Online Debiasing for Adaptively Collected High-dimensional Data

Yash Deshpande* Adel Javanmard† Mohammad Mehrabi‡§

December 19, 2019

Abstract

Adaptive collection of data is commonplace in applications throughout science and engineering. From the point of view of statistical inference however, adaptive data collection induces memory and correlation in the sample, and poses significant challenge.

We consider the high-dimensional linear regression, where the sample is collected adaptively, and the sample size n can be smaller than p, the number of covariates. In this setting, there are two distinct sources of bias: the first due to regularization imposed for consistent estimation, e.g. using the LASSO, and the second due to adaptivity in collecting the sample. We propose *'online debiasing'*, a general procedure for estimators such as the LASSO, which addresses both sources of bias. In two concrete contexts (i) batched data collection and (ii) time series analysis, we demonstrate that online debiasing optimally debiases the LASSO estimate when the underlying parameter θ_0 has sparsity of order $o(\sqrt{n}/\log p)$. In this regime, the debiased estimator can be used to compute p-values and confidence intervals of optimal size.

1 Introduction

Modern data collection, experimentation and modeling are often adaptive in nature. For example, clinical trials are run in phases, wherein the data from a previous phase inform and influences the design of future phases. In commercial recommendation engines, algorithms collect data by eliciting feedback from their users; data which is ultimately used to improve the algorithms underlying the recommendations. In such applications, adaptive data collection is often carried out for objectives correlated to, but distinct from statistical inference. In clinical trials, an ethical experimenter might prefer to assign more patients a treatment that they might benefit from, instead of the control treatment. In e-commerce, recommendation engines aim to minimize lost revenue to pure experimentation. In other applications, collecting data is potentially costly, and practitioners may choose to collect samples that are a priori deemed most informative. Since such objectives are intimately related to statistical estimation, it is not surprising that adaptively collected data can be used to derive statistically consistent estimates, often using standard estimators. The question of statistical inference however, is more subtle: on the one hand, consistent estimation indicates

*Institute for Data, Systems and Society, Massachusetts Institute of Technology, Email: yash@mit.edu

†Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Email: ajavanma@usc.edu

‡Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Email: mehrabim@usc.edu

§The names of the authors are in alphabetical order.

that the collected sample is informative enough. On the other hand, adaptive collection induces endogenous correlation in the sample, resulting in bias in the estimates. In this paper, we address the following natural question raised by this dichotomy:

Can adaptively collected data be used for ex post statistical inference?

We will focus on the linear model, where the sample $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$ satisfy:

$$y_i = \langle x_i, \theta_0 \rangle + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2). \tag{1}$$

Here $\theta_0 \in \mathbb{R}^p$ is an unknown parameter vector relating the covariates x_i to the response y_i, and the noise ε_i are i.i.d. $\mathsf{N}(0, \sigma^2)$ random variables. In vector form, we write Eq.(1) as

$$y = X\theta_0 + \varepsilon, \tag{2}$$

where $y = (y_1, y_2, \dots, y_n)$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ and the design matrix $X \in \mathbb{R}^{n\times p}$ has rows $x_1^{\mathsf{T}}, \dots, x_n^{\mathsf{T}}$. When the sample is adaptively collected, the data point (y_i, x_i) is obtained *after viewing the previous data points* $(y_1, x_1), \dots, (y_{i-1}, x_{i-1})$[1].

In the 'sample-rich' regime when $p < n$, the standard approach would be to compute the least squares estimate $\widehat{\theta}^{\mathsf{LS}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}y$, and assess the uncertainty in $\widehat{\theta}^{\mathsf{LS}}$ using a central limit approximation $(X^{\mathsf{T}}X)^{1/2}(\widehat{\theta}^{\mathsf{LS}} - \theta_0) \approx \mathsf{N}(0, \mathrm{I}_p)$ [LW82]. However, while the estimator $\widehat{\theta}^{\mathsf{LS}}$ is consistent under fairly weak conditions, adaptive data collection complicates the task of characterizing its distribution. One hint for this is the observation that, in stark contrast with the non-adaptive setting, $\widehat{\theta}^{\mathsf{LS}} = \theta_0 + (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\varepsilon$ is, in general, *a biased estimate* of θ_0. Adaptive data collection creates correlation between the responses y_i (therefore ε_i) and covariate vectors $x_{i+1}, x_{i+2}, \dots, x_n$ observed in the future. In the context of multi-armed bandits, where the estimator $\widehat{\theta}^{\mathsf{LS}}$ for model (1) reduces to sample averages, [XQL13, VBW15] observed such bias empirically, and [NXTZ17, SRR19] characterized and developed upper bounds on the bias. While bias is an important problem, we remark here that the estimate also shows higher-order distributional defects that complicate inferential tasks.

This phenomenon is exacerbated in the high-dimensional or 'feature-rich' regime when $p > n$. Here the design matrix X becomes rank-deficient, and consistent parameter estimation requires (i) additional structural assumptions on θ_0 and (ii) regularized estimators beyond $\widehat{\theta}^{\mathsf{LS}}$, such as the LASSO [Tib96]. Such estimators are non-linear, non-explicit and, consequently it is difficult to characterize their distribution even with strong random design assumptions [BM12, JM14b]. In analogy to the low-dimensional regime, it is relatively easier to develop consistency guarantees for estimation using the LASSO when $p > n$. Given the sample $(y_1, x_1), \dots (y_n, x_n)$ one can compute the LASSO estimate $\widehat{\theta}^{\mathsf{L}} = \widehat{\theta}^{\mathsf{L}}(Y, X; \lambda)$

$$\widehat{\theta}^{\mathsf{L}} = \arg\min_{\theta} \Big\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \Big\}, \tag{3}$$

If θ_0 is sparse with at most $s_0 \ll p$ non-zero entries and the design X satisfies some technical conditions, the LASSO estimate, for an appropriate choice of λ_n has mean squared error $\mathbb{E}\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_2^2$ of order $\sigma^2 s_0 (\log p)/n$ [BM+15, BB15]. In particular the estimate is consistent provided the sparsity

[1] Formally, we assume a filtration $(\mathfrak{F}_i)_{i\le n}$ to which the sequence $(y_i, x_i)_{i\le n}$ is adapted, and with respect to which the sequence $(x_i)_{i\le n}$ is predictable

satisfies $s_0 = o(n/\log p)$, This estimator is biased because of two distinct reasons. The first is the regularization imposed in Eq.(3), which disposes $\widehat{\theta}^{\mathsf{L}}$ to have small ℓ_1 norm. The second is the correlation induced between X and ε due to adaptive data collection. To address the first source, [ZZ11, JM14a, VdGBR+14] proposed a debiased estimate

$$\widehat{\theta}^{\mathsf{off}} = \widehat{\theta}^{\mathsf{L}} + \frac{1}{n} M X^{\mathsf{T}}(y - X\widehat{\theta}^{\mathsf{L}}), \tag{4}$$

where M is chosen as an 'approximate inverse' of the sample covariance $\widehat{\Sigma} = X^{\mathsf{T}}X/n$. The intuition for this idea is the following decomposition that follows directly from Eqs.(1), (4):

$$\widehat{\theta}^{\mathsf{off}} - \theta_0 = (I_p - M\widehat{\Sigma})(\widehat{\theta}^{\mathsf{L}} - \theta_0) + \frac{1}{n} M X^{\mathsf{T}}\varepsilon. \tag{5}$$

When the data collection is non-adaptive, X and ε are independent and therefore, conditional on the design X, $MX^{\mathsf{T}}\varepsilon/n$ is distributed as $\mathsf{N}(0, \sigma^2 Q/n)$ where $Q = M\widehat{\Sigma}M^{\mathsf{T}}$. Further, the bias in $\widehat{\theta}^{\mathsf{off}}$ is isolated to the first term, which intuitively should be of smaller order than the second term, provided both $\widehat{\theta}^{\mathsf{L}} - \theta_0$ and $M\widehat{\Sigma} - \mathrm{I}_p$ are small in an appropriate sense. This intuition suggests that, if the second term dominates the first term in $\widehat{\theta}^{\mathsf{off}}$, we can produce confidence intervals for θ_0 in the usual fashion using the debiased estimate $\widehat{\theta}^{\mathsf{off}}$ [JM14a, JM14b, VdGBR+14]. For instance, with $Q = M\widehat{\Sigma}M^{\mathsf{T}}$, the interval $\big[\widehat{\theta}^{\mathsf{off}}_1 - 1.96\sigma\sqrt{Q_{11}/n}, \widehat{\theta}^{\mathsf{off}}_1 + 1.96\sigma\sqrt{Q_{11}/n}\big]$ forms a standard 95% confidence interval for the parameter $\theta_{0,1}$. In the so-called 'random design' setting –when the rows of X are drawn i.i.d. from a broad class of distributions– this approach to inference via the debiased estimate $\widehat{\theta}^{\mathsf{off}}$ enjoys several optimality guarantees: the resulting confidence intervals have minimax optimal size [Jav14, JM14a, CG17], and are semi-parametrically efficient [VdGBR+14].

This line of argument breaks down when the sample is adaptively collected as the debiased estimate $\widehat{\theta}^{\mathsf{off}}$ still suffers the second source of bias. Indeed, this is exactly analogous to $\widehat{\theta}^{\mathsf{LS}}$ in low dimensions. Since M, X and the noise ε are correlated, we can no longer assert that the term $MX^{\mathsf{T}}\varepsilon/n$ is unbiased. Indeed, characterizing its distribution in general may be quite difficult, given the intricate correlation between M, X and ε induced by the data collecting policy and the procedure choosing M. As we will see in specific examples in Sections 2 and 3, this bias can have a dramatic influence on the validity of $\widehat{\theta}^{\mathsf{off}}$ to compute confidence intervals and p-values.

Online debiasing Our approach builds on the insight in [DMST18] who propose the following *online debiased* estimator $\widehat{\theta}^{\mathsf{on}} = \widehat{\theta}^{\mathsf{on}}(y, X; (M_i)_{i\le n}, \lambda)$ of the form

$$\widehat{\theta}^{\mathsf{on}} \equiv \widehat{\theta}^{\mathsf{L}} + \frac{1}{n}\sum_{i=1}^{n} M_i x_i (y_i - x_i^{\mathsf{T}}\widehat{\theta}^{\mathsf{L}}). \tag{6}$$

The term 'online' is from the crucial constraint of *predictability* imposed on the sequence $(M_i)_{i\le n}$.

Definition 1.1. *(Predictability) Without loss of generality, there exists a filtration $(\mathfrak{F}_i)_{i\ge 0}$ so that, for $i = 1, 2, \dots, n$, (i) ε_i are adapted to $\mathfrak{F}_i$ and ε_i is independent of $\mathfrak{F}_j$ for $j < i$. We assume that the sequences $(x_i)_{i\ge 1}$ and $(M_i)_{i\ge 1}$ are* predictable *with respect to $\mathfrak{F}_i$, i.e. for each i, x_i and M_i are measurable with respect to $\mathfrak{F}_{i-1}$.*

With predictability, the data points (y_i, x_i) are adapted to the filtration $(\mathfrak{F}_i)_{i\le n}$ and, moreover, the covariates x_i are predictable with respect to $\mathfrak{F}_i$. Intuitively, the σ-algebra $\mathfrak{F}_i$ contains all information in the data, as well as potential external randomness, that is used to query the new data point (y_{i+1}, x_{i+1}). Predictability ensures that only this information may be used to construct the matrix M_{i+1}. Analogous to Eq.(5) we can decompose $\widehat{\theta}^{\mathsf{on}}$ into two components:

$$\widehat{\theta}^{\mathsf{on}} = \theta_0 + \frac{1}{\sqrt{n}}\big(B_n(\widehat{\theta}^{\mathsf{L}} - \theta_0) + W_n\big) \tag{7}$$

$$\text{where } B_n \equiv \sqrt{n}\Big(I_p - \frac{1}{n}\sum_i M_i x_i x_i^{\mathsf{T}}\Big),$$

$$\text{and } W_n \equiv \frac{1}{\sqrt{n}}\sum_i M_i x_i \varepsilon_i.$$

Predictability of $(M_i)_{i\le n}$ ensures that W_n is unbiased and the bias in $\widehat{\theta}^{\mathsf{on}}$ is contained entirely in the first term $B_n(\widehat{\theta}^{\mathsf{L}} - \theta_0)$. In fact, the sequence $\sqrt{n}W_n = \sum_i M_i x_i \varepsilon_i$ is a martingale and, assuming a martingale central limit behavior, we might expect that W_n is approximately Gaussian. With this isolation achieved, the main idea is to minimize the (conditional) variance of the term W_n, while keeping the bias, quantified by B_n, of stochastically smaller order. In [DMST18], this is done using ridge regression to construct the debiasing sequence $(M_i)_{i\le n}$, in the low-dimensional setting (p fixed and n diverging). However, that approach yields strictly sub-optimal results in the high-dimensional regime of $p > n$. An important contribution of this paper is to devise an online debiasing approach for high-dimensional regime.

1.1 Contributions

In this paper, we will develop online debiasing for high-dimensional regression with adaptive data collection, which can be used for statistical inference. We focus on two canonical scenarios of adaptive data collection: (i) batched data collection and (ii) time series analysis.

Batched data collection: We model a data collecting process that operates in two batches or phases. Data points collected in the first batch and, in particular, estimates computed on them influence the data collection in the second batch.

Autoregressive time series: In this setting, the data is a high-dimensional time series $z_1, z_2, \dots$. We consider the standard autoregressive model (AR) of bounded order d wherein the data $z_t \in \mathbb{R}^N$ at time point t is modeled, up to exogenous variation, as a linear function of the previous d time points $z_{t-1}, z_{t-2}, \dots, z_{t-d}$.

The rest of the paper is organized as follows. Sections 2 and 3 develop concrete debiasing schemes $(M_i)_{i\le n}$ for batched data collection and time series. In each case we (i) provide concrete recipes to construct the debiasing sequence $(M_i)_{i\le n}$ and (ii) prove asymptotic distributional characterizations of online debiased estimators $\widehat{\theta}^{\mathsf{on}}$ under reasonable assumptions. Section 4 shows how the distributional characterizations obtained may be used to construct valid confidence intervals, and p-values. Section 5 contains numerical experiments that demonstrate the validity our proposals on both synthetic and real data. In Section 6 we provide simple iterative algorithms based on projected gradient and coordinate descent to compute the debiasing sequence $(M_i)_{i\le n}$ from data.

Notation Henceforth, we use the shorthand $[p] \equiv \{1, \dots, p\}$ for an integer $p \geq 1$, and $a \wedge b \equiv \min(a, b)$, $a \vee b \equiv \max(a, b)$. We also indicate the matrices in upper case letters and use lower case letters for vectors and scalars. We write $\|v\|_p$ for the standard ℓ_p norm of a vector v, $\|v\|_p = (\sum_i |v_i|^p)^{1/p}$ and $\|v\|_0$ for the number of nonzero elements of v. We also denote by $\text{supp}(v)$, the support of v that is the positions of its nonzero entries. For a matrix A, $\|A\|_p$ represents its ℓ_p operator norm and $\|A\|_\infty = \max_{i,j} |A_{ij}|$ denotes the maximum absolute value of its entries. For two matrices A, B, we use the shorthand $\langle A, B\rangle \equiv \text{trace}(A^\mathsf{T} B)$. In addition $\phi(x)$ and $\Phi(x)$ respectively represents the probability density function and the cumulative distribution function of standard normal variable. Also, we use the term *with high probability* to imply that the probability converges to one as $n \to \infty$.

2 Batched data collection

As a prototypical example of adaptive data collection in practice, we will consider a stylized model wherein the experimenter (or analyst) collects data in two phases or batches. In the first phase, the experimenter collects an initial sample $(y_1, x_1), \dots, (y_{n_1}, x_{n_1})$ of size $n_1 < n$ where the responses follow Eq.(1) and the covariates are i.i.d. from a distribution $\mathbb{P}_x$. Following this, she computes an intermediate estimate $\widehat{\theta}^1$ of θ_0 and then collects an additional sample $(y_{n_1+1}, x_{n_1+1}), \dots, (y_n, x_n)$ of size $n_2 = n - n_1$, where the covariates x_i are drawn independently from the law of x_1, conditional on the event $\{\langle x_1, \widehat{\theta}^1\rangle \geq \varsigma\}$, where ς is a threshold, that may be data-dependent. This is a reasonable model in scenarios where the response y_i represents an instantaneous reward that the experimenter wishes to maximize, as in multi-armed bandits [LR85, BCB+12]. The experimenter then faces the classic exploration-exploitation dilemma: she has to trade-off learning θ_0 well, which is necessary to maximize her long-term reward, and exploiting what is known about θ_0 to immediately accrue reward. As an example, clinical trials may be designed to be response-adaptive and allocate patients to treatments that they are likely to benefit from based on prior data [ZLK+08, KHW+11]. The multi-armed bandit problem is a standard formalization of this trade-off, and a variety of bandit algorithms are designed to operate in distinct phases of 'explore–then exploit'[RT10, DM12, BB15, PRC+16]. The model we describe above is a close approximation of data collected from one arm in a run of such an algorithm.

With the full samples $(y_1, x_1), \dots, (y_n, x_n)$ at hand, the experimenter would like to perform inference on a fixed coordinate $\theta_{0,a}$ of the underlying parameter. It might still be reasonable to expect $\widehat{\theta}^\mathsf{L} = \widehat{\theta}^\mathsf{L}(y, X; \lambda)$ to have small estimation error. Indeed, this can be shown to hold provided the sample covariance $\widehat{\Sigma} = (1/n)\sum_i x_i x_i^\mathsf{T}$ satisfies the *compatibility condition* [BVDG11].

Definition 2.1 (Compatibility). *Fix a subset $S \subseteq [p]$ and a number $\phi_0 > 0$, a matrix $\widehat{\Sigma} \in \mathbb{R}^{p\times p}$ satisfies the (ϕ_0, S)-compatibility condition, or is (ϕ_0, S)-compatible, if for every non-zero vector v with $\|v_{S^c}\|_1 \leq 3\|v_S\|_1$:*

$$\frac{|S|\langle v, \widehat{\Sigma} v\rangle}{\|v\|_1^2} \geq \phi_0.$$

The following theorem is a version of Theorem 6.1 in [BVDG11] and is proved in an analogous manner. We refer to Appendix A.1 for its proof.

Theorem 2.2 ([BVDG11, Theorem 6.1]). *Suppose that the true parameter θ_0 is s_0-sparse and the distribution $\mathbb{P}_x$ is such that with probability one the following two conditions hold: (i) the covariance $\mathbb{E}\{xx^{\mathsf{T}}\}$ and $\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}^1\rangle \geq \varsigma\}$ are $(\phi_0, \mathrm{supp}(\theta_0))$-compatible and (ii) x as well as $x|_{\langle x,\widehat{\theta}^1\rangle\geq\varsigma}$ are κ-subgaussian. Suppose that $n \geq 400^2(\kappa^4/\phi_0^2)s_0^2 \log p$ the LASSO estimate $\widehat{\theta}^{\mathsf{L}}(y, X; \lambda_n)$ with $\lambda_n = 40\kappa\sigma\sqrt{(\log p)/n}$ satisfies, with probability exceeding $1 - p^{-3}$,*

$$\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \leq \frac{3s_0\lambda_n}{\phi_0} = \frac{120\kappa\sigma}{\phi_0} s_0\sqrt{\frac{\log p}{n}}.$$

Remark 2.3. *(Estimating the noise variance) For the correct estimation rate using the LASSO, Theorem 2.2 requires knowledge of the noise level σ, which is used to calibrate the regularization λ_n. Other estimators like the scaled LASSO [SZ12] or the square-root LASSO [BCW11] allows to estimate σ consistently when it is unknown. This can be incorporated into the present setting, as done in [JM14a]. For simplicity, we focus on the case when the noise level is known. However, the results hold as far as a consistent estimate of σ is used. Formally, an estimate $\widehat{\sigma} = \widehat{\sigma}(y, X)$ of the noise level satisfying, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \sup_{\|\theta_0\|_0\leq s_0} \mathbb{P}\left(\left|\frac{\widehat{\sigma}}{\sigma} - 1\right| \geq \varepsilon\right) = 0\,. \tag{8}$$

Remark 2.4. *At the expense of increasing the absolute constants in Theorem 2.2, the probability $1 - p^{-3}$ can be made $1 - p^{-C}$ for any arbitrary constant $C > 1$.*

As an example, non-degenerate Gaussian distributions satisfy the conditions required of $\mathbb{P}_x$ in Theorem 2.2. The proof of next Example is deferred to Section A.4 (See Lemmas A.12 and A.13)

Example 2.5. *(Compatibility for Gaussian designs) Suppose that $\mathbb{P}_x = \mathsf{N}(0, \Sigma)$ for a positive definite covariance Σ. Then, for any vector $\widehat{\theta}$ and subset $S \subseteq [p]$, the second moments $\mathbb{E}\{xx^{\mathsf{T}}\}$ and $\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\}$ are (ϕ_0, S)-compatible with $\phi_0 = \lambda_{\min}(\Sigma)/16$.*

Furthermore, both x and $x|_{\langle x,\widehat{\theta}\rangle\geq\varsigma}$ are κ-subgaussian with $\kappa = 2\lambda_{\max}^{1/2}(\Sigma)$.

Theorem 2.2 shows that, under an appropriate compatibility condition, the LASSO estimate admits ℓ_1 error at a rate of $s_0\sqrt{\log p/n}$. Importantly, despite the adaptivity introduced by the sampling of data, the error of LASSO estimate has the same asymptotic rate as expected without adaptivity. With slightly stronger restricted-eigenvalue conditions on the covariances $\mathbb{E}\{xx^{\mathsf{T}}\}$ and $\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}^1\rangle \geq \varsigma\}$, it is also possible to extend Theorem 2.2 to show ℓ_2 error of order $s_0 \log p/n$, analogous to the non-adaptive setting. However, since the ℓ_2 error rate will not be used for our analysis of online debiasing, we do not pursue this direction here.

Offline debiasing: a numerical illustration A natural strategy is to simply debias the estimate $\widehat{\theta}^{\mathsf{L}}$ using the methods of [JM14b, JM14a, VdGBR+14], which we refer to as 'offline' debiasing. It is instructive to see the weakness of offline debiasing on a concrete example to motivate online debiasing. Consider a simple setting where $\theta_0 \in \{0, 1\}^{600}$ with exactly $s_0 = 10$ non-zero entries. We obtain the first batch $(y_1, x_1), \dots, (y_{500}, x_{500})$ of observations with $y_i = \langle x_i, \theta_0\rangle + \varepsilon_i$, $x_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \Sigma)$

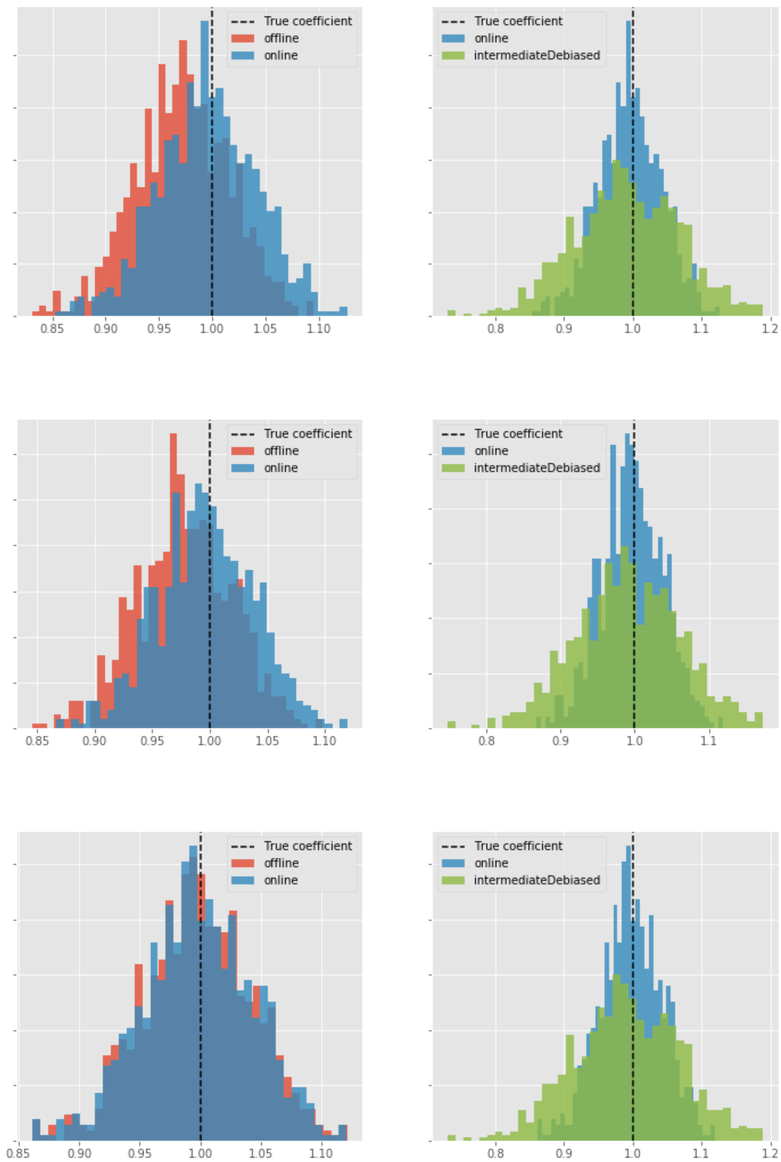


Figure 1: (Left) Histograms of the offline debiased estimate $\widehat{\theta}^{\mathsf{off}}$ and online debiased estimate $\widehat{\theta}^{\mathsf{on}}$ restricted to the support of θ_0. (Right) Histograms of the offline debiased estimate *only using the first batch* $\widehat{\theta}^1$ and the online debiased estimate $\widehat{\theta}^{\mathsf{on}}$. The dashed line indicates the true coefficient size. (Top) $\widehat{\theta}^1$ is debiased LASSO on first batch, (Middle) $\widehat{\theta}^1$ is ridge estimate on first batch, (Bottom) $\widehat{\theta}^1$ is LASSO estimate on the first batch. Offline debiasing works well once restricted to the first batch (called intermediate debiased in the plots), but then loses power in comparison. Online debiasing is cognizant of the adaptivity and debiases without losing power even in the presence of adaptivity.

and $\varepsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$ where we use the covariance Σ as below:

$$\Sigma_{a,b} = \begin{cases} 1 & \text{if } a = b, \\ 0.1 & \text{if } |a-b| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Based on this data, we construct an intermediate estimator $\widehat{\theta}^1$ on $(y^{(1)}, X_1)$ using three different strategies: (i) LASSO with oracle choice of the regularization λ, (ii) debiased LASSO and (ii) ridge regression with cross-validation.

With this estimate we now sample new covariates $x_{501}, \dots, x_{1000}$ independently from the law of $x|_{\langle x, \widehat{\theta}^1 \rangle \geq \langle \widehat{\theta}^1, \Sigma \widehat{\theta}^1 \rangle^{1/2}}$ and the corresponding outcomes $y_{501}, \dots, y_{1000}$. Unconditionally $\langle x, \widehat{\theta}^1 \rangle \sim \mathsf{N}(0, \langle \widehat{\theta}^1, \Sigma \widehat{\theta}^1 \rangle)$, so this choice of threshold corresponds to sampling covariates that correlate with $\widehat{\theta}^1$ at least one standard deviation higher than expected unconditionally.

This procedure yields two batches of data, each of $n_1 = n_2 = 500$ data points, combining to give a sample of 1000 data points. From the full dataset $(y_1, x_1), \dots, (y_{1000}, x_{1000})$ we compute the LASSO estimate $\widehat{\theta}^{\mathsf{L}} = \widehat{\theta}^{\mathsf{L}}(y, X; \lambda)$ with $\lambda = 2.5 \lambda_{\max}(\Sigma) \sqrt{(\log p)/n}$.

Offline debiasing [JM14b, JM14a, JM18] gives the following prescription to debias $\widehat{\theta}^{\mathsf{L}}$:

$$\widehat{\theta}^{\mathsf{off}} = \widehat{\theta}^{\mathsf{L}} + \frac{1}{n} \Omega(\widehat{\theta}^1) X^{\mathsf{T}} (y - X \widehat{\theta}^{\mathsf{L}}),$$

where $\Omega(\widehat{\theta})$ is the population precision:

$$\Omega(\widehat{\theta})^{-1} = \frac{1}{2} \mathbb{E}\{x x^{\mathsf{T}}\} + \frac{1}{2} \mathbb{E} \left\{ x x^{\mathsf{T}} \middle| \langle x, \widehat{\theta}^1 \rangle \geq \|\Sigma^{1/2} \widehat{\theta}^1\| \right\}.$$

To compute this, if $x \sim \mathsf{N}(0, \Sigma)$ then we use the following identity (see Lemma A.11 for the proof):

$$x|_{\langle x, \widehat{\theta}^1 \rangle \geq \|\Sigma^{1/2} \widehat{\theta}^1\|} \overset{\mathrm{d}}{=} \frac{\Sigma \widehat{\theta}^1}{\|\Sigma^{1/2} \widehat{\theta}^1\|} \xi_1 + \Sigma'^{1/2} \xi_2, \tag{9}$$

where $\xi_2 \sim \mathsf{N}(0, I_p)$, $\Sigma' = \Sigma - \Sigma \widehat{\theta}^1 (\widehat{\theta}^1)^{\mathsf{T}} \Sigma / \langle \widehat{\theta}^1, \Sigma \widehat{\theta}^1 \rangle$, and ξ_1 is independent of ξ_2 and satisfies the truncated normal distribution with density:

$$\frac{\mathrm{d}\mathbb{P}_{\xi_1}(u)}{\mathrm{d}u} = \frac{1}{\sqrt{2\pi} \Phi(-1)} \exp(-u^2/2) \mathbb{I}(u \geq 1).$$

Therefore

$$\Omega(\widehat{\theta}^1)^{-1} = \Sigma + \frac{1}{2} (\mathbb{E}\{\xi_1^2\} - 1) \frac{\Sigma \widehat{\theta}^1 (\widehat{\theta}^1)^{\mathsf{T}} \Sigma}{\langle \widehat{\theta}^1, \Sigma \widehat{\theta}^1 \rangle},$$

$$\Omega(\widehat{\theta}^1) = \Sigma^{-1} + \Big(\frac{2}{1 + \mathbb{E}\{\xi_1^2\}} - 1 \Big) \frac{\widehat{\theta}^1 (\widehat{\theta}^1)^{\mathsf{T}}}{\langle \widehat{\theta}^1, \Sigma \widehat{\theta}^1 \rangle},$$

where the second equation is an application of Sherman–Morrison formula.

As we will see in the next subsection, online debiasing instead proposes the following construction:

$$\widehat{\theta}^{\mathsf{on}} = \widehat{\theta}^{\mathsf{L}} + \frac{1}{n} \Sigma^{-1} X_1^{\mathsf{T}} (y^{(1)} - X_1 \widehat{\theta}^{\mathsf{L}}) + \frac{1}{n} \Omega^{(2)}(\widehat{\theta}^1) X_2^{\mathsf{T}} (y^{(2)} - X_2 \widehat{\theta}^{\mathsf{L}}).$$

Here $y^{(1)}, y^{(2)}, X_1, X_2$ are the outcomes (resp. covariates) from the first and second batches and $\Omega^{(2)}(\widehat{\theta}) = \mathbb{E}\{xx^{\mathsf{T}} | \langle x, \widehat{\theta}^1 \rangle \geq \|\Sigma^{1/2}\widehat{\theta}\|\}$ is the population precision on the second batch. Similar to $\Omega(\widehat{\theta}^1)$ we can use the distributional identity (9) to obtain the closed form formula

$$\Omega^{(2)}(\widehat{\theta}^1) = \Sigma^{-1} + \big(\mathbb{E}\{\xi_1^2\}^{-1} - 1\big) \frac{\widehat{\theta}^1 (\widehat{\theta}^1)^{\mathsf{T}}}{\langle \widehat{\theta}^1, \Sigma \widehat{\theta}^1 \rangle}.$$

We generate the dataset for 100 Monte Carlo iterations and compute the offline debiased estimate $\widehat{\theta}^{\mathsf{off}}$ and the online debiased estimate $\widehat{\theta}^{\mathsf{on}}$ for each iteration. Figure 1 (left panel) shows the histogram of the entries $\widehat{\theta}^{\mathsf{off}}$ on the support of θ_0 (red). In the same panel, we also plot the corresponding histogram of entries of our online debiased entries $\widehat{\theta}^{\mathsf{on}}$ (blue). For all three choices of $\widehat{\theta}^1$, the online debiased estimate $\widehat{\theta}^{\mathsf{on}}$ is appropriately centered around the true coefficient. However this is only true for the offline debiased estimate $\widehat{\theta}^{\mathsf{off}}$ when $\widehat{\theta}^1$ is chosen to be the LASSO estimate on the first batch.

One can also split samples in the following way. Since the second batch of data was adaptively collected while the first batch was not, we can compute a debiased estimate using only the first, non-adaptive batch:

$$\widehat{\theta}^{\mathsf{off},1} \equiv \widehat{\theta}^{\mathsf{L}}(y^{(1)}, X_1) + \frac{1}{n} \Sigma^{-1} X_1^{\mathsf{T}} (y^{(1)} - X_1 \widehat{\theta}^{\mathsf{L}}(y^{(1)}, X_1)).$$

Figure 1 (right panel) shows the histogram of the entries of $\widehat{\theta}^{\mathsf{off},1}$ restricted to the support of θ_0, and the comparison with $\widehat{\theta}^{\mathsf{on}}$. As can be expected, both $\widehat{\theta}^{\mathsf{off},1}$ and $\widehat{\theta}^{\mathsf{on}}$ are appropriately centered around the true coefficient 1. However, as is common with sample-splitting, $\widehat{\theta}^{\mathsf{off},1}$ displays a larger variance and correspondingly loses power in comparison with $\widehat{\theta}^{\mathsf{on}}$ since it uses only half of the data. The power loss becomes even more pronounced when there are more than two phases of data collection, or if the phases are particularly imbalanced.

In this illustration, we also assumed the knowledge of $\mathbb{P}_x$, i.e. the laws of the covariates in each batch. This was necessary to compute the precisions $\Omega(\widehat{\theta}^1)$ and $\Omega^2(\widehat{\theta}^1)$, which figured in the estimates $\widehat{\theta}^{\mathsf{off}}$ and $\widehat{\theta}^{\mathsf{on}}$ respectively. When there are $\Omega(p^2)$ unlabeled data points available, these precisions can be estimated accurately from the data. This places a stringent requirement on the sample size, especially in the high-dimensional setting when p is large. In the following, we will describe a general construction of the online debiased estimate $\widehat{\theta}^{\mathsf{on}}$ that avoids oracle knowledge of $\mathbb{P}_x$ and does not require reconstructing the population precisions accurately.

2.1 Constructing the online debiased estimator

The samples naturally separate into two batches: the first n_1 data points and the remaining n_2 points. Let X_1 and X_2 denote the design matrices of the two batches and, similarly, $y^{(1)}$ and $y^{(2)}$ the two responses vectors. We propose an online debiased estimator as follows:

$$\widehat{\theta}^{\mathsf{on}} = \widehat{\theta}^{\mathsf{L}} + \frac{1}{n} M^{(1)} X_1^{\mathsf{T}} (y^{(1)} - X_1 \widehat{\theta}^{\mathsf{L}}) + \frac{1}{n} M^{(2)} X_2^{\mathsf{T}} (y^{(2)} - X_2 \widehat{\theta}^{\mathsf{L}}), \tag{10}$$

where we will construct $M^{(1)}$ as a function of X_1 and $M^{(2)}$ as a function of X_1 as well as X_2. The proposal Eq.(10) follows from the general recipe in Eq.(6) by setting

- $M_i = M^{(1)}$ for $i = [n_1]$ and $M_i = M^{(2)}$ for $i = n_1 + 1, \ldots, n$.

- Filtrations $\mathfrak{F}_i$ constructed as follows. For $i < n_1$, $y_1, \dots, y_i$, $x_1, \dots x_{n_1}$ and $\varepsilon_1, \dots, \varepsilon_i$ are measurable with respect to $\mathfrak{F}_i$. For $i \geq n_1$, $y_1, \dots, y_i$, $x_1, \dots, x_n$ and $\varepsilon_1, \dots \varepsilon_i$ are measurable with respect to $\mathfrak{F}_i$.

By construction, this choice satisfies the predictability condition.

Before we describe how we construct the matrices $M^{(1)}$ and $M^{(2)}$, we note that Eq.(10) nests an intuitive 'sample splitting' approach. Indeed, debiasing $\widehat{\theta}^{\mathsf{L}}$ using exactly one of the two batches is equivalent to setting one of $M^{(1)}$ or $M^{(2)}$ to 0. While sample splitting can be shown to work under appropriate conditions, our approach is more efficient with use of the data and gains power in comparison. Moreover, as we will see in Section 3 it can also be adapted to settings like time series where sample splitting is not immediately applicable.

We construct $M^{(1)}$ and $M^{(2)}$ using a modification of the program used in [JM14a]. Let $\widehat{\Sigma}^{(1)} = (1/n_1)X_1^{\mathsf{T}} X_1$ and $\widehat{\Sigma}^{(2)} = (1/n_2)X_2^{\mathsf{T}} X_2$ be the sample covariances of each batch; let $M^{(1)}$ have rows $(m_a^{(1)})_{1 \leq a \leq p}$ and similarly for $M^{(2)}$. Using parameters $\mu_\ell, L > 0$ that we set later, we choose $m_a^{(\ell)}$, the a^{th} row of $M^{(\ell)}$, as a solution to the program

$$\begin{aligned} &\text{minimize} \quad \langle m, \widehat{\Sigma}^{(\ell)} m \rangle \\ &\text{subject to} \quad \|\widehat{\Sigma}^{(\ell)} m - e_a\|_\infty \leq \mu_\ell, \;\; \|m\|_1 \leq L. \end{aligned} \tag{11}$$

Here e_a is the a^{th} basis vector: a vector which is one at the a^{th} coordinate and zero everywhere else. The program (11) differs from that in [JM14a] by the ℓ_1 constraint on m. The idea for the program (11) is simple: the first constraint ensures that $\widehat{\Sigma}^{(\ell)} m$ is close, in ℓ_∞ sense to the e_a, the a^{th} basis vector and as we will see in Theorem 2.8 it controls the bias term Δ of $\widehat{\theta}^{\mathsf{on}}$. The objective is a multiple of the variance of the martingale term W in $\widehat{\theta}^{\mathsf{on}}$ (cf. Eq. (15)). We wish to minimize this as it directly affects the power of the test statistic or the length of valid confidence intervals constructed based on $\widehat{\theta}^{\mathsf{on}}$.

The additional ℓ_1 constraint is to ensure that the value of the program $\langle m_a^{(\ell)}, \widehat{\Sigma}^{(\ell)} m_a^{(\ell)} \rangle$ does not fluctuate much from sample to sample (this is further discussed as 'stability condition' in Lemmas A.9 and 3.8). This ensures that the martingale part of the residual displays a central limit behavior. In the non-adaptive setting, inference can be performed conditional on design X, and fluctuation in $\langle m_a^{(\ell)}, \widehat{\Sigma}^{(\ell)} m_a^{(\ell)} \rangle$ is conditioned out. In the adaptive setting, this is not possible: one effectively cannot condition on the design without conditioning on the noise realization, and therefore we perform inference unconditionally on X.

The program (11) is a convex optimization problem and thus standard interior point methods can be used to solve it in polynomial time [BV04]. We also show in Section 6, simple iterative schemes based on coordinate descent and projected gradient as alternate fast methods to solve program (10).

2.2 Online debiasing: a distributional characterization

We begin the analysis of the online debiased estimator $\widehat{\theta}^{\sf on}$ by a decomposition that mimics the classical debiasing.

$$\widehat{\theta}^{\sf on} = \widehat{\theta}_0 + \frac{1}{\sqrt{n}}\big(B_n(\widehat{\theta}^{\sf L} - \theta_0) + W_n\big), \tag{12}$$

$$B_n = \sqrt{n}\Big(I_p - \frac{n_1}{n}M^{(1)}\widehat{\Sigma}^{(1)} - \frac{n_2}{n}M^{(2)}\widehat{\Sigma}^{(2)}\Big) \tag{13}$$

$$W_n = \frac{1}{\sqrt{n}}\sum_{i\le n_1} M^{(1)}x_i\varepsilon_i + \frac{1}{\sqrt{n}}\sum_{n_1<i\le n} M^{(2)}x_i\varepsilon_i. \tag{14}$$

Assumption 2.6. *(Requirements of design) Suppose that the distribution $\mathbb{P}_x$ and the intermediate estimate $\widehat{\theta}^1$, that is used in collecting the second batch, satisfy the following:*

1. *There exists a constant $\Lambda_0 > 0$ so that the eigenvalues of $\mathbb{E}\{xx^{\sf T}\}$ and $\mathbb{E}\{xx^{\sf T}|\langle x, \widehat{\theta}^1\rangle \ge \varsigma\}$ are bounded below by Λ_0.*
2. *There exists a constant $\phi_0 > 0$ so that, $\mathbb{E}\{xx^{\sf T}\}$ and $\mathbb{E}\{xx^{\sf T}|\langle x, \widehat{\theta}^1\rangle \ge 0\}$ are $(\phi_0, \mathrm{supp}(\theta_0))$-compatible.*
3. *The laws of x and $x|_{\langle x,\widehat{\theta}^1\rangle\ge\varsigma}$ are κ-subgaussian for a constant $\kappa > 0$.*
4. *The precision matrices $\Omega = \mathbb{E}\{xx^{\sf T}\}^{-1}$ and $\Omega^{(2)}(\widehat{\theta}^1) = \mathbb{E}\{xx^{\sf T}|\langle x, \widehat{\theta}^1\rangle \ge \varsigma\}^{-1}$ satisfy $\|\Omega\|_1 \vee \|\Omega^{(2)}(\widehat{\theta}^1)\|_1 \le L$.*
5. *The conditional covariance $\Sigma^{(2)}(\theta) = \mathbb{E}\{xx^{\sf T}|\langle x, \theta\rangle \ge \varsigma\}$ is K-Lipschitz in its argument θ, i.e. $\|\Sigma^{(2)}(\theta') - \Sigma^{(2)}(\theta)\|_\infty \le K\|\theta - \theta'\|_1$.*

The first two conditions of Assumption 2.6 are for ensuring that the base LASSO estimator $\widehat{\theta}^{\sf L}$ has small estimation error. In addition, our debiasing makes use of the third and fourth constraints on the precision matrices of the sampling distributions. In the above, we will typically allow $L = L_n$ to diverge with n.

In the following Example we show that Gaussian random designs satisfy all the conditions of Assumption 2.6. We refer to Section A.4 for its proof.

Example 2.7. *Let $\mathbb{P}_x = \mathsf{N}(0, \Sigma)$ and $\widehat{\theta}$ be any vector such that $\|\widehat{\theta}\|_1\|\widehat{\theta}\|_\infty \le L_\Sigma\lambda_{\min}(\Sigma)\|\widehat{\theta}\|/2$ and $\|\Sigma^{-1}\|_1 \le L_\Sigma/2$. Then the distributions of x and $x|_{\langle x,\widehat{\theta}\rangle\ge\varsigma}$, with $\varsigma = \bar{\varsigma}\langle\widehat{\theta}, \Sigma\widehat{\theta}\rangle^{1/2}$ for a constant $\bar{\varsigma} \ge 0$ satisfy the conditions of Assumption 2.6 with*

$$\Lambda_0 = \lambda_{\min}(\Sigma),\ \phi_0 = \frac{\lambda_{\min}(\Sigma)}{16},\ \kappa = 2\lambda_{\max}^{1/2}(\Sigma),\ K = \sqrt{8}(1+\bar{\varsigma}^2)\frac{\lambda_{\max}(\Sigma)^{3/2}}{\lambda_{\min}(\Sigma)^{1/2}}, L = L_\Sigma.$$

Under Assumption 2.6 we provide a *non-asymptotic* bound on the bias of the online debiased estimator $\widehat{\theta}^{\sf on}$.

Theorem 2.8. *(Non-asymptotic bound on bias) Under Assumption 2.6 and the sample size condition that $n \ge 400^2\kappa^4 s_0^2\log p/\phi_0^2$ and $n_1 \wedge n_2 \ge (2\Lambda_0/\kappa^2 + (30\kappa)^2/\Lambda_0)\log p$, we have that*

$$\sqrt{n}(\widehat{\theta}^{\sf on} - \theta_0) = W_n + \Delta_n, \tag{15}$$

where $\mathbb{E}\{W_n\} = 0$ *and*

$$\mathbb{P}\Big\{\|\Delta_n\|_\infty \geq 4000\frac{\kappa^2\sigma}{\sqrt{\Lambda_0}\phi_0}\frac{s_0\log p}{\sqrt{n}}\Big\} \leq p^{-3}. \tag{16}$$

Further, assuming $\|\theta_0\|_1 \leq cs_0p^3(\log p)/n$*, we have*

$$\|\mathbb{E}\{\widehat{\theta}^{\mathsf{on}} - \theta_0\}\|_\infty \leq 10\left(400\frac{\kappa^2\sigma}{\sqrt{\Lambda_0}\phi_0} + c\right)\frac{s_0\log p}{n}\,. \tag{17}$$

The proof of Theorem 2.8 is given in Appendix A.2. When the parameters $\Lambda_0, \phi_0, \sigma, \kappa$ are of order one, the theorem shows that the bias of the online debiased estimator is of order $s_0\log p/n$. This may be compared with the LASSO estimator $\widehat{\theta}^{\mathsf{L}}$ whose bias is typically of order $\lambda \asymp \sigma\sqrt{\log p/n}$. In particular, in the regime when $s_0 = o(\sqrt{n/\log p})$, this bias is asymptotically dominated by the variance, which is of order $\sigma/\sqrt{n}$.

In order to establish asymptotic Gaussian behavior of the online debiased estimate $\widehat{\theta}^{\mathsf{on}}$, we consider a specific asymptotic regime for the problem instances.

Assumption 2.9. *(Asymptotic regime) We consider problem instances indexed by the sample size* n*, where* n, p, s_0 *satisfy the following:*

1. $\liminf_{n\to\infty}\frac{n_1\wedge n_2}{n} \geq c$*, for a positive universal constant* $c \in (0,1]$*. In other words, both batches contain at least a fixed fraction of data points.*
2. *The parameters satisfy:*

$$\lim_{n\to\infty}\frac{1}{\phi_0}s_0\sqrt{\frac{\log p}{n}}\left(L^2K \vee \sqrt{\frac{\log p}{\Lambda_0}}\right) \to 0\,. \tag{18}$$

The following proposition establishes that in the asymptotic regime, the unbiased component W_n has a Gaussian limiting distribution. The key underlying technical idea is to ensure that the martingale sum in W_n is stable in an appropriate sense.

Proposition 2.10. *Suppose that Assumption 2.6 holds and consider the asymptotic regime of Assumption 2.9. Let* $a = a(n) \in [p]$ *be a fixed sequence of coordinates. Define the conditional variance* $V_{n,a}$ *of the* a^{th} *coordinate as*

$$V_{n,a} = \sigma^2\Big(\frac{n_1}{n}\langle m_a^{(1)}, \widehat{\Sigma}^{(1)}m_a^{(1)}\rangle + \frac{n_2}{n}\langle m_a^{(2)}, \widehat{\Sigma}^{(2)}m_a^{(2)}\rangle\Big)\,. \tag{19}$$

Then, for any bounded continuous $\varphi : \mathbb{R} \to \mathbb{R}$

$$\lim_{n\to\infty}\mathbb{E}\Big\{\varphi\Big(\frac{W_{n,a}}{\sqrt{V_{n,a}}}\Big)\Big\} = \mathbb{E}\{\varphi(\xi)\},$$

where $\xi \sim \mathsf{N}(0,1)$*. The same holds for* φ *being a step function* $\varphi(z) = \mathbb{I}(z \leq x)$ *for any* $x \in \mathbb{R}$*. In particular,*

$$\lim_{n\to\infty}\mathbb{P}\Big\{\frac{W_{n,a}}{\sqrt{V_{n,a}}} \leq x\Big\} = \Phi(x),$$

where Φ *is the standard Gaussian cdf.*

The proof of Proposition 2.10 is deferred to Appendix A.3. The combination of Theorem 2.8 and Proposition 2.10 immediately yields the following distributional characterization for $\widehat{\theta}^{\mathsf{on}}$.

Theorem 2.11. *Under Assumptions 2.6 and 2.9, the conclusion of Proposition 2.10 holds with* $\sqrt{n}(\widehat{\theta}^{\mathsf{on}}_a - \theta_0)$ *in place of* W_n*. In particular,*

$$\lim_{n\to\infty} \mathbb{P}\Big\{\sqrt{\frac{n}{V_{n,a}}}(\widehat{\theta}^{\mathsf{on}}_a - \theta_{0,a}) \le x\Big\} = \Phi(x), \tag{20}$$

where $V_{n,a}$ *is defined as in Proposition 2.10.*

To compare the sample size requirements made for ℓ_1-consistent estimation, i.e. Theorem 2.2 and those in Assumption 2.9, it is instructive to simplify to the case when $\kappa, \phi_0, \Lambda_0$ are of order one. Then Theorem 2.2 requires, for ℓ_1-consistency, that $n_1 \vee n_2 = \Omega(s_0^2 \log p)$, i.e. at least one of the batches is larger than $s_0^2 \log p$. However, Theorem 2.11 makes the same assumption on $n_1 \wedge n_2$, or both batches exceed $s_0^2 \log p$ in size. For online debiasing, this is the case of interest. Indeed if $n_1 \gg n_2$ (or vice versa), we can apply offline debiasing to the larger batch to obtain a debiased estimate. Conversely, when n_1 and n_2 are comparable as in Assumption 2.9, this 'sample-splitting' approach leads to loss of power corresponding to a constant factor reduction in the sample size. This is the setting addressed in Theorem 2.11 via online debiasing.

3 Online debiasing for high-dimensional time series

Consider the standard *vector autoregressive model* of order d (or $\mathsf{VAR}(d)$ for short) [SS06]. In this model the data point z_t linearly evolve according to the dynamics:

$$z_t = \sum_{\ell=1}^{d} A^{(\ell)} z_{t-\ell} + \zeta_t, \tag{21}$$

where $A^{(\ell)} \in \mathbb{R}^{p\times p}$ and $\zeta_t \overset{\text{iid}}{\sim} \mathsf{N}(0, \Sigma_\zeta)$. VAR models are extensively used across science and engineering (see [FSGM+07, SW01, HENR88, SBB15] for notable examples in macroeconomics, genomics and neuroscience). Given the data $z_1, \dots, z_T$, a fundamental task is to estimate the parameters of the VAR model, viz. the matrices $A^{(1)}, \dots A^{(d)}$. The estimates of the parameters can be used in a variety of ways depending on the context: to detect or test for stationarity, forecast future data, or suggest causal links. Since each matrix is $p \times p$, this forms a putative total of dp^2 parameters, which we estimate from a total of $(T-d)p$ linear equations (Eq.(21) with $t = d+1, \dots, T$). For the i^{th} coordinate of z_t, Eq.(21) reads

$$z_{t,i} = \sum_{\ell=1}^{d} \langle z_{t-\ell}, A_i^{(\ell)} \rangle + \zeta_{t,i}, \tag{22}$$

where $A_i^{(\ell)}$ denotes the i^{th} row of the matrix $A^{(\ell)}$. This can be interpreted in the linear regression form Eq.(1) in dimension dp with $\theta_0 \in \mathbb{R}^{dp}, X \in \mathbb{R}^{(T-d)\times dp}$, $y, \varepsilon \in \mathbb{R}^{T-d}$ identified as:

$$\begin{aligned}
\theta_0 &= (A_i^{(1)}, A_i^{(2)}, \dots, A_i^{(d)})^{\mathsf{T}}, \\
X &= \begin{bmatrix} z_d^{\mathsf{T}} & z_{d-1}^{\mathsf{T}} & \cdots & z_1^{\mathsf{T}} \\ z_{d+1}^{\mathsf{T}} & z_d^{\mathsf{T}} & \cdots & z_2^{\mathsf{T}} \\ \vdots & \vdots & \ddots & \vdots \\ z_{T-1}^{\mathsf{T}} & z_{T-2}^{\mathsf{T}} & \cdots & z_{T-d}^{\mathsf{T}} \end{bmatrix}, \\
y &= (z_{d+1,i}, z_{d+2,i}, \dots, z_{T,i}), \\
\varepsilon &= (\zeta_{d+1,i}, \zeta_{d+2,i}, \dots, \zeta_{T,i}).
\end{aligned} \tag{23}$$

We omit the dependence on the coordinate i, and also denote the rows of X by $x_1, \dots, x_n \in \mathbb{R}^{dp}$, with $n = T - d$. Given sufficient data, or when T is large in comparison with dp, it is possible to estimate the parameters using least squares [SS06, LW82]. In [BM+15], Basu and Michailidis consider the problem of estimating the parameters when number of time points T is small in comparison with the total number of parameters dp. In order to make the estimation problem tractable, they assume that the matrices $A^{(\ell)}$ are sparse, and prove an estimation result. These build on similar ideas as Theorem 2.2, and prove appropriate restricted eigenvalue property for the design $X^{\mathsf{T}}X/n$. As anticipated, this result hinges on the stationary properties of the model (21), which we summarize prior to stating the estimation result.

Definition 3.1. *(Restricted Eigenvalue Property* (RE)*) For a positive semidefinite matrix $S \in \mathbb{R}^{m\times m}$ and numbers $\alpha, \phi > 0$, the matrix S satisfies the restricted eigenvalue property, denoted by $S \in \mathrm{RE}(\alpha, \tau)$, if for any vector $v \in \mathbb{R}^m$:*

$$\langle v, Sv\rangle \geq \alpha\|v\|_2^2 - \alpha\tau\|v\|_1^2 \,. \tag{24}$$

In [BM+15], the authors provide conditions on the autocovariance of the data points, which imply the restricted eigenvalue property for the sample covariance, with high probability. Formally, assuming that the covariates and noise terms are generated according to centered Gaussian stationary processes, [BM+15] introduce a notion of stability of the processes based on their spectral density. To be concrete, for the stationary process $x_t = (z_{t+d-1}^{\mathsf{T}}, \dots, z_t^{\mathsf{T}})^{\mathsf{T}}$ (rows of X), let $\Gamma_x(s) = \mathrm{Cov}(x_t, x_{t+s})$, for $t, s \in \mathbb{Z}$ and define the spectral density $f_x(r) \equiv 1/(2\pi)\sum_{\ell=-\infty}^{\infty}\Gamma_X(\ell)e^{-j\ell r}$, for $r \in [-\pi, \pi]$. The measure of stability of the process is defined as the maximum eigenvalue of the density

$$\mathcal{M}(f_x) \equiv \sup_{r\in[-\pi,\pi]} \sigma_{\max}(f_x(r))\,. \tag{25}$$

Likewise, the minimum eigenvalue of the spectrum is defined as $m(f_x) \equiv \inf_{r\in[-\pi,\pi]} \sigma_{\min}(f_x(r))$, which captures the dependence among the covariates. (Note that for the case of i.i.d. samples, $\mathcal{M}(f_x)$ and $m(f_x)$ reduce to the maximum and minimum eigenvalue of the population covariance.)

Definition 3.2 (Stability and invertibility of $\mathsf{VAR}(d)$ Process [BM+15])**.** *A* $\mathsf{VAR}(d)$ *process with* *an associated reverse characteristic polynomial*

$$\mathcal{A}(\gamma) = I - \sum_{\ell=1}^{d} A^{(\ell)}\gamma^{\ell}\,, \tag{26}$$

is called stable and invertible if $\det(\mathcal{A}(\gamma)) \neq 0$ *for all* $\gamma \in \mathbb{C}$ *with* $|\gamma| = 1$.

An important contribution in [BM+15] is to show that, for a stable VAR process $m(f_x) > 0$ with sample size $n \gtrsim s_0 \log p$, the sample covariance $X^{\mathsf{T}}X/n$ satisfies the $\mathrm{RE}(\alpha, \tau)$ condition holds for proper numbers α, τ, with high probability.

Define

$$\mu_{\min}(\mathcal{A}) = \min_{|\gamma|=1} \Lambda_{\min}(\mathcal{A}^*(\gamma)\mathcal{A}(\gamma))\,,$$
$$\mu_{\max}(\mathcal{A}) = \max_{|\gamma|=1} \Lambda_{\max}(\mathcal{A}^*(\gamma)\mathcal{A}(\gamma))\,.$$

By bounding $m(f_x)$ and $M(f_x)$ in terms of $\mu_{\min}(\mathcal{A})$, $\mu_{\max}(\mathcal{A})$ and $\Lambda_{\min}(\Sigma_\epsilon)$, $\Lambda_{\max}(\Sigma_\epsilon)$, we have the following result:

Proposition 3.3. *Let* $\{z_1, \dots, z_T\}$ *be generated according to the (stable)* $\mathsf{VAR}(d)$ *process* (21) *and let* $n = T - d$. *Then there exist constants* $c \in (0, 1)$ *and* $C > 1$ *such that for all* $n \geq C \max\{\omega^2, 1\} \log(dp)$, *with probability at least* $1 - \exp(-cn \min\{\omega^{-2}, 1\})$, *we have* $X^{\mathsf{T}}X/n \in \mathrm{RE}(\alpha, \tau)$. *Here,* α, ω *and* τ *are given by:*

$$\begin{aligned}
\omega &= \frac{d\Lambda_{\max}(\Sigma_\varepsilon)\mu_{\max}(\mathcal{A})}{\Lambda_{\min}(\Sigma_\varepsilon)\mu_{\min}(\mathcal{A})}\,,\\
\alpha &= \frac{\Lambda_{\min}(\Sigma_\varepsilon)}{2\mu_{\max}(\mathcal{A})}\,,\\
\tau &= \alpha(\omega^2 \vee 1)\frac{\log(dp)}{n}\,.
\end{aligned} \tag{27}$$

Proposition 3.3 can be proved along the same lines as [BM+15, Proposition 4.2]. However, our proofs differ slightly as follows:

1. [BM+15] writes the $\mathsf{VAR}(d)$ model as a $\mathsf{VAR}(1)$ model and then vectorize the obtained equation to get a linear regression form (cf. Section 4.1 of [BM+15]). This way, they prove $I \otimes (X^{\mathsf{T}}X/n) \in \mathrm{RE}(\alpha, \tau)$. But in their proof, as a first step they show that $X^{\mathsf{T}}X/n \in \mathrm{RE}(\alpha, \tau)$ as we need here.

2. [BM+15, Proposition 4.2] assumes $n \geq Ck \max\{\omega^2, 1\} \log(dp)$, with $k = \sum_{\ell=1}^{d} \|\mathrm{vec}(A^{(\ell)})\|_0$, the total number of nonzero entries of matrices A_ℓ and then it is later used to get $\tau \leq 1/(Ck)$. However, note that the definition of RE condition is independent of the sparsity of matrices $A^{(\ell)}$. So, we can use their result with $k = 1$.

3. The proof of RE condition involves upper bounding $\mathcal{M}(f_x)$. We bound $\mathcal{M}(f_x)$ in a different way than [BM+15, Proposition 4.2]. We refer to Appendix B.1 for more details.

With the restricted eigenvalue property in place for the sample covariance $X^{\mathsf{T}}X/n$, there is a standard argument to obtain estimation error for the ℓ_1-regularized estimator, which has also been followed in other work, e.g., [BVDG11, BRT09, LW12].

Proposition 3.4 (Estimation Bound)**.** *Recall the relation $y = X\theta_0 + \varepsilon$, where X, y, θ_0 are given by (23) and let $\widehat{\theta}^{\mathsf{L}}$ be the Lasso estimator*

$$\widehat{\theta}^{\mathsf{L}} = \arg\min_{\theta\in\mathbb{R}^{dp}} \left\{ \frac{1}{2n}\|y - X\theta\|_2^2 + \lambda_n\|\theta\|_1 \right\}. \tag{28}$$

Assume that $|\text{supp}(\theta_0)| \le s_0$. There exist constants $c, C > 0$ such that the following happens. For $n \ge C\alpha(\omega^2 \vee 1)s_0\log(dp)$, and $\lambda_n = \lambda_0\sqrt{\log(dp)/n}$, with $\lambda_0 \ge \lambda_* \equiv 4\Lambda_{\max}(\Sigma_\varepsilon)(1 \vee \mu_{\max}(\mathcal{A}))/\mu_{\min}(\mathcal{A})$, with probability at least $1 - \exp(-c\log(dp^2)) - \exp(-cn(1\wedge\omega^{-2}))$, we have

$$\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \le C\sigma\frac{s_0\lambda_n}{\alpha},$$

where α, ω are defined in Proposition 3.3.

3.1 Constructing the online debiased estimator

We partition the time indices $[n]$ into K episodes $E_0, \dots, E_{K-1}$, with E_ℓ of length r_ℓ, so that $\sum_{\ell=0}^{K-1} r_\ell = n$. We also let $n_\ell = r_0 + \ldots + r_\ell$, for $\ell = 0, \dots, K-1$; hence, $n_{K-1} = n$. Define

$$\widehat{\Sigma}^{(\ell)} = \frac{1}{n_\ell}\sum_{t\in E_0\cup\ldots\cup E_\ell} x_t x_t^{\mathsf{T}},$$

be the sample covariance of the features in the first $\ell+1$ episodes. For each coordinate $a \in [dp]$, we construct the decorrelating vector $m_a^\ell \in \mathbb{R}^{dp}$ at step $\ell \ge 0$ by solving the following optimization:

$$\begin{aligned} &\text{minimize} \quad m^{\mathsf{T}}\widehat{\Sigma}^{(\ell)}m \\ &\text{subject to} \quad \|\widehat{\Sigma}^{(\ell)}m - e_a\|_\infty \le \mu_\ell, \quad \|m\|_1 \le L, \end{aligned} \tag{29}$$

for some appropriate values of $\mu_\ell, L > 0$ which will be determined later from our analysis of the debiased estimator.

We then construct the online debiased estimator for coordinate a of θ_0 as follows:

$$\widehat{\theta}_a^{\mathsf{on}} = \widehat{\theta}_a^{\mathsf{L}} + \frac{1}{n}\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} \langle m_a^\ell, x_t\rangle(y_t - \langle x_t, \widehat{\theta}^{\mathsf{L}}\rangle) \tag{30}$$

Defining $M^{(\ell)} \in \mathbb{R}^{dp\times dp}$ as the matrix with rows $(m_a^\ell)^{\mathsf{T}}$ for $a \in [dp]$, we can write $\widehat{\theta}^{\mathsf{on}}$ as:

$$\widehat{\theta}^{\mathsf{on}} = \widehat{\theta}^{\mathsf{L}} + \frac{1}{n}\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} M^{(\ell)}x_t(y_t - \langle x_t, \widehat{\theta}^{\mathsf{L}}\rangle). \tag{31}$$

In Section 3.2, we show that the constructed online debiased estimator $\widehat{\theta}^{\mathsf{on}}$ is asymptotically unbiased and admits a normal distribution. To do that we provide a high probability bound on the bias of $\widehat{\theta}^{\mathsf{on}}$ (See Lemma B.1). This bound is in terms of the batch sizes r_ℓ, from which we propose the following guideline for choosing them: $r_0 \sim \sqrt{n}$ and $r_\ell \sim \beta^\ell$, for a constant $\beta > 1$, and $\ell \geq 1$.

Note that the programs constructing $M^{(\ell)}$ closely resemble the optimization (11) proposed in Section 2. However, we define the matrices $\widehat{\Sigma}^{(\ell)}$ (and consequently the sizes $n_1, n_2, \dots$) differently for both. The reason for this is that, by assumption, the time series data are stationary, while the batched data of Section 2 are non-stationary. Therefore, in time series, we can use all past data points in Optimization (29) to form an approximate inverse. On the other hand, in non-stationary settings like Section 2, it is better to restrict the samples included in the sample covariance $\widehat{\Sigma}^{(\ell)}$ to a smaller window.

Before proceeding into the distributional characterization of the online debiased estimator for θ_0 (coefficients of A_i matrices), we discuss a simple numerical example in which the (offline) debiased estimator of [JM14a] does not undergo an unbiased normal distribution, while the constructed online debiased estimator admits such distribution.

A numerical example. Consider the linear time series model described in 21 with $p = 15$, $d = 5$, $T = 60$, and diagonal $A^{(i)}$ matrices with value $b = 0.15$ on their diagonals. Note that this a high-dimensional setting as the number of parameters dp exceeds $n = T - d$, as the model (23). The covariance matrix Σ_ζ of the noise terms ζ_t is chosen as $\Sigma_\zeta(i,j) = \rho^{\mathbb{I}(i\neq j)}$ with $\rho = 0.5$ and $i, j \in [p]$. The population covariance matrix of vector $x_t = (z_{t+d-1}^{\mathsf{T}}, \dots, z_t^{\mathsf{T}})^{\mathsf{T}}$ is a dp by dp matrix Σ consisting of d^2 blocks of size $p \times p$ with $\Gamma_z(r-s)$ as block (r, s). The analytical formula to compute $\Gamma_z(\ell)$ is given by [BM+15]:

$$\Gamma_z(\ell) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{A}^{-1}(e^{-j\theta})\Sigma_\zeta(\mathcal{A}^{-1}(e^{-j\theta}))^* e^{j\ell\theta} d\theta \,,$$

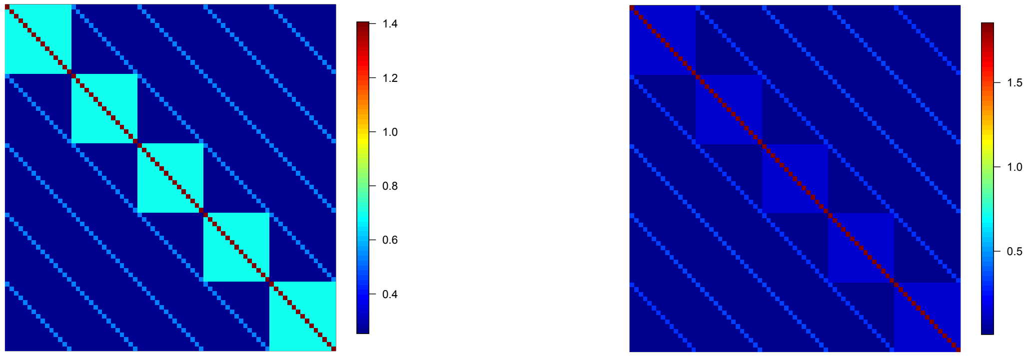
where $\mathcal{A}(\gamma)$ is given in equation (26). Figure 2 shows the heat maps of magnitudes of the elements of Σ and the precision matrix $\Omega = \Sigma^{-1}$ for the on hand $\mathsf{VAR}(5)$ process. We focus on the noise component of both online and offline debiased estimators, i.e.,

$$W^{\text{on}} = \frac{1}{\sqrt{n}}\sum_{\ell=0}^{K-2} M^{(\ell)} \sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} x_t\varepsilon_t \,, \tag{32}$$

$$W^{\text{off}} = \frac{1}{\sqrt{n}} M^{\text{off}}\sum_{t=1}^{n} x_t\varepsilon_t \,, \tag{33}$$

with $M^{(\ell)}$ constructed from the solutions to optimization (29) for $\ell = 0, \dots, K-2$, and M^{off} is also constructed by optimization (29), considering all the covariates (i.e., setting $\ell = K-1$). Also, recall that $\varepsilon = (\zeta_{d+1,i}, \zeta_{d+2,i} \dots, \zeta_{T,i})$ by equation (23).

In Figure 3, we show the QQ-plot, PP-plot and histogram of W_1^{on} and W_1^{off} (corresponding to the entry $(1,1)$ of matrix A_1) for 1000 different realizations of the noise ζ_t. As we observe, even the noise component W^{off} is biased because the offline construction of M depends on all features x_t and hence endogenous noise ζ_t. Recall that for the setting with an i.i.d sample, the noise component is zero mean gaussian for any finite sample size n. However, the online construction of decorrelating matrices $M^{(\ell)}$, makes the noise term a martingale and hence W^{on} converges in distribution to a zero mean normal vector, allowing for a distributional characterization of the online debiased estimator.



(a) Heat map of Σ (b) Heat map of Ω

Figure 2: Heat maps of magnitudes of elements of covariance matrix $\Sigma \equiv \mathbb{E}(x_i x_i^T)$ (left plot), and precision matrix $\Omega = \Sigma^{-1}$ (right plot). In this example. x_i's are generated from a $\mathsf{VAR}(d)$ model with covariance matrix of noise $\Sigma_\zeta(i,j) = \rho^{|i-j|}$ with values $d = 5$, $p = 15$, $T = 60$, $\rho = 0.5$, and diagonal $A^{(i)}$ matrices with $b = 0.15$ on diagonals.

3.2 Distributional characterization of online debiased estimator for time series

Similar to the case of batched data collection, we start our analysis of the online debiased estimator $\widehat{\theta}^{\mathsf{on}}$ by considering a bias-variance decomposition of it. Using $y_t = \langle x_t, \theta_0 \rangle + \varepsilon_t$ in the definition (31):

$$
\begin{aligned}
\widehat{\theta}^{\mathsf{on}} - \theta_0 &= \widehat{\theta}^{\mathsf{L}} - \theta_0 + \frac{1}{n} \sum_{\ell=0}^{K-2} \sum_{t \in E_{\ell+1}} M^{(\ell)} x_t x_t^{\mathsf{T}} (\theta_0 - \widehat{\theta}^{\mathsf{L}}) + \frac{1}{n} \sum_{\ell=0}^{K-2} \sum_{t \in E_{\ell+1}} M^{(\ell)} x_t \varepsilon_t \\
&= \Big(I - \frac{1}{n} \sum_{\ell=0}^{K-2} \sum_{t \in E_{\ell+1}} M^{(\ell)} x_t x_t^{\mathsf{T}} \Big) (\widehat{\theta}^{\mathsf{L}} - \theta_0) + \frac{1}{n} \sum_{\ell=0}^{K-2} \sum_{t \in E_{\ell+1}} M^{(\ell)} x_t \varepsilon_t \,. \qquad (34)
\end{aligned}
$$

We adopt the shorthand $R^{(\ell)} = (1/r_\ell) \sum_{t \in E_\ell} x_t x_t^{\mathsf{T}}$ for the sample covariance of features in episode ℓ. Letting

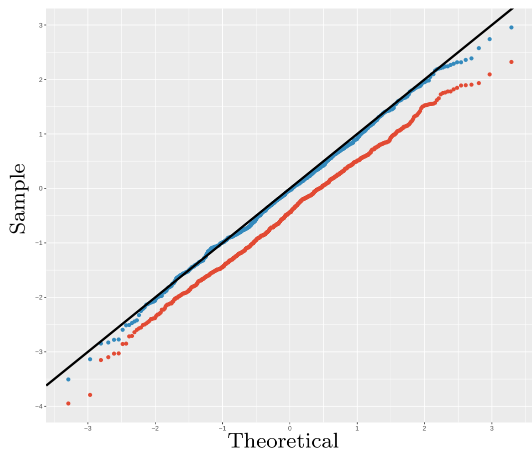
$$
B_n \equiv \sqrt{n} \Big(I - \frac{1}{n} \sum_{\ell=0}^{K-2} r_{\ell+1} M^{(\ell)} R^{(\ell+1)} \Big) \,, \qquad (35)
$$

$$
W_n \equiv \frac{1}{\sqrt{n}} \sum_{\ell=0}^{K-2} M^{(\ell)} \Big(\sum_{t \in E_{\ell+1}} x_t \varepsilon_t \Big) \,, \qquad (36)
$$

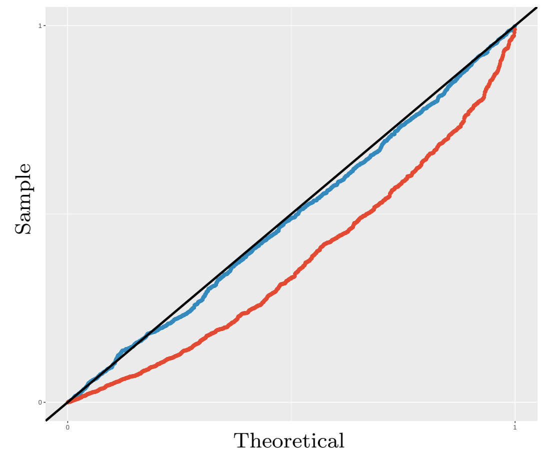
we arrive at the following decomposition

$$
\widehat{\theta}^{\mathsf{on}} = \theta_0 + \frac{1}{\sqrt{n}} \big(B_n (\widehat{\theta}^{\mathsf{L}} - \theta_0) + W_n \big) \,. \qquad (37)
$$

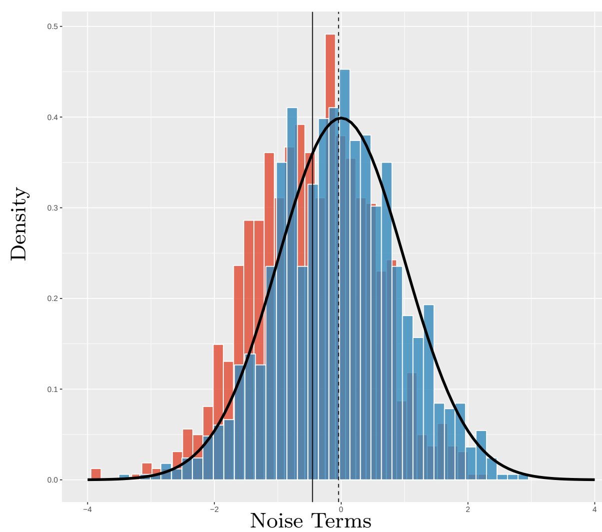
Similar to the case of batched data collection, here by constructing the decorrelating matrices $M^{(\ell)}$ over episodes, we ensure that the noise part in the debiased estimator W_n is indeed a martingale and using the martingale CLT it admits an asymptotically gaussian distribution. To see why, recall the notation $x_t = (z_{t+d-1}^{\mathsf{T}}, \dots, z_t^{\mathsf{T}})^{\mathsf{T}}$ (row of X). Therefore, x_t is independent of $\{\zeta_{r,a}$:



(a)



(b)



(c)

Figure 3: Different empirical behavior of noise terms associated with online and offline debiased fixed coordinate of a $\mathsf{VAR}(d)$ model with $\Sigma_\zeta(i,j) = \rho^{|i-j|}$. In this example, $d = 50, p = 15, T = 60, \rho = 0.5$, and $A^{(i)}$ matrices are diagonal with value $b = 0.15$ on their diagonals. Plots 3a, 3b, and 3c show the QQ plots, PP plots, and the histogram of online debiased noise terms (blue) and offline debiased noise terms (red) of 1000 independent experiments, respectively and black curve/lines denote the ideal standard normal distribution. Deviation of offline debiased noise terms from standard normal distribution implies the failure of offline debiasing method for statistical inference purposes for cases when samples are correlated. On the other hand, as we were expecting from theoretical arguments in subsection 3.2, online debiased noise terms are highly aligned with standard normal distribution. The solid and dash lines indicate the location of the mean of offline and online debiased noise terms, respectively. A significant distance of the average of offline data points (solid line) from zero can be seen in this figure.

$r \geq t+d-1\} \equiv \{\varepsilon_r : r \geq t-1\}$. That said, since $M^{(\ell)} = (m_1^\ell, \dots, m_{dp}^\ell)^\mathsf{T}$ is a function of $\{x_t : t \in E_0 \cup \ldots \cup E_\ell\} = \{z_t : t \in [n_\ell]\}$, it is independent of $\{\varepsilon_t : t \geq n_\ell\}$. Finally ε_t are zero mean which implies that W_n is a martingale, with respect to the natural filtration $\mathfrak{F}_j = \{\varepsilon_1, \dots, \varepsilon_j\}$, $j \in \mathbb{N}$.

Assumption 3.5. *(Requirements of design) Suppose that*

1. $\Lambda_{\min}(\Sigma_\epsilon) > c_1 > 0$ *and* $\mu_{\max}(\mathcal{A}) < c_2 < \infty$.
2. *Denoting by* $\Sigma \equiv \mathbb{E}(x_t x_t^\mathsf{T}) \in \mathbb{R}^{dp \times dp}$ *the population covariance of the data points* $\{x_t\}$. *Define*

$$L_\Sigma \equiv \max_{i \in [dp]} \|\Sigma^{-1} e_i\|_1 , \quad D_\Sigma \equiv \max_{i \in [dp]} \Sigma_{i,i} . \tag{38}$$

Assume that $L_\Sigma^2 D_\Sigma = o_P(n/\log^3(dpn))$.

As we show in the proof of Proposition 3.3, $\Lambda_{\min}(\Sigma) > \Lambda_{\min}(\Sigma_\epsilon)/\mu_{\max}(\mathcal{A})$. *Therefore, Assumption* (1) *above implies that* $\Lambda_{\min}(\Sigma) > c > 0$, *for some constant* $c > 0$.

As we will see from the analysis of the term B_n (and similar to the case of batched data collection), the parameters μ_ℓ in the optimization (29) play a key role in construction of the decorrelating matrices $M^{(\ell)}$ and controlling the term B_n (which directly controls the bias of $\widehat{\theta}^{\mathsf{on}}$). We would like to choose μ_ℓ small enough to reduce the bias, but large enough so that the optimization (29) is yet feasible. This brings us to upper bounding

$$\mu_\ell^{\min}(\widehat{\Sigma}^{(\ell)}) \equiv \min_{M \in \mathbb{R}^{dp \times dp}} |M \widehat{\Sigma}^{(\ell)} - I|_\infty .$$

Our next lemma establishes such bound that hold with probability converging rapidly to one as $n, p \to \infty$.

Lemma 3.6. *For* $a > 0$, *let* $\mathcal{G}_n = \mathcal{G}_n(\tau)$ *be the following probability event*

$$\mathcal{G}_n(\tau) \equiv \Big\{ \widehat{\Sigma}^{(\ell)} \in \mathbb{R}^{dp \times dp} : \ \mu_\ell^{\min}(\widehat{\Sigma}^{(\ell)}) < \tau \sqrt{\frac{\log(dp)}{n_\ell}} \Big\} . \tag{39}$$

Then there exists a constant $c > 0$ *such that letting*

$$c_0 \equiv c \Big(\frac{\tau}{3d}\Big)^2 \Big(\frac{\mu_{\min}(\mathcal{A})}{\mu_{\max}(\mathcal{A})}\Big)^2 \Big(\frac{\Lambda_{\min}(\Sigma_\varepsilon)}{\Lambda_{\max}(\Sigma_\varepsilon)}\Big)^2 - 2 ,$$

and for $n_\ell \geq \frac{c_0+2}{c} \log(dp)$, *the following holds true.*

$$\mathbb{P}(\widehat{\Sigma}^{(\ell)} \in \mathcal{G}_n) \geq 1 - 6(dp)^{-c_0'+2} , \quad c_0 \equiv c \Big(\frac{\tau}{3d}\Big)^2 \Big(\frac{\mu_{\min}(\mathcal{A})}{\mu_{\max}(\mathcal{A})}\Big)^2 \Big(\frac{\Lambda_{\min}(\Sigma_\varepsilon)}{\Lambda_{\max}(\Sigma_\varepsilon)}\Big)^2 - 2 . \tag{40}$$

The proof of Lemma 3.6 is given in Appendix B.2.

Theorem 3.7. *(Bias control) Consider the* $\mathsf{VAR}(d)$ *model* (21) *for time series and let* $\widehat{\theta}^{\mathsf{on}}$ *be the debiased estimator* (31) *with* $\mu_\ell = \tau\sqrt{(\log p)/n_\ell}$ *and* $L \leq L_\Sigma$, *with* L_Σ *defined by* (38). *Then, under*

Assumption 3.5(1), the sample size condition $n \ge C(\omega^2 \vee 1)s_0 \log(dp)$, and for $\lambda = \lambda_0\sqrt{\log(dp)/n}$ with $\lambda_0 \ge \lambda_* \equiv 4\Lambda_{\max}(\Sigma_\varepsilon)(1 \vee \mu_{\max}(\mathcal{A}))/\mu_{\min}(\mathcal{A}))$, we have that

$$\sqrt{n}(\widehat{\theta}^{\sf on} - \theta_0) = W_n + \Delta_n, \tag{41}$$

where $\mathbb{E}\{W_n\} = 0$ and

$$\mathbb{P}\Big\{\|\Delta_n\|_\infty \ge C_0\sigma\frac{s_0\log(dp)}{\sqrt{n}}\Big\} \le 12p^{-c_2} + \exp(-c\log(dp^2)) + \exp(-cn(1\wedge\omega^{-2}))\,, \tag{42}$$

for some constants $C, C_0, c, c_2 > 0$, and ω given by (27). In particular

$$\|\mathbb{E}\{\widehat{\theta}^{\sf on} - \theta_0\}\|_\infty \le 10C\sigma s_0 \log(dp)/n\,.$$

We refer to Appendix B.3 for the proof of Theorem 3.7.

Assuming the quantities σ, $C_{\Sigma_\epsilon,A}$ are of order one, the theorem shows that the bias of the online debiased estimator is of order $L_\Sigma s_0(\log p)/\sqrt{n}$. On the other hand, recall the filtration $\mathcal{F}_t$ generated by $\{\varepsilon_1,\dots,\varepsilon_t\}$ and rewrite (36) as $W_n = \sum_t v_t\varepsilon_t$, where $v_t = M^{(\ell)}x_t/\sqrt{n}$ (Sample t belongs to episode $\ell+1$). As shown in Lemma 3.8 below, for each coordinate $i \in [dp]$, the conditional variance $\sum_{t=1}^n \mathbb{E}(\varepsilon_t^2 v_{t,i}^2|\mathcal{F}_{t-1}) = (\sigma^2/n)\sum_{t=1}^n\langle m_a^\ell, z_t\rangle^2$ is of order one. Hence $\|\Delta_n\|_\infty$ is asymptotically dominated by the noise variance, in the regime that $s_0 = o(\sqrt{n}/(L_\Sigma\log(dp)))$.

We next proceed to characterize the distribution of the noise term W_n. To derive this, we apply the martingale CLT (e.g., see [HH14, Corollary 3.2]) to show that the unbiased component W_n admits a Gaussian limiting distribution. A key technical step for this end is to show that the martingale sum W_n is stable in an appropriate sense.

Lemma 3.8. *(Stability of martingale W_n) Let $\widehat{\theta}^{\sf on}$ be the debiased estimator (31) with $\mu_\ell = \tau\sqrt{(\log p)/n_\ell}$ and $L = L_\Sigma$, with L_Σ defined by (38). Under Assumption 3.5(2), and for any fixed sequence of integers $a(n) \in [dp]$,[2] we have*

$$V_{n,a} \equiv \frac{\sigma^2}{n}\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}}\langle m_a^\ell, x_t\rangle^2 = \sigma^2\Omega_{a,a} + o_P(1), \tag{43}$$

with $\Omega \equiv \Sigma^{-1}$ the precision matrix. In addition, we have

$$\max\Big\{\frac{1}{\sqrt{n}}|\langle m_a^\ell, x_t\rangle\varepsilon_t| : \ell\in[K-2],\, t\in[n-1]\Big\} = o_P(1). \tag{44}$$

We refer to Appendix B.4 for the proof of Lemma 3.8. With Lemma 3.8 in place, we can apply a martingale central limit theorem [HH14, Corollary 3.2] to obtain the following result.

Corollary 3.9. *Consider the $\mathsf{VAR}(d)$ model (21) for time series and let $\widehat{\theta}^{\sf on}$ be the debiased estimator (31) with $\mu_\ell = \tau\sqrt{(\log p)/n_\ell}$ and $L \le L_\Sigma$, with L_Σ defined by (38). For an arbitrary but fixed sequence of integers $a(n) \in [dp]$, define the conditional variance V_n as*

$$V_{n,a} \equiv \frac{\sigma^2}{n}\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}}\langle m_a^\ell, x_t\rangle^2\,.$$

[2] We index the sequence with the sample size n that is diverging. Since we are in high-dimensional setting $p \ge n$ is also diverging.

Under Assumption 3.5, for any fixed coordinate $a \in [dp]$, and for all $x \in \mathbb{R}$ we have

$$\lim_{n\to\infty} \mathbb{P}\Big\{\frac{W_{n,a}}{\sqrt{V_{n,a}}} \le x\Big\} = \Phi(x)\,, \tag{45}$$

where Φ is the standard Gaussian cdf.

For the task of statistical inference, Theorem 3.7 and Lemma 3.8 suggest to consider the scaled residual $\sqrt{n}(\widehat{\theta}^{\sf on}_a - \theta_{0,a})/(\sigma\sqrt{V_{n,a}})$ as the test statistics. Our next proposition characterizes its distribution. The proof is straightforward given the result of Theorem 3.7 and Corollary 3.9 and is deferred to Appendix B.5. In its statement we omit explicit constants that can be easily derived from Theorem 3.7.

Proposition 3.10. *Consider the $\mathsf{VAR}(d)$ model (21) for time series and let $\widehat{\theta}^{\sf on}$ be the debiased estimator (31) with $\mu_\ell = \tau\sqrt{(\log p)/n_\ell}$, $\lambda = \lambda_0\sqrt{\log(dp)/n}$, $L \le L_\Sigma$, with L_Σ defined by (38) and λ being the regularization parameter in the Lasso estimator $\widehat{\theta}^{\sf L}$, for τ, λ_0 large enough constants.*

Suppose that Assumption 3.5 holds and $s_0 = o(\sqrt{n}/\log(dp))$, then the following holds true for any fixed sequence of integers $a(n) \in [dp]$. For all $x \in \mathbb{R}$, we have

$$\lim_{n\to\infty}\sup_{\|\theta_0\|_0\le s_0}\left|\mathbb{P}\Big\{\frac{\sqrt{n}(\widehat{\theta}^{\sf on}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} \le x\Big\} - \Phi(x)\right| = 0\,. \tag{46}$$

4 Statistical inference

An immediate use of distributional characterizations (20) or (46) is to construct confidence intervals and also provide valid p-values for hypothesis testing regarding the model coefficients. Throughout, we make the sparsity assumption $s_0 = o(\sqrt{n}/\log p_0)$, with p_0 the number of model parameters (for the batched data collection setting $p_0 = p$, and for the $\mathsf{VAR}(d)$ model $p_0 = dp$).

Confidence intervals: For fixed coordinate $a \in [p_0]$ and significance level $\alpha \in (0,1)$, we let

$$\begin{aligned} J_a(\alpha) &\equiv [\widehat{\theta}^{\sf on}_a - \delta(\alpha,n), \widehat{\theta}^{\sf on}_a + \delta(\alpha,n)]\,,\\ \delta(\alpha,n) &\equiv \Phi^{-1}(1-\alpha/2)\sqrt{V_{n,a}/n}\,, \end{aligned} \tag{47}$$

where $V_{n,a}$ is defined by Equation (19) for the batched data collection setting and by Equation (43) for the $\mathsf{VAR}(d)$ model.

As a result of Proposition 3.10, the confidence interval $J_a(\alpha)$ is asymptotically valid because

$$\begin{aligned} \lim_{n\to\infty}\mathbb{P}(\theta_{0,a}\in J_a(\alpha)) &= \lim_{n\to\infty}\mathbb{P}\Big\{\frac{\sqrt{n}(\widehat{\theta}^{\sf on}_a - \theta_{0,a})}{\sqrt{V_{n,a}/n}} \le \Phi^{-1}(1-\alpha/2)\Big\}\\ &\quad - \lim_{n\to\infty}\mathbb{P}\Big\{\frac{\sqrt{n}(\widehat{\theta}^{\sf on}_a - \theta_{0,a})}{\sqrt{V_{n,a}n}} \le \Phi^{-1}(1-\alpha/2)\Big\}\\ &= \Phi(\Phi^{-1}(1-\alpha/2)) - \Phi(-\Phi^{-1}(1-\alpha/2)) = 1-\alpha\,. \end{aligned} \tag{48}$$

Further, note that the length of confidence interval $J_a(\alpha)$ is of order $O(\sigma/\sqrt{n})$ (using Lemma A.9 for the batched data collection setting and Lemma 3.8 for the time series). It is worth noting that this is the minimax optimal rate [JM14b, Jav14] and is of the same order of the length of confidence intervals obtained by the least-square estimator for the classical regime $n > p$ with i.i.d samples.

Hypothesis testing: Another consequence of Proposition 3.10 is that it allows for testing hypothesis of form $H_0 : \theta_{0,a} = 0$ versus the alternative $H_A : \theta_{0,a} \neq 0$ and provide valid p-values. Recall that θ_0 denotes the model parameters, either for the batched data collection setting or the $\mathsf{VAR}(d)$ model (which encodes the entries $A^{(\ell)}_{i,j}$ in model (21)). Such testing mechanism is of crucial importance in practice as it allows to diagnose the significantly relevant covariates to the outcome. In case of time series, it translates to understanding the effect of a covariate $z_{t-\ell,j}$ on a covariate $z_{t,i}$, and to provide valid statistical measures (p-values) for such associations. We construct two-sided p-values for testing H_0, using our test statistic as follows:

$$P_a = 2\left(1 - \Phi\left(\frac{\sqrt{n}|\widehat{\theta}^{\mathsf{on}}_a|}{\sqrt{V_{n,a}}}\right)\right) . \tag{49}$$

Our testing (rejection) rule given the p-value P_a is:

$$R(a) = \begin{cases} 1 & \text{if } P_a \le \alpha \quad (\text{reject } H_0)\,, \\ 0 & \text{otherwise} \quad (\text{fail to reject } H_0)\,. \end{cases} \tag{50}$$

Employing the distributional characterizations (20) or (46), it is easy to verify that the constructed p-value P_a is valid in the sense that under the null hypothesis it admits a uniform distribution: $\mathbb{P}_{\theta_{0,a}=0}(P_a \le u) = u$ for all $u \in [0,1]$.

Group inference In many applications, one may want to do inference for a group of model parameters, $\theta_{0,G} \equiv (\theta_{0,a})_{a\in G}$ simultaneously, rather than the individual inference. This is the case particularly, when the model covariates are highly correlated with each other or they are likely to affect the outcome (in time series application, the future covariate vectors) jointly.

To address group inference, we focus on the time series setting. The setting of batched data collection can be handled in a similar way. We first state a simple generalization of Proposition 3.10 to a group of coordinates with finite size as $n, p \to \infty$. The proof is very similar to the proof of Proposition 3.10 and is omitted.

Lemma 4.1. *Let $G = G(n)$ be a sequence of sets $G(n) \subset [dp]$ with $|G(n)| = k$ fixed as $n, p \to \infty$. Also, let the conditional variance $V_n \in \mathbb{R}^{dp\times dp}$ be defined by (43) for the $\mathsf{VAR}(d)$ model, that is:*

$$V_n \equiv \frac{\sigma^2}{n} \sum_{\ell=0}^{K-2} \sum_{t\in E_{\ell+1}} (M^{(\ell)}x_t)(M^{(\ell)}x_t)^{\mathsf{T}} . \tag{51}$$

Under the assumptions of Proposition 3.10, for all $u = (u_1, \dots, u_k) \in \mathbb{R}^k$ we have

$$\lim_{n\to\infty} \left|\mathbb{P}\left\{\sqrt{n}(V_{n,G})^{-1/2}(\widehat{\theta}^{\mathsf{on}}_G - \theta_{0,G}) \le u\right\} - \Phi_k(u)\right| = 0\,, \tag{52}$$

where $V_{n,G} \in \mathbb{R}^{k\times k}$ is the submatrix obtained by restricting V_n to the rows and columns in G. Here $(a_1, \dots, a_k) \le (b_1, \dots, b_k)$ indicates that $a_i \le b_i$ for $i \in [k]$ and $\Phi_k(u) = \Phi(u_1)\dots\Phi_k(u)$.

Much in the same way as individual inference, we can use Lemma 4.1 for simultaneous inference on a group of parameters. Concretely, let $\mathcal{S}_{k,\alpha} \subseteq \mathbb{R}^k$ be any Borel set with k-dimensional Gaussian measure at least $1-\alpha$. Then for a group $G \subset [dp]$, with size $|G| = k$, we construct the confidence set $J_G(\alpha) \subseteq \mathbb{R}^k$ as follows

$$J_G(\alpha) \equiv \widehat{\theta}^{\mathsf{on}}_G + (V_{n,R})^{1/2} \mathcal{S}_{k,\alpha} \,. \tag{53}$$

Then, using Lemma 4.1 (along the same lines in deriving (48)), we conclude that $J_G(\alpha)$ is a valid confidence region, namely

$$\lim_{n\to\infty} \mathbb{P}(\theta_{0,G} \in J_G(\alpha)) = 1-\alpha \,. \tag{54}$$

5 Numerical experiments

In this section, we evaluate the performance of online debiasing framework on synthetic data. Consider the $\mathsf{VAR}(d)$ time series model (21). In the first setting, we let $p = 20$, $d = 3$, $T = 50$ and construct the covariance matrix of noise terms Σ_ζ by putting 1 on its diagonal and $\rho = 0.3$ on its off-diagonal. To make it closer to the practice, instead of considering sparse coefficient matrices , we work with approximately sparse matrices. Specifically, the entries of $A^{(i)}$ are generated independently from a Bernoulli distribution with success probability $q = 0.1$, multiplied by $b \cdot \text{Unif}(\{+1,-1\})$ with $b = 0.1$, and then added to a Gaussian matrix with mean 0 and standard error $1/p$. In formula, each entry is generated independently from

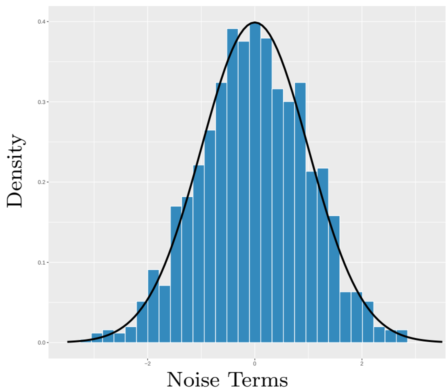
$$b \cdot \text{Bern}(q) \cdot \text{Unif}(\{+1,-1\}) + \mathcal{N}(0, 1/p^2) \,.$$

We used $r_0 = 6$ (length of first episode E_0) and $\beta = 1.3$ for lengths of other episodes $E_\ell \sim \beta^\ell$. For each $i \in [p]$ we do the following. Let $\theta_0 = (A^{(1)}_i, A^{(2)}_i, \dots, A^{(d)}_i)^\mathsf{T} \in \mathbb{R}^{dp}$ encode the i^{th} rows of the matrices $A^{(\ell)}$ and compute the noise component of $\widehat{\theta}^{\mathsf{on}}$ as

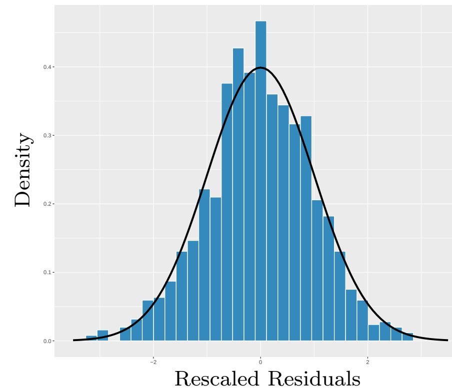
$$W_n \equiv \frac{1}{\sqrt{n}} \sum_{\ell=0}^{K-2} M^{(\ell)} \Big(\sum_{t \in E_{\ell+1}} x_t \varepsilon_t \Big) \,, \tag{55}$$

and rescaled residual $T_n \in \mathbb{R}^{dp}$ with $T_{n,a} = \sqrt{\frac{n}{V_{n,a}}} (\widehat{\theta}^{\mathsf{on}}_a - \theta_{0,a})$ and $V_{n,a}$ given by Equation (43) and $\sigma = 1$. Left and right plots of Figure 4 denote the QQ-plot, PP-plot and histogram of noise terms and rescaled residuals of *all coordinates* (across all $i \in [p]$ and $a \in [dp]$) stacked together, respectively.

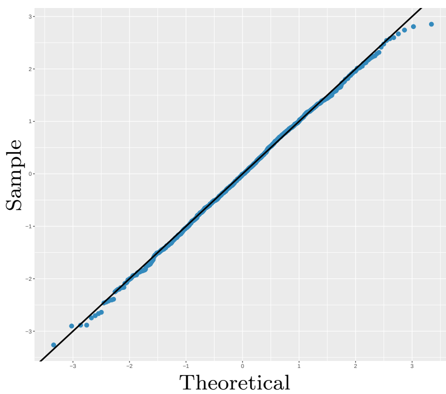
True and False Positive Rates. Consider the linear time-series model (21) with $A^{(i)}$ matrices having entries drawn independently from the distribution $b \cdot \text{Bern}(q) \cdot \text{Unif}(\{+1,-1\})$ and noise terms be gaussian with covariance matrix Σ_ζ. In this example, we evaluate the performance of our proposed online debiasing method for constructing confidence intervals and hypothesis testing as discussed in Section 4. We consider four metrics: True Positive Rate (TPR) , False Positive Rate (FPR), Average length of confidence intervals (Avg CI length), and coverage rate of confidence intervals. Tables 1 and 2 summarize the results for various configurations of the Var(d) processes and significance level $\alpha = 0.05$. Table 1 corresponds to the cases where noise covariance has the structure $\Sigma_\zeta(i,j) = 0.1^{|i-j|}$ and Table 2 corresponds to the case of $\Sigma_\zeta(i,j) = 0.1^{\mathbb{I}(i\neq j)}$. The reported



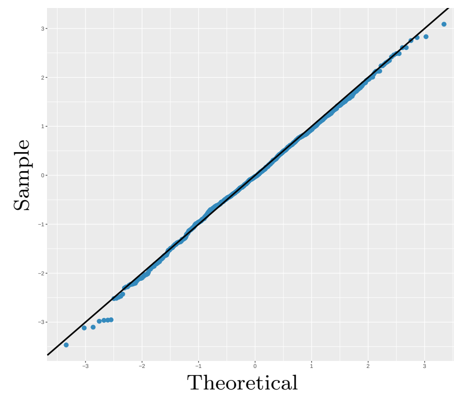
(a) Histogram of Noise Terms W_n



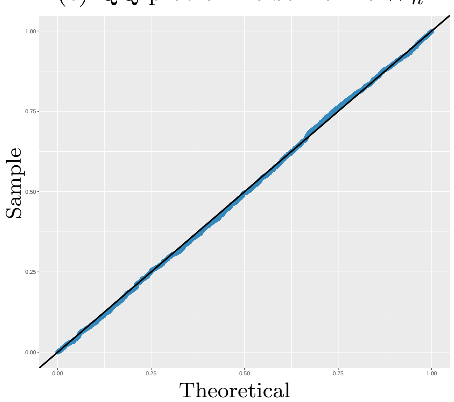
(b) Histogram of Residuals T_n



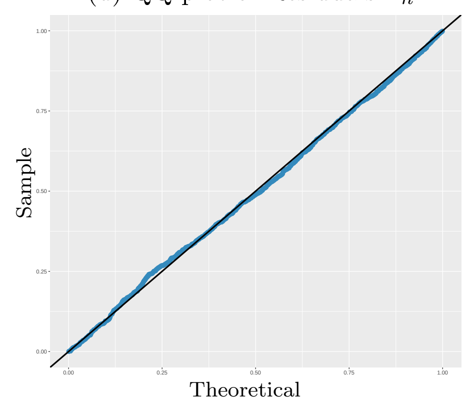
(c) QQ plot of Noise Terms W_n



(d) QQ plot of Residuals T_n



(e) PP plot of Noise Terms W_n



(f) PP plot of Residuals T_n

Figure 4: A simple example of an online debiased Var(3) process with dimension $p = 20$ and $T = 50$ sample data points. Plots 4a, 4c, 4e demonstrate respectively the histogram, QQ-plot, and PP plot of noise values of all $dp^2 = 1200$ entries of A_i matrices in linear time series model (21). Plots 4b, 4d, 4f are histogram, QQ-plot, and PP-plot of rescaled residuals of all coordinates as well. Alignment of data points in these plots with their corresponding standard normal $(0, 1)$ line corroborates our theoretical results on the asymptotic normal behavior of noise terms and rescaled residuals discussed in corollary 3.9 and proposition 3.10, respectively.

Table 1: Evaluation of the online debiasing approach for statistical inference on the coefficients of a $\mathsf{VAR}(d)$ model under different configurations. Here the noise terms ζ_i are gaussian with covariance matrix $\Sigma_\zeta(i,j) = 0.1^{|i-j|}$. The results are reported in terms of four metrics: FPR (False Positive Rate), TPR (True Positive Rate), Coverage rate and Average length of confidence intervals (Avg CI length) at significance level $\alpha = 0.05$

d \ Parameters	p	T	q	b	FPR	TPR	Avg CI length	Coverage rate
$d = 1$	40	30	0.01	2	0.0276	1	3.56	0.9725
	35	30	0.01	2	0.0354	0.9166	3.7090	0.9648
	60	55	0.01	0.9	0.0314	0.7058	2.5933	0.9686
$d = 2$	55	100	0.01	0.8	0.0424	0.8000	1.9822	0.9572
	40	75	0.01	0.9	0.0343	0.9166	2.5166	0.9656
	50	95	0.01	0.7	0.0368	0.6182	2.4694	0.963
$d = 3$	45	130	0.005	0.9	0.0370	0.6858	2.070	0.9632
	40	110	0.01	0.7	0.0374	0.6512	2.1481	0.9623
	50	145	0.005	0.85	0.0369	0.6327	2.2028	0.9631

measures for each configuration (each row of the table) are average over 20 different realizations of the $\mathsf{VAR}(d)$ model.

Table 2: Evaluation of the online debiasing approach for statistical inference on the coefficients of a $\mathsf{VAR}(d)$ model under different configurations. Here the noise terms ζ_i are gaussian with covariance matrix $\Sigma_\zeta(i,j) = 0.1^{\mathbb{I}(i \neq j)}$. The results are reported in terms of four metrics: FPR (False Positive Rate), TPR (True Positive Rate), Coverage rate and Average length of confidence intervals (Avg CI length) at significance level $\alpha = 0.05$

d \ Parameters	p	T	q	b	FPR	TPR	Avg CI length	Coverage rate
$d = 1$	40	30	0.01	2	0.0402	1	3.5835	0.96
	40	35	0.02	1.2	0.0414	0.8125	2.6081	0.9575
	50	40	0.015	0.9	0.0365	0.7435	2.0404	0.9632
$d = 2$	35	65	0.01	0.9	0.0420	0.8077	2.4386	0.9580
	45	85	0.01	0.9	0.0336	0.7298	2.5358	0.9655
	50	70	0.01	0.95	0.0220	0.8333	2.4504	0.9775
$d = 3$	40	115	0.01	0.9	0.0395	0.7906	1.6978	0.9598
	45	130	0.005	0.95	0.0359	0.7714	2.1548	0.9641
	50	145	0.005	0.85	0.0371	0.5918	2.1303	0.9624

5.1 Real data experiments: a marketing application

Retailers often offer sales of various categories of products and for an effective management of the business, they need to understand the cross-category effect of products on each other, e.g., how does price, promotion or sale of category A will effect the sales of category B after some time.

We used data of sales, prices and promotions of Chicago-area grocery store chain Dominick's that can be found publicly from `https://research.chicagobooth.edu/kilts/marketing-databases/dominicks`. The same data set has been used in [WBBM17] where a VARX model is employed to estimate the demand effects. In this experiment, we use the proposed online debiasing approach to provide p-values for the cross-category effects.

We consider 25 categories of products over 71 weeks, so for each week i, we have information x_i for sales, prices and promotions of 25 categories (as no promotion was considered for cigarette during our observed time interval, x_i's have dimension 74). For more details regarding calculating sales, prices and promotions see [SPHD04] and [GWC16]. We consider $\mathsf{VAR}(1)$ and $\mathsf{VAR}(2)$ models as generating process for covariates x_i and then apply our proposed online debiasing method to calculate p-values (see Eq. (49)) for the null hypothesis of form $H_0: \theta_{0,a} = 0$ with $\theta_{0,a}$ an entry in the VAR models as discussed earlier in Section 4. We refer to Appendix E for the reports of the p-values. Here we highlight some of the significant associations using $\mathsf{VAR}(2)$ model: promotion of soaps on sales of dish-detergent after one week with p-value $= 0.0011$; promotion of shampoos on sales of laundry detergent after one week with p-value $= 0.0093$; promotion of front-end candies on sales of soft drinks after one week with p-val $= 0.0257$.

6 Implementation and extensions

6.1 Iterative schemes to implement online debiasing

The online debiased estimator (31) involves the decorrelating matrices $M^{(\ell)}$, whose rows $(m_a^\ell)_{a\in[dp]}$ are constructed by the optimization (29). For the sake of computational efficiently, it is useful to work with a Lagrangian equivalent version of this optimization. Consider the following optimization

$$\text{minimize}_{\|m\|_1\le L} \quad \frac{1}{2} m^\mathsf{T} \widehat{\Sigma}^{(\ell)} m - \langle m, e_a\rangle + \mu_\ell \|m\|_1\,, \tag{56}$$

with μ_ℓ and L taking the same values as in Optimization (29).

The next result, from [Jav14, Chapter 5] is on the connection between the solutions of the unconstrained problem (56) and (29). For the reader's convenience, the proof is also given in Appendix C.1.

Lemma 6.1. *A solution of optimization* (56) *is also a solution of the optimization problem* (29)*. Also, if problem* (29) *is feasible then problem* (56) *has bounded solution.*

Using the above lemma, we can instead work with the Lagrangian version (56) for constructing the decorrelating vector m_a^ℓ.

Here, we propose to solve optimization problem (56) using iterative method. Note the objective function evolves slightly at each episode and hence we expect the solutions m_a^ℓ and $m_a^{\ell+1}$ to be close to each other. An appealing property of iterative methods is that we can leverage this observation by setting m_a^ℓ as the initialization for the iterations that compute $m_a^{\ell+1}$, yielding shorter convergence time. In the sequel we discuss two of such iterative schemes.

6.1.1 Coordinate descent algorithms

In this method, at each iteration we update one of the coordinates of m, say m_j, while fixing the other coordinates. We write the objective function of (56) by separating m_j from the other coordinates:

$$\frac{1}{2}\widehat{\Sigma}^{(\ell)}_{j,j} m_j^2 + \sum_{r,s\neq j} \widehat{\Sigma}^{(\ell)}_{r,s}\, m_r m_s - m_a + \mu_\ell \|m_{\sim j}\|_1 + \mu_\ell |m_j|\,, \tag{57}$$

where $\widehat{\Sigma}^{(\ell)}_{j,\sim j}$ denotes the $j^{\rm th}$ row (column) of $\widehat{\Sigma}^{(\ell)}$ with $\widehat{\Sigma}^{(\ell)}_{j,j}$ removed. Likewise, $m_{\sim j}$ represents the restriction of m to coordinates other than j. Minimizing (57) with respect to m_j gives

$$m_j + \frac{1}{\widehat{\Sigma}^{(\ell)}_{j,j}} \Big(\widehat{\Sigma}^{(\ell)}_{j,\sim j} m_{\sim j} - \mathbb{I}(a=j) + \mu_\ell \,\mathrm{sign}(m_j) \Big) = 0\,.$$

It is easy to verify that the solution of the above is given by

$$m_j = \frac{1}{\widehat{\Sigma}^{(\ell)}_{j,j}} \eta\Big(-\widehat{\Sigma}^{(\ell)}_{j,\sim j} m_{\sim j} + \mathbb{I}(a=j); \mu_\ell \Big)\,, \tag{58}$$

with $\eta(\cdot;\cdot) : \mathbb{R}\times\mathbb{R}_+ \to \mathbb{R}$ denoting the soft-thresholding function defined as

$$\eta(z,\mu) = \begin{cases} z-\mu & \text{if } z>\mu\,,\\ 0 & \text{if } -\mu \le z \le \mu\,,\\ z+\mu & \text{otherwise}\,. \end{cases} \tag{59}$$

For a vector u, $\eta(u;\mu)$ is perceived entry-wise.

This brings us to the following update rule to compute $m_a^\ell \in \mathbb{R}^{dp}$ (solution of (56)). Th notation Π_L, in line 5 below, denotes the Euclidean projection onto the ℓ_1 ball of radius L and can be computed in $O(dp)$ times using the procedure of [DSSSC08].

1: (initialization): $m(0) \leftarrow m_a^{(\ell-1)}$
2: **for** iteration $h = 1,\dots,H$ **do**
3:     **for** $j = 1,2,\dots,dp$ **do**
4:         $m_j(h) \leftarrow \frac{1}{\widehat{\Sigma}^{(\ell)}_{j,j}}\eta\Big(-\widehat{\Sigma}^{(\ell)}_{j,\sim j} m_{\sim j}(h-1) + \mathbb{I}(a=j);\mu_\ell\Big)$
5:     $m(h) \leftarrow \Pi_L(m(h))$
6: **return** $m_a^\ell \leftarrow m(H)$

In our experiments we implemented the same coordinate descent iterations explained above to solve for the decorrelating vectors m_a^ℓ.

6.1.2 Gradient descent algorithms

Letting $\mathcal{L}(m) = (1/2) m^{\mathsf{T}} \widehat{\Sigma}^{(\ell)} m - \langle m, e_a\rangle$, we can write the objective of (56) as $\mathcal{L}(m) + \mu_\ell \|m\|_1$. Projected gradient descent, applied to this constrained objective, results in a sequence of iterates $m(h)$, with $h = 0,1,2,\dots$ the iteration number, as follows:

$$m(h+1) = \arg\min_{\|m\|_1\le L} \Big\{ \mathcal{L}(m(h)) + \langle \nabla\mathcal{L}(m(h)), m - m(h)\rangle + \frac{\eta}{2}\|m - m(h)\|_2^2 + \mu_\ell \|m\|_1 \Big\}\,. \tag{60}$$

In words, the next iterate $m(h+1)$ is obtained by constrained minimization of a first order approximation to $\mathcal{L}(m)$, combined with a smoothing term that keeps the next iterate close to the current one. Since the objective function is convex ($\widehat{\Sigma}^{(\ell)} \succeq 0$), iterates (60) are guaranteed to converge to the global minimum of (56).

Plugging for $\mathcal{L}(m)$ and dropping the constant term $\mathcal{L}(m(h))$, update (60) reads as

$$\begin{aligned} m(h+1) &= \arg\min_{\|m\|_1 \le L} \Big\{ \langle \widehat{\Sigma}^{(\ell)} m(h) - e_a, m - m(h)\rangle + \frac{\eta}{2}\|m - m(h)\|_2^2 + \mu_\ell \|m\|_1 \Big\} \\ &= \arg\min_{\|m\|_1 \le L} \Big\{ \frac{\eta}{2}\Big(m - m(h) + \frac{1}{\eta}(\widehat{\Sigma}^{(\ell)} m(h) - e_a)\Big)^2 + \mu_\ell \|m\|_1 \Big\}. \end{aligned} \tag{61}$$

To compute the update (61), we first solve the unconstrained problem which has a closed form solution given by $\eta\Big(m(h) - \frac{1}{\eta}(\widehat{\Sigma}^{(\ell)} m(h) - e_a); \frac{\mu_\ell}{\eta}\Big)$, with η the soft thresholding function given by (59). The solution is then projected onto the ball of radius L.

In the following box, we summarize the projected gradient descent update rule for constructing the decorrelating vectors m_a^ℓ.

1: (initialization): $m(0) \leftarrow m_a^{(\ell-1)}$
2: **for** iteration $h = 1, \dots, H$ **do**
3: $\quad m(h) \leftarrow \eta\Big(m(h) - \frac{1}{\eta}(\widehat{\Sigma}^{(\ell)} m(h) - e_a); \frac{\mu_\ell}{\eta}\Big)$
4: $\quad m(h) \leftarrow \Pi_L(m(h))$
5: **return** $m_a^\ell \leftarrow m(H)$

6.2 Sparse inverse covariance

In Section 3.1 (Figure 3) we provided a numerical example wherein the offline debiasing does not admit an asymptotically normal distribution. As we see from the heat map in Figure 2b, the precision matrix Ω has $\sim 20\%$ non-negligible entries per row. The goal of this section is to show that when Ω is sufficiently sparse, the offline debiased estimator has an asymptotically normal distribution and can be used for valid inference on model parameters.

The idea is to show that the decorrelating matrix M is sufficiently close to the precision matrix Ω. Since Ω is deterministic, this helps with controlling the statistical dependence between M and ε. Formally, starting from the decomposition (5) we write

$$\begin{aligned} \widehat{\theta}^{\sf off} &= \theta_0 + (I - M\widehat{\Sigma}^{(K-1)})(\widehat{\theta}^{\sf L} - \theta_0) + \frac{1}{n} M X^{\sf T} \varepsilon \\ &= \theta_0 + (I - M\widehat{\Sigma}^{(K-1)})(\widehat{\theta}^{\sf L} - \theta_0) + \frac{1}{n}(M - \Omega) X^{\sf T} \varepsilon + \frac{1}{n}\Omega X^{\sf T}\varepsilon, \end{aligned} \tag{62}$$

where we recall that $\widehat{\Sigma}^{(K-1)}$ is the empirical covariance of all the covariate vectors (episodes $E_0, \dots, E_{K-1}$). Therefore, we can write

$$\begin{aligned} \sqrt{n}(\widehat{\theta}^{\sf off} - \theta_0) &= \Delta_1 + \Delta_2 + \frac{1}{\sqrt{n}}\Omega X^{\sf T}\varepsilon, \\ \Delta_1 &= \sqrt{n}(I - M\widehat{\Sigma}^{(K-1)})(\widehat{\theta}^{\sf L} - \theta_0), \\ \Delta_2 &= \frac{1}{\sqrt{n}}(M - \Omega) X^{\sf T}\varepsilon. \end{aligned} \tag{63}$$

The term $\Omega X^\mathsf{T}\varepsilon/\sqrt{n}$ is gaussian with $O(1)$ variance at each coordinate. For bias term Δ_1, we show that $\Delta_1 = O(s_0(\log p)/\sqrt{n})$ by controlling $|I - M\widehat{\Sigma}^{(K-1)}|$. To bound the bias term Δ_2 we write

$$\|\Delta_2\|_\infty \le \left(\max_{i\in[p]} \|(M-\Omega)e_i\|_1\right)\left(\frac{1}{\sqrt{n}}\|X^\mathsf{T}\varepsilon\|_\infty\right). \tag{64}$$

By using [BM+15, Proposition 3.2], we have $\|X^\mathsf{T}\varepsilon\|_\infty/\sqrt{n} = O_P(\sqrt{\log(dp)})$. Therefore, to bound Δ_2 we need to control $M - \Omega$ (in the maximum ℓ_1 distance of the rows). We provide such bound in our next lemma, under the sparsity assumption on the rows of Ω.

Define

$$s_\Omega \equiv \max_{i\in[dp]} \Big|j \in [dp]: \ \ \Omega_{i,j} \ne 0\Big|,$$

the maximum sparsity of rows of Ω. In addition, let the (offline) decorrelating vectors m_a be defined as follows, for $a \in [dp]$:

$$m_a \in \arg\min_{m\in\mathbb{R}^{dp}} \quad \frac{1}{2} m^\mathsf{T}\widehat{\Sigma}^{(K-1)}m - \langle m, e_a\rangle + \mu_n\|m\|_1. \tag{65}$$

Lemma 6.2. *Consider the decorrelating vectors m_a, $a \in [dp]$, given by optimization (65) with $\mu_n = 2\tau\sqrt{\frac{\log(dp)}{n}}$. Then, for some proper constant $c > 0$ and the sample size condition $n \ge 32\alpha(\omega^2 \vee 1)s_\Omega \log(dp)$, the following happens with probability at least $1-\exp(-c\log(dp^2))-\exp(-cn(1\wedge\omega^{-2}))$:*

$$\max_{i\in[dp]} \|m_a - \Omega e_a\|_1 \le \frac{192\tau}{\alpha} s_\Omega\sqrt{\frac{\log(dp)}{n}},$$

where α and ω are defined in Proposition 3.3.

The proof of Lemma 6.2 is deferred to Section C.2.

By employing this lemma, if Ω is sufficiently sparse, that is $s_\Omega = o(\sqrt{n}/\log(dp))$, then the bias term $\|\Delta_2\|_\infty$ also vanishes asymptotically and the (offline) debiased estimator $\widehat{\theta}^{\mathsf{off}}$ admits an unbiased normal distribution. We formalize such distributional characterization in the next theorem.

Theorem 6.3. *Consider the $\mathsf{VAR}(d)$ model (21) for time series and let $\widehat{\theta}^{\mathsf{off}}$ be the (offline) debiased estimator (4), with the decorrelating matrix $M = (m_1, \dots, m_{dp})^\mathsf{T} \in \mathbb{R}^{dp\times dp}$ constructed as in (65), with $\mu_n = 2\tau\sqrt{(\log p)/n}$. Also, let $\lambda = \lambda_0\sqrt{\log(dp)/n}$ be the regularization parameter in the Lasso estimator $\widehat{\theta}^{\mathsf{L}}$, with τ, λ_0 large enough constants.*

Suppose that $s_0 = o(\sqrt{n}/\log(dp))$ and $s_\Omega = o(\sqrt{n}/\log(dp))$, then the following holds true for any fixed sequence of integers $a(n) \in [dp]$: For all $x \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \sup_{\|\theta_0\|_0\le s_0} \left|\mathbb{P}\left\{\frac{\sqrt{n}(\widehat{\theta}^{\mathsf{off}}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} \le x\right\} - \Phi(x)\right| = 0\,, \tag{66}$$

where $V_{n,a} \equiv \sigma^2(M\widehat{\Sigma}^{(n)}M^\mathsf{T})_{a,a}$.

We refer to Section C.3 for the proof of Theorem 6.3.

A Numerical Example. Consider a $\mathsf{VAR}(d)$ model with parameters $p = 25, d = 3, T = 70$, and Gaussian noise terms with covariance matrix Σ_ζ satisfying $\Sigma_\zeta(i,j) = \rho^{|i-j|}$ for $\rho = 0.1$. Let A_i matrices have entries generated independently from $b \cdot \mathrm{Bern}(q).\mathrm{Unif}(\{+1,-1\})$ formula with parameters $b = 0.15, q = 0.05$. Figure 5a shows the magnitudes of entries of elements of sparse precision matrix $\Omega = \mathbb{E}(x_i x_i^T)^{-1}$. Figures 5b, 5c, and 5d demonstrate normality of rescaled residuals of offline debiased estimator built by decorrelating matrix M with rows coming from optimization described in (65).

After this paper was posted, we learned of simultaneous work (an updated version of [BDMP17]) that also studies the performance of the (offline) debiased estimator for time series with sparse precisions. We would like to highlight some of differences between the two papers: 1) [BDMP17] considers decorrelating matrix M constructed by an optimization of form (29), using the entire sample covariance $\widehat{\Sigma}^{(K-1)}$, while we work with the Lagrangian equivalent (65). 2) [BDMP17] considers $\mathsf{VAR}(1)$ model, while we work with $\mathsf{VAR}(d)$ models. 3) [BDMP17] assumes a stronger notion of sparsity, viz. the sparsity of the entire precision matrix as well as the transition matrix to scale as $o(\sqrt{n}/\log p)$. Our results only require the *row-wise sparsity* of the precision matrix to scale as $o(\sqrt{n}/\log p)$, cf. Theorem 6.3.

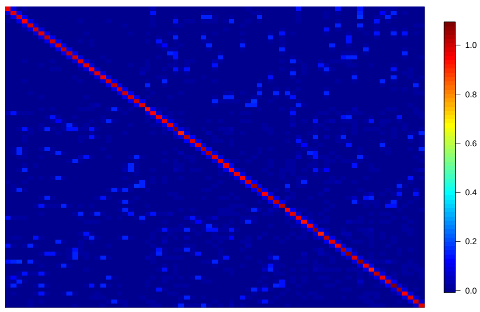
6.3 Concluding remarks

In this work we devised the 'online debiasing' approach for the high-dimensional regression and showed that it asymptotically admits an unbiased Gaussian distribution, even when the samples are collected adaptively. Also through numerical examples we demonstrated that the (offline) debiased estimator suffers from the bias induced by the correlation in the samples and cannot be used for valid statistical inference in these settings (unless the precision matrix is sufficiently sparse).

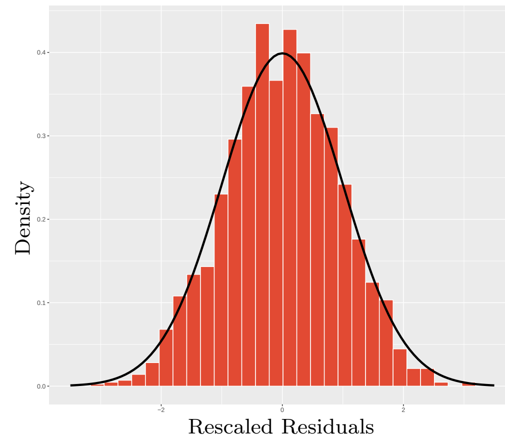
Since its proposal, the (offline) debiasing approach has been used as a tool to address a variety of problems such as estimating average treatment effect and casual inference in high-dimension [AIW16], precision matrix estimation [JvdG17], distributed multitask learning, and studying neuronal functional network dynamics [SML+18], hierarchical testing [GRBC19], to name a few. It has also been used for different statistical aims such as controlling FDR in high-dimensions [JJ+19], estimation of the prediction risk [JM18], inference on predictions [CG17, JL17] and explained variance [CG18, JL17], to testing more general hypotheses regarding the model parameters, like testing membership in a convex cone, testing the parameter strength, and testing arbitrary functions of the parameters [JL17]. We anticipate that the online debiasing approach and analysis can be used to tackle similar problems under adaptive data collection. We leave this for future work.

Acknowledgements

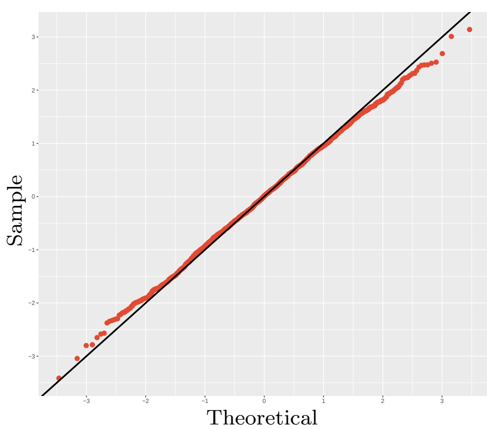
A. Javanmard was partially supported by an Outlier Research in Business (iORB) grant from the USC Marshall School of Business, a Google Faculty Research Award and the NSF CAREER Award DMS-1844481.



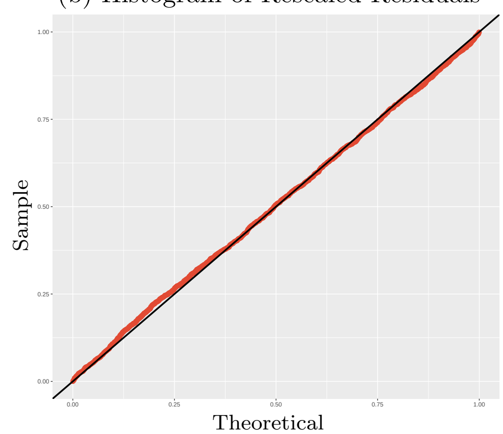
(a) Heat map of magnitudes of entries of $\Omega = \mathbb{E}(x_i x_i^T)^{-1}$



(b) Histogram of Rescaled Residuals



(c) QQ plot of Rescaled Residuals



(d) PP plot of Rescaled Residuals

Figure 5: A Simple example of a $\mathsf{VAR}(d)$ process with parameters $p = 25, d = 3, T = 70$, and noise term covariance matrix Σ_ζ s.t $\Sigma_\zeta(i,j) = \rho^{|i-j|}$ with $\rho = 0.1$. A_i matrices have independent elements coming from $b \cdot \text{Bern}(q).\text{Unif}(\{+1,-1\})$ formula with $b = 0.15, q = 0.05$. Normality of rescaled residuals (figures 5b, 5c, and 5d) validates the successful performance of offline debiasing estimator under sparsity of precision matrix Ω (figure 5a) as we discussed in theorem 6.3.

References

[AIW16] Susan Athey, Guido W Imbens, and Stefan Wager, *Approximate residual balancing: De-biased inference of average treatment effects in high dimensions*, arXiv preprint arXiv:1604.07125 (2016). 31

[BB15] Hamsa Bastani and Mohsen Bayati, *Online decision-making with high-dimensional covariates*, Available at SSRN 2661896 (2015). 2, 5

[BCB+12] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al., *Regret analysis of stochastic and non-stochastic multi-armed bandit problems*, Foundations and Trends® in Machine Learning **5** (2012), no. 1, 1–122. 5

[BCW11] Alexandre Belloni, Victor Chernozhukov, and Lie Wang, *Square-root lasso: pivotal recovery of sparse signals via conic programming*, Biometrika **98** (2011), no. 4, 791–806. 6

[BDMP17] Sumanta Basu, Sreyoshi Das, George Michailidis, and Amiyatosh K Purnanandam, *A system-wide approach to measure connectivity in the financial sector*, Available at SSRN 2816137 (2017). 31

[BM12] M. Bayati and A. Montanari, *The LASSO risk for gaussian matrices*, IEEE Trans. on Inform. Theory **58** (2012), 1997–2017. 2

[BM+15] Sumanta Basu, George Michailidis, et al., *Regularized estimation in sparse high-dimensional time series models*, The Annals of Statistics **43** (2015), no. 4, 1535–1567. 2, 14, 15, 17, 30, 48, 49, 50, 57

[BRT09] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Amer. J. of Mathematics **37** (2009), 1705–1732. 16

[BV04] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004. 10

[BVDG11] Peter Bühlmann and Sara Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media, 2011. 5, 6, 16, 37

[CG17] T Tony Cai and Zijian Guo, *Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity*, The Annals of statistics **45** (2017), no. 2, 615–646. 3, 31

[CG18] ———, *Semi-supervised inference for explained variance in high-dimensional linear regression and its applications*, arXiv preprint arXiv:1806.06179 (2018). 31

[DM12] Yash Deshpande and Andrea Montanari, *Linear bandits in high dimension and recommendation systems*, Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on, IEEE, 2012, pp. 1750–1754. 5

[DMST18] Yash Deshpande, Lester Mackey, Vasilis Syrgkanis, and Matt Taddy, *Accurate inference for adaptive linear models*, International Conference on Machine Learning, 2018, pp. 1202–1211. 3, 4

[DSSSC08] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra, *Efficient projections onto the l 1-ball for learning in high dimensions*, Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 272–279. 28

[FSGM+07] André Fujita, Joao R Sato, Humberto M Garay-Malpartida, Rui Yamaguchi, Satoru Miyano, Mari C Sogayar, and Carlos E Ferreira, *Modeling gene expression regulatory networks with the sparse vector autoregressive model*, BMC systems biology **1** (2007), no. 1, 39. 13

[GRBC19] Zijian Guo, Claude Renaux, Peter Bühlmann, and T Tony Cai, *Group inference in high dimensions with applications to hierarchical testing*, arXiv preprint arXiv:1909.01503 (2019). 31

[GWC16] Sarah Gelper, Ines Wilms, and Christophe Croux, *Identifying demand effects in a large network of product categories*, Journal of Retailing **92** (2016), no. 1, 25–39. 27

[HENR88] Douglas Holtz-Eakin, Whitney Newey, and Harvey S Rosen, *Estimating vector autoregressions with panel data*, Econometrica: Journal of the Econometric Society (1988), 1371–1395. 13

[HH14] Peter Hall and Christopher C Heyde, *Martingale limit theory and its application*, Academic press, 2014. 21, 44, 58

[HTW15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright, *Statistical learning with sparsity: the lasso and generalizations*, Chapman and Hall/CRC, 2015. 38

[Jav14] Adel Javanmard, *Inference and estimation in high-dimensional data analysis*, Ph.D. thesis, PhD Thesis, Stanford University, 2014. 3, 23, 27

[JJ+19] Adel Javanmard, Hamid Javadi, et al., *False discovery rate control via debiased lasso*, Electronic Journal of Statistics **13** (2019), no. 1, 1212–1253. 31

[JL17] Adel Javanmard and Jason D Lee, *A flexible framework for hypothesis testing in high-dimensions*, arXiv preprint arXiv:1704.07971 (2017). 31

[JM14a] Adel Javanmard and Andrea Montanari, *Confidence intervals and hypothesis testing for high-dimensional regression.*, Journal of Machine Learning Research **15** (2014), no. 1, 2869–2909. 3, 6, 8, 10, 17

[JM14b] ———, *Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory*, IEEE Transactions on Information Theory **60** (2014), no. 10, 6522–6554. 2, 3, 6, 8, 23

[JM18] ———, *Debiasing the lasso: Optimal sample size for gaussian designs*, The Annals of Statistics **46** (2018), no. 6A, 2593–2622. 8, 31

[JvdG17] Jana Janková and Sara van de Geer, *Honest confidence regions and optimality in high-dimensional precision matrix estimation*, Test **26** (2017), no. 1, 143–162. 31

[KHW+11] Edward S Kim, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus, Sanjay Gupta, et al., *The battle trial: personalizing therapy for lung cancer*, Cancer discovery **1** (2011), no. 1, 44–53. 5

[LR85] Tze Leung Lai and Herbert Robbins, *Asymptotically efficient adaptive allocation rules*, Advances in applied mathematics **6** (1985), no. 1, 4–22. 5

[LW82] Tze Leung Lai and Ching Zong Wei, *Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems*, The Annals of Statistics (1982), 154–166. 2, 14

[LW12] Po-Ling Loh and Martin J. Wainwright, *High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity*, Ann. Statist. **40** (2012), no. 3, 1637–1664. 16

[NXTZ17] Xinkun Nie, Tian Xiaoying, Jonathan Taylor, and James Zou, *Why adaptively collected data have negative bias and how to correct for it*. 2

[PRC+16] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, Erik Snowberg, et al., *Batched bandit problems*, The Annals of Statistics **44** (2016), no. 2, 660–681. 5

[RT10] Paat Rusmevichientong and John N Tsitsiklis, *Linearly parameterized bandits*, Mathematics of Operations Research **35** (2010), no. 2, 395–411. 5

[SBB15] Anil K Seth, Adam B Barrett, and Lionel Barnett, *Granger causality analysis in neuroscience and neuroimaging*, Journal of Neuroscience **35** (2015), no. 8, 3293–3297. 13

[SML+18] Alireza Sheikhattar, Sina Miran, Ji Liu, Jonathan B Fritz, Shihab A Shamma, Patrick O Kanold, and Behtash Babadi, *Extracting neuronal functional network dynamics via adaptive granger causality analysis*, Proceedings of the National Academy of Sciences **115** (2018), no. 17, E3869–E3878. 31

[SPHD04] Shuba Srinivasan, Koen Pauwels, Dominique M Hanssens, and Marnik G Dekimpe, *Do promotions benefit manufacturers, retailers, or both?*, Management Science **50** (2004), no. 5, 617–629. 27

[SRR19] Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo, *On the bias, risk and consistency of sample means in multi-armed bandits*, arXiv preprint arXiv:1902.00746 (2019). 2

[SS06] Robert H Shumway and David S Stoffer, *Time series analysis and its applications: with r examples*, Springer Science & Business Media, 2006. 13, 14

[SW01] James H Stock and Mark W Watson, *Vector autoregressions*, Journal of Economic perspectives **15** (2001), no. 4, 101–115. 13

[SZ12] Tingni Sun and Cun-Hui Zhang, *Scaled sparse linear regression*, Biometrika **99** (2012), no. 4, 879–898. 6

[Tib96] R. Tibshirani, *Regression shrinkage and selection with the Lasso*, J. Royal. Statist. Soc B **58** (1996), 267–288. 2

[VBW15] Sofia Villar, Jack Bowden, and James Wason, *Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges*, Statistical science: a review journal of the Institute of Mathematical Statistics **30** (2015), no. 2, 199. 2

[VdGBR+14] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, Ruben Dezeure, et al., *On asymptotically optimal confidence regions and tests for high-dimensional models*, The Annals of Statistics **42** (2014), no. 3, 1166–1202. 3, 6

[Ver12] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, Compressed Sensing: Theory and Applications (Y.C. Eldar and G. Kutyniok, eds.), Cambridge University Press, 2012, pp. 210–268. 60

[WBBM17] Ines Wilms, Sumanta Basu, Jacob Bien, and David S Matteson, *Interpretable vector autoregressions with exogenous time series*, arXiv preprint arXiv:1711.03623 (2017). 26

[XQL13] Min Xu, Tao Qin, and Tie-Yan Liu, *Estimation bias in multi-armed bandit algorithms for search advertising*, Advances in Neural Information Processing Systems, 2013, pp. 2400–2408. 2

[ZLK+08] Xian Zhou, Suyu Liu, Edward S Kim, Roy S Herbst, and J Jack Lee, *Bayesian adaptive design for targeted therapy development in lung cancer—a step toward personalized medicine*, Clinical Trials **5** (2008), no. 3, 181–193. 5

[ZZ11] C.-H. Zhang and S.S. Zhang, *Confidence Intervals for Low-Dimensional Parameters in High-Dimensional Linear Models*, arXiv:1110.2563, 2011. 3

A Proofs of Section 2

A.1 Proof of Theorem 2.2

The important technical step is to prove that, under the conditions specified in Theorem 2.2, the *sample covariance* $\widehat{\Sigma} = (1/n)\sum_i x_i x_i^{\mathsf{T}}$ is $(\phi_0/4, \mathrm{supp}(\theta_0))$ compatible.

Proposition A.1. *With probability exceeding $1 - p^{-4}$ the sample covariance $\widehat{\Sigma}$ is $(\phi_0/4, \mathrm{supp}(\theta_0))$ compatible when $n_1 \vee n_2 \geq 2^{32}(\kappa^4/\phi_0^2)s_0^2 \log p$, for an absolute constant C.*

Let $\widehat{\Sigma}^{(1)}$ and $\widehat{\Sigma}^{(2)}$ denote the sample covariances of each batch, i.e. $\widehat{\Sigma}^{(1)} = (1/n_1)\sum_{i\leq n_1} x_i x_i^{\mathsf{T}}$ and similarly $\widehat{\Sigma}^{(2)} = (1/n_2)\sum_{i>n_1} x_i x_i^{\mathsf{T}}$. We also let $\Sigma^{(2)}$ be the conditional covariance $\Sigma^{(2)} = \Sigma^{(2)}(\widehat{\theta}^1) = \mathbb{E}\{xx^{\mathsf{T}} | \langle x, \widehat{\theta}^1\rangle \geq \varsigma\}$. We first prove that at least one of the sample covariances $\widehat{\Sigma}^{(1)}$ and $\widehat{\Sigma}^{(2)}$ closely approximate their population counterparts, and that this implies they are $(\phi_0/2, \mathrm{supp}(\theta_0))$-compatible.

Lemma A.2. *With probability at least $1 - p^{-4}$*

$$\|\widehat{\Sigma}^{(1)} - \Sigma\|_\infty \wedge \|\widehat{\Sigma}^{(2)} - \Sigma^{(2)}\|_\infty \leq 12\kappa^2\sqrt{\frac{\log p}{n}},$$

Proof. Since $n = n_1 + n_2 \leq 2\max(n_1, n_2)$, at least one of n_1 and n_2 exceeds $n/2$. We assume that $n_2 \geq n/2$, and prove that $\|\widehat{\Sigma}^{(2)} - \Sigma^{(2)}\|_\infty$ satisfies the bound in the claim. The case $n_1 \geq n/2$ is similar. Since we are proving the case $n_2 \geq n/2$, for notational convenience, we assume probabilities and expectations in the rest of the proof are conditional on the first batch $(y_1, x_1), \dots (y_{n_1}, x_{n_1})$, and omit this in the notation.

For a fixed pair $(a, b) \in [p] \times [p]$:

$$\widehat{\Sigma}^{(2)}_{a,b} - \Sigma^{(2)}_{a,b} = \frac{1}{n_2}\sum_{i>n_1} x_{i,a}x_{i,b} - \mathbb{E}\{x_{i,a}x_{i,b}\}$$

Using Lemma D.4 we have that $\|x_{i,a}x_{i,b}\|_{\psi_1} \leq 2\|x_i\|_{\psi_2}^2 \leq 2\kappa^2$ almost surely. Then using the tail inequality Lemma D.5 we have for any $\varepsilon \leq 2e\kappa^2$

$$\mathbb{P}\Big\{|\widehat{\Sigma}^{(2)}_{a,b} - \Sigma^{(2)}_{a,b}| \geq \varepsilon\Big\} \leq 2\exp\Big\{-\frac{n_2\varepsilon^2}{6e\kappa^4}\Big\}$$

With $\varepsilon = \varepsilon(p, n_2, \kappa) = 12\kappa^2\sqrt{\log p/n_2} \leq 20\kappa^2\sqrt{\log p/n}$ we have that $\mathbb{P}\{|\widehat{\Sigma}^{(2)}_{a,b} - \Sigma^{(2)}_{a,b}| \geq \varepsilon(p, n_2, \kappa)\} \leq p^{-8}$, whence the claim follows by union bound over pairs (a, b). □

Lemma A.3 ([BVDG11, Corollary 6.8])**.** *Suppose that Σ is (ϕ_0, S)-compatible. Then any matrix Σ' such that $\|\Sigma' - \Sigma\|_\infty \leq \phi_0/(32|S|)$ is $(\phi_0/2, S)$-compatible.*

We can now prove Proposition A.1.

Proof of Proposition A.1. Combining Lemmas A.2 and A.3 yields that, with probability $1 - p^{-4}$, at least one of $\widehat{\Sigma}^{(1)}$ and $\widehat{\Sigma}^{(2)}$ are $(\phi_0/2, \mathrm{supp}(\theta_0))$-compatible provided

$$12\kappa^2 \sqrt{\frac{\log p}{n}} \le \frac{\phi_0}{32 s_0},$$

$$\text{which is implied by } n \ge \Big(\frac{400\kappa^2 s_0}{\phi_0}\sqrt{\log p}\Big)^2.$$

Since $\widehat{\Sigma} = (n_1/n)\widehat{\Sigma}^{(1)} + (n_2/n)\widehat{\Sigma}^{(2)}$ and at least one of n_1/n and n_2/n exceed $1/2$, this implies that $\widehat{\Sigma}$ is $(\phi_0/4, \mathrm{supp}(\theta_0))$-compatible with probability exceeding $1 - p^{-4}$. □

The following lemma shows that $X^{\mathsf{T}}\varepsilon$ is small entrywise.

Lemma A.4. *For any $\lambda_n \ge 40\kappa\sigma\sqrt{(\log p)/n}$, with probability at least $1 - p^{-4}$, $\|X^{\mathsf{T}}\varepsilon\|_\infty \le n\lambda_n/2$.*

Proof. The a^{th} coordinate of the vector $X^{\mathsf{T}}\varepsilon$ is $\sum_i x_{ia}\varepsilon_i$. As the rows of X are uniformly κ-subgaussian and $\|\varepsilon_i\|_{\psi_2} = \sigma$, Lemma D.4 implies that the sequence $(x_{ia}\varepsilon_i)_{1\le i\le n}$ is uniformly $2\kappa\sigma$-subexponential. Applying the Bernstein-type martingale tail bound Lemma D.6, for $\varepsilon \le 12e\kappa\sigma$:

$$\mathbb{P}\Big\{\Big|\sum_i x_{ia}\varepsilon_i\Big| \ge \varepsilon n\Big\} \le 2\exp\Big\{-\frac{n\varepsilon^2}{24e\kappa^2\sigma^2}\Big\}$$

Set $\varepsilon = \varepsilon(p, n, \kappa, \sigma) = 20\kappa\sigma\sqrt{(\log p)/n}$, the exponent on the right hand side above is at least $5\log p$, which implies after union bound over a that

$$\begin{aligned}
\mathbb{P}\{\|X^{\mathsf{T}}\varepsilon\|_\infty \ge \varepsilon n\} &= \mathbb{P}\Big\{\max_a \Big|\sum_i x_{ia}\varepsilon_i\Big| \ge \varepsilon n\Big\} \\
&\le \sum_a \mathbb{P}\Big\{\Big|\sum_i x_{ia}\varepsilon_i\Big| \ge \varepsilon n\Big\} \\
&\le 2p^{-6}.
\end{aligned}$$

This implies the claim for p large enough. □

The rest of the proof is standard, cf. [HTW15] and is given below for the reader's convenience.

Proof of Theorem 2.2. Throughout we condition on the intersection of good events in Proposition A.1 and Lemma A.4, which happens with probability at least $1 - 2p^{-4}$. On this good event, the sample covariance $\widehat{\Sigma}$ is $(\phi_0/4, \mathrm{supp}(\theta_0))$-compatible and $\|X^{\mathsf{T}}\varepsilon\|_\infty \le 20\kappa\sigma\sqrt{n\log p} \le n\lambda_n/2$.

By optimality of $\widehat{\theta}^{\mathsf{L}}$:

$$\frac{1}{2}\|y - X\widehat{\theta}^{\mathsf{L}}\|^2 + \lambda_n\|\widehat{\theta}^{\mathsf{L}}\|_1 \le \frac{1}{2}\|y - X\theta_0\|^2 + \lambda_n\|\theta_0\|_1.$$

Using $y = X\theta_0 + \varepsilon$, the shorthand $\nu = \widehat{\theta}^{\mathsf{L}} - \theta_0$ and expanding the squares leads to

$$\begin{aligned}
\frac{1}{2}\langle \nu, \widehat{\Sigma}\nu\rangle &\le \frac{1}{n}\langle X^{\mathsf{T}}\varepsilon, \nu\rangle + \lambda_n(\|\theta_0\|_1 - \|\widehat{\theta}^{\mathsf{L}}\|_1) \\
&\le \frac{1}{n}\|\nu\|_1\|X^{\mathsf{T}}\varepsilon\|_\infty + \lambda_n(\|\theta_0\|_1 - \|\widehat{\theta}^{\mathsf{L}}\|_1) \\
&\le \lambda_n\Big\{\frac{1}{2}\|\nu\|_1 + \|\theta_0\|_1 - \|\widehat{\theta}^{\mathsf{L}}\|_1\Big\}.
\end{aligned} \tag{67}$$

First we show that the error vector ν satisfies $\|\nu_{S_0^c}\|_1 \le 3\|\nu_{S_0}\|_1$, where $S_0 \equiv \mathrm{supp}(\theta_0)$. Note that $\|\widehat{\theta}^{\mathsf{L}}\|_1 = \|\theta_0 + \nu\|_1 = \|\theta_0 + \nu_{S_0}\|_1 + \|\nu_{S_0^c}\|_1$. By triangle inequality, therefore:

$$\begin{aligned}\|\theta_0\|_1 - \|\widehat{\theta}^{\mathsf{L}}\|_1 &= \|\theta_0\|_1 - \|\theta_0 + \nu_{S_0}\|_1 - \|\nu_{S_0^c}\|_1 \\ &\le \|\nu_{S_0}\|_1 - \|\nu_{S_0^c}\|_1.\end{aligned}$$

Combining this with the basic lasso inequality Eq.(67) we obtain

$$\begin{aligned}\frac{1}{2}\langle \nu, \widehat{\Sigma}\nu\rangle &\le \lambda_n\Big\{\frac{1}{2}\|\nu\|_1 + \|\nu_{S_0}\|_1 - \|\nu_{S_0^c}\|_1\Big\} \\ &= \frac{\lambda_n}{2}\Big\{3\|\nu_{S_0}\|_1 - \|\nu_{S_0^c}\|.\Big\}\end{aligned}$$

As $\widehat{\Sigma}$ is positive-semidefinite, the LHS above is non-negative, which implies $\|\nu_{S_0^c}\|_1 \le 3\|\nu_{S_0}\|_1$. Now, we can use the fact that $\widehat{\Sigma}$ is $(\phi_0/4, S_0)$-compatible to lower bound the LHS by $\|\nu\|_1^2\phi_0/2s_0$. This leads to

$$\frac{\phi_0\|\nu\|_1^2}{2s_0} \le \frac{3\lambda_n\|\nu_{S_0}\|_1}{2} \le \frac{3\lambda_n\|\nu\|_1}{2}.$$

Simplifying this results in $\|\nu\|_1 = \|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \le 3s_0\lambda_n/\phi_0$ as required.

□

A.2 Bias control: Proof of Theorem 2.8

Recall the decomposition (12) from which we obtain:

$$\begin{aligned}\Delta_n &= B_n(\widehat{\theta}^{\mathsf{L}} - \theta_0), \\ B_n &= \sqrt{n}\Big(I_p - \frac{n_1}{n}M^{(1)}\widehat{\Sigma}^{(1)} - \frac{n_2}{n}M^{(2)}\widehat{\Sigma}^{(2)}\Big), \\ W_n &= \frac{1}{\sqrt{n}}\sum_{i\le n_1} M^{(1)}x_i\varepsilon_i + \frac{1}{\sqrt{n}}\sum_{n_1<i\le n} M^{(2)}x_i\varepsilon_i.\end{aligned}$$

By construction $M^{(1)}$ is a function of X_1 and hence is independent of $\varepsilon_1, \dots, \varepsilon_{n_1}$. In addition, $M^{(2)}$ is independent of $\varepsilon_{n_1+1}, \dots, \varepsilon_n$. Therefore $\mathbb{E}\{W_n\} = 0$ as required. The key is to show the bound on $\|\Delta_n\|_\infty$. We start by using Hölder inequality

$$\|\Delta_n\|_\infty \le \|B_n\|_\infty\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1.$$

Since the ℓ_1 error of $\widehat{\theta}^{\mathsf{L}}$ is bounded in Theorem 2.2, we need only to show the bound on B_n. For this, we use triangle inequality and that $M^{(1)}$ and $M^{(2)}$ are feasible for the online debiasing program:

$$\begin{aligned}\|B_n\|_\infty &= \sqrt{n}\Big\|\frac{n_1}{n}(I_p - M^{(1)}\widehat{\Sigma}^{(1)}) + \frac{n_2}{n}(I_p - M^{(2)}\widehat{\Sigma}^{(2)})\Big\|_\infty \\ &\le \sqrt{n}\Big(\frac{n_1}{n}\|I_p - M^{(1)}\widehat{\Sigma}^{(1)}\|_\infty + \frac{n_2}{n}\|I_p - M^{(2)}\widehat{\Sigma}(2)\|_\infty\Big) \\ &\le \sqrt{n}\Big(\frac{n_1\mu_1}{n} + \frac{n_2\mu_2}{n}\Big).\end{aligned}$$

The following lemma shows that, with high probability, we can take μ_1, μ_2 so that the resulting bound on B_n is of order $\sqrt{\log p}$.

Lemma A.5. *Denote by $\Omega = (\mathbb{E}\{xx^{\mathsf{T}}\})^{-1}$ and $\Omega^{(2)}(\widehat{\theta}) = (\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\})^{-1}$ be the population precision matrices for the first and second batches. Suppose that $n_1 \wedge n_2 \geq 2\Lambda_0/\kappa^2 \log p$. Then, with probability at least $1 - p^{-4}$*

$$\|I_p - \Omega\widehat{\Sigma}^{(1)}\|_\infty \leq 15\kappa\Lambda_0^{-1/2}\sqrt{\frac{\log p}{n_1}},$$
$$\|I_p - \Omega^{(2)}\widehat{\Sigma}^{(2)}\|_\infty \leq 15\kappa\Lambda_0^{-1/2}\sqrt{\frac{\log p}{n_2}}.$$

In particular, with the same probability, the online debiasing program (10) is feasible with $\mu_\ell = 15\kappa^2\Lambda_0^{-1}\sqrt{(\log p)/n_\ell} < 1/2$.

It follows from the lemma, Theorem 2.2 and the previous display that, with probability at least $1 - 2p^{-3}$

$$\begin{aligned}
\|\Delta_n\|_\infty &\leq \|B_n\|_\infty \|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \\
&\leq 15\kappa\Lambda_0^{-1/2}\sqrt{n}\Big(\frac{n_1}{n}\sqrt{\frac{\log p}{n_1}} + \frac{n_2}{n}\sqrt{\frac{\log p}{n_2}}\Big) \cdot 120\kappa\sigma\phi_0^{-1}s_0\sqrt{\frac{\log p}{n}}, \\
&\leq 2000\frac{\kappa^2\sigma}{\sqrt{\Lambda_0}\phi_0}\frac{s_0\log p}{n}(\sqrt{n_1} + \sqrt{n_2}) \\
&\leq 4000\frac{\kappa^2\sigma}{\sqrt{\Lambda_0}\phi_0}\frac{s_0\log p}{\sqrt{n}}.
\end{aligned} \tag{68}$$

This implies the first claim that, with probability rapidly converging to one, $\Delta_n/\sqrt{n}$ is of order $s_0 \log p/n$.

We should also expect $\|\mathbb{E}\{\widehat{\theta}^{\mathsf{on}} - \theta_0\}\|_\infty$ to be of the same order. To prove this, however, we need some control (if only rough) on $\widehat{\theta}^{\mathsf{on}}$ in the exceptional case when the LASSO error is large or the online debiasing program is infeasible. Let G_1 denote the good event of Lemma A.4 and G_2 denote the good event of Theorem 2.2 as below:

$$G_1 = \Big\{ \text{For } \ell = 1, 2 : \|I_p - \Omega^{(\ell)}\widehat{\Sigma}^{(\ell)}\|_\infty \leq 15\kappa\Lambda_0^{-1/2}\sqrt{\frac{\log p}{n_\ell}} \Big\},$$
$$G_2 = \Big\{\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \leq \frac{3s_0\lambda_n}{\phi_0} = \frac{120\kappa\sigma}{\phi_0}s_0\sqrt{\frac{\log p}{n}}.\Big\}.$$

On the intersection $G = G_1 \cap G_2$, Δ_n satisfies the bound (68). For the complement: we will use the following rough bound on the LASSO error:

Lemma A.6 (Rough bound on LASSO error)**.** *For LASSO estimate $\widehat{\theta}^{\mathsf{L}}$ with regularization λ_n the following bound holds:*

$$\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 \leq \frac{\|\varepsilon\|^2}{2n\lambda_n} + 2\|\theta_0\|_1\,.$$

Now, since W_n is unbiased:

$$\begin{aligned}
\|\mathbb{E}\{\widehat{\theta}^{\sf on} - \theta_0\}\|_\infty &= \Big\|\frac{\mathbb{E}\{\Delta_n\}}{\sqrt{n}}\Big\|_\infty \\
&= \Big\|\frac{\mathbb{E}\{\Delta_n \mathbb{I}(G)\}}{\sqrt{n}}\Big\|_\infty + \Big\|\frac{\mathbb{E}\{\Delta_n \mathbb{I}(G^c)\}}{\sqrt{n}}\Big\|_\infty \\
&\le 4000 \frac{\kappa^2 \sigma}{\sqrt{\Lambda_0}\phi_0} \frac{s_0 \log p}{n} + \mathbb{E}\{\|\widehat{\theta}^{\sf L} - \theta_0\|_1 \mathbb{I}(G^c)\}.
\end{aligned}$$

For the second term, we can use Lemma A.6, Cauchy Schwarz and that $\mathbb{P}\{G^c\} \le 4p^{-3}$ to obtain:

$$\begin{aligned}
\mathbb{E}\{\|\widehat{\theta}^{\sf L} - \theta_0\|_1 \mathbb{I}(G^c)\} &\le \mathbb{E}\Big\{\frac{\|\varepsilon\|^2 \mathbb{I}(G^c)}{2n\lambda_n} + 2\|\theta_0\|_1 \mathbb{I}(G^c)\Big\} \\
&\le \frac{\mathbb{E}\{\|\varepsilon\|^4\}^{1/2} \mathbb{P}(G^c)^{1/2}}{2n\lambda_n} + 2\|\theta_0\|_1 \mathbb{P}\{G^c\} \\
&\le \frac{\sqrt{3}\sigma^2}{\sqrt{n} p^{1.5} \lambda_n} + 8\|\theta_0\|_1 p^{-3} \le 10c \frac{s_0 \log p}{n},
\end{aligned}$$

for n, p large enough . This implies the claim on the bias.

It remains only to prove the intermediate Lemmas A.5 and A.6.

Proof of Lemma A.5. We prove the claim for the second batch, and in the rest of the proof, we assume that all probabilities and expectations are conditional on the first batch (in particular, the intermediate estimate $\widehat{\theta}^1$). The (a, b) entry of $I_p - \Omega^{(2)}\widehat{\Sigma}^{(2)}$ reads

$$\begin{aligned}
(I_p - \Omega^{(2)}\widehat{\Sigma}^{(2)})_{a,b} &= \mathbb{I}(a = b) - \langle \Omega^{(2)} e_a, \widehat{\Sigma}^{(2)} e_b \rangle \\
&= \frac{1}{n_2} \sum_{i > n_1} \mathbb{I}(a = b) - \langle e_a, \Omega^{(2)} x_i \rangle x_{ib}.
\end{aligned}$$

Now, $\mathbb{E}\{\langle e_a, \Omega^{(2)} x_i \rangle x_{i,b} \rangle\} = \mathbb{I}(a = b)$ and $\langle e_a, \Omega^{(2)} x_i \rangle$ is $(\|\Omega^{(2)}\|_2 \kappa)$-subgaussian. Since $\Sigma^{(2)} \succcurlyeq \Lambda_0 I_p$, we have that $\|\Omega^{(2)}\|_2 \le {\Lambda_0}^{-1}$. This observation, coupled with Lemma D.4, yields $\langle e_a, \Omega^{(2)} x_i \rangle x_{i,b}$ is $2\kappa^2/\Lambda_0$-subexponential. Then we may apply Lemma D.5 for $\varepsilon \le 12\kappa^2/\Lambda_0$ as below:

$$\mathbb{P}\{(I_p - \Omega^{(2)}\widehat{\Sigma}^{(2)})_{a,b} \ge \varepsilon\} \le \exp\Big(-\frac{n_2 \varepsilon^2}{36\kappa^2 {\Lambda_0}^{-1}}\Big).$$

Keeping $\varepsilon = \varepsilon(p, n_2, \kappa, \Lambda_0) = 15\kappa {\Lambda_0}^{-1/2} \sqrt{(\log p)/n_2}$ we obtain:

$$\mathbb{P}\Big\{(I_p - \Omega^{(2)}\widehat{\Sigma}^{(2)})_{a,b} \ge 15\kappa {\Lambda_0}^{-1/2} \sqrt{\frac{\log p}{n_2}}\Big\} \le p^{-6}.$$

Union bounding over the pairs (a, b) yields the claim. The requirement $n_2 \ge 2(\Lambda_0/\kappa^2) \log p$ ensures that the choice ε above satisfies $\varepsilon \le 12\kappa^2/\Lambda_0$.

□

Proof of Lemma A.6. We first bound the size of $\widehat{\theta}^{\mathsf{L}}$. By optimality of $\widehat{\theta}^{\mathsf{L}}$:

$$\begin{aligned}\lambda_n\|\widehat{\theta}^{\mathsf{L}}\|_1 &\le \frac{1}{2n}\|\varepsilon\|_2^2 + \lambda_n\|\theta_0\|_1 - \frac{1}{2n}\|y - X\widehat{\theta}^{\mathsf{L}}\|_2^2 \\ &\le \frac{1}{2n}\|\varepsilon\|_2^2 + \lambda_n\|\theta_0\|_1 .\end{aligned}$$

We now use triangle inequality, the bound above and that $\|\theta_0\|_1 \le p^c$ as in Assumption 2.6:

$$\begin{aligned}\|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1 &\le \|\widehat{\theta}^{\mathsf{L}}\|_1 + \|\theta_0\|_1 \\ &\le \frac{1}{2n\lambda_n}\|\varepsilon\|^2 + 2\|\theta_0\|_1 .\end{aligned}$$

as required. □

A.3 Central limit asymptotics: proofs of Proposition 2.10 and Theorem 2.11

Our approach is to apply a martingale central limit theorem to show that $W_{n,a}$ is approximately normal. An important first step is to show that the conditional covariance $V_{n,a}$ is stable, or approximately constant. Recall that $V_{n,a}$ is defined as

$$V_{n,a} = \sigma^2\Big(\frac{n_1}{n}\langle m_a^{(1)}, \widehat{\Sigma}^{(1)} m_a^{(1)}\rangle + \frac{n_2}{n}\langle m_a^{(2)}, \widehat{\Sigma}^{(2)} m_a^{(2)}\rangle\Big).$$

We define its deterministic equivalent as follows. Consider the function $f : \mathbb{S}^n \to \mathbb{R}$ by:

$$f(\Sigma) = \{\min\langle m, \Sigma m\rangle : \|\Sigma m - e_a\|_\infty \le \mu\,,\ \ \|m\|_1 \le L\}.$$

We begin with two lemmas about the stability of the optimization program used to obtain the online debiasing matrices.

Lemma A.7. *On its domain (and uniformly in μ, e_a), f is L^2-Lipschitz with respect to the $\|\cdot\|_\infty$ norm.*

Proof. For two matrices Σ, Σ' in the domain, let m, m' be the respective optimizers (which exist by compactness of the set $\{m : \|\Sigma m - v\|_\infty \le \mu, \|m\|_1 \le L\}$. We prove that $|f(\Sigma) - f(\Sigma')| \le L^2\|\Sigma - \Sigma'\|_\infty$.

$$\begin{aligned}f(\Sigma) - f(\Sigma') &= \langle \Sigma, mm^{\mathsf{T}}\rangle - \langle \Sigma', m'(m')^{\mathsf{T}}\rangle \\ &\le \langle \Sigma, m'(m')^{\mathsf{T}}\rangle - \langle \Sigma', m'(m')^{\mathsf{T}}\rangle \\ &= \langle (\Sigma - \Sigma')m', m'\rangle \\ &\le \|(\Sigma - \Sigma')m'\|_\infty \|m'\|_1 \\ &\le \|\Sigma - \Sigma'\|_\infty \|m'\|_1^2 \le L^2\|\Sigma - \Sigma'\|_\infty .\end{aligned}$$

Here the first inequality follows from optimality of m and the last two inequalities are Hölder inequality. The reverse inequality $f(\Sigma) - f(\Sigma') \ge -L^2\|\Sigma - \Sigma'\|_\infty$ is proved in the same way. □

Lemma A.8. *We have the following lower bound on the optimization value reached to compute $f(\Sigma)$:*

$$\frac{(1-\mu)^2}{\lambda_{\max}(\Sigma)} \le f(\Sigma) \le \frac{1}{\lambda_{\min}(\Sigma)}.$$

Proof. We first prove the lower bound for $f(\Sigma)$. Suppose m is is an optimizer for the program. Then

$$\|\Sigma m\|_2 \geq \|\Sigma m\|_\infty \geq \|e_a\|_\infty - \mu = 1 - \mu.$$

On the other hand, the value is given by

$$\langle m, \Sigma m\rangle = \langle \Sigma m, \Sigma^{-1}(\Sigma m)\rangle \geq \lambda_{\min}(\Sigma^{-1})\|\Sigma m\|_2^2 = \|\Sigma m\|_2^2 \lambda_{\max}(\Sigma)^{-1}.$$

Combining these gives the lower bound.

For the upper bound, it suffices to consider any feasible point; we choose $m = \Sigma^{-1}e_a$, which is feasible since $\|\Sigma^{-1}\|_1 \leq L$. The value is then $\langle e_a, \Sigma^{-1}e_a\rangle \leq \lambda_{\max}(\Sigma^{-1})$ which gives the upper bound. □

Lemma A.9. *(Stability of $W_{n,a}$) Define $\Sigma^{(2)}(\theta) = \mathbb{E}\{xx^{\mathsf{T}}|\langle x_1, \theta\rangle \geq \varsigma\}$. Then, under Assumptions 2.6 and 2.9*

$$\lim_{n\to\infty}\Big|V_{n,a} - \sigma^2\Big(\frac{n_1 f(\Sigma)}{n} + \frac{n_2 f(\Sigma^2(\theta_0))}{n}\Big)\Big| = 0, \qquad \textit{in probability.}$$

Proof. Using Lemma A.7:

$$\begin{aligned}
&\Big|V_{n,a} - \sigma^2\Big(\frac{n_1}{n}f(\Sigma) + \frac{n_2}{n}f(\Sigma(\theta_0))\Big)\Big| \\
&= \frac{\sigma^2 n_1}{n}(f(\widehat{\Sigma}^{(1)}) - f(\Sigma)) + \frac{\sigma^2 n_2}{n}(f(\widehat{\Sigma}^{(2)} - f(\Sigma(\theta_0)))) \\
&\leq L^2\frac{\sigma^2 n_1}{n}\|\Sigma - \widehat{\Sigma}^{(1)}\|_\infty + L^2\frac{\sigma^2 n_2}{n}\|\Sigma^{(2)}(\theta_0) - \widehat{\Sigma}^{(2)}\|_\infty \\
&\leq L^2\frac{\sigma^2 n_1}{n}\|\Sigma - \widehat{\Sigma}^{(1)}\|_\infty + L^2\frac{\sigma^2 n_2}{n}\big(\|\Sigma^{(2)}(\theta_0) - \Sigma^{(2)}(\widehat{\theta}^1)\|_\infty + \|\Sigma^{(2)}(\widehat{\theta}^1) - \widehat{\Sigma}^{(2)}\|_\infty\big) \\
&\leq \sigma^2 L^2\|\Sigma - \widehat{\Sigma}^{(1)}\|_\infty + \sigma^2 L^2\big(K\|\widehat{\theta}^1 - \theta_0\|_1 + \|\Sigma^{(2)}(\widehat{\theta}^1) - \widehat{\Sigma}^{(2)}\|_\infty\big)\,.
\end{aligned}$$

Using Lemma A.2 the first and third term vanish in probability. It is straightforward to apply Theorem 2.2 to the intermediate estimate $\widehat{\theta}^1$; indeed Assumption 2.9 guarantees that $n_1 \geq cn$ for a universal c. Therefore the intermediate estimate has an error $\|\widehat{\theta}^1 - \theta_0\|_1$ of order $\kappa\sigma\phi_0^{-1}\sqrt{(s_0^2\log p)/n}$ with probability converging to one. In particular, the second term is, with probability converging to one, of order $KL^2\sigma^3\kappa\phi_0^{-1}\sqrt{s_0^2(\log p)/n} = o(1)$ by Assumption 2.9. □

Lemma A.10. *Under Assumptions 2.6 and 2.9, with probability at least $1 - p^{-2}$*

$$\max_i|\langle m_a, x_i\rangle| \leq 10L\kappa\sqrt{\log p},$$

In particular $\lim_{n\to\infty}\max_i|\langle m_a, x_i\rangle| = 0$ *in probability.*

Proof. By Hölder inequality, $\max_i\langle|\langle m_a, x_i\rangle| \leq \max_i\|m_a\|_1\|x_i\|_\infty \leq L\max_i\|x_i\|_\infty$. Therefore, it suffices to prove that, with the required probability $\max_{i,a}|x_{i,a}| \leq 10\kappa\sqrt{\log p}$. Let $u = 10\kappa\sqrt{\log p}$. Since x_i are uniformly κ-subgaussian, we obtain for $q > 0$:

$$\begin{aligned}
\mathbb{P}\{|x_{i,a}| \geq u\} &\leq u^{-q}\mathbb{E}\{|x_{i,a}|^q\} \leq (\sqrt{q}\kappa/u)^q \\
&= \exp\Big(-\frac{q}{2}\log\frac{u^2}{\kappa^2 q}\Big) \leq \exp\Big(-\frac{u^2}{2\kappa^2}\Big) \leq p^{-5}\,,
\end{aligned}$$

where the last line follows by choosing $q = u^2/e\kappa^2$. By union bound over $i \in [n], a \in [p]$, we obtain:

$$\mathbb{P}\{\max_{i,a} |x_{i,a}| \geq u\} \leq \sum_{i,a} \mathbb{P}\{|x_{i,a}| \geq u\} \leq p^{-3},$$

which implies the claim (note that $p \geq n$ as we are focusing on the high-dimensional regime). □

With these in hand we can prove Proposition 2.10 and Theorem 2.11.

Proof of Proposition 2.10. Consider the minimal filtration $\mathfrak{F}_i$ so that

1. For $i < n_1$, $y_1, \dots, y_i$, $x_1, \dots x_{n_1}$ and $\varepsilon_1, \dots, \varepsilon_i$ are measurable with respect to $\mathfrak{F}_i$.
2. For $i \geq n_1$ $y_1, \dots, y_i$, $x_1, \dots, x_n$ and $\varepsilon_1, \dots \varepsilon_i$ are measurable with respect to $\mathfrak{F}_i$.

The martingale W_n (and therefore, its a^{th} coordinate $W_{n,a}$) is adapted to the filtration $\mathfrak{F}_i$. We can now apply the martingale central limit theorem [HH14, Corollary 3.1] to $W_{n,a}$ to obtain the result. From Lemmas A.8 and A.9 we know that $V_{n,a}$ is bounded away from 0, asymptotically. The stability and conditional Lindeberg conditions of [HH14, Corollary 3.1] are verified by Lemmas A.9 and A.10. □

Proof of Theorem 2.11. This is a straightforward corollary of the bias bound of 2.8 and Proposition 2.10. We will show that:

$$\lim_{n\to\infty} \mathbb{P}\Big\{\sqrt{\frac{n}{V_{n,a}}}(\widehat{\theta}^{\mathsf{on}}_a - \theta_{0,a}) \leq x\Big\} \leq \Phi(x).$$

The reverse inequality follows using the same argument.

Fix a $\delta > 0$. We decompose the difference above as:

$$\sqrt{\frac{n}{V_{n,a}}}(\widehat{\theta}^{\mathsf{on}}_a - \theta_{0,a}) = \frac{W_{n,a}}{\sqrt{V_{n,a}}} + \frac{\Delta_{n,a}}{\sqrt{V_{n,a}}}.$$

Therefore,

$$\mathbb{P}\Big\{\sqrt{\frac{n}{V_{n,a}}}(\widehat{\theta}^{\mathsf{on}}_a - \theta_{0,a}) \leq x\Big\} \leq \mathbb{P}\Big\{\frac{W_{n,a}}{\sqrt{V_{n,a}}} \leq x + \delta\Big\} + \mathbb{P}\{|\Delta_{n,a}| \geq \sqrt{V_{n,a}}\delta\}.$$

By Proposition 2.10 the first term converges to $\Phi(x+\delta)$. To see that the second term vanishes, observe first that Lemma A.8 and Lemma A.9, imply that $V_{n,a}$ is bounded away from 0 in probability. Using this:

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\{|\Delta_{n,a}| \geq \sqrt{V_{n,a}}\delta\} &\leq \lim_{n\to\infty} \mathbb{P}\{\|\Delta_n\|_\infty \geq \sqrt{V_{n,a}}\delta\} \\ &\leq \lim_{n\to\infty} \mathbb{P}\Big\{\|\Delta_n\|_\infty \geq 4000\frac{\kappa^2\sigma}{\sqrt{\Lambda_0}\phi_0}\frac{s_0\log p}{\sqrt{n}}\Big\} = 0\end{aligned}$$

by applying Theorem 2.8 and that for n large enough, $\sqrt{V_{n,a}}\delta$ exceeds the bound on $\|\Delta_n\|_\infty$ used. Since δ is arbitrary, the claim follows. □

A.4 Proofs for Gaussian designs

In this Section we prove that Gaussian designs of Example 2.5 satisfy the requirements of Theorem 2.2 and Theorem 2.8.

The following distributional identity will be important.

Lemma A.11. *Consider the parametrization* $\varsigma = \bar{\varsigma}\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle^{1/2}$*. Then*

$$x|_{\langle x, \widehat{\theta}\rangle \geq \varsigma} \stackrel{\mathrm{d}}{=} \frac{\Sigma\widehat{\theta}}{\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle^{1/2}}\xi_1 + \Big(\Sigma - \frac{\Sigma\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\Sigma}{\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle}\Big)^{1/2}\xi_2,$$

where ξ_1, ξ_2 *are independent,* $\xi_2 \sim \mathsf{N}(0, I_p)$ *and* ξ_1 *has the density:*

$$\frac{\mathrm{d}\mathbb{P}_{\xi_1}}{\mathrm{d}u}(u) = \frac{1}{\sqrt{2\pi}\Phi(-\bar{\varsigma})}\exp(-u^2/2)\mathbb{I}(u \geq \bar{\varsigma}).$$

Proof. This follows from the distribution of $x|\langle x, \widehat{\theta}\rangle$ being $\mathsf{N}(\mu', \Sigma')$ with

$$\mu' = \frac{\Sigma\widehat{\theta}}{\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle}\langle x, \widehat{\theta}\rangle, \quad \Sigma' = \Sigma - \frac{\Sigma\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\Sigma}{\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle}.$$

□

The following lemma shows that they satisfy compatibility.

Lemma A.12. *Let* $\mathbb{P}_x = \mathsf{N}(0, \Sigma)$ *for a positive definite covariance* Σ*. Then, for any vector* $\widehat{\theta}$ *and subset* $S \subseteq [p]$*, the second moments* $\mathbb{E}\{xx^{\mathsf{T}}\}$ *and* $\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\}$ *are* (ϕ_0, S)*-compatible with* $\phi_0 = \lambda_{\min}(\Sigma)/16$*.*

Proof. Fix an $S \subseteq [p]$. We prove that $\Sigma = \mathbb{E}\{x_1 x_1^{\mathsf{T}}\}$ is (ϕ_0, S)-compatible with $\phi_0 = \lambda_{\min}(\Sigma)/16$. Note that, for any v satisfying $\|v_{S^c}\|_1 \leq 3\|v_S\|$, its ℓ_1 norm satisfies $\|v\|_1 \leq 4\|v_S\|_1$. Further $\Sigma \succcurlyeq \lambda_{\min}(\Sigma)I_p$ implies:

$$\frac{|S|\langle v, \Sigma v\rangle}{\|v\|_1^2} \geq \lambda_{\min}(\Sigma)\frac{|S|\|v\|^2}{\|v\|_1^2} \geq \lambda_{\min}(\Sigma)\frac{|S|\|v_S\|^2}{16\|v_S\|_1^2} \geq \frac{\lambda_{\min}(\Sigma)}{16}.$$

For $\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\}$, we use Lemma A.11 to obtain

$$\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\} = \Sigma + (\mathbb{E}\{\xi_1^2\} - 1)\frac{\Sigma\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\Sigma}{\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle},$$

where ξ_1 is as in Lemma A.11. Since $\mathbb{E}\{\xi_1^2\} = 1 + \bar{\varsigma}\varphi(\bar{\varsigma})/\Phi(-\bar{\varsigma}) \geq 1 + \bar{\varsigma}^2$ whenever $\bar{\varsigma} \geq 0$:

$$\mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\} \geq \Sigma + \bar{\varsigma}^2\frac{\Sigma\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\Sigma}{\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle} \succcurlyeq \lambda_{\min}(\Sigma)I_p\,.$$

The rest of the proof is as for Σ.

□

Lemma A.13. *Let $\mathbb{P}_x = \mathsf{N}(0, \Sigma)$ for a positive definite covariance Σ. Then, for any vector $\widehat{\theta}$ and subset $S \subseteq [p]$, the random vectors x and $x|_{\langle x,\widehat{\theta}\rangle \geq \varsigma}$ are κ-subgaussian with $\kappa = 2\lambda_{\max}^{1/2}(\Sigma)$.*

Proof. By definition, $\langle x, v\rangle \sim \mathsf{N}(0, v^{\mathsf{T}}\Sigma v)$ is $\sqrt{v^{\mathsf{T}}\Sigma v}$-subGaussian. Optimizing over all unit vectors v, x is $\lambda_{\max}^{1/2}(\Sigma)$-subgaussian.

For $x|_{\langle x,\widehat{\theta}\rangle \geq \varsigma}$, we use the decomposition of Lemma A.11:

$$x|_{\langle x,\widehat{\theta}\rangle \geq \varsigma} \stackrel{d}{=} \frac{\Sigma\widehat{\theta}}{\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle^{1/2}}\xi_1 + \Big(\Sigma - \frac{\Sigma\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\Sigma}{\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle}\Big)^{1/2}\xi_2.$$

Clearly, ξ_2 is 1-subgaussian, which means the second term is $\lambda_{\max}^{1/2}(\Sigma)$-subgaussian. For the first term, we claim that ξ_1 is 1-subgaussian and therefore the first term is $\lambda_{\max}^{1/2}(\Sigma)$-subgaussian. To show this, we start with the moment generating function of ξ_1. Recall that $\bar{\varsigma} = \varsigma/\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle^{1/2}$:

$$\mathbb{E}\{e^{\lambda\xi_1}\} = \int_{\bar{\varsigma}}^{\infty} e^{\lambda u}e^{-u^2/2}\frac{\mathrm{d}u}{\sqrt{2\pi}\Phi(-\bar{\varsigma})} = e^{\lambda^2/2}\frac{\Phi(\lambda - \bar{\varsigma})}{\Phi(-\bar{\varsigma})}.$$

Here φ and Φ are the density and c.d.f. of the standard normal distribution. It follows that:

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\log\mathbb{E}\{e^{\lambda\xi_1}\} &= \frac{1}{2} + \frac{(\lambda-\bar{\varsigma})\varphi(\lambda-\bar{\varsigma})}{\Phi(\lambda-\bar{\varsigma})} - \frac{\varphi(\lambda-\bar{\varsigma})^2}{\Phi(\lambda-\bar{\varsigma})^2} \\ &\leq -\frac{1}{2} + \sup_{\lambda\geq\bar{\varsigma}}\frac{(\lambda-\bar{\varsigma})\varphi(\lambda-\bar{\varsigma})}{\Phi(\lambda-\bar{\varsigma})} \\ &\leq \frac{1}{2} + \sup_{\lambda\geq 0}\frac{\lambda\varphi(\lambda)}{\Phi(\lambda)} < 1\,.\end{aligned}$$

Therefore, by integration, ξ_1 is 1-subgaussian. The claim then follows. □

For Example 2.7, it remains only to show the constraint on the approximate sparsity of the inverse covariance. We show this in the following

Lemma A.14. *Let $\mathbb{P}_x = \mathsf{N}(0, \Sigma)$ and $\widehat{\theta}$ be any vector such that $\|\widehat{\theta}\|_1\|\widehat{\theta}\|_\infty \leq L\lambda_{\min}(\Sigma)\|\widehat{\theta}\|^2/2$ and $\|\Sigma^{-1}\|_1 \leq L/2$. Then, with $\Omega = \mathbb{E}\{xx^{\mathsf{T}}\}^{-1}$ and $\Omega^{(2)}(\widehat{\theta}) = \mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\}^{-1}$:*

$$\|\Omega\|_1 \vee \|\Omega^{(2)}\|_1 \leq L.$$

Proof. By assumption $\|\Omega\|_1 \leq L/2$, so we only require to prove the claim for $\Omega^{(2)} = \mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\}^{-1}$. Using Lemma A.11, we can compute the precision matrix:

$$\begin{aligned}\Omega^{(2)} &= \mathbb{E}\{xx^{\mathsf{T}}|\langle x, \widehat{\theta}\rangle \geq \varsigma\}^{-1} \\ &= \Big(\Sigma + (\mathbb{E}\{\xi_1^2\} - 1)\frac{\Sigma\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\Sigma}{\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle}\Big)^{-1} \\ &= \Omega + (\mathbb{E}\{\xi_1^2\}^{-1} - 1)\frac{\widehat{\theta}\widehat{\theta}^{\mathsf{T}}}{\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle}\,,\end{aligned}$$

where the last step follows by an application of Sherman–Morrison formula. Since $\mathbb{E}\{\xi_1^2\} = 1 + \bar{\varsigma}\varphi(\bar{\varsigma})/\Phi(-\bar{\varsigma})$, where $\bar{\varsigma} = \varsigma/\langle\widehat{\theta}, \Sigma\widehat{\theta}\rangle^{1/2}$ this yields:

$$\Omega^{(2)} = \Omega - \frac{\bar{\varsigma}\varphi(\bar{\varsigma})}{\Phi(-\bar{\varsigma}) + \bar{\varsigma}\varphi(\bar{\varsigma})}\frac{\widehat{\theta}\widehat{\theta}^{\mathsf{T}}}{\langle\widehat{\theta}, \Sigma\widehat{\theta}\rangle}.$$

By triangle inequality, for any $\bar{\varsigma} \geq 0$:

$$\begin{aligned}\|\Omega^{(2)}\|_1 &\leq \|\Omega\|_1 + \frac{\|\widehat{\theta}\widehat{\theta}^{\mathsf{T}}\|_1}{\langle\widehat{\theta}, \Sigma\widehat{\theta}\rangle} \\ &\leq \frac{L}{2} + \frac{\|\widehat{\theta}\|_1\|\widehat{\theta}\|_\infty}{\lambda_{\min}(\Sigma)\|\widehat{\theta}\|^2} \leq L.\end{aligned}$$

□

Next we show that the conditional covariance of x is appropriately Lipschitz.

Lemma A.15. *Suppose $\varsigma = \bar{\varsigma}\langle\theta, \Sigma\theta\rangle^{1/2}$ for a constant $\bar{\varsigma} \geq 0$. Then The conditional covariance function $\Sigma^{(2)}(\theta) = \mathbb{E}\{xx^{\mathsf{T}}|\langle x, \theta\rangle \geq \varsigma\}$ satisfies:*

$$\|\Sigma^{(2)}(\theta') - \Sigma^{(2)}(\theta)\|_\infty \leq K\|\theta' - \theta\|,$$

where $K = \sqrt{8}(1 + \bar{\varsigma}^2)\lambda_{\max}(\Sigma)^{3/2}/\lambda_{\min}(\Sigma)^{1/2}$.

Proof. Using Lemma A.11,

$$\Sigma^{(2)}(\theta) = \Sigma + (\mathbb{E}\{\xi_1^2\} - 1)\frac{\Sigma\theta\theta^{\mathsf{T}}\Sigma}{\langle\theta, \Sigma\theta\rangle}.$$

Let $v = \Sigma^{1/2}\theta/\|\Sigma^{1/2}\theta\|$ and $v' = \Sigma^{1/2}\theta'/\|\Sigma^{1/2}\theta'\|$. With this,

$$\begin{aligned}\|\Sigma^{(2)}(\theta') - \Sigma^{(2)}(\theta)\|_\infty &= (\mathbb{E}\{\xi_1^2\} - 1)\|\Sigma^{1/2}(vv^{\mathsf{T}} - v'v'^{\mathsf{T}})\Sigma^{1/2}\|_\infty \\ &\leq (\mathbb{E}\{\xi_1^2\} - 1)\,\lambda_{\max}(\Sigma)\|vv^{\mathsf{T}} - v'v'^{\mathsf{T}}\|_2 \\ &\leq (\mathbb{E}\{\xi_1^2\} - 1)\lambda_{\max}(\Sigma)\|vv^{\mathsf{T}} - v'v'^{\mathsf{T}}\|_F \\ &\overset{(a)}{\leq} \sqrt{2}(\mathbb{E}\{\xi_1^2\} - 1)\lambda_{\max}(\Sigma)\|v - v'\| \\ &\overset{(b)}{\leq} \frac{\sqrt{8}\lambda_{\max}(\Sigma)^{3/2}}{\lambda_{\min}(\Sigma)^{1/2}}(\mathbb{E}\{\xi_1^2\} - 1)\|\theta - \theta'\| \\ &\overset{(c)}{\leq} \frac{\sqrt{8}\lambda_{\max}(\Sigma)^{3/2}}{\lambda_{\min}(\Sigma)^{1/2}}(\bar{\varsigma}^2 + 1)\|\theta - \theta'\|\,.\end{aligned}$$

Here, (a) follows by noting that for two unit vectors v, v', we have

$$\|vv^{\mathsf{T}} - v'v'^{\mathsf{T}}\|_F^2 = 2 - 2(v^{\mathsf{T}}v')^2 = 2(1 - v^{\mathsf{T}}v')(1 + v^{\mathsf{T}}v') \leq 2\|v - v'\|^2\,.$$

Also, (b) holds using the following chain of triangle inequalities

$$\begin{aligned}
\|v - v'\| &= \Big\| \frac{\Sigma^{1/2}\theta}{\|\Sigma^{1/2}\theta\|} - \frac{\Sigma^{1/2}\theta'}{\|\Sigma^{1/2}\theta'\|} \Big\| \\
&\le \frac{\|\Sigma^{1/2}(\theta - \theta')\|}{\|\Sigma^{1/2}\theta\|} + \|\Sigma^{1/2}\theta'\| \Big| \frac{1}{\|\Sigma^{1/2}\theta\|} - \frac{1}{\|\Sigma^{1/2}\theta'\|} \Big| \\
&\le 2\frac{\|\Sigma^{1/2}(\theta - \theta')\|}{\|\Sigma^{1/2}\theta\|} \le 2\sqrt{\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}}\, \|\theta - \theta'\|
\end{aligned}$$

Finally (c) holds since

$$\mathbb{E}\{\xi_1^1\} - 1 = \bar{\varsigma}\varphi(\bar{\varsigma})/\Phi(-\bar{\varsigma}) \le \bar{\varsigma}^2 + 1\,,$$

using standard tail bound $\varphi(\bar{\varsigma})\frac{\bar{\varsigma}}{\bar{\varsigma}^2+1} \le \Phi(-\bar{\varsigma})$. □

B Proofs of Section 3

B.1 Remarks on the proof of Proposition 3.3

The p-dimensional $\mathsf{VAR}(d)$ model (21) can be represented as a dp-dimensional $\mathsf{VAR}(1)$ model. Recall our notation $x_t = (z_{t+d-1}^{\mathsf{T}}, \dots, z_t^{\mathsf{T}})^{\mathsf{T}}$ (rows of X in (23)). Then (21) can be written as

$$x_t = \widetilde{A} x_{t-1} + \tilde{\zeta}_t\,, \tag{69}$$

with

$$\widetilde{A} = \left(\begin{array}{cccc|c} A_1 & A_2 & \dots & A_{d-1} & A_d \\ \hline & & I_{(d-1)p} & & 0 \end{array}\right), \qquad \tilde{\zeta}_t = \begin{pmatrix} \zeta_{t+d-1} \\ 0 \end{pmatrix}. \tag{70}$$

The reverse characteristic polynomial for the $\mathsf{VAR}(1)$ model reads as $\tilde{\mathcal{A}} = I - \tilde{A}z$.

The proof of RE condition [BM+15, Proposition 4.2] requires upper bounding $\Lambda_{\max}(\Sigma)$, and lower bounding $\Lambda_{\min}(\Sigma)$ which in conjunction with some concentration bounds implies the RE condition for the sample covariance. Specifically, for bounding $\Lambda_{\max}(\Sigma)$, by definition $\Lambda_{\max}(\Sigma) \le 2\pi\mathcal{M}(f_x)$, which along with [BM+15, Equation (4.1)] gives

$$\Lambda_{\max}(\Sigma) \le 2\pi\mathcal{M}(f_x) \le \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\tilde{\mathcal{A}})}\,. \tag{71}$$

The lower bound on $\Lambda_{\min}(\Sigma)$ is shown to be

$$\Lambda_{\min}(\Sigma) \ge \frac{\Lambda_{\min}(\Sigma_\epsilon)}{\mu_{\max}(\mathcal{A})}\,. \tag{72}$$

The analogous claim 3.3 can be proved by following the same lines of the proof of [BM+15, Proposition 4.2] and we omit the details. However, the bound (71) involves $\tilde{\mathcal{A}}$ while bound (72) is in terms of $\mathcal{A}$. Here, we derive an upper bound on $\Lambda_{\max}(\Sigma)$ that is also in terms of $\mathcal{A}$, which results in (27).

We use the notation $\Gamma_x(\ell) = \mathbb{E}[x_t x_{t+\ell}^\mathsf{T}]$ to refer the autocovariance of the dp-dimensional process x_t. Therefore $\Sigma = \Gamma_x(0)$. Likewise, the autocovariance $\Gamma_z(\ell)$ is defined for the p-dimensional process z_t. We represent $\Gamma_x(\ell)$ in terms of d^2 blocks, each of which is a $p \times p$ matrix. The block in position (r, s) is $\Gamma_z(\ell + r - s)$. Now, for a vector $v \in \mathbb{R}^{dp}$ with unit ℓ_2 norm, decompose it as d blocks of p dimensional vectors $v = (v_1^\mathsf{T}, v_2^\mathsf{T}, \dots, v_d^\mathsf{T})^\mathsf{T}$, by which we have

$$v^\mathsf{T} \Gamma_z(\ell) v = \sum_{1 \le r,s \le d} v_r^\mathsf{T} \Gamma_x(\ell + r - s) v_s \,. \tag{73}$$

Since the spectral density $f_z(\theta)$ is the Fourier transform of the autocorrelation function, we have by Equation (73),

$$\begin{aligned}
\langle v, f_z(\theta) v\rangle &= \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \langle v, \Gamma_z(\ell) e^{-j\ell\theta} v\rangle \\
&= \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \sum_{1 \le r,s \le d} \langle v_r, \Gamma_z(\ell + r - s) e^{-j\ell\theta} v_s\rangle \\
&= \sum_{1 \le r,s \le d} \langle v_r, \Big(\frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Gamma_x(\ell + r - s) e^{-j(\ell+r-s)\theta}\Big) v_s e^{j(r-s)\theta}\rangle \\
&= \sum_{1 \le r,s \le d} \langle v_r, f_x(\theta) e^{j(r-s)\theta} v_s\rangle \\
&= V(\theta)^* f_x(\theta) V(\theta),
\end{aligned}$$

with $V(\theta) = \sum_{r=1}^{d} e^{-jr\theta} v_r$. Now, we have:

$$\|V(\theta)\|_2 \le \sum_{r=1}^{d} \|v_r\|_2 \le \Big(d \sum_{r=1}^{d} \|v_r\|_2^2\Big)^{1/2} \le \sqrt{d}.$$

Combining this with the Rayleigh quotient calculation above, yields $\mathcal{M}(f_x) \le d\mathcal{M}(f_z)$. Now, by using [BM+15, Equation (4.1)] for the process z_t, with reverse characteristic polynomial $\mathcal{A}$, we obtain

$$\Lambda_{\max}(\Sigma) \le 2\pi \mathcal{M}(f_x) \le 2\pi d \mathcal{M}(f_z) \le \frac{d \Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})} \,. \tag{74}$$

B.2 Proof of Lemma 3.6

Define

$$\tilde{\mathcal{G}}_n(\tau) \equiv \left\{ \widehat{\Sigma}^{(\ell)} \in \mathbb{R}^{dp \times dp} : \ |\Omega \widehat{\Sigma}^{(\ell)} - I|_\infty < \tau \sqrt{\frac{\log(dp)}{n_\ell}} \right\}.$$

By definition, we have

$$\mu_\ell^{\min}(\widehat{\Sigma}^{(\ell)}) = \min_{M \in \mathbb{R}^{dp \times dp}} |M \widehat{\Sigma}^{(\ell)} - I|_\infty \le |\Omega \widehat{\Sigma}^{(\ell)} - I|_\infty \,, \tag{75}$$

and therefore $\tilde{\mathcal{G}}_n(\tau) \subseteq \mathcal{G}_n$. So it suffices to lower bound the probability of event $\tilde{\mathcal{G}}_n(\tau)$.

By deploying [BM+15, Proposition 2.4] and the bound (74), we have the following in place. There exists a constant $c > 0$, such that for any vectors $u, v \in \mathbb{R}^{dp}$ with $\|u\| \le 1$, $\|v\| \le 1$, and any $\eta \ge 0$,

$$\mathbb{P}\left(|u^{\mathsf{T}}(\widehat{\Sigma}^{(\ell)} - \Sigma)v| > \frac{3d\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})}\eta\right) \le 6\exp\left(-cn_\ell \min\{\eta^2, \eta\}\right) . \tag{76}$$

Now fix $i, j \in [dp]$ and let $u = \frac{\Omega e_i}{\|\Omega e_i\|}$ and $v = e_j$ to get

$$\begin{aligned}
&\mathbb{P}\Big(|(\Omega\widehat{\Sigma}^{(\ell)} - I)_{ij}| \ge \tau\sqrt{\frac{\log(dp)}{n_\ell}}\Big)\\
&= \mathbb{P}\Big(|u^{\mathsf{T}}(\widehat{\Sigma}^{(\ell)} - \Sigma)v| \ge \frac{\tau}{\|\Sigma^{-1}e_i\|}\sqrt{\frac{\log(dp)}{n_\ell}}\Big)\\
&\le \mathbb{P}\Big(|u^{\mathsf{T}}(\widehat{\Sigma}^{(\ell)} - \Sigma)v| \ge \frac{\tau\Lambda_{\min}(\Sigma_\varepsilon)}{\mu_{\max}(\mathcal{A})}\sqrt{\frac{\log(dp)}{n_\ell}}\Big)\\
&\le 6p^{-c_0'}\,, \qquad \text{with } c_0' \equiv c\Big(\frac{\tau}{3d}\Big)^2\Big(\frac{\mu_{\min}(\mathcal{A})}{\mu_{\max}(\mathcal{A})}\Big)^2\Big(\frac{\Lambda_{\min}(\Sigma_\varepsilon)}{\Lambda_{\max}(\Sigma_\varepsilon)}\Big)^2,
\end{aligned}$$

where in the first inequality we used that $\|\Omega e_i\| \le \Lambda_{\min}(\Sigma)^{-1} \le \mu_{\max}(\mathcal{A})/\Lambda_{\min}(\Sigma_\varepsilon)$. In the second inequality, we used that $\eta = (\frac{\tau}{3d})(\frac{\Lambda_{\min}(\Sigma_\epsilon)}{\Lambda_{\max}(\Sigma_\epsilon)})(\frac{\mu_{\min}(\mathcal{A})}{\mu_{\max}(\mathcal{A})})\sqrt{\frac{\log(dp)}{n_\ell}} < 1$ and hence $\min(\eta, \eta^2) = \eta^2$. Then, by union bounding over $i, j \in [dp]$, we get

$$\mathbb{P}(\widehat{\Sigma}^{(\ell)} \in \tilde{\mathcal{G}}_n(\tau)) \ge 1 - 6(dp)^{-c_0'+2}\,,$$

which completes the proof.

B.3 Proof of Theorem 3.7

Starting from the decomposition (37), we have

$$\sqrt{n}(\widehat{\theta}^{\mathsf{on}} - \theta_0) = \Delta_n + W_n\,,$$

with $\Delta_n = B_n(\widehat{\theta}^{\mathsf{L}} - \theta_0)$. As explained below (37), W_n is a martingale with respect to filtration $\mathcal{F}_j = \{\varepsilon_1, \dots, \varepsilon_j\}$, $j \in \mathbb{N}$ and hence $\mathbb{E}(W_n) = 0$.

We also note that $\|\Delta_n\|_\infty \le \|B_n\|_\infty \|\widehat{\theta}^{\mathsf{L}} - \theta_0\|_1$. Our next lemma bounds $\|B_n\|_\infty$.

Lemma B.1. *Suppose that the Optimization problem (29) is feasible for all $i \in [dp]$. Then, there exists a constant $c_1 > 0$, such that*

$$\|B_n\|_\infty \le (\tau + Lc_1)\sqrt{\frac{\log(dp)}{n}}\left(r_0 + \sum_{\ell=1}^{K-1}\frac{r_\ell}{\sqrt{n_{\ell-1}}} + \sum_{\ell=1}^{K-1}\sqrt{r_\ell}\right), \tag{77}$$

with probability at least $1 - 12(dp)^{-c_2}$, where $c_2 = c(c_1\mu_{\min}(A))^2/(3d\Lambda_{\max}(\Sigma_\epsilon))^2 - 2$.

The bound provided in Lemma B.1 holds for general batch sizes $\{r_0, \dots, r_{K-1}\}$. We choose the batch lengths as $r_\ell = \beta^\ell$ for some $\beta > 1$ and $\ell = 1, \dots, K-2$. We also let $r_0 = \sqrt{n}$ and choose r_{K-1} so that the total lengths of batches add up to n (that is $r_0 + r_1 + \dots + r_{K-1} = n$). Therefore, $K = O(\log_\beta(n))$. Following this choice, bound (77) simplifies to the following bound:

$$\|B_n\|_\infty \le (\tau + Lc_1) C_\beta \sqrt{\log p}\,, \tag{78}$$

for some constant $C_\beta > 0$ that depends on the constant β.

Next by combining Proposition 3.4 and Lemma B.1 we obtain

$$\begin{aligned}\|\Delta_n\|_\infty &\le (\tau + Lc_1) C_\beta \sqrt{\log p}\, C\sigma \frac{s_0 \lambda_n}{\alpha} \\ &\le \lambda_0 (\tau + Lc_1) C_\beta \frac{C\sigma}{\alpha} s_0 \sqrt{\log p} \sqrt{\frac{\log(dp)}{n}} \le C_0 \sigma s_0 \frac{\log(dp)}{\sqrt{n}}\,,\end{aligned} \tag{79}$$

with probability at least $1 - 12p^{-c_2} - \exp(-c\log(dp^2)) - \exp(-cn/(1 \vee \omega^2))$, where

$$c_2 = c\Big(\frac{c_1 \mu_{\min}(A)}{3d\Lambda_{\max}(\Sigma_\varepsilon)}\Big)^2, \quad \omega = \frac{d\Lambda_{\max}(\Sigma_\varepsilon)\mu_{\max}(\mathcal{A})}{(\Lambda_{\min}(\Sigma_\varepsilon)\mu_{\min}(\mathcal{A}))}\,,$$

and constant C is given in the statement of Proposition 3.4. In the last step we absorbed all the constants in $C_0 = C_0(\alpha, \lambda_0, a, L)$.

Next note that

$$\begin{aligned}\|\mathbb{E}\{\widehat{\theta}^{\mathsf{on}} - \theta_0\}\|_\infty &= \frac{1}{\sqrt{n}} \|\mathbb{E}\{\Delta_n\}\|_\infty \\ &\le \frac{1}{\sqrt{n}} \mathbb{E}\{\|\Delta_n\|_\infty\} \\ &= \frac{1}{\sqrt{n}} \int_0^\infty \mathbb{P}\{\|\Delta_n\|_\infty \ge u\} \mathrm{d}u \le 10 C\sigma s_0 \frac{\log(dp)}{n}\,,\end{aligned} \tag{80}$$

by using the tail bound (79).

B.3.1 Proof of Lemma B.1

Fix $a \in [dp]$ and define $B_{n,a} \equiv \sqrt{n} e_a - \frac{1}{\sqrt{n}} \sum_{\ell=0}^{K-2} r_{\ell+1} R^{(\ell+1)} m_a^\ell$. We then have

$$B_{n,a} = \sqrt{n} e_a - \frac{1}{\sqrt{n}} \sum_{\ell=0}^{K-2} r_{\ell+1} R^{(\ell+1)} m_a^\ell = \frac{r_0}{\sqrt{n}} e_a + \sum_{\ell=0}^{K-2} \frac{r_{\ell+1}}{\sqrt{n}} \Big(e_a - R^{(\ell+1)} m_a^\ell\Big)\,, \tag{81}$$

where we used that $\sum_{\ell=0}^{K-1} r_\ell = n$. To bound $|B_{n,a}|$, we go through the following steps:

1. By the construction of decorrelating vectors m_a^ℓ as in optimization (29), we have

$$\|\widehat{\Sigma}^{(\ell)} m_a^\ell - e_a\|_\infty \le \mu_\ell\,, \quad \ell = 0, \dots, K-1\,. \tag{82}$$

2. We write

$$
\begin{aligned}
\|\Sigma m_a^\ell - e_a\|_\infty &\le \|\widehat{\Sigma}^{(\ell)} m_a^\ell - e_a\|_\infty + \|(\widehat{\Sigma}^{(\ell)} - \Sigma) m_a^\ell\|_\infty \\
&\le \mu_\ell + \left|\widehat{\Sigma}^{(\ell)} - \Sigma\right|_\infty \times \|m_a^\ell\|_1 \\
&\le \mu_\ell + Lc_1\sqrt{\frac{\log(dp)}{n_\ell}}\,, \qquad (83)
\end{aligned}
$$

where the first step follows from triangle inequality; the second one holds by (82); the third inequality follows from the constraint $\|m_a^\ell\|_1 \le L$ in optimization (29) along with Equation (76). Specifically, we apply Equation (76) with $v = e_i$, $u = e_j$ and union bound over $i, j \in [dp]$, from which we obtain that the last step above holds with probability at least $1 - 6(dp)^{-c_2}$, with $c_2 = c(c_1 \mu_{\min}(A))^2/(3d\Lambda_{\max}(\Sigma_\epsilon))^2 - 2$.

3. By a similar argument as in (83), we have $|R^{(\ell+1)} - \Sigma|_\infty \le c_1\sqrt{(\log p)/r_{\ell+1}}$, with probability at least $1 - 12(dp)^{-c_2}$. Therefore,

$$
\begin{aligned}
\|e_a - R^{(\ell+1)} m_a^\ell\|_\infty &\le \|e_a - \Sigma m_a^\ell\|_\infty + \|(\Sigma - R^{(\ell+1)}) m_a^\ell\|_\infty \\
&\le \mu_\ell + Lc_1\sqrt{\frac{\log(dp)}{n_\ell}} + \left|\Sigma - R^{(\ell+1)}\right|_\infty \times \|m_a^\ell\|_1 \\
&\le \mu_\ell + Lc_1\sqrt{\frac{\log(dp)}{n_\ell}} + Lc_1\sqrt{\frac{\log(dp)}{r_{\ell+1}}}\,. \qquad (84)
\end{aligned}
$$

Using (84) in (81), we obtain

$$
\begin{aligned}
\|B_{n,a}\|_\infty &\le \frac{r_0}{\sqrt{n}} + \frac{1}{\sqrt{n}}\sum_{\ell=0}^{K-2} r_{\ell+1}\|e_a - R^{(\ell+1)} m_a^\ell\|_\infty \\
&\le \frac{r_0}{\sqrt{n}} + \frac{1}{\sqrt{n}}\sum_{\ell=0}^{K-2} r_{\ell+1}\left(\mu_\ell + Lc_1\sqrt{\frac{\log p}{n_\ell}} + Lc_1\sqrt{\frac{\log p}{r_{\ell+1}}}\right) \\
&\le \frac{r_0}{\sqrt{n}} + \frac{\sqrt{\log p}}{\sqrt{n}}\sum_{\ell=0}^{K-2} r_{\ell+1}\left(\frac{\tau + Lc_1}{\sqrt{n_\ell}} + \frac{Lc_1}{\sqrt{r_{\ell+1}}}\right)\,. \qquad (85)
\end{aligned}
$$

Simplifying the above bound and since it holds for all $i \in [dp]$, we obtain that with probability at least $1 - 12(dp)^{-c_2}$,

$$
\|B_n\|_\infty \le (\tau + Lc_1)\sqrt{\frac{\log p}{n}}\left(r_0 + \sum_{\ell=1}^{K-1}\frac{r_\ell}{\sqrt{n_{\ell-1}}} + \sum_{\ell=1}^{K-1}\sqrt{r_\ell}\right)\,, \qquad (86)
$$

which concludes the proof.

B.4 Proof of Lemma 3.8

We start by proving Claim (43). Let $m_a = \Omega e_a$ be the first column of the inverse (stationary) covariance. Using the fact that $\mathbb{E}\{x_t x_t^{\mathsf{T}}\} = \Sigma$ we have $\langle m_a, \mathbb{E}\{x_t x_t^{\mathsf{T}}\} m_a\rangle = \Omega_{a,a}$, which is claimed to be the dominant term in the conditional variance $V_{n,a}$. Therefore, we decompose the difference as follows:

$$
\begin{aligned}
V_{n,a} - \Omega_{a,a} &= \frac{\sigma^2}{n}\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} \Big[\langle m_a^\ell, x_t\rangle^2 - \Omega_{a,a}\Big] - \frac{r_0\sigma^2}{n}\Omega_{a,a}\\
&= \frac{\sigma^2}{n}\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} \Big[\langle m_a^\ell, x_t\rangle^2 - \langle m_a, \mathbb{E}\{x_t x_t^{\mathsf{T}}\} m_a\rangle\Big] - \frac{r_0\sigma^2}{n}\Omega_{a,a}\\
&= \frac{\sigma^2}{n}\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} [\langle m_a^\ell, x_t\rangle^2 - \langle m_a, x_t\rangle^2]\\
&\quad + \frac{1}{n}\sum_{t=0}^{n-1}\langle m_a, (x_t x_t^{\mathsf{T}} - \mathbb{E}\{x_t x_t^{\mathsf{T}}\}) m_a\rangle - \frac{r_0\sigma^2}{n}\Omega_{a,a}\,. \qquad (87)
\end{aligned}
$$

We teat each of these three terms separately. Write

$$
\begin{aligned}
\Big|\frac{1}{n}\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} [\langle m_a^\ell, x_t\rangle^2 - \langle m_a, x_t\rangle^2]\Big| &= \frac{1}{n}\Big|\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} [\langle m_a^\ell - m_a, x_t\rangle\langle m_a^\ell + m_a, x_t\rangle]\Big|\\
&\le \frac{1}{n}\Big\|\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} \langle m_a^\ell - m_a, x_t\rangle x_t\Big\|_\infty \|m_a^\ell + m_a\|_1\\
&\le \frac{2L}{n}\Big\|\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} \langle m_a^\ell - m_a, x_t\rangle x_t\Big\|_\infty\,. \qquad (88)
\end{aligned}
$$

To bound the last quantity, note that

$$
\begin{aligned}
\frac{1}{n}\Big\|\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} \langle m_a^\ell - m_a, x_t\rangle x_t\Big\|_\infty &\le \Big\|e_a - \frac{1}{n}\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} \langle m_a^\ell, x_t\rangle x_t\Big\|_\infty\\
&\quad + \Big\|e_a - \frac{1}{n}\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} \langle m_a, x_t\rangle x_t\Big\|_\infty\\
&= \Big\|e_a^{\mathsf{T}} - \frac{1}{n}\sum_{\ell=0}^{K-2} r_{\ell+1}(m_a^\ell)^{\mathsf{T}} R^{(\ell+1)}\Big\|_\infty + \Big\|e_a^{\mathsf{T}} - m_a^{\mathsf{T}}\widehat{\Sigma}^{(K-1)}\Big\|_\infty\\
&\le C_1\sqrt{\frac{\log p}{n}} + a\sqrt{\frac{\log p}{n}} = C_2\sqrt{\frac{\log p}{n}}\,, \qquad (89)
\end{aligned}
$$

for some constant $C_1 > 0$ and $C_2 = C_1 + a$. The last inequality follows from the positive events of Lemma B.1 and Lemma 3.6. Combining Equations (88) and (89), we obtain

$$
\Big|\frac{1}{n}\sum_{\ell=0}^{K-2}\sum_{t\in E_{\ell+1}} [\langle m_a^\ell, x_t\rangle^2 - \langle m_a, x_t\rangle^2]\Big| = O_P\Big(L_\Sigma\sqrt{\frac{\log p}{n}}\Big). \qquad (90)
$$

For the second term in (87), we can use Equation 76 with $v = u = m_a/\|m_a\|, \eta = (\log p)/n$ to obtain

$$\begin{aligned}\Big|\frac{1}{n}\sum_{t=0}^{n-1}\langle m_a, (x_t x_t^{\mathsf{T}} - \mathbb{E}\{x_t x_t^{\mathsf{T}}\})m_a\rangle\Big| &= \big|\langle m_a, (\widehat{\Sigma}^{(K-1)} - \Sigma)m_a\rangle\big| \\ &\le \frac{3d\Lambda_{\max}(\Sigma_\varepsilon)}{\mu_{\min}(\mathcal{A})}\|m_a\|^2\sqrt{\frac{\log p}{n}} \\ &\le \frac{3d\Lambda_{\max}(\Sigma_\varepsilon)}{\mu_{\min}(\mathcal{A})\Lambda_{\min}(\Sigma)^2}\sqrt{\frac{\log p}{n}} = O_P\Big(\sqrt{\frac{\log p}{n}}\Big)\,, \end{aligned} \tag{91}$$

where we used that $\|m_a\| = \|\Omega e_a\| \le \Lambda_{\max}(\Omega) = \Lambda_{\min}(\Sigma)^{-1}$. For the third term, we have $r_0 = \sqrt{n}$. Also, $\Omega_{a,a} \le \|\Omega e_a\|_1 \le L_\Sigma$. Therefore, this term is $O(L_\Sigma/\sqrt{n})$. Combining this bound with (90) and (91) in Equation (87) we get the Claim (43).

We next prove Claim (44). Note that $|\varepsilon_t| = |\zeta_{t+d}|$ is bounded with $\sigma\sqrt{2\log(n)}$, with high probability for $t \in [n]$, by tail bound for Gaussian variables. In addition, $\max_\ell|\langle m_a^\ell, x_t\rangle| \le \|m_a^\ell\|_1\|x_t\|_\infty \le L\|x_t\|_\infty \le L_\Sigma|X|_\infty$. Note that variance of each entry $x_{t,i}$ is bounded by D_Σ. Hence, by tail bound for Gaussian variables and union bounding we have

$$\mathbb{P}\left(|X|_\infty < \sqrt{2D_\Sigma\log(dpn)}\right) \ge 1 - (pdn)^{-2}\,, \tag{92}$$

Putting these bounds together we get

$$\begin{aligned}&\max\Big\{\frac{1}{\sqrt{n}}|\langle m_a^\ell, x_t\rangle\varepsilon_t| \,:\, \ell\in[K-2],\, t\in[n-1]\Big\} \\ &\le \frac{1}{\sqrt{n}}L_\Sigma\sqrt{2D_\Sigma\log(dpn)}\sqrt{2\log(n)}\sigma\sqrt{2\log(n)} \\ &\le \sigma L_\Sigma\sqrt{D_\Sigma}\left(8\frac{\log^3(dpn)}{n}\right)^{1/2} = o(1)\,,\end{aligned}$$

using Assumption 3.5 (2).

B.5 Proof of Proposition 3.10

We prove that for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty}\sup_{\|\theta_0\|_0\le s_0}\mathbb{P}\Big\{\frac{\sqrt{n}(\widehat{\theta}_a^{\mathsf{on}} - \theta_{0,a})}{\sqrt{V_{n,a}}} \le x\Big\} \le \Phi(x)\,. \tag{93}$$

We can obtain a matching lower bound by a similar argument which implies the result.

Invoking the decomposition (41) we have

$$\frac{\sqrt{n}(\widehat{\theta}_a^{\mathsf{on}} - \theta_{0,a})}{\sqrt{V_{n,a}}} = \frac{W_n}{\sqrt{V_{n,a}}} + \frac{\Delta_n}{\sqrt{V_{n,a}}}\,.$$

By Corollary 3.9, we have that $\widetilde{W}_n \equiv W_n/\sqrt{V_{n,a}} \to \mathsf{N}(0,1)$ in distribution. Fix an arbitrary $\varepsilon > 0$ and write

$$\mathbb{P}\Big\{\frac{\sqrt{n}(\widehat{\theta}^{\sf on}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} \le x\Big\} = \mathbb{P}\Big\{\widetilde{W}_n + \frac{\Delta_n}{\sqrt{V_{n,a}}} \le x\Big\}$$
$$\le \mathbb{P}\{\widetilde{W}_n \le x+\varepsilon\} + \mathbb{P}\Big\{\frac{|\Delta_a|}{\sqrt{V_{n,a}}} \ge \varepsilon\Big\}$$

By taking the limit and deploying Equation (41), we get

$$\lim_{n\to\infty} \sup_{\|\theta_0\|_0\le s_0} \mathbb{P}\Big\{\frac{\sqrt{n}(\widehat{\theta}^{\sf on}_a - \theta_{0,a})}{\sqrt{V_{n,a}}} \le x\Big\} \le \Phi(x+\varepsilon) + \lim_{n\to\infty} \sup_{\|\theta_0\|_0\le s_0} \mathbb{P}\Big\{\frac{|\Delta_a|}{\sqrt{V_{n,a}}} \ge \varepsilon\Big\} \tag{94}$$

We show that the limit on the right hand side vanishes for any $\varepsilon > 0$. By virtue of Lemma 3.8 (Equation (43)), we have

$$\begin{aligned}
\lim_{n\to\infty} \sup_{\|\theta_0\|_0\le s_0} \mathbb{P}\Big\{\frac{|\Delta_a|}{\sqrt{V_{n,a}}} \ge \varepsilon\Big\} &\le \lim_{n\to\infty} \sup_{\|\theta_0\|_0\le s_0} \mathbb{P}\Big\{\frac{|\Delta_a|}{\sigma\sqrt{\Omega_{a,a}}} \ge \varepsilon\Big\}\\
&\le \lim_{n\to\infty} \sup_{\|\theta_0\|_0\le s_0} \mathbb{P}\Big\{|\Delta_a| \ge \varepsilon\sigma\sqrt{\Omega_{a,a}}\Big\}\\
&\le \lim_{n\to\infty} \Big\{12p^{-c_2} + d^{-c}p^{-2c} + \exp(-cn(1\wedge\omega^{-2}))\Big\} = 0\,.
\end{aligned} \tag{95}$$

Here, in the last inequality we used that $s_0 = o(\sqrt{n}/\log(dp))$ and therefore, for large enough n, we have $C_0 s_0 \log(dp)/\sqrt{n} < \varepsilon\sigma\sqrt{\Omega_{a,a}}$ and hence we can apply bound (42).

Using (95) in bound (94) and since ε was arbitrary, we obtain (93).

C Proofs of Section 6

C.1 Proof of Lemma 6.1

Rewrite the optimization problem (29) as follows:

$$\begin{aligned}
&\text{minimize} \quad m^{\sf T}\widehat{\Sigma}^{(\ell)}m\\
&\text{subject to} \quad \langle z, \widehat{\Sigma}^{(\ell)}m - e_a\rangle \le \mu_\ell, \quad \|m\|_1 \le L, \quad \|z\|_1 = 1\,,
\end{aligned} \tag{96}$$

The Lagrangian is given by

$$\mathcal{L}(m,z,\lambda) = m^{\sf T}\widehat{\Sigma}^{(\ell)}m + \lambda(\langle z, \widehat{\Sigma}^{(\ell)}m - e_a\rangle - \mu_\ell), \quad \|z\|_1 = 1, \quad \|m\|_1 \le L\,, \tag{97}$$

If $\lambda \le 2L$, minimizing Lagrangian over m is equivalent to $\frac{\partial\mathcal{L}}{\partial m} = 0$ and we get $m_* = -\lambda z_*/2$. The dual problem is then given by

$$\begin{aligned}
&\text{maximize} \quad -\frac{\lambda^2}{4} z^{\sf T}\widehat{\Sigma}^{(\ell)}z - \lambda\langle z, e_a\rangle - \lambda\mu_\ell\\
&\text{subject to} \quad \frac{\lambda}{2} \le L, \quad \|z\|_1 = 1\,,
\end{aligned} \tag{98}$$

As $\|z\|_1 = 1$, by introducing $\beta = -\frac{\lambda}{2}z$, we get $\|\beta\|_1 = \frac{\lambda}{2}$. Rewrite the dual optimization problem in terms of β to get

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\beta^\mathsf{T}\widehat{\Sigma}^{(\ell)}\beta - \langle \beta, e_a\rangle + \mu_\ell\|\beta\|_1 \\ \text{subject to} \quad & \|\beta\|_1 \le L\,, \end{aligned} \tag{99}$$

Given β_* as the minimizer of the above optimization problem, from the relation of β and z we realize that $m_* = \beta_*$.

Also note that since optimization (99) is the dual to problem (96), we have that if (96) is feasible then the problem (99) is bounded.

C.2 Proof of Lemma 6.2

By virtue of Proposition 3.3, the sample covariance $\widehat{\Sigma}^{(K-1)}$ satisfies RE condition, $\widehat{\Sigma}^{(K-1)} \sim \text{RE}(\alpha, \tau)$, where

$$\alpha = \frac{\Lambda_{\min}(\Sigma_\varepsilon)}{2\mu_{\max}(\mathcal{A})}\,, \qquad \tau = \alpha(\omega^2 \vee 1)\frac{\log(dp)}{n}\,, \tag{100}$$

and by the sample size condition we have $\tau < 1/(32 s_\Omega)$.

Hereafter, we use the shorthand $m_a^* = \Omega e_a$ and let $\mathcal{L}(m)$ be the objective function in the optimization (65). By optimality of m_a, we have $\mathcal{L}(m_a^*) \le \mathcal{L}(m_a)$. Defining the error vector $\nu \equiv m_a - m_a^*$ and after some simple algebraic calculation we obtain the equivalent inequality

$$\frac{1}{2}\nu^\mathsf{T}\widehat{\Sigma}^{(K-1)}\nu \le \langle \nu, e_a - \widehat{\Sigma}^{(K-1)}m_a^*\rangle + \mu_n(\|m_a^*\|_1 - \|m_a^* + \nu\|_1)\,. \tag{101}$$

In the following we first upper bound the right hand side. By employing Lemma 3.6 (for $\ell = K-1$ and $n_{K-1} = n$), we have that with high probability

$$\langle \nu, e_a - \widehat{\Sigma}^{(K-1)}m_a^*\rangle \le \|\nu\|_1 a\sqrt{\frac{\log(dp)}{n}} = (\|\nu_S\|_1 + \|\nu_{S^c}\|_1)\frac{\mu_n}{2}\,,$$

where $S = \text{supp}(\Omega e_a)$ and hence $|S| \le s_\Omega$. On the other hand,

$$\|m_a + \nu\|_1 - \|m_a^*\|_1 \ge (\|m_{a,S}^*\|_1 - \|\nu_S\|_1) + \|\nu_{S^c}\|_1 - \|m_a^*\|_1 = \|\nu_{S^c}\|_1 - \|\nu_S\|_1\,.$$

Combining these pieces we get that the right-hand side of (101) is upper bounded by

$$(\|\nu_S\|_1 + \|\nu_{S^c}\|_1)\frac{\mu_n}{2} + \mu_n\left(\|\nu_S\|_1 - \|\nu_{S^c}\|_1\right) = \frac{3}{2}\mu_n\|\nu_S\|_1 - \frac{1}{2}\mu_n\|\nu_{S^c}\|_1\,, \tag{102}$$

Given that $\widehat{\Sigma}^{(K-1)} \succeq 0$, the left hand side of (101) is non-negative, which implies that $\|\nu_{S^c}\|_1 \le 3\|\nu_S\|_1$ and hence

$$\|\nu\|_1 \le 4\|\nu_S\|_1 \le 4\sqrt{s_\Omega}\|\nu_S\|_2 \le 4\sqrt{s_\Omega}\|\nu\|_2\,. \tag{103}$$

Next by using the RE condition for $\widehat{\Sigma}^{(K-1)}$ we write

$$\nu^\mathsf{T}\widehat{\Sigma}^{(K-1)}\nu \ge \alpha\|\nu\|_2^2 - \alpha\tau\|\nu\|_1^2 \ge \alpha(1 - 16 s_\Omega\tau)\|\nu\|_2^2 \ge \frac{\alpha}{2}\|\nu\|_2^2\,, \tag{104}$$

where we used $\tau \le 1/(32 s_\Omega)$ in the final step.

Putting (101), (102) and (104) together, we obtain

$$\frac{\alpha}{4}\|\nu\|_2^2 \le \frac{3}{2}\mu_n\|\nu_S\|_1 \le 6\sqrt{s_\Omega}\mu_n\|\nu\|_2\,.$$

Simplifying the bound and using equation 103, we get

$$\|\nu\|_2 \le \frac{24}{\alpha}\sqrt{s_\Omega}\mu_n\,,$$
$$\|\nu\|_1 \le \frac{96}{\alpha}s_\Omega\mu_n\,,$$

which completes the proof.

C.3 Proof of Theorem 6.3

Continuing from the decomposition (63) we have

$$\sqrt{n}(\widehat{\theta}^{\mathsf{off}} - \theta_0) = \Delta_1 + \Delta_2 + Z\,, \tag{105}$$

with $Z = \Omega X^{\mathsf{T}}\varepsilon/\sqrt{n}$. By using Lemma 3.6 (for $\ell = n$) and recalling the choice of $\mu_n = \tau\sqrt{(\log p)/n}$ we have that the following optimization is feasible, with high probability:

$$\begin{aligned} &\text{minimize} \quad m^{\mathsf{T}}\widehat{\Sigma}^{(n)}m \\ &\text{subject to} \quad \|\widehat{\Sigma}^{(n)}m - e_a\|_\infty \le \mu_n\,. \end{aligned}$$

Therefore, optimization (65) (which is shown to be its dual in Lemma (6.1)) has bounded solution. Hence, its solution should satisfy the KKT condition which reads as

$$\widehat{\Sigma}^{(n)}m_a - e_a + \mu_n\text{sign}(m_a) = 0\,, \tag{106}$$

which implies $\|\widehat{\Sigma}^{(n)}m_a - e_a\|_\infty \le \mu_n$. Invoking the estimation error bound of Lasso for time series (Proposition 3.4), we bound Δ_1 as

$$\|\Delta_1\|_\infty \le C\sqrt{n}\mu_n s_0\sqrt{\frac{\log p}{n}} = O_P\Big(s_0\frac{\log(dp)}{\sqrt{n}}\Big)\,. \tag{107}$$

We next bound the bias term Δ_2. By virtue of [BM+15, Proposition 3.2] we have the deviation bound $\|X^{\mathsf{T}}\varepsilon\|_\infty/\sqrt{n} = O_P(\sqrt{\log(dp)})$, which in combination with Lemma 6.2 gives us the following bound

$$\|\Delta_2\|_\infty \le \left(\max_{i\in[dp]}\|(M-\Omega)e_i)\|\right)\left(\frac{1}{\sqrt{n}}\|X^{\mathsf{T}}\varepsilon\|_\infty\right) = O_P\Big(s_\Omega\frac{\log(dp)}{\sqrt{n}}\Big)\,. \tag{108}$$

Therefore, letting $\Delta = \Delta_1 + \Delta_2$, we have $\|\Delta\|_\infty = o_P(1)$, by recalling our assumption $s_0 = o(\sqrt{n}/\log(dp))$ and $s_\Omega = o(\sqrt{n}/\log(dp))$.

Our next lemma is analogous to Lemma 3.8 for the covariance of the noise component in the offline debiased estimator, and its proof is deferred to Section C.1.

Lemma C.1. *Assume that* $s_\Omega = o(\sqrt{n}/\log(dp))$ *and* $\Lambda_{\min}(\Sigma_\epsilon)/\mu_{\max}(\mathcal{A}) > c_{\min} > 0$ *for some constant* $c_{min} > 0$*. For* $\mu_n = \tau\sqrt{(\log p)/n}$ *and the decorrelating vectors* m_i *constructed by* (65)*, the following holds. For any fixed sequence of integers* $a(n) \in [dp]$*, we have*

$$m_a^\mathsf{T} \widehat{\Sigma}^{(n)} m_a = \Omega_{a,a} + o_P(1/\sqrt{\log(dp)})\,. \tag{109}$$

We are now ready to prove the theorem statement. We show that

$$\lim_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\left\{ \frac{\sqrt{n}(\widehat{\theta}_a^{\mathsf{off}} - \theta_{0,a})}{\sqrt{V_{n,a}}} \le u \right\} \le \Phi(u)\,. \tag{110}$$

A similar lower bound can be proved analogously. By the decomposition (105) we have

$$\frac{\sqrt{n}(\widehat{\theta}_a^{\mathsf{off}} - \theta_{0,a})}{\sqrt{V_{n,a}}} = \frac{\Delta_a}{\sqrt{V_{n,a}}} + \frac{Z_a}{\sqrt{V_{n,a}}}\,.$$

Define

$$\widetilde{Z}_a \equiv \frac{Z_a}{\sigma\sqrt{\Omega_{a,a}}} = \frac{1}{\sigma\sqrt{n\Omega_{a,a}}}(\Omega X^\mathsf{T} \varepsilon)_a = \frac{1}{\sigma\sqrt{n\Omega_{a,a}}} \sum_{i=1}^n e_a^\mathsf{T} \Omega x_i \varepsilon_i\,.$$

Since ε_i is independent of x_i, the summand $\sum_{i=1}^n e_a^\mathsf{T} \Omega x_i \varepsilon_i$ is a martingale. Furthermore, $\mathbb{E}[(e_a^\mathsf{T} \Omega x_i \varepsilon_i)^2] = \sigma^2 \Omega_{a,a}$. Hence, by a martingale central limit theorem [HH14, Corollary 3.2], we have that $\widetilde{Z}_a \to \mathsf{N}(0,1)$ in distribution. In other words,

$$\lim_{n\to\infty} \mathbb{P}\{\widetilde{Z}_a u\} = \Phi(u)\,. \tag{111}$$

Next, fix $\delta \in (0,1)$ and write

$$\begin{aligned}
\mathbb{P}\left\{ \frac{\sqrt{n}(\widehat{\theta}_a^{\mathsf{off}} - \theta_{0,a})}{\sqrt{V_{n,a}}} \le u \right\} &= \mathbb{P}\left\{ \frac{\sqrt{\Omega_{a,a}}}{\sqrt{V_{n,a}}} \widetilde{Z}_a + \frac{\Delta_a}{\sqrt{V_{n,a}}} \le u \right\} \\
&\le \mathbb{P}\left\{ \frac{\sqrt{\Omega_{a,a}}}{\sqrt{V_{n,a}}} \widetilde{Z}_a \le u + \delta \right\} + \mathbb{P}\left\{ \frac{\Delta_a}{\sqrt{V_{n,a}}} \ge \delta \right\} \\
&\le \mathbb{P}\left\{ \widetilde{Z}_a \le u + 2\delta + \delta|u| \right\} + \mathbb{P}\left\{ \left| \frac{\sqrt{\Omega_{a,a}}}{\sqrt{V_{n,a}}} - 1 \right| \ge \delta \right\} \\
&\quad + \mathbb{P}\left\{ \frac{\Delta_a}{\sqrt{V_{n,a}}} \ge \delta \right\}\,.
\end{aligned}$$

Now by taking the limit of both sides and using (111) and Lemma C.1, we obtain

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\left\{ \frac{\sqrt{n}(\widehat{\theta}_a^{\mathsf{off}} - \theta_{0,a})}{\sqrt{V_{n,a}}} \le u \right\} \le \\
&\qquad \Phi(u + 2\delta + \delta|u|) + \limsup_{n\to\infty} \sup_{\|\theta_0\|_0 \le s_0} \mathbb{P}\left\{ \frac{\Delta_a}{\sqrt{V_{n,a}}} \ge \delta \right\}\,.
\end{aligned} \tag{112}$$

Since $\delta \in (0,1)$ was chosen arbitrarily, it suffices to show that the limit on the right hand side vanishes. To do that, we use Lemma C.1 again to write

$$\begin{aligned}\lim_{n\to\infty} \sup_{\|\theta_0\|_0\le s_0} \mathbb{P}\Big\{\frac{|\Delta_a|}{\sqrt{V_{n,a}}} \ge \delta\Big\} &\le \lim_{n\to\infty} \sup_{\|\theta_0\|_0\le s_0} \mathbb{P}\Big\{\frac{|\Delta_a|}{\sigma\sqrt{(\Omega_{a,a}}} \ge \delta\Big\} \\ &\le \lim_{n\to\infty} \sup_{\|\theta_0\|_0\le s_0} \mathbb{P}\Big\{|\Delta_a| \ge \delta\sigma\sqrt{\Omega_{a,a}}\Big\} = 0\,,\end{aligned}$$

where the last step follows since we showed $\|\Delta\|_\infty = o_P(1)$. The proof is complete.

C.3.1 Proof of Lemma C.1

By invoking bound (72) on minimum eigenvalue of the population covariance, we have

$$\Lambda_{\min}(\Sigma) \ge \frac{\Lambda_{\min}(\Sigma_\epsilon)}{\mu_{\max}(\mathcal{A})} > c_{\min}\,. \tag{113}$$

by our assumption. Therefore, $\Lambda_{\max}(\Omega) = (\Lambda_{\min}(\Sigma))^{-1} < 1/c_{\min}$. Since $\Omega \succeq 0$, we have $|\Omega_{a,b}| \le \sqrt{\Omega_{a,a}\Omega_{b,b}}$ for any two indices $a,b \in [dp]$. Hence, $|\Omega|_\infty \le 1/c_{\min}$. This implies that $\|\Omega e_a\|_1 \le s_\Omega/c_{\min}$. Using this observation along with the bound established in Lemma 6.2, we obtain

$$\|m_a\|_1 \le \|\Omega e_a\| + \|m_a - \Omega e_a\|_1 \le \frac{s_\Omega}{c_{\min}} + \frac{192\tau}{\alpha} s_\Omega \sqrt{\frac{\log(dp)}{n}} = O(s_\Omega)\,. \tag{114}$$

We also have

$$\|m_a - \Omega e_a\|_\infty \le \|m_a - \Omega e_a\|_1 = O\Big(s_\Omega\sqrt{\frac{\log(dp)}{n}}\Big)\,. \tag{115}$$

In addition, by the KKT condition (106) we have

$$\|\widehat{\Sigma}^{(n)} m_a - e_a\|_\infty \le \mu_n\,. \tag{116}$$

Combining bounds (114), (115) and (116), we have

$$\begin{aligned}|m_a^{\mathsf{T}}\widehat{\Sigma}^{(n)} m_a - \Omega_{a,a}| &\le |(m_a^{\mathsf{T}}\widehat{\Sigma}^{(n)} - e_a^{\mathsf{T}})m_a| + |e_a^{\mathsf{T}} m_a - \Omega_{a,a}| \\ &\le \|m_a^{\mathsf{T}}\widehat{\Sigma}^{(n)} - e_a^{\mathsf{T}}\|_\infty \|m_a\|_1 + \|m_a - \Omega e_a\|_\infty \\ &= O\Big(s_\Omega\sqrt{\frac{\log(dp)}{n}}\Big) = o(1/\sqrt{\log(dp)})\,,\end{aligned}$$

which completes the proof.

D Technical preliminaries

Definition D.1. *(Subgaussian norm) The subgaussian norm of a random variable X, denoted by $\|X\|_{\psi_2}$, is defined as*

$$\|X\|_{\psi_2} \equiv \sup_{q\ge 1} q^{-1/2}\mathbb{E}\{|X|^q\}^{1/q}.$$

For a random vector X the subgaussian norm is defined as

$$\|X\|_{\psi_2} \equiv \sup_{\|v\|=1} \|\langle X, v\rangle\|_{\psi_2}.$$

Definition D.2. *(Subexponential norm) The subexponential norm of a random variable X is defined as*

$$\|X\|_{\psi_1} \equiv \sup_{q\geq 1} q^{-1}\mathbb{E}\{|X|^q\}^{1/q}.$$

For a random vector X the subexponential norm is defined by

$$\|X\|_{\psi_1} \equiv \sup_{\|v\|=1} \|\langle X, v\rangle\|_{\psi_1}.$$

Definition D.3. *(Uniformly subgaussian/subexponential sequences) We say a sequence of random variables $\{X_i\}_{i\geq 1}$ adapted to a filtration $\{\mathcal{F}_i\}_{i\geq 0}$ is* uniformly K-subgaussian *if, almost surely:*

$$\sup_{i\geq 1}\sup_{q\geq 1} q^{-1/2}\mathbb{E}\{|X_i|^q|\mathcal{F}_{i-1}\}^{1/q} \leq K.$$

A sequence of random vectors $\{X_i\}_{i\geq 1}$ is uniformly K-subgaussian if, almost surely,

$$\sup_{i\geq 1}\sup_{\|v\|=1}\sup_{q\geq 1} \mathbb{E}\{|\langle X_i, v\rangle|^q|\mathcal{F}_{i-1}\}^{1/q} \leq K.$$

Subexponential sequences are defined analogously, replacing the factor $q^{-1/2}$ with q^{-1} above.

Lemma D.4. *For a pair of random variables X, Y, $\|XY\|_{\psi_1} \leq 2\|X\|_{\psi_2}\|Y\|_{\psi_2}$.*

Proof. By Cauchy Schwarz:

$$\begin{aligned}
\|XY\|_{\psi_1} &= \sup_{q\geq 1} q^{-1}\mathbb{E}\{|XY|^q\}^{1/q} \\
&\leq \sup_{q\geq 1} q^{-1}\mathbb{E}\{|X|^{2q}\}^{1/2q}\mathbb{E}\{|Y|^{2q}\}^{1/2q} \\
&\leq 2\big(\sup_{q\geq 2}(2q)^{-1/2}\mathbb{E}\{|X|^{2q}\}^{1/2q}\big)\cdot\big(\sup_{q\geq 2}(2q)^{-1/2}\mathbb{E}\{|Y|^{2q}\}^{1/2q}\big) \\
&\leq 2\|X\|_{\psi_2}\|Y\|_{\psi_2}.
\end{aligned}$$

□

The following lemma from [Ver12] is a Bernstein-type tail inequality for sub-exponential random variables.

Lemma D.5 ([Ver12, Proposition 5.16])**.** *Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables with* $\max_i\|X_i\|_{\psi_1} \leq K$*. Then for any $\varepsilon \geq 0$:*

$$\mathbb{P}\Big\{\Big|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}\{X_i\}\Big| \geq \varepsilon\Big\} \leq 2\exp\Big\{-\frac{n\varepsilon}{6eK}\min\Big(\frac{\varepsilon}{eK}, 1\Big)\Big\} \tag{117}$$

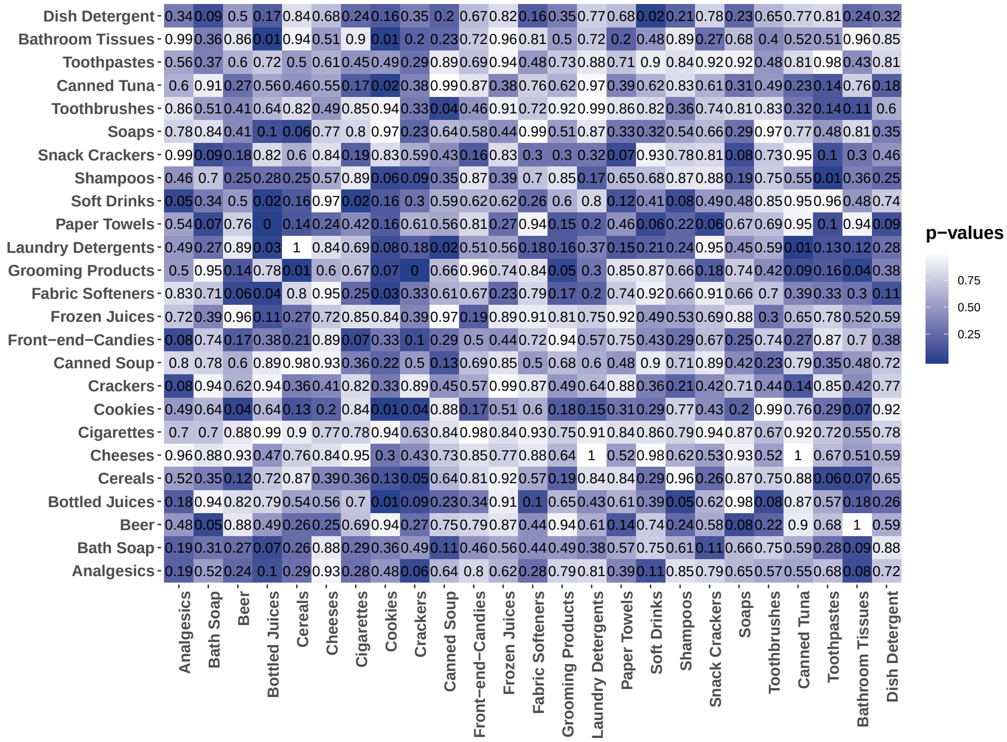
We also use a martingale generalization of [Ver12, Proposition 5.16], whose proof is we omit.

Lemma D.6. *Suppose $(\mathcal{F}_i)_{i\geq 0}$ is a filtration, $X_1, X_2, \ldots, X_n$ is a uniformly K-subexponential sequence of random variables adapted to $(\mathcal{F}_i)_{i\geq 0}$ such that almost surely $\mathbb{E}\{X_i|\mathcal{F}_{i-1}\} = 0$. Then for any $\varepsilon \geq 0$:*

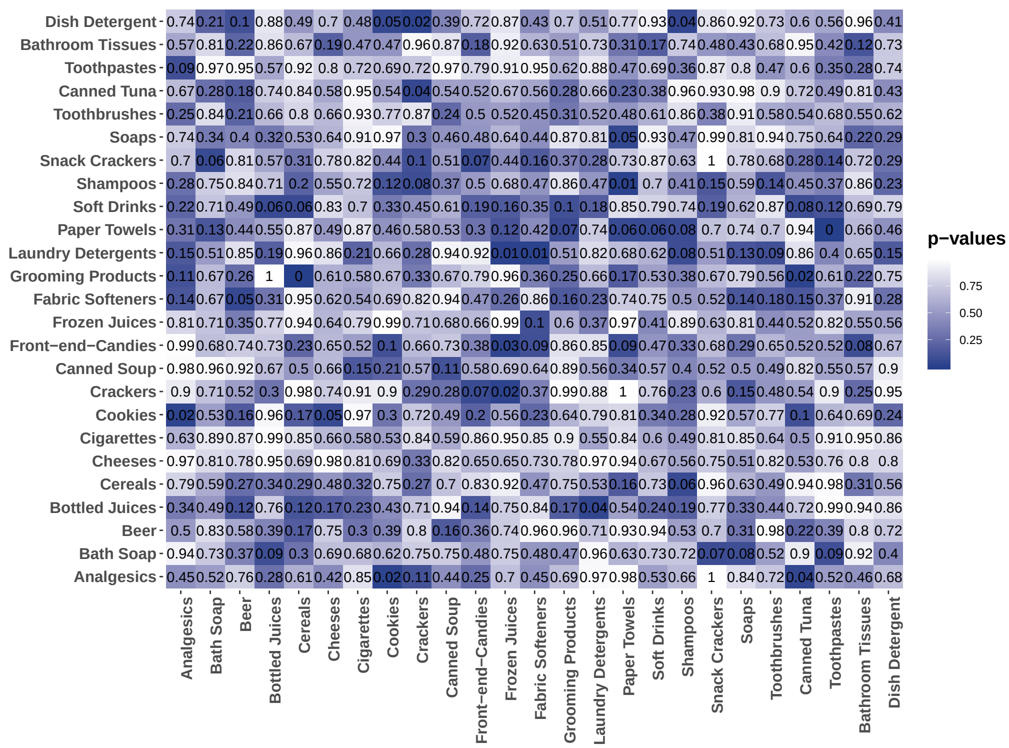
$$\mathbb{P}\Big\{\Big|\frac{1}{n}\sum_{i=1}^n X_i\Big| \geq \varepsilon\Big\} \leq 2\exp\Big\{-\frac{n\varepsilon}{6eK}\min\Big(\frac{\varepsilon}{eK}, 1\Big)\Big\} \tag{118}$$

E Simulation results for the Dominick's data set

In this section we report the p-values obtained by the online debiasing for the cross-category effects. Figures 6, 7, 8 provide the p-values corresponding to the effect of price, sale, and promotions of different categories on the other categories, after one week ($d = 1$) and two weeks ($d = 2$). The darker cells indicate smaller p-values and hence more significant associations.

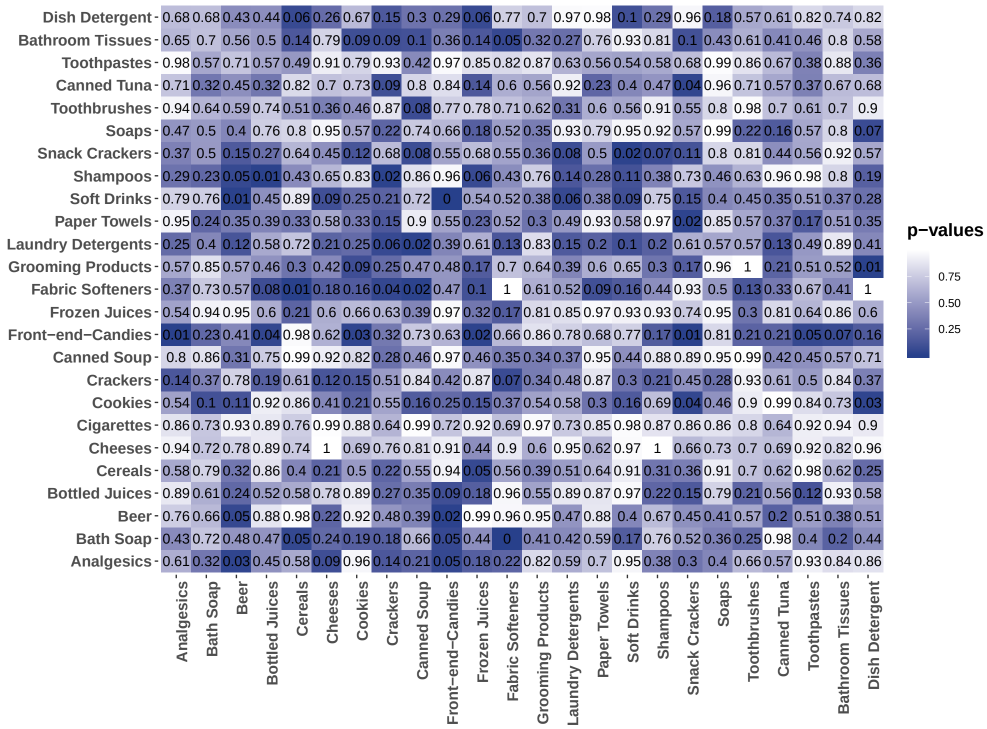


(a) 1-Week effect of sales of $x-$axis categories on sales of $y-$axis categories

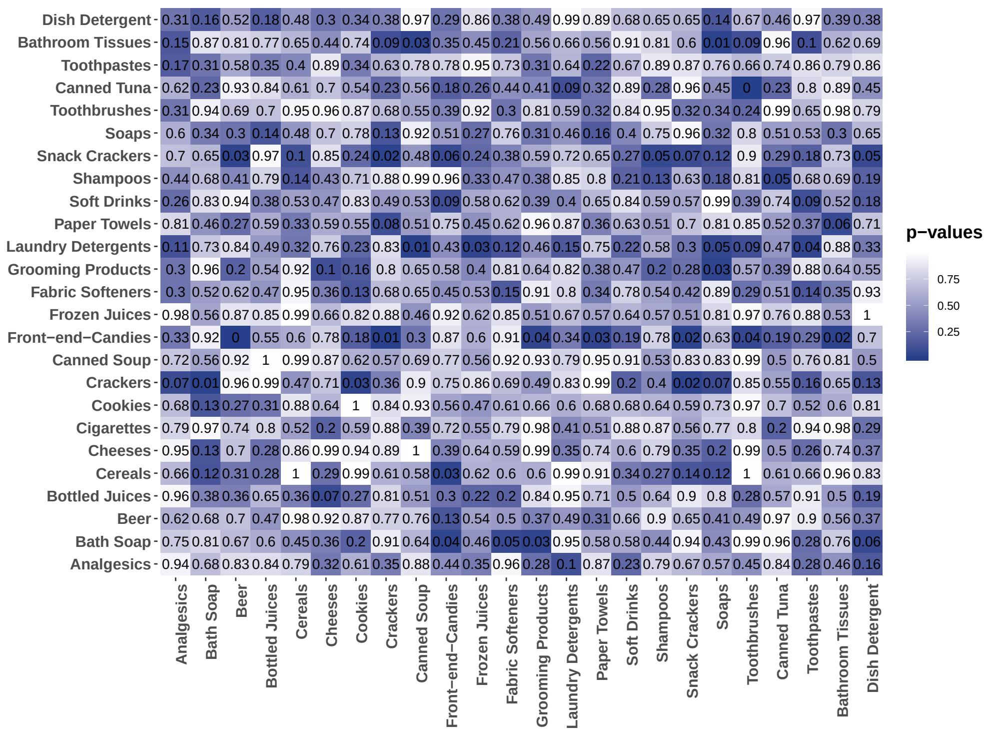


(b) 1-Week effect of prices of $x-$axis categories on sales of $y-$axis categories

Figure 6: Figures 6a, and 6b respectively show p-values for cross-category effects of sales, prices of x-axis categories on sales of $y-$axis categories after one week.

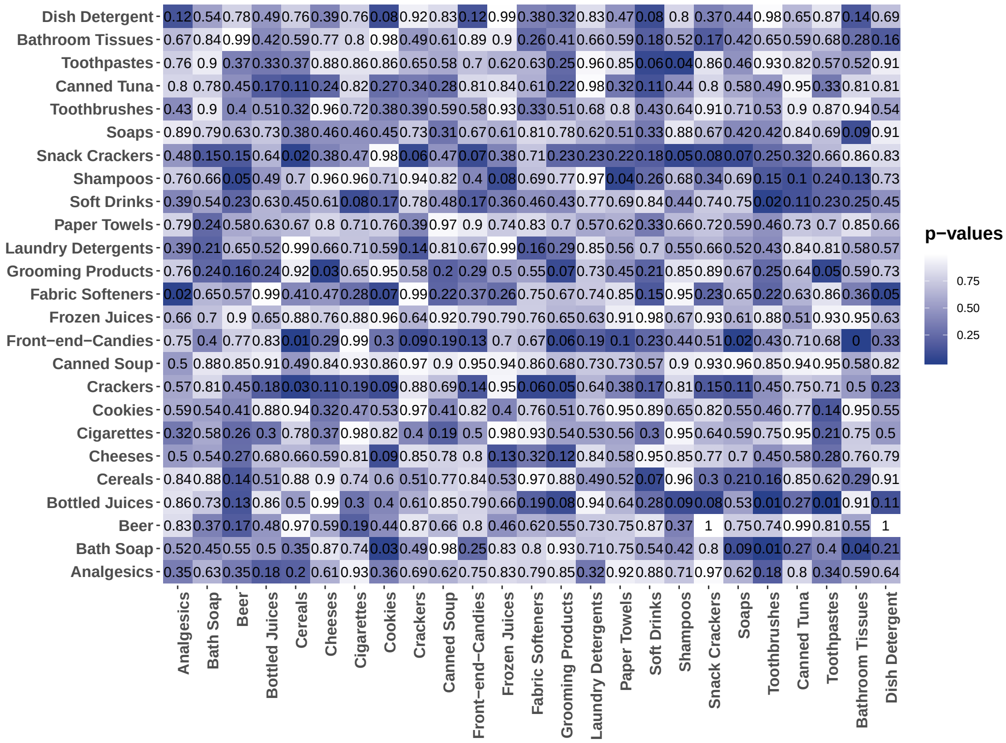


(a) 1-Week effect of promotions of $x-$axis categories on sales of $y-$axis categories

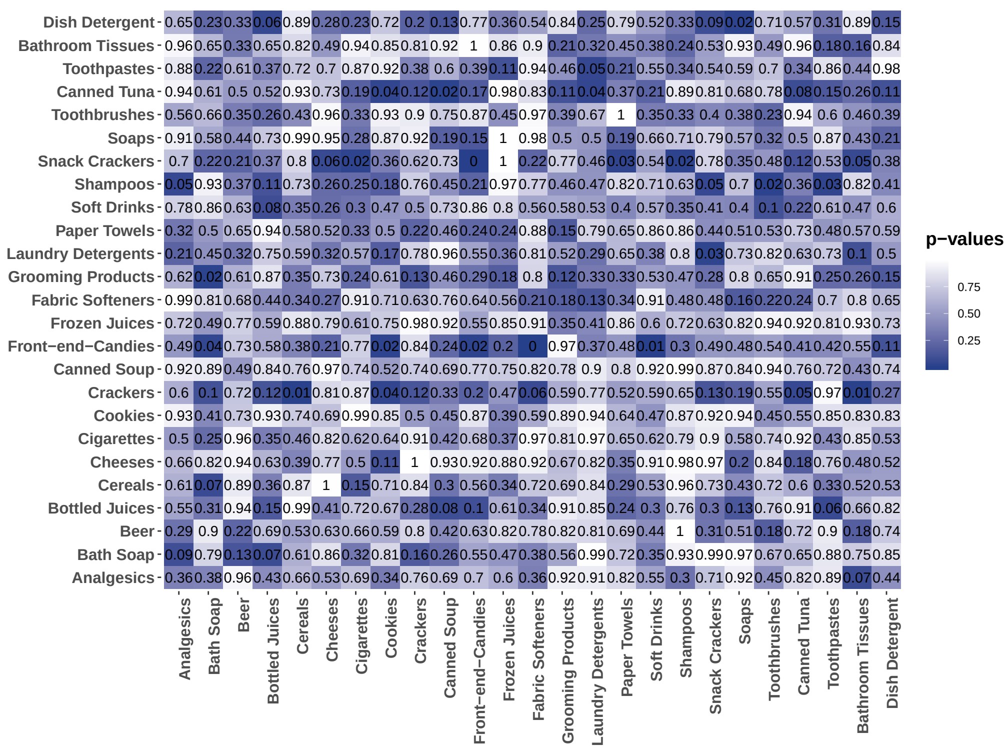


(b) 2-Week effect of promotions of $x-$axis categories on sales of $y-$axis categories

Figure 7: Figures 7a, and 7b show $p-$values for cross-category effects of promotions of $x-$axis categories on sales of $y-$axis categories after one week and two weeks.



(a) 2-Week effect of sales of $x-$axis categories on sales of $y-$axis categories



(b) 2-Week effect of prices of $x-$axis categories on sales of $y-$axis categories

Figure 8: Figures 8a, and 8b respectively show p-values for cross-category effects of sales, and prices of x-axis categories on sales of $y-$axis categories after two weeks.