A Model Free Perspective for Linear Regression: Uniform-in-model Bounds for Post Selection Inference

Arun K. Kuchibhotla, Lawrence D. Brown, Andreas Buja, Edward I. George, and Linda Zhao

University of Pennsylvania
e-mail: arunku@wharton.upenn.edu

Abstract: For the last two decades, high-dimensional data and methods have proliferated throughout the literature. The classical technique of linear regression, however, has not lost its touch in applications. Most high-dimensional estimation techniques can be seen as variable selection tools which lead to a smaller set of variables where classical linear regression technique applies. In this paper, we prove estimation error and linear representation bounds for the linear regression estimator uniformly over (many) subsets of variables. Based on deterministic inequalities, our results provide "good" rates when applied to both independent and dependent data. These results are useful in correctly interpreting the linear regression estimator obtained after exploring the data and also in post model-selection inference. All the results are derived under no model assumptions and are non-asymptotic in nature.

1. Introduction and Motivation

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. Although results exist for model misspecification, it is often not clear just what is being estimated. Suppose the statistician is unwilling to assume sparsity of the parameter or even just a linear model? Minimax lower bounds for this problem imply the impossibility of consistent estimation of the parameter vector without structural constraints; see Raskutti et al. (2011). Thus, for consistent estimation with sparsity as the structural constraint, the number of non-zero elements of the parameter vector must be less than the sample size n. Now consider the following popular procedure in applied statistics and data science. High dimensional data is first explored either in a principled way (e.g., lasso or best subset selection) or even in an unprincipled way to select a manageable set of variables, and then linear regression is applied to the reduced data. For practical purposes, this final set of variables is often much smaller than the sample size and the total number of initial variables, and yet treated as a high-dimensional linear regression. By definition, this procedure makes use of all the data to come up with a "significant" subset of variables and no sparsity constraints are required. The current article is about understanding what is being estimated by this procedure in a model-free high-dimensional framework.

Variable selection plays a central role in data analysis when data on too many variables is available. This could be for logistical reasons or to obtain a parsimonious set of variables for interpretation purposes. As described above, it has been common practice to explore the data to first select a subset of variables, and then ignore the selection process for estimation and inference. The implications of such a method of data analysis have been recognized for a long time and can often be disastrous in terms of providing misleading conclusions, see Berk et al. (2013) and the references therein for a discussion. These considerations have led to the recent field of post-selection inference. Regression applications, in particular the structure of response and covariates, do not play any special role in this general problem, and the exploration methodology above is typically practiced whenever there are too many variables to consider for a final statistical analysis. In this paper, however, we focus on linear regression, as it leads to tractable closed form estimation yielding a more transparent analysis. We should mention that a similar analysis can be done for other M-estimation problems; see Section 6 for more details.

In addressing the problem of post model-selection inference, perhaps the main question needing an answer is "what is being estimated by the estimator from the data analysis?" A major thrust of the present article is to provide an answer to this question in a very general setting for linear regression, an answer that will be seen to lead to a valid interpretation of the post selection linear regression estimator. This question was answered in a very restrictive setting in Berk et al. (2013) from an intuitive point of view. In particular, Berk et al. (2013) assumed that the covariates in the data are fixed (non-random) and the response is normally distributed. This distributional assumption allows for a simple explanation of what is being estimated by the least squares linear regression estimator on a subset of covariates. We are not aware of other work that treats this question in the fully general setting we consider. However, in a related vein, Belloni and Chernozhukov (2013) established the rate of convergence results for the least squares linear regression estimator post lasso type model selection, see their Theorem 4, comparing its behavior with respect to the sparse oracle estimator.

Before answering the main question posed above one must clarify "what does it mean to say $\hat{\theta}$ is estimating θ?" It is natural to answer this by showing that $\hat{\theta}$ is consistent for θ, however, because we are in a high-dimensional setting, the norm underlying this consistency must be made precise. To then answer our main question in full generality, we establish various deterministic inequalities that are uniform over subsets of covariates. Finally, we apply these inequalities to both independent and dependent data to obtain concrete rates of convergence. We use the dependence structure of data introduced by Wu (2005), which is based on the idea of coupling and covers the dependence structure of many linear and non-linear time series. In the process of applying our results to dependent observations, we prove a tail bound for zero mean dependent sums that extends the results of Wu and Wu (2016).

Our main results include uniform-in-model estimation error of the least squares linear regression estimator in terms of the $\|\cdot\|_2$-, $\|\cdot\|_1$-norms, and also uniform-in-model asymptotic linear representation of the estimator in terms of the $\|\cdot\|_2$-norm. Each model here corresponds to a distinct subset of covariates. These results are established for both independent and dependent observations. All of our results are non-asymptotic in nature and allow for the total number of covariates to grow almost exponentially in the sample size when the observations have exponential tails. The rates we obtain are comparable to the ones obtained by Portnoy (1988), see also Portnoy (1984, 1985) and He and Shao (1996) for more results, though there are many differences in the settings considered. Portnoy assumes a true linear model with fixed covariates, but deals with a more general class of loss functions and his results are not uniform-in-model.

There is a rich literature on uniform asymptotic linear representations, which have been used in optimal M-estimation problems. See Section 4 of Arcones (2005) and Sections 10.2, 10.3, Equation (10.25) of Dodge and Jurevckova (2000) for examples where uniform asymptotic linear representations are established for a large class of M-estimators indexed by a subset of $\mathbb{R}$. The main focus there is to choose a tuning parameter that asymptotically leads to an estimator with "smallest" variance, and to take into account this randomness in proving that the final estimator with the tuning parameter estimate has an asymptotic normal distribution with "smallest" variance. It is possible to derive some of our results by viewing the problem of least squares linear regression of the response on a subset of covariates as a parametrized M-estimation problem indexed by the set of subsets, and then applying the general results of Arcones (2005).

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 dependent observations is considered Section 5. An extension of our results to a class of general M-estimators is given in Section 6. Proofs of the results in this section along with several examples will be provided in a future paper. A discussion of our results along with their implications is given in Section 7. Some auxiliary probability results for sums of independent and functionally dependent random variables are given in Appendix A and Appendix 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 "model" is used to specify the subset of covariates used in the regression and does not refer to any probability model. We do <u>not</u> assume a linear model (in any sense) to be true anywhere for any choice of covariates in this or in the subsequent sections. 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 j-th coordinate of v. For any non-empty model M given by a subset of $\{1, 2, \ldots, q\}$, let $v(M)$ denote a sub-vector 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 non-empty model $M \subseteq \{1, 2, \ldots, q\}$ and any symmetric matrix $A \in \mathbb{R}^{q\times q}$, let $A(M)$ denote the sub-matrix of A with indices in $M \times M$ and for $1 \leq j, k \leq q$, let $A(j, k)$ denotes the value at the j-th row and the k-th 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 non-zero entries in v (note this is not a norm). For any 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 inequalities will be used throughout without any special mention. For any matrix $A \in \mathbb{R}^{q\times q}$ and $v \in \mathbb{R}^q$,

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

For any $1 \le k \le p$, define the set of models

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

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

Traditionally, it is common to include an intercept term when fitting the linear regression. 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 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

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

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

$$\hat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top \in \mathbb{R}^{p\times p}, \quad \text{and} \quad \hat{\Gamma}_n := \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \in \mathbb{R}^p. \tag{3}$$

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

$$\hat{\beta}_{n,M} := \underset{\theta\in\mathbb{R}^{|M|}}{\arg\min}\ \hat{R}_n(\theta; M). \tag{4}$$

Based on the quadratic expansion (2) of the empirical objective $\hat{R}_n(\theta; M)$, the estimator $\hat{\beta}_{n,M}$ is given by the closed form expression

$$\hat{\beta}_{n,M} = \left[\hat{\Sigma}_n(M)\right]^{-1}\hat{\Gamma}_n(M), \tag{5}$$

assuming non-singularity of $\hat{\Sigma}_n(M)$. It is worth mentioning 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 Equation (5) that $\hat{\beta}_{n,M}$ is a (non-linear) 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 independent and identically distributed (iid) with finite fourth moments, it follows that $\hat{\Sigma}_n(M)$ and $\hat{\Gamma}_n(M)$ converge in probability to their expectations. The iid assumption here can be relaxed to weak dependence and non-identically distributed random vectors. Define the "expected" matrix and vector as

$$\Sigma_n := \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i X_i^\top\right] \in \mathbb{R}^{p\times p}, \quad \text{and} \quad \Gamma_n := \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i Y_i\right] \in \mathbb{R}^p. \tag{6}$$

If the convergence (in probability) of $(\hat{\Sigma}_n(M), \hat{\Gamma}_n(M))$ to $(\Sigma_n(M), \Gamma_n(M))$ holds, then by a Slutsky type argument, it follows that $\hat{\beta}_{n,M}$ converges to $\beta_{n,M}$, where

$$\begin{aligned}\beta_{n,M} &:= \operatorname*{arg\,min}_{\theta\in\mathbb{R}^{|M|}} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\{Y_i - X_i^\top(M)\theta\right\}^2\right] \\ &= \operatorname*{arg\,min}_{\theta\in\mathbb{R}^{|M|}} \theta^\top\Sigma_n(M)\theta - 2\theta^\top\Gamma_n(M) \\ &= (\Sigma_n(M))^{-1}\,\Gamma_n(M).\end{aligned} \tag{7}$$

These convergence statements are only about a single model M and not uniform. By uniform-in-model $\|\cdot\|_2$-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 $\Gamma_n(M)$. The specific structure of these matrices being the average of random matrices and random vectors is not required. 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

$$\beta_M(\Sigma,\Gamma) = (\Sigma(M))^{-1}\,\Gamma(M), \tag{8}$$

assuming the existence of the inverse of $\Sigma(M)$. We call this $\beta_M(\cdot,\cdot)$ the linear regression map. In the next section, we shall bound

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

in terms of the differences $\Sigma_1 - \Sigma_2$ and $\Gamma_1 - \Gamma_2$. In this regard, 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. The following error norms will be very useful for these results:

$$\begin{aligned}\mathrm{RIP}(k,\Sigma_1-\Sigma_2) &:= \sup_{M\in\mathcal{M}(k)} \|\Sigma_1(M) - \Sigma_2(M)\|_{op}\,, \\ \mathcal{D}(k,\Gamma_1-\Gamma_2) &= \sup_{M\in\mathcal{M}(k)} \|\Gamma_1(M) - \Gamma_2(M)\|_2\,.\end{aligned} \tag{9}$$

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 restricted isometry property used in compressed sensing and high-dimensional linear regression literature. Also, define the k-sparse minimum eigenvalue of a matrix $A \in \mathbb{R}^{p\times p}$ as

$$\Lambda(k;A) = \inf_{\theta\in\mathbb{R}^p, \|\theta\|_0\leq k} \frac{\theta^\top A\theta}{\|\theta\|_2^2}. \tag{10}$$

Even though all the results in the next section are written in terms of the linear regression map (8), our main focus is still related to the matrices and vectors defined in (3) and (6).

3. Deterministic Results for Linear Regression

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 (9). These are, respectively, the maximal k-sparse eigenvalue of $\Sigma_1 - \Sigma_2$ and the maximal k-sparse

$\|\cdot\|_2$-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 $k \geq 1$,*

$$\sup_{M\in\mathcal{M}(k)} \|\Sigma_1(M) - \Sigma_2(M)\|_{op} \leq k|||\Sigma_1 - \Sigma_2|||_\infty,$$
$$\sup_{M\in\mathcal{M}(k)} \|\Gamma_1(M) - \Gamma_2(M)\|_2 \leq k^{1/2} \|\Gamma_1 - \Gamma_2\|_\infty$$

Proof. It is easy to see that

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

Here we have used inequalities (1). A similar proof implies the second result. □

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 sub-optimal dependence in the dimension. For the special cases of independent and dependent random vectors discussed in Sections 4 and 5, we directly control RIP and $\mathcal{D}$.

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

$$S_{r,k}(\Sigma, \Gamma) := \sup_{M\in\mathcal{M}(k)} \|\beta_M(\Sigma, \Gamma)\|_r = \sup_{M\in\mathcal{M}(k)} \left\|(\Sigma(M))^{-1}\,\Gamma(M)\right\|_r. \tag{11}$$

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

$$RIP(k, \Sigma_1 - \Sigma_2) \leq \Lambda(k; \Sigma_2). \tag{12}$$

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 (8) 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,

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

To control Δ_1, note that

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

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

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

Using the triangle inequality of $\|\cdot\|_2$-norm and assumption (12), 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 sub-optimal 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 (11)) 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 model 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 in control of statistician. So, in the classical analysis, an growth rate on $\|\beta\|_2$ is imposed as an assumption.

From the viewpoint taken in this paper under misspecification the nature picks the whole distribution sequence of random vectors and the quantity $\beta_M(\cdot,\cdot)$ came up in the analysis. In the full generality of linear regression maps, we do not know of any techniques to bound the norm of this vector. It is, however, possible to bound this, if $\beta_M(\cdot,\cdot)$ is defined by a least squares linear regression problem. Recall the definition of Σ_n, Γ_n from (6) and $\beta_{n,M}$ from (7). Observe that by definition of $\beta_{n,M}$,

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

Hence for every $M \in \mathcal{M}(p)$,

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

Therefore, using the definitions of $\Lambda(k;\Sigma_n)$ and $S_{r,k}$ in (10) and (11),

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

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. $\diamond$

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

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

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

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

$$\begin{aligned} &\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\|_1 \\ &\qquad \le |M|^{1/2}\frac{\mathcal{D}\left(k, \Gamma_1-\Gamma_2\right) + \mathit{RIP}(k,\Sigma_1-\Sigma_2)\,\|\beta_M(\Sigma_2,\Gamma_2)\|_2}{\Lambda(k;\Sigma_2) - \mathit{RIP}(k,\Sigma_1-\Sigma_2)}. \end{aligned}$$

Proof. The proof follows by using the first inequality in (1). $\square$

The results above only prove a rate of convergence which gives uniform consistency. These results are not readily applicable for inference. From classical asymptotic theory, we know that for inference about a parameter an asymptotic distribution result is required. It is also well-known that asymptotic normality of an estimator is usually proved by proving an asymptotic linear representation. In what follows we prove uniform-in-model linear representation for the linear regression map. The result in terms of the regression map itself can be too abstract. For this reason, it might be helpful to revisit the usual estimators $\hat{\beta}_{n,M}$ and $\beta_{n,M}$ from (4) and (7) 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

$$\left\|\hat{\beta}_{n,M} - \beta_{n,M} - \left[\Sigma_n(M)\right]^{-1}\left(\hat{\Gamma}_n(M) - \hat{\Sigma}_n(M)\beta_{n,M}\right)\right\|_2 \approx 0. \tag{13}$$

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 the left hand side quantity in (13) is called the linear representation error. Now using essentially the same argument and letting Σ_1 (and Σ_2) take place of $\hat{\Sigma}_n$ (and Σ_n), we get the following result. Recall the notation $S_{2,k}(\cdot,\cdot)$ and $\Lambda(\cdot,\cdot)$ from Equations (11) and (10).

Theorem 3.3. *(Uniform Linear Representation) Let $k \ge 1$ be any integer such that*

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

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

$$\begin{aligned}\Big\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2) - [\Sigma_2(M)]^{-1}\left(\Gamma_1(M) - \Sigma_1(M)\beta_M(\Sigma_2,\Gamma_2)\right)\Big\|_2 \\ \le \frac{\mathit{RIP}(k,\Sigma_1-\Sigma_2)}{\Lambda(k;\Sigma_2)}\left\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\right\|_2.\end{aligned} \tag{14}$$

Furthermore, using Theorem 3.1, we get

$$\begin{aligned}\sup_{M\in\mathcal{M}(k)}\Big\|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2) - [\Sigma_2(M)]^{-1}\left(\Gamma_1(M) - \Sigma_1(M)\beta_M(\Sigma_2,\Gamma_2)\right)\Big\|_2 \\ \le \frac{\mathit{RIP}(k,\Sigma_1-\Sigma_2)}{\Lambda(k;\Sigma_2)}\frac{\mathcal{D}\left(k,\Gamma_1-\Gamma_2\right) + \mathit{RIP}(k,\Sigma_1-\Sigma_2)S_{2,k}(\Sigma_2,\Gamma_2)}{\Lambda(k;\Sigma_2) - \mathit{RIP}(k,\Sigma_1-\Sigma_2)}\end{aligned} \tag{15}$$

Proof. From the definition (8) of $\beta_M(\Sigma,\Gamma)$, we have

$$\begin{aligned}\Sigma_1(M)\beta_M(\Sigma_1,\Gamma_1) - \Gamma_1(M) = 0, \\ \Sigma_2(M)\beta_M(\Sigma_2,\Gamma_2) - \Gamma_2(M) = 0.\end{aligned} \tag{16}$$

Adding and subtracting $\beta_M(\Sigma_2,\Gamma_2)$ from $\beta_M(\Sigma_1,\Gamma_1)$ in (16), 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

$$\begin{aligned}&(\Sigma_2(M) - \Sigma_1(M))\left(\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\right) \\ &\quad = \Sigma_2(M)\left(\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\right) - \left[\Gamma_1(M) - \Sigma_1(M)\beta_M(\Sigma_2,\Gamma_2)\right].\end{aligned} \tag{17}$$

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

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

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 (14) only proves an upper bound. It can, however, be seen from Equation (17) that for any $M \in \mathcal{M}(k)$,

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

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 (9) and (10), that

$$\mathrm{RIP}(k,\Sigma_2) = \sup_{M\in\mathcal{M}(k)} \|\Sigma_2(M)\|_{op} \quad \text{and} \quad \Lambda(k,\Sigma_1-\Sigma_2) = \inf_{\theta\in\mathbb{R}^p,\|\theta\|_0\le k} \frac{\theta^\top(\Sigma_1-\Sigma_2)\theta}{\|\theta\|_2^2}.$$

If the minimal and maximal k-sparse eigenvalues 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 (14) and (15) prove more than just linear representation. These bounds allow us to improve the bounds provided for uniform L_2-consistency. Bound (14) is of the form

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

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

$$\begin{aligned}\frac{1}{2}&\left\|[\Sigma_2(M)]^{-1}(\Gamma_1(M) - \Sigma_1(M)\beta_M(\Sigma_2,\Gamma_2))\right\|_2 \\ &\le \|\beta_M(\Sigma_1,\Gamma_1) - \beta_M(\Sigma_2,\Gamma_2)\|_2 \\ &\quad \le 2\left\|[\Sigma_2(M)]^{-1}(\Gamma_1(M) - \Sigma_1(M)\beta_M(\Sigma_2,\Gamma_2))\right\|_2.\end{aligned} \tag{18}$$

This is a more precise result than informed by Theorem 3.1 since here we characterize the estimation error exactly up to a factor of 2. Also, note that in case of $\hat{\beta}_{n,M}$ 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 non-linear functionals such as $\hat{\beta}_{n,M}$.

If $\mathrm{RIP}(k,\Sigma_1-\Sigma_2)$ is converging to zero, then the right hand side of bound (14) is of smaller order than both the terms appearing on the left hand side (which are the same as those appearing in (18)). ⋄

Remark 3.4 (Alternative to RIP) A careful inspection of the proof of Theorem 3.3 and Theorem 3.1 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 bounded condition number of Σ_2, however, we only deal with $\mathrm{RIP}(k,\Sigma_1-\Sigma_2)$ in the following sections. ⋄

Summarizing all the results in this section it is enough to control

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

to derive uniform-in-model 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. So, if one wants to use shrinkage estimators 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-model type results.

Before proceeding to the rates of convergence of these error norms for independent and dependence data, we describe the importance of defining the linear regression map with general matrices instead of just gram matrices. Of course, it is more general now but it would be worthless in case no interesting applications exist. 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 models of size less than k and so the supremum is over

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

   models. Note that this bound is polynomial in the total number of covariates but is exponential in the size of models 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

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

   and so, RIP in this case is a supremum of at least $(ep/k)^k$ 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 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 the sample mean gram matrix estimator 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 the estimator and its properties. It should be noted that they do not study the estimator accuracy with respect to the RIP-norm. We do not prove it here and will be explored in the future.
2. *Outlier Contamination:* Real data, more often than not, is contaminated with outliers and it is a hard problem to remove/classify observations in case contamination is present. Robust statistics provide estimators that can ignore or down-weigh the observations suspected to be outliers and behave comparably when there is no contamination present in the data. Some simple examples include entry-wise median, or trimmed mean. See Minsker (2015) and reference 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 error. Chen et al. (2013) provide a simple estimator of the gram matrix under adversarial corruption and case-wise 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 in page 4 of Loh and Wainwright (2012), this setting includes some interesting cases like missing data and noisy covariates. A brief hint of the settings is given below:
   - If $Z_i = X_i + W_i$ with W_i independent of X_i. Also, W_i is assumed to be of 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.
   - If $Z_i = X_i \odot u_i$, where $u_i \in \mathbb{R}^p$ is again a random vector independent of Xi and $\odot$ is the Hadamard 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 (3). 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. So, all our results in this section apply to these settings.

In the following two sections, we prove finite sample non-asymptotic bounds for $\text{RIP}(k, \Sigma_1 - \Sigma_2)$ and $\mathcal{D}(k, \Gamma_1 - \Gamma_2)$ for
$$\Sigma_1 = \hat{\Sigma}_n, \quad \Sigma_2 = \Sigma_n \quad \text{and} \quad \Gamma_1 = \hat{\Gamma}_n, \quad \Gamma_2 = \Gamma_n.$$
See Equations (3) and (6). For convenience, we rewrite Theorem 3.3 for this setting. Also, for notational simplicity, let
$$\Lambda_n(k) := \Lambda(k, \Sigma_n), \ \text{RIP}_n(k) := \text{RIP}(k, \hat{\Sigma}_n - \Sigma_n) \quad \text{and} \quad \mathcal{D}_n(k) := \mathcal{D}(k, \hat{\Gamma}_n - \Gamma_n). \tag{19}$$
Recall the definition of $\hat{\beta}_{n,M}$, $\beta_{n,M}$ and $S_{2,k}$ from (5), (7) and (11).

Theorem 3.4. *Let $k \ge 1$ be any integer such that $RIP_n(k) \le \Lambda_n(k)$. Then for all models $M \in \mathcal{M}(k)$,*
$$\sup_{M \in \mathcal{M}(k)} \left\| \hat{\beta}_{n,M} - \beta_{n,M} - \frac{1}{n} \sum_{i=1}^{n} \left[\Sigma_n(M)\right]^{-1} X_i(M) \left(Y_i - X_i^\top(M)\beta_{n,M}\right) \right\|_2 \le \frac{RIP_n(k)}{\Lambda_n(k)} \left(\frac{\mathcal{D}_n(k) + RIP_n(k) S_{2,k}(\Sigma_n, \Gamma_n)}{\Lambda_n(k) - RIP_n(k)} \right).$$

It is worth recalling here that Γ_n and Σ_n are **non-random** matrices given in (6). So, Theorem 3.4 proves an asymptotic linear representation.

Remark 3.5 (Non-uniform Bounds) The bound above applies for any k satisfying the assumption $\mathrm{RIP}_n(k) \le \Lambda_n(k)$ and noting that for $M \in \mathcal{M}(k)$, $\mathrm{RIP}_n(|M|) \le \mathrm{RIP}_n(k)$ as well as $\Lambda_n(|M|) \ge \Lambda_n(k)$, Theorem 3.4 implies that

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

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

Remark 3.6 (Post-selection Consistency) One of the main importance of our results is in proving consistency of the least squares linear regression estimator after data exploration. Suppose a random model $\hat{M}$ chosen based on data satisfies $|\hat{M}| \le k$ with probability converging to one. Then with probability converging to one,

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

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

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

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

which under the assumption of bounded second moment of Y_i's is of the order $\Lambda_n^{-1/2}(k)$. Because of this, we do not further discuss $S_{2,k}$ separately. ⋄

4. Application to Independent Observations

In this section, we derive bounds for $\mathrm{RIP}_n(k)$ and $\mathcal{D}_n(k)$ defined in (19) under the assumption of independence and 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 exists positive reals $\alpha > 0$, and $K_{n,p} > 0$ such that

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

(JExp) Assume that there exists positive reals $\alpha > 0$, 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 j-th coordinate of X_i. The notation $\|\cdot\|_{\psi_\alpha}$ refers to a quasi-norm 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 since $\|W\|_{\psi_\alpha} < \infty$ implies that for all $t \ge 0$,

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

and the right hand side resembles the survival function of a Weibull random variable of order $\alpha > 0$. The special cases $\alpha = 1, 2$ are very much used in high-dimensional literature as an assumed tail behavior. A random variable W satisfying $\|W\|_{\psi_\alpha} < \infty$ with $\alpha = 2$ are called sub-Gaussian and with $\alpha = 1$ are called sub-exponential; 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 a much stronger assumption than (MExp) since (JExp) implies that the coordinates of X_i should be "almost" independent; see Chapter 3 of Vershynin (2018) for further discussion.

The following results bound $\mathcal{D}_n(k)$ and $\mathrm{RIP}_n(k)$ based on Theorem A.1 in Appendix A. Before stating the results, we need the following preliminary calculations and notation. For any set K with metric $d(\cdot,\cdot)$, a set $\mathcal{N}$ is called an 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

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

Let $\mathcal{N}_d(\varepsilon)$ represent the 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

$$\Theta_k := \{\theta \in \mathbb{R}^p :\, \|\theta\|_0 \le k,\, \|\theta\|_2 \le 1\}. \tag{20}$$

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 $1/4$-net $\mathcal{N}(\varepsilon, \Theta_k)$ of Θ_k with respect to the Euclidean norm on $\mathbb{R}^p$ can be chosen to satisfy

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

and so, can be bounded in cardinality as

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

Applying Lemma 4.1 of Pollard (1990) it follows that

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

(Lemma 4.1 of Pollard (1990) provides the bound on the covering number to be $(3/\varepsilon)^d$ but it can be improved from the proof to $(1+1/\varepsilon)^d$.) Here one can chose 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

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

Therefore,

$$\mathcal{D}_n(k) \le 2 \sup_{\theta\in\mathcal{N}(1/2,\Theta_k)} \left|\frac{1}{n}\sum_{i=1}^n \left\{\theta^\top X_iY_i - \mathbb{E}\left[\theta^\top X_iY_i\right]\right\}\right|. \tag{21}$$

By a similar argument, it can be derived that

$$\mathrm{RIP}_n(k) \le 2 \sup_{\theta\in\mathcal{N}(1/4,\Theta_k)} \left|\frac{1}{n}\sum_{i=1}^n \left\{\left(X_i^\top\theta\right)^2 - \mathbb{E}\left[\left(X_i^\top\theta\right)^2\right]\right\}\right|. \tag{22}$$

See Lemma 2.2 of Vershynin (2012) for a derivation. It is important to realize here that 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.

Using Theorem A.1 of Appendix A, we get the following results under independence.

Theorem 4.1. *Fix $n, k \ge 1$ and let $t \ge 0$ be any real number. Define*

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

Then the following probability statements hold true:

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

$$\begin{aligned}
\mathcal{D}_n(k) \le 14\sqrt{\frac{\Upsilon_{n,k}^{\Gamma}(t + k\log(3ep/k))}{n}} \\
+ C_\alpha K_{n,p}^2 \frac{k^{1/2}(\log(2n))^{2/\alpha}(t + k\log(3ep/k))^{1/T_1(\alpha/2)}}{n},
\end{aligned}$$

and

$$\begin{aligned}
RIP_n(k) \le 14\sqrt{\frac{\Upsilon_{n,k}^{\Sigma}(t + k\log(5ep/k))}{n}} \\
+ C_\alpha K_{n,p}^2 \frac{k(\log(2n))^{2/\alpha}(t + k\log(5ep/k))^{1/T_1(\alpha/2)}}{n}.
\end{aligned}$$

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

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

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

and

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

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 (21) and (22). □

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

Corollary 4.1. *Under the setting and notation of Theorem 4.1, we have the following rates of convergence if $K_{n,p} = O(1)$.*

(a) *Under Assumption (MExp),*

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

(b) *Under Assumption (JExp),*

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

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

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

We believe these to be optimal since if X and Y are independent and are 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. ◇

A direct application of Corollary 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 Equation (19) and also $\hat{\beta}_{n,M}, \beta_{n,M}$ from Equations (5), (7).

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

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

(a) 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

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

Remark 4.2 The result can be made much more precise by giving the exact tail bound for all the quantities using the exact result Theorem 4.1. We leave it to the reader to derive these results. 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 the minimax optimal rate of convergence for high-dimensional linear regression estimator under a true linear model with 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 (Case of Fixed Covariates) Note that the results in this section are under absolutely no model assumptions but are derived under independence and exponential tails. Also, our results only require independence of random vectors and no assumptions was made about the distributions of individual random vectors in terms of being identical.

It is worth mentioning a special case of our setting that is popular in classical as well as modern linear regression literature: the setting of fixed covariates. One of the classical assumptions in linear regression is non-stochastic covariates. As explained in Buja et al. (2014), this assumption has its roots in the ancilarity theory under a true linear model. If the covariates are non-stochastic, 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. 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 (11). $\diamond$

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

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

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

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

So, 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 $1/n$; see Serfling (1980, Chapter 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 $(5ep/k)^k$ many degenerate U-statistics of order 2, which is expected to be order $(\log(5ep/k)^k)/n = (k\log(5ep/k))/n$. See de la Peña and Giné (1999) for results about supremum of degenerate U-statistics. $\diamond$

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 case the dimension grows, Assumption (JExp) cannot be satisfied with non-zero mean of X_i's unless $\|\mathbb{E}[X_i]\|_2 = O(1)$. Under Assumption (MExp), these quantities can grow with k and it is hard to pinpoint their growth rate. In many cases, it is reasonable to assume bounded operator norm of the covariance matrix instead of the second moment 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 covariates and response, respectively. Without the assumption of equality of $\mathbb{E}[X_i]$, $\hat{\Sigma}_n^*$ is not consistent for the covariance matrix of $\bar{X}$. Define

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

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

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

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

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

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

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

From the calculations presented above, it follows that

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

The right hand side terms above can be controlled using Theorem A.1. Thus, the linear representation changes when using the sample covariance matrix. ◇

5. Application to 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 too, 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 et al. (2013) and Wu and Wu (2016). For a study of dependent processes under a similar framework, 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 independent and identically distributed random variables on some measurable space $(\mathcal{E}, \mathcal{B})$. Define the q-dimensional process W_i with causal representation as

$$W_i = G_i(\ldots, \varepsilon_{i-1}, \varepsilon_i) \in \mathbb{R}^q, \tag{23}$$

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 non-decreasing filtration

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

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

$$\delta_{s,r,j} := \max_{1 \le i \le n} \left\|W_i(j) - W_{i,s}(j)\right\|_r, \quad \text{and} \quad \Delta_{m,r,j} := \sum_{s=m}^{\infty} \delta_{s,r,j}, \tag{24}$$

where

$$W_{i,s}(j) := g_{ij}(\mathcal{F}_{i,i-s}) \quad \text{with} \quad \mathcal{F}_{i,i-s} := \sigma\left(\ldots, \varepsilon_{i-s-1}, \varepsilon_{i-s}', \varepsilon_{i-s+1}, \ldots, \varepsilon_{i-1}, \varepsilon_i\right). \tag{25}$$

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 j-th coordinate is given by

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

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

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

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 fits 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 (25) that

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

Here W_i' represents an independent and identically distributed copy of W_i. Hence,

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

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

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

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. $\diamond$

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

$$s(\lambda) := (1/2 + 1/\lambda)^{-1}, \quad \text{and} \quad T_1(\lambda) := \min\{\lambda, 1\} \quad \text{for all} \quad \lambda > 0. \tag{26}$$

Theorem 5.1. *Suppose $Z_1, \ldots, Z_n$ are random vectors in $\mathbb{R}^q$ with a causal representation such as (23) with mean zero. Assume that for some $\alpha > 0$, and $\nu > 0$,*

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

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 $t \ge 0$, with probability at least $1 - 8e^{-t}$,

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

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-model 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 iid 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

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

The coupled random variable $W_{i,s}$ has the same expectation as W_i. Since the quantities we need to bound involve product 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 4.1 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 (21) and (22) that

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

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 (20))

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

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

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

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

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

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

and

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

Here $T_1(\alpha)$ and $s(\alpha)$ are functions given in (26) 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 Lemma B.3 and Lemma B.4, it follows that

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

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

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 (19) and also $\hat\beta_{n,M}, \beta_{n,M}$ from Equations (5), (7).

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 $n \to \infty$.*

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

and

$$\begin{aligned}
&\sup_{M\in\mathcal{M}(k)}\left\|\hat{\beta}_{n,M}-\beta_{n,M}-\frac{1}{n}\sum_{i=1}^{n}\left[\Sigma_n(M)\right]^{-1}X_i(M)\left(Y_i-X_i^{\top}(M)\beta_{n,M}\right)\right\|_2\\
&\quad= O_p\left(\frac{\max\{\Upsilon_{n,k}^{\Gamma},\Upsilon_{n,k}^{\Sigma}\}k\log(ep/k)}{n}\right)\\
&\qquad+ K_{n,p}^4 O_p\left(\frac{k^2(\log n)^{2/s(\alpha/2)}(k\log(ep/k))^{2/T_1(s(\alpha/2))}\Omega_n^2(\nu)}{n^2}\right).
\end{aligned}$$

In comparison to Theorem 4.2, the rates attained here are very similar expect 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 the 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 dimensional 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

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

Our assumption, however, is written as

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

So, our parameters (α,ν) corresponds 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 non-stationary. Some useful examples verifying the bounds on the functional dependence measure are also provided in Zhang and Cheng (2014). ◇

6. Extensions to a General Class of M-estimators

The results derived here can be extended to a very general class of M-estimators again using deterministic inequalities. We present an example result in this section which includes many canonical generalized linear models. The proof and verification for generalized linear models will be given in a future paper. For this result, suppose $(X_i, Y_i), 1 \le i \le n$ are random vectors in $\mathbb{R}^p \times \mathbb{R}$. Consider for some loss function $L : \mathbb{R}^2 \to \mathbb{R}$, the estimator

$$\hat{\beta}_{n,M} := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{|M|}} \frac{1}{n} \sum_{i=1}^{n} h_M(X_i(M)) L(Y_i, X_i^\top(M)\theta).$$

The corresponding target of estimation is then given by

$$\beta_{n,M} := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{|M|}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[h_M(X_i(M)) L(Y_i, X_i^\top(M)\theta)\right].$$

Here $h_M(\cdot)$ is a non-negative real-valued function defined on $\mathbb{R}^{|M|}$. Assume that $L(\cdot,\cdot)$ is convex in the second argument (for all values of first argument) and is twice differentiable with respect to the second argument and let

$$L'(y,u) = \frac{\partial}{\partial t} L(y,t)\bigg|_{t=u} \quad \text{and} \quad L''(y,u) = \frac{\partial}{\partial t} L'(y,t)\bigg|_{t=u}.$$

Define

$$C_+(y,u) := \sup_{|s-t| \le u} \frac{L''(y,s)}{L''(y,t)} \quad (\ge 1).$$

Define for $M \subseteq \{1, 2, \ldots, p\}$ and for $\theta \in \mathbb{R}^{|M|}$,

$$\begin{aligned}
\hat{\mathcal{Z}}_{n,M}(\theta) &:= \frac{1}{n} \sum_{i=1}^{n} h_M(X_i(M)) L'(Y_i, X_i^\top(M)\theta) X_i(M), \\
\hat{\mathcal{J}}_{n,M}(\theta) &:= \frac{1}{n} \sum_{i=1}^{n} h_M(X_i(M)) L''(Y_i, X_i^\top(M)\theta) X_i(M) X_i^\top(M), \\
\mathcal{J}_{n,M}(\theta) &:= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[h_M(X_i(M)) L''(Y_i, X_i^\top(M)\theta) X_i(M) X_i^\top(M)\right].
\end{aligned}$$

The following theorem provides **deterministic** version of uniform-in-model $\|\cdot\|_2$-consistency and uniform-in-model linear representation for $\hat{\beta}_{n,M}$.

Theorem 6.1. *For any $n, k \ge 1$, define for $M \subseteq \{1, 2, \ldots, p\}$,*

$$\delta_{n,M} := \left\| \left[\hat{\mathcal{J}}_{n,M}(\beta_{n,M})\right]^{-1} \hat{\mathcal{Z}}_{n,M}(\beta_{n,M}) \right\|_2,$$

and for any set of models $\mathcal{M}$, set the event

$$\mathcal{E}_{\mathcal{M},n} := \left\{ \max_{M \in \mathcal{M}} \max_{1 \le i \le n} C_+\left(Y_i, 2\|X_i(M)\|_2 \delta_{n,M}\right) \le 2 \right\}.$$

On the event $\mathcal{E}_{\mathcal{M},n}$, simultaneously for all models $M \in \mathcal{M}$, there exists $\hat{\beta}_{n,M} \in \mathbb{R}^{|M|}$ satisfying

$$\hat{\mathcal{Z}}_{n,M}(\hat{\beta}_{n,M}) = 0 \quad \textit{and} \quad \frac{\delta_{n,M}}{2} \le \left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_2 \le 2\delta_{n,M}.$$

Furthermore on the event $\mathcal{E}_{\mathcal{M},n}$, simultaneously for all models $M \in \mathcal{M}$, the estimators $\hat{\beta}_{n,M}$ satisfy

$$\left\| \hat{\beta}_{n,M} - \beta_{n,M} + \left[\mathcal{J}_{n,M}(\beta_{n,M})\right]^{-1} \hat{\mathcal{Z}}_{n,M}(\beta_{n,M}) \right\|_2 \le \Delta_{n,M} \delta_{n,M},$$

where

$$\Delta_{n,M} := \frac{\left\|\hat{\mathcal{J}}_{n,M}(\beta_{n,M}) - \mathcal{J}_{n,M}(\beta_{n,M})\right\|_{op}}{\lambda_{\min}\left(\mathcal{J}_{n,M}(\beta_{n,M})\right)} + \left[\max_{1\le i\le n} C_+(Y_i, 2\left\|X_i(M)\right\|_2 \delta_{n,M}) - 1\right].$$

For generalized linear models with canonical link function, $h_M \equiv 1$ and $C_+(y,u)$ is independent of y. An application of Theorems 6.1 for $\mathcal{M} = \mathcal{M}(k)$ proceeds by verifying the following steps:

- For some $k \ge 1$ satisfying certain rate constraints
$$\sup_{|M|\le k} \left\|\hat{\mathcal{J}}_{n,M}(\beta_M) - \mathcal{J}_{n,M}(\beta_M)\right\|_{op} = o_p(1);$$
- For the same $k \ge 1$,
$$\sup_{|M|\le k} \max_{1\le i\le n} \left\|X_i(M)\right\|_2 \delta_{n,M} = o_p(1).$$

Under some mild conditions $C_+(y,u)$ converges to 1 as $u \to 0$ so that event $\mathcal{E}_{\mathcal{M}(k),n}$ holds with probability converging to 1 as $n \to \infty$. For Poisson regression and logistic regression, it is easy to verify that $C_+(y,u) \le \exp(3u)$ for all $u \ge 0$. Observe that the total number of covariates p converging to infinity as $n \to \infty$ is allowed in this analysis but is not a requirement.

7. Discussion and Conclusions

In this paper, we have proved uniform-in-model results for the least squares linear regression estimator under no model assumptions 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 exponential tails to provide logarithmic dependence on the dimension p.

From the results in Section 6, it is clear that uniform-in-models consistency and linear representation hold for a large class of M-estimators. The implications of these results is that one can use all the information from all the covariates to build a model (subset of covariates) and apply a general M-estimation technique on the final model selected. The results of Section 6 can be extended to non-differentiable loss functions using techniques from empirical process theory, in particular, the stochastic uniform equicontinuity assumption.

All of our results are free of true probability model assumptions. 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 $|M| \le k$. This implication follows from uniform-in-model feature of the results. The conclusion here is that if the statistician has a target in mind, then all he/she needs to check is if $\beta_{n,M}$ is close to the target he/she is 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 post-selection inference. The results in this paper allows for a valid post-selection inference procedure using high-dimensional central limit theorem and multiplier bootstrap. The related exploration will be provided in a future manuscript.

Acknowledgments

The authors thank Abhishek Chakrabortty for discussions that led to Remark 4.5.

Appendix 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 (2018) for a clear exposition.

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

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

Define

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

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

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

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

Proof. Fix $1 \le j \le p$ and apply Theorem 4 of Adamczak (2008) with $\mathcal{F} = \{f\}$ where $f(W_i) = W_i(j)$ for $1 \le i \le n$. 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. □

Appendix 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

$$Z_i = g_i(\ldots, \varepsilon_{i-1}, \varepsilon_i), \tag{27}$$

for some real valued function g_i. We write $\delta_{k,r} = \|Z_i - Z_{i,k}\|_r$. The following proposition bounds the r-th moment of Z_i in terms of $\|\{Z\}\|_{r,\nu}$. 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*

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

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 non-decreasing 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 $\beta \ge 0$ and $p \ge 2$,

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

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

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

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

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

and so,

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

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

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

If $\beta > 1/2$, then

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

Hence the result follows. ◻

Define the functions

$$s(\lambda) := (1/2 + 1/\lambda)^{-1}, \quad \text{and} \quad T_1(\lambda) := \min\{\lambda, 1\} \quad \text{for all} \quad \lambda > 0. \tag{28}$$

Theorem B.1. *Suppose $Z_1, \ldots, Z_n$ are elements of the causal process (27) with mean zero. If for some $\alpha > 0$, and $\nu > 0$,*

$$\|\{Z\}\|_{\psi_\alpha,\nu} = \sup_{p\ge 2} \sup_{m\ge 0} p^{-1/\alpha}(m+1)^\nu \Delta_{m,p} < \infty. \tag{29}$$

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 $p \ge 2$,

$$\left\|\sum_{i=1}^n Z_i\right\|_p \le \sqrt{pn}\,\|\{Z\}\|_{\psi_\alpha,\nu} B_\nu + C_\alpha \|\{Z\}\|_{\psi_\alpha,\nu} (\log n)^{1/s(\alpha)} p^{1/T_1(s(\alpha))}\Omega_n(\nu), \tag{30}$$

where C_α *is a constant depending only on* α, 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* (30).

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

$$S_n = S_{n,0} + (S_n - S_{n,n}) + \sum_{\ell=1}^L \left(S_{n,\xi_\ell} - S_{n,\xi_{\ell-1}}\right) := \mathbf{I} + \mathbf{II} + \mathbf{III}. \tag{31}$$

We prove the moment bound (30) by bounding the moments of each term in the decomposition (31).

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 (29). Hence, we get that for any $p \geq 1$,

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

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, $\|Z_i\|_2 \leq \|\{Z\}\|_{2,\nu}$ and hence

$$\|S_{n,0}\|_p \leq \sqrt{6pn}\,\|\{Z\}\|_{2,\nu} + C_\alpha \|\{Z\}\|_{\psi_\alpha,\nu} p^{1/T_1(\alpha)} (\log n)^{1/\alpha}. \tag{32}$$

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 non-decreasing filtration

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

It is easy to see that

$$\mathbb{E}\left[D_{n-k+1,m+1} \middle| \mathcal{G}_{k-1,m+1}\right] = 0. \tag{33}$$

Therefore, $\{(D_{n-k+1,m+1}, \mathcal{G}_{k,m+1}) : 1 \leq k \leq n\}$ 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 $p \geq 2$,

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

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

$$\|D_{k,m+1}\|_p = \left\|\mathbb{E}\left[Z_k - g(\ldots, \varepsilon'_{k-m-1}, \varepsilon_{k-m}, \ldots, \varepsilon_k) \middle| \varepsilon_k, \ldots, \varepsilon_{k-m-1}\right]\right\|_p \leq \delta_{m+1,p}. \tag{34}$$

Hence, for $p \geq 2$,

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

and

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

Under assumption (29), we obtain

$$\|\mathbf{II}\|_p = \|S_n - S_{n,n}\|_p \le \|\{Z\}\|_{\psi_\alpha,\nu} \frac{n^{1/2}p^{1/2+1/\alpha}}{(n+2)^\nu} = \|\{Z\}\|_{\psi_\alpha,\nu} n^{1/2-\nu} p^{1/2+1/\alpha}. \tag{35}$$

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

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

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

$$\begin{aligned}
A_{n,\ell} &= \sum_{i=1}^{\left\lfloor \frac{n}{2\xi_\ell} \right\rfloor} \left(\sum_{j=1}^{\xi_\ell} \mathcal{D}_{2\xi_\ell i + j, \ell}\right) = \sum_{i=1}^{\left\lfloor \frac{n}{2\xi_\ell} \right\rfloor} \left(\sum_{k=2\xi_\ell i+1}^{2\xi_\ell i + \xi_\ell} \left(Z_k^{(\xi_\ell)} - Z_k^{(\xi_{\ell-1})}\right)\right) \\
&= \sum_{i=1}^{\left\lfloor \frac{n}{2\xi_\ell} \right\rfloor} \left(M_{2\xi_\ell i + \xi_\ell, \ell} - M_{2\xi_\ell i, \ell}\right).
\end{aligned} \tag{36}$$

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

$$\|M_{j,\ell} - M_{i,\ell}\|_p \le \sum_{t=1+\xi_{\ell-1}}^{\xi_\ell} \left\| \sum_{k=i+1}^{j} \left(Z_k^{(t)} - Z_k^{(t-1)}\right)\right\|_p. \tag{37}$$

As proved in (33), 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 (34). Substituting this in inequality (37) and using $\xi_{\ell-1} \ge \xi_\ell/2$, we get

$$\begin{aligned} \|M_{j,\ell} - M_{i,\ell}\|_p &\le p^{1/2}(j-i)^{1/2} \sum_{t=1+\xi_{\ell-1}}^{\xi_\ell} \delta_{t,p} \le p^{1/2}(j-i)^{1/2}\Delta_{1+\xi_{\ell-1},p} \\ &\le \|\{Z\}\|_{p,\nu}\, p^{1/2}(j-i)^{1/2}(2+\xi_{\ell-1})^{-\nu} \\ &\le 2^\nu \|\{Z\}\|_{p,\nu}\, p^{1/2}(j-i)^{1/2}\xi_\ell^{-\nu}. \end{aligned} \tag{38}$$

Under assumption (29), we get

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

See (28) for the definition of $s(\alpha)$. Thus, for all $1 \le i \le \lfloor \frac{n}{2\xi_\ell} \rfloor$,

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

So, the summands of $A_{n,\ell}$ in the final representation in (36) are independent and satisfy the hypothesis of Theorem A.1 with $\beta = s(\alpha)$. Therefore, for $p \ge 2$,

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

Here the second inequality follows from (38).

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

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

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

$$\|M_{n,\ell}\|_p \le \frac{2^{1+\nu}}{\xi_\ell^\nu}\left[\|\{Z\}\|_{2,\nu}\sqrt{6pn} + C_\alpha \|\{Z\}\|_{\psi_\alpha,\nu} p^{1/T_1(s(\alpha))} (\log n)^{1/s(\alpha)} \xi_\ell^{1/2}\right]. \tag{39}$$

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 H́older'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 (39), we get

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

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

$$\begin{aligned}&\left\|\sum_{\ell=1}^L M_{n,\ell}\right\|_p \\ &\quad \le \frac{5\pi^3 2^2}{3\sqrt{3}}\left[\frac{2^\nu \|\{Z\}\|_{2,\nu}\sqrt{6pn}}{\nu^3} + C_\alpha \|\{Z\}\|_{\psi_\alpha,\nu}(\log n)^{1/s(\alpha)}\Omega_n(\nu) p^{1/T_1(s(\alpha))}\right].\end{aligned} \tag{40}$$

Combining the moment bounds (32), (35) and (40), it follows that for $p \ge 2$,

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

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 (20).

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 (24) 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 non-zero coordinates $\theta(j)$ of θ. Since the functional dependence measure is a norm, it follows that

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

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 $\nu > 0$,

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

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. □

References

Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13:no. 34, 1000–1034.

Arcones, M. A. (2005). Convergence of the optimal M-estimator over a parametric family of M-estimators. *Test*, 14(1):281–315.

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

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.*, 41(2):802–837.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhan, K., and Zhao, L. (2014). Models as Approximations, Part I: A Conspiracy of Nonlinearity and Random Regressors in Linear Regression. *ArXiv e-prints*.

Cai, T. T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.*, 40(4):2014–2042.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185.

Catoni, O. and Giulini, I. (2017). Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *ArXiv e-prints*.

Chen, Y., Caramanis, C., and Mannor, S. (2013). Robust sparse regression under adversarial corruption. In *Proceedings of The 30th International Conference on Machine Learning*, volume 28, pages 774–782.

de la Peña, V. H. and Giné, E. (1999). *Decoupling*. Probability and its Applications (New York). Springer-Verlag, New York. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.

Dodge, Y. and Jurevckova, J. (2000). *Adaptive regression*. Springer-Verlag, New York.

Foygel, R. and Srebro, N. (2011). Fast-rate and optimistic-rate error bounds for L1-regularized regression. *ArXiv e-prints*.

He, X. and Shao, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.*, 24(6):2608–2630.

Kuchibhotla, A. K. and Chakrabortty, A. (2018). Unified exposition of concentration bounds for sub-weibull processes with applications in high-dimensional statistics. *In Preparation*.

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

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

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *Ann. Statist.*, 40(3):1637–1664.

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

Plan, Y. and Vershynin, R. (2013). One-bit compressed sensing by linear programming. *Comm. Pure Appl. Math.*, 66(8):1275–1297.

Pollard, D. (1990). *Empirical processes: theory and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, Hayward, CA; American Statistical Association, Alexandria, VA.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.*, 12(4):1298–1309.

Portnoy, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.*, 13(4):1403–1417.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.*, 16(1):356–366.

Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic nonlinear econometric models*. Springer-Verlag, Berlin. Asymptotic theory.

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

Rio, E. (2009). Moment inequalities for sums of dependent random variables under projective conditions. *J. Theoret. Probab.*, 22(1):146–163.

Romano, J. P. and Wolf, M. (2000). A more general central limit theorem for m-dependent random variables with unbounded m. *Statist. Probab. Lett.*, 47(2):115–124.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons, Inc., New York. Wiley Series in Probability and Mathematical Statistics.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

van Zwet, W. R. (1984). A Berry-Esseen bound for symmetric statistics. *Z. Wahrsch. Verw. Gebiete*, 66(3):425–440.

Vershynin, R. (2012). How close is the sample covariance matrix to the actual covariance matrix? *J. Theoret. Probab.*, 25(3):655–686.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*.

Wei, X. and Minsker, S. (2017). Estimation of the covariance structure of heavy-tailed distributions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and

Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 2855–2864. Curran Associates, Inc.

Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA*, 102(40):14150–14154.

Wu, W. B. and Mielniczuk, J. (2010). A new look at measuring dependence. In *Dependence in probability and statistics*, volume 200 of *Lect. Notes Stat.*, pages 123–142. Springer, Berlin.

Wu, W.-B. and Wu, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Stat.*, 10(1):352–379.

Zhang, D. and Wu, W. B. (2017). Gaussian approximation for high dimensional time series. *Ann. Statist.*, 45(5):1895–1919.

Zhang, X. and Cheng, G. (2014). Bootstrapping High Dimensional Time Series. *ArXiv e-prints*.