2)} \right\|_{r/2} \\
&= \left\| W_i^{(1)} \left[W_i^{(2)} - W_{i,s}^{(2)} \right] + W_{i,s}^{(2)} \left[W_i^{(1)} - W_{i,s}^{(1)} \right] \right\|_{r/2} \\
&\le \left\| W_i^{(1)} \left[W_i^{(2)} - W_{i,s}^{(2)} \right] \right\|_{r/2} + \left\| W_{i,s}^{(2)} \left[W_i^{(1)} - W_{i,s}^{(1)} \right] \right\|_{r/2} \\
&\le \left\| W_i^{(1)} \right\|_r \left\| W_i^{(2)} - W_{i,s}^{(2)} \right\|_r + \left\| W_{i,s}^{(2)} \right\|_r \left\| W_i^{(1)} - W_{i,s}^{(1)} \right\|_r \\
&\le \left\| W_i^{(1)} \right\|_r \delta_{k,r}^{(2)} + \left\| W_{i,s}^{(2)} \right\|_r \delta_{k,r}^{(1)}.
\end{aligned}$$

Since ε'_{i-k} is identically distributed as ε_{i-k}, $\left\| W_{i,s}^{(2)} \right\|_r = \left\| W_i^{(2)} \right\|_r$. So, an upper bound on the dependence adjusted norm can be obtained as

$$\begin{aligned}
\Delta_{m,r/2} = \sum_{k=m}^{\infty} \max_{1 \le i \le n} \varphi_{k,r/2,i} &\le \max_{1 \le i \le n} \left\| W_i^{(1)} \right\|_r \sum_{k=m}^{\infty} \delta_{k,r}^{(2)} + \max_{1 \le i \le n} \left\| W_i^{(2)} \right\|_r \sum_{k=m}^{\infty} \delta_{k,r}^{(1)} \\
&\le \max_{1 \le i \le n} \left\| W_i^{(1)} \right\|_r \Delta_{m,r}^{(2)} + \max_{1 \le i \le n} \left\| W_i^{(2)} \right\|_r \Delta_{m,r}^{(1)},
\end{aligned}$$

and thus,

$$\begin{aligned}
\|\{W\}\|_{r/2,\nu} &\le \max_{1 \le i \le n} \left\| W_i^{(1)} \right\|_r \left\| \{W^{(2)}\} \right\|_{r,\nu} + \max_{1 \le i \le n} \left\| W_i^{(2)} \right\|_r \left\| \{W^{(1)}\} \right\|_{r,\nu} \\
&\le \left\| \{W^{(1)}\} \right\|_{r,0} \left\| \{W^{(2)}\} \right\|_{r,\nu} + \max_{1 \le i \le n} \left| \mathbb{E}\left[W_i^{(1)} \right] \right| \left\| \{W^{(2)}\} \right\|_{r,\nu} \\
&\quad + \left\| \{W^{(2)}\} \right\|_{r,0} \left\| \{W^{(1)}\} \right\|_{r,\nu} + \max_{1 \le i \le n} \left| \mathbb{E}\left[W_i^{(2)} \right] \right| \left\| \{W^{(1)}\} \right\|_{r,\nu},
\end{aligned}$$

proving the result. ∎

C. Proof of Proposition 3.1

Proof. It is easy to see that

$$\begin{aligned}
\mathrm{RIP}(k, \Sigma_1 - \Sigma_2) &= \sup_{\substack{\theta \in \mathbb{R}^p, \|\theta\|_0 \le k, \\ \|\theta\|_2 \le 1}} \left| \theta^\top (\Sigma_1 - \Sigma_2) \theta \right| \\
&\le \sup_{\substack{\theta \in \mathbb{R}^p, \\ \|\theta\|_0 \le k, \|\theta\|_2 \le 1}} \|\theta\|_1^2 \, |||\Sigma_1 - \Sigma_2|||_\infty \le k |||\Sigma_1 - \Sigma_2|||_\infty.
\end{aligned}$$

Here, we have used inequalities (3). A similar proof implies the second result. ∎

REFERENCES

Adamczak, R. (2008) A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability* 13(34), 1000–1034.

Bachoc, F., G. Blanchard, P. Neuvial (2018) On the post selection inference constant under restricted isometry properties. *Electronic Journal of Statistics* 12(2), 3736–3757.

Bachoc, F., H. Leeb, B. M. Pötscher (2019a) Valid confidence intervals for post-model-selection predictors. *Annals of Statistics* 47(3), 1475–1504.

Bachoc, F., D. Preinerstorfer, & L. Steinberger (2019b) Uniformly valid confidence intervals post-model-selection. *Annals of Statistics* 48(1), 440–463.

Belloni, A. & V. Chernozhukov (2013) Least squares after model selection in high-dimensional sparse models. *Bernoulli* 19(2), 521–547.

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, K. Kato (2018) High-dimensional econometrics and regularized GMM. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Buja, A., R. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, K. Zhan & L. Zhao (2019) Models as approximations, part I: A conspiracy of random regressors and model deviations against classical inference in regression. *Statistical Science* 34(4), 523–544.

Cai, T. T. & M. Yuan (2012) Adaptive covariance matrix estimation through block thresholding. *Annals of Statistics* 40(4), 2014–2042.

Catoni, O. (2012) Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques* 48(4), 1148–1185.

Catoni, O. & I. Giulini (2017) Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. Preprint, arXiv:1712.02747.

Chakrabortty, A., P. Nandy, & H. Li (2021) Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models. Preprint, arXiv:1809.10652.

Chen, Y., C. Caramanis & S. Mannor (2013) Robust sparse regression under adversarial corruption. In S. Dasgupta & D. McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, pp. 774–782. Proceedings of Machine Learning Research.

de la Peña, V.H. & E. Giné (1999) *Decoupling*. Probability and Its Applications. Springer-Verlag.

Foygel, R. & N. Srebro (2011) Fast-rate and optimistic-rate error bounds for L1-regularized regression. Preprint, arXiv:1108.0373.

Giessing, A. (2018) On high-dimensional misspecified quantile regression. PhD thesis, University of Michigan.

Javanmard, A. & A. Montanari (2018) Debiasing the lasso: Optimal sample size for Gaussian designs. *Annals of Statistics* 46(6A), 2593–2622.

Kuchibhotla, A.K. (2018). Deterministic inequalities for smooth M-estimators. Preprint, arXiv:1809.05172.

Kuchibhotla, A.K., L.D. Brown, A. Buja, J. Cai, E.I. George, & L. Zhao (2019) Valid post-selection inference in model-free linear regression. *Annals of Statistics* 48(5), 2953–2981.

Kuchibhotla, A.K., L.D. Brown, A. Buja, E.I. George, & L. Zhao (2018) A model free perspective for linear regression: Uniform-in-model bounds for post selection inference. Preprint, arXiv:1802.05801v1.

Kuchibhotla, A.K. & A. Chakrabortty (2020) Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. Preprint, arXiv:1804.02605.

Kuchibhotla, A. K., S. Mukherjee, D. Banerjee (2021) High-dimensional CLT: Improvements, non-uniform extensions and large deviations. *Bernoulli* 27(1), 192–217.

Leeb, H. & B. M. Pötscher (2005) Model selection and inference: Facts and fiction. *Econometric Theory* 21(1), 21–59.

Leeb, H. & B. M. Pötscher (2006a) Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* 34(5), 2554–2591.

Leeb, H. & B. M. Pötscher (2006b) Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory* 22(1), 69–97.

Leeb, H. & B. M. Pötscher (2008) Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* 24(2), 338–376.

Liu, W. & W. B. Wu (2010) Asymptotics of spectral density estimates. *Econometric Theory* 26(4), 1218–1245.

Liu, W., H. Xiao & W. B. Wu (2013) Probability and moment inequalities under dependence. *Statistica Sinica* 23(3), 1257–1272.

Loh, P.-L. & M. J. Wainwright (2012) High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Annals of Statistics* 40(3), 1637–1664.

Minsker, S. (2015) Geometric median and robust estimation in Banach spaces. *Bernoulli* 21(4), 2308–2335.

Monahan, J. F. (2008) *A Primer on Linear Models*. CRC Press.

Plan, Y. & R. Vershynin (2013) One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics* 66(8), 1275–1297.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 2. Institute of Mathematical Statistics and American Statistical Association.

Pötscher, B. M. & I. R. Prucha (1997) *Dynamic Nonlinear Econometric Models: Asymptotic Theory*. Springer-Verlag.

Raskutti, G., M. J. Wainwright & B. Yu (2011) Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory* 57(10), 6976–6994.

Rinaldo, A., L. Wasserman, M. G'Sell, J. Lei, & R. Tibshirani (2019) Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *Annals of Statistics* 47(6), 3438–3469.

Rio, E. (2009) Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability* 22(1), 146–163.

Romano, J. P. & M. Wolf (2000) A more general central limit theorem for m-dependent random variables with unbounded m. *Statistics & Probability Letters* 47(2), 115–124.

Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Mathematical Statistics. Wiley.

van der Vaart, A.W. & J.A. Wellner (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag.

van Zwet, W. R. (1984) A Berry–Esseen bound for symmetric statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 66(3), 425–440.

Vershynin, R. (2012) How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability* 25(3), 655–686.

Vershynin, R. (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wei, X. & S. Minsker (2017) Estimation of the covariance structure of heavy-tailed distributions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (eds.), *Advances in Neural Information Processing Systems*, vol. 30, pp. 2855–2864. Curran Associates, Inc.

White, H. (1980a) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society* 48(4), 817–838.

White, H. (1980b) Using least squares to approximate unknown regression functions. *International Economic Review* 21(1), 149–170.

White, H. (2001) *Asymptotic Theory for Econometricians*. Economic Theory, Econometrics, and Mathematical Economics. Academic Press.

Wu, W. B. (2005) Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America* 102(40), 14150–14154.

Wu, W. B. & J. Mielniczuk (2010) A new look at measuring dependence. In ed. P. Doukhan, G. Lang, D. Surgailis, & G. Teyssiere, *Dependence in Probability and Statistics*, pp. 123–142. Springer.
Wu, W.-B. & Y. N. Wu (2016) Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics* 10(1), 352–379.
Zhang, D. & W. B. Wu (2017) Gaussian approximation for high dimensional time series. *Annals of Statistics* 45(5), 1895–1919.
Zhang, X. & G. Cheng (2014) Bootstrapping high dimensional time series. Preprint, arXiv:1406.1037.