

A note on the improved sparse Hanson-Wright inequalities

Guozheng Dai[‡], Yiyun He^{*}, Ke Wang[†] and Yizhe Zhu[§]

Abstract

We establish sparse Hanson–Wright inequalities for quadratic forms of sparse α -sub-exponential random vectors with exponent parameter $\alpha \in (0, 2]$. In the regime $0 < \alpha \leq 1$ we derive a refined inequality that is optimal in several canonical models. These results extend the classical Hanson–Wright bound to the sparse setting. Illustrative applications include covariance matrix estimation with incomplete observations, low-rank matrix approximation under the maximum norm with sparsified sketches, and concentration inequalities for sparse α -sub-exponential random vectors.

1 Introduction

The Hanson-Wright inequality quantifies the concentration of a random quadratic form around its mean. Introduced in the 1970s [14], it is now a standard device for covariance estimation, random-matrix analysis, empirical-process bounds and high-dimensional regression.

Let $\xi = (\xi_1, \dots, \xi_n)^\top$ be a random vector with independent centered entries and $A = (a_{ij})_{n \times n}$ be a fixed matrix. The Hanson-Wright inequality is to explore the concentration properties of the following quadratic random variables

$$S_A(\xi) := \xi^\top A \xi = \sum_{i,j} a_{ij} \xi_i \xi_j.$$

A well-known result, proved by Hanson and Wright [14], claims that if ξ_i are independent centered 2-sub-exponential (sub-gaussian) random variables such that $\max_i \|\xi_i\|_{\Psi_2} \leq L$, then for $t \geq 0$ (the following version was provided in [30])

$$\mathbb{P}\{|S_A(\xi) - \mathbb{E}S_A(\xi)| \geq t\} \leq 2 \exp\left(-c \min\left\{\frac{t^2}{L^4 \|A\|_F^2}, \frac{t}{L^2 \|A\|_{l_2 \rightarrow l_2}}\right\}\right), \quad (1.1)$$

[‡]Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, guozhengdai@ust.hk.

^{*}Department of Mathematics, University of California San Diego, La Jolla, CA 92093, USA, yih@ucsd.edu.

[†]Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, kewang@ust.hk.

[§]Department of Mathematics, University of Southern California, 3620 Vermont Avenue Los Angeles, CA 90089, USA, yizhezhu@usc.edu.

where $\|\cdot\|_F$ and $\|\cdot\|_{l_2 \rightarrow l_2}$ are the Frobenius norm and the spectral norm of a matrix respectively. Recall that a random variable η_1 is α -sub-exponential if satisfying

$$\mathbb{P}\{|\eta_1| \geq t\} \leq 2 \exp\left(-\frac{t^\alpha}{K^\alpha}\right), \quad t \geq 0.$$

Its α -sub-exponential norm is defined as

$$\|\eta_1\|_{\Psi_\alpha} := \inf\left\{t > 0 : \mathbb{E} \exp\left(\frac{|\eta_1|^\alpha}{t^\alpha}\right) \leq 2\right\}.$$

There is a lot of work dedicated to extending the classical Hanson-Wright inequality (1.1) to more general cases. For example, assume that ξ_1, \dots, ξ_n are independent centered α -sub-exponential variables and satisfy $\max_i \|\xi_i\|_{\Psi_\alpha} \leq L(\alpha)$. When $1 \leq \alpha \leq 2$, Adamczak and Latała [2] proved for $t \geq 0$ (see also [3])

$$\mathbb{P}\{|S_A(\xi) - \mathbb{E}S_A(\xi)| \geq L(\alpha)^2 t\} \leq 2 \exp(-c(\alpha)f_1(t)), \quad (1.2)$$

where

$$f_1(t) = \min\left\{\left(\frac{t}{\|A\|_F}\right)^2, \frac{t}{\|A\|_{l_2 \rightarrow l_2}}, \left(\frac{t}{\|A\|_{l_{\alpha^*}(l_2)}}\right)^\alpha, \left(\frac{t}{\|A\|_{l_2 \rightarrow l_{\alpha^*}}}\right)^{\frac{2\alpha}{2+\alpha}}, \left(\frac{t}{\|A\|_{l_{\alpha^*} \rightarrow l_{\alpha^*}}}\right)^{\frac{\alpha}{2}}\right\}. \quad (1.3)$$

One can refer to (1.9) below for the definitions of the matrix's norms and $\alpha^* = \alpha/(\alpha - 1)$. When $0 < \alpha \leq 1$, Kolesko and Latała [21] proved for $t \geq 0$ (see also [13])

$$\mathbb{P}\{|S_A(\xi) - \mathbb{E}S_A(\xi)| \geq L(\alpha)^2 t\} \leq 2 \exp(-c(\alpha)f_2(t)), \quad (1.4)$$

where

$$f_2(t) = \min\left\{\left(\frac{t}{\|A\|_F}\right)^2, \frac{t}{\|A\|_{l_2 \rightarrow l_2}}, \left(\frac{t}{\|A\|_{l_2 \rightarrow l_\infty}}\right)^{\frac{2\alpha}{2+\alpha}}, \left(\frac{t}{\|A\|_\infty}\right)^{\frac{\alpha}{2}}\right\}. \quad (1.5)$$

As we have seen above, the Hanson-Wright inequalities (1.2) and (1.4) are quite intricate. By evaluating the family of norms used therein, Sambale [31] obtained a simplified version. This simplified version is not only easily calculable but also sufficient for many applications (see [31] for details). In particular, Sambale proved for $0 < \alpha \leq 2$ and $t \geq 0$

$$\begin{aligned} & \mathbb{P}\{|S_A(\xi) - \mathbb{E}S_A(\xi)| \geq t\} \\ & \leq 2 \exp\left(-c(\alpha) \min\left\{\frac{t^2}{L(\alpha)^4 \|A\|_F^2}, \left(\frac{t}{L(\alpha)^2 \|A\|_{l_2 \rightarrow l_2}}\right)^{\alpha/2}\right\}\right). \end{aligned} \quad (1.6)$$

Motivated by the covariance estimation problem in the matrix variate model, Zhou [36] showed a sparse Hanson-Wright type inequality for sub-gaussian random variables. In particular, assume that $\{\xi_i = \delta_i \cdot \zeta_i, 1 \leq i \leq n\}$ is a sequence of independent random variables, where

$\delta_i \sim \text{Bernoulli}(p_i)$ (that is, taking values 0, 1 with probability $1 - p_i$ and p_i respectively) and ζ_i is a centered sub-gaussian variable independent of δ_i . Zhou proved for $t \geq 0$

$$\mathbb{P}\{|S_A(\xi) - \mathbb{E}S_A(\xi)| \geq t\} \leq 2 \exp\left(-c \min\left\{\frac{t^2}{L^4 \gamma_1(A)}, \frac{t}{L^2 \|A\|_{l_2 \rightarrow l_2}}\right\}\right), \quad (1.7)$$

where $L = \max_i \|\zeta_i\|_{\Psi_2}$ and $\gamma_1(A) = \sum_k a_{kk}^2 p_k + \sum_{i \neq j} a_{ij}^2 p_i p_j$.

Recently, He, Wang and Zhu [15] established sparse Hanson-Wright type inequalities in a more general setting. Specifically, let $\{\xi_i = \delta_i \cdot \zeta_i, 1 \leq i \leq n\}$ be a sequence of independent random variables, where $\delta_i \sim \text{Bernoulli}(p_i)$ and ζ_i is a centered α -sub-exponential variable independent of δ_i , $1 \leq i \leq n$. [15, Corollary 1] showed that for $\alpha > 0$ and $t \geq 0$

$$\begin{aligned} & \mathbb{P}\{|S_A(\xi) - \mathbb{E}S_A(\xi)| \geq t\} \\ & \leq 2 \exp\left(-c(\alpha) \min\left\{\frac{t^2}{L(\alpha)^4 \gamma_1(A)}, \frac{t}{L(\alpha)^2 \gamma_2(A)}, \left(\frac{t}{L(\alpha)^2 \|A\|_\infty}\right)^{\min\{\frac{\alpha}{2}, \frac{1}{2}\}}\right\}\right), \end{aligned} \quad (1.8)$$

where $L(\alpha) = \max_i \|\zeta_i\|_{\Psi_\alpha}$, $\|A\|_\infty = \max_{ij} |a_{ij}|$, and

$$\gamma_2(A) = \max_i \left\{ \sum_{j:j \neq i} |a_{ij}| p_j, \sum_{j:j \neq i} |a_{ji}| p_j, |a_{ii}| \right\}.$$

The work in [15] further extends to sparse versions of Bernstein and Bennett type inequalities.

In addition to the works mentioned above, there are numerous other papers that delve into the Hanson-Wright inequality with sparse structure. We refer interested readers to [8, 12, 16, 28, 32, 33, 35, 37] and the references therein for further exploration.

Notation Given a fixed vector $x = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$, denote by $\|x\|_r = (\sum_i |x_i|^r)^{1/r}$ the l_r norm. Use $\|\xi\|_{L_r} = (\mathbb{E}|\xi|^r)^{1/r}$ as the L_r norm of a random variable ξ . Let $A = (a_{ij})$ be an $m \times n$ matrix. We use the following notations of the matrix norms:

$$\begin{aligned} \|A\|_F &= \sqrt{\sum_{i,j} |a_{ij}|^2}, \quad \|A\|_\infty = \max_{i,j} |a_{ij}|, \\ \|A\|_{l_r(l_2)} &= \left(\sum_{i \leq m} \left(\sum_{j \leq n} |a_{ij}|^2\right)^{r/2}\right)^{1/r}, \\ \|A\|_{l_{r_1} \rightarrow l_{r_2}} &= \sup\left\{\left|\sum_{i,j} a_{ij} x_j y_i\right| : \|x\|_{r_1} \leq 1, \|y\|_{r_2^*} \leq 1\right\}, \end{aligned} \quad (1.9)$$

where $1 \leq r_1, r_2 < \infty$ and $r_2^* = r_2/(r_2 - 1)$. Indeed, $\|A\|_{l_2 \rightarrow l_2}$ is the spectral norm of A and $\|A\|_{l_1 \rightarrow l_\infty} = \|A\|_\infty$.

Let ξ_1, \dots, ξ_n be a sequence of random variables and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a measurable function. Then, $\mathbb{E}_{\xi_{i_1}, \dots, \xi_{i_s}} f(\xi_1, \dots, \xi_n)$ means only taking expectation with respect to random variables $\xi_{i_1}, \dots, \xi_{i_s}$, where $\{i_1, \dots, i_s\} \subset \{1, \dots, n\}$. Denote $\xi \sim \mathcal{W}_s(\alpha)$ when ξ is a symmetric Weibull variable with the scale parameter 1 and the shape parameter α . Specifically, $-\log \mathbb{P}\{|\xi| > x\} = x^\alpha, x \geq 0$.

Unless otherwise stated, denote by C, C_1, c, c_1, \dots the universal constants independent of any parameters and the dimension n . Besides, let $C(\delta), c(\delta), \dots$ be the constants depending only on

the parameter δ . Their values can change from line to line. For convenience, we say $f \lesssim g$ if $f \leq Cg$ for some universal constant C . We write $f \lesssim_\delta g$ if $f \leq C(\delta)g$ for some constant $C(\delta)$. Besides, we say $f \asymp g$ if $f \lesssim g$ and $g \lesssim f$, so does $f \asymp_\delta g$.

Organization of the paper The rest of the paper is organized as follows. In the remaining part of Section 1, we will first present our main results. Then, we will draw comparisons between our results and existing ones to highlight the novelties and improvements. To conclude Section 1, we will pose two open questions that warrant further exploration.

In Section 2, we give three applications of our concentration inequalities. The applications include: (i) covariance matrix estimation from partially observed data; (ii) low-rank matrix approximation with sparsification; and (iii) concentration phenomena for sparse α -sub-exponential random vectors.

Section 3 is dedicated to introducing several key lemmas. These lemmas cover aspects such as the relationship between tails and moments, the decoupling inequality, the contraction principle, and concentration inequalities. They serve as fundamental tools for the subsequent theoretical derivations.

We prove our main results in Section 4 and our applications in Section 5. In the appendix, we offer proofs for Proposition 1.1, Theorem 5, Corollary 1, and Lemma 3.8.

1.1 Main results

This paper concentrates on the sparse Hanson-Wright inequalities in α -sub-exponential random variables, $0 < \alpha \leq 2$. Our first main result in this regime reads as follows:

THEOREM 1. *Assume that $\{\delta_i, 1 \leq i \leq n\}$ and $\{\zeta_i, 1 \leq i \leq n\}$ are two independent sequences of random variables, where $\delta_i \sim \text{Bernoulli}(p_i)$ are independent and ζ_i are independent centered α -sub-exponential random variables. Consider a random vector $\xi = (\xi_1, \dots, \xi_n)$ with $\xi_i = \delta_i \cdot \zeta_i$. Let $A = (a_{ij})_{n \times n}$ be a symmetric fixed matrix. Then for $0 < \alpha \leq 2$ and $t \geq 0$*

$$\begin{aligned} & \mathbb{P}\{|\xi^\top A \xi - \mathbb{E} \xi^\top A \xi| \geq L(\alpha)^2 t\} \\ & \leq 2 \exp \left(-c(\alpha) \min \left\{ \frac{t^2}{\sum_k a_{kk}^2 p_k + \sum_{i \neq j} a_{ij}^2 p_i p_j}, \left(\frac{t}{\|A\|_{l_2 \rightarrow l_2}} \right)^{\alpha/2} \right\} \right), \end{aligned}$$

where $L(\alpha) = \max_i \|\zeta_i\|_{\Psi_\alpha}$ and $c(\alpha)$ is a positive constant depending only on α .

REMARK 1.1 (Comparison of Theorem 1 with Hanson-Wright inequalities in [31, 36]). (i). When $p_1 = \dots = p_n = 1$, i.e., $\delta_i = 1$ for $1 \leq i \leq n$, Theorem 1 recovers Sambale's result [31], see (1.6). (ii). When $\alpha = 2$, Theorem 1 recovers the sparse Hanson-Wright inequality for sparse sub-Gaussian random vectors obtained by Zhou in [36], see (1.7).

Moreover, when only considering the case $0 < \alpha \leq 1$, we have the following more refined result.

THEOREM 2. *In the same setting of Theorem 1, we have for $0 < \alpha \leq 1$ and $t \geq 0$*

$$\mathbb{P}\{|\xi^\top A \xi - \mathbb{E} \xi^\top A \xi| \geq L(\alpha)^2 t\} \leq 2 \exp(-c(\alpha)f(t)),$$

where $L(\alpha) = \max_i \|\zeta_i\|_{\Psi_\alpha}$, $c(\alpha)$ is a positive constant depending only on α , and

$$f(t) = \min \left\{ \frac{t^2}{\sum_k a_{kk}^2 p_k + \sum_{i \neq j} a_{ij}^2 p_i p_j}, \frac{t}{\|(a_{ij} \sqrt{p_i p_j})_{n \times n}\|_{l_2 \rightarrow l_2}}, \left(\frac{t}{\max_i \left(\sum_j a_{ij}^2 p_j \right)^{1/2}} \right)^{\frac{2\alpha}{2+\alpha}}, \left(\frac{t}{\|A\|_\infty} \right)^{\frac{\alpha}{2}} \right\}.$$

REMARK 1.2 (Theorem 2 improves Theorem 1). *Theorem 1 for the case $0 < \alpha \leq 1$ can be derived from Theorem 2. Indeed, a direct integration of the tail probability in Theorem 2 yields that*

$$\begin{aligned} \left\| \xi^\top A \xi - \mathbb{E} \xi^\top A \xi \right\|_{L_r} &\lesssim_\alpha r^{2/\alpha} \|A\|_\infty + r^{1/2+1/\alpha} \max_i \left(\sum_j a_{ij}^2 p_j \right)^{1/2} \\ &\quad + r \|(a_{ij} \sqrt{p_i p_j})\|_{l_2 \rightarrow l_2} + \sqrt{r} \left(\sum_k a_{kk}^2 p_k + \sum_{i \neq j} a_{ij}^2 p_i p_j \right)^{1/2} \\ &\lesssim_\alpha r^{2/\alpha} \|A\|_{l_2 \rightarrow l_2} + \sqrt{r} \left(\sum_k a_{kk}^2 p_k + \sum_{i \neq j} a_{ij}^2 p_i p_j \right)^{1/2}, \end{aligned}$$

where we use the following fact

$$\begin{aligned} \max_i \left(\sum_j a_{ij}^2 p_j \right)^{1/2} &\leq \|A\|_{l_2 \rightarrow l_\infty} \leq \|A\|_{l_2 \rightarrow l_2}, \\ \|(a_{ij} \sqrt{p_i p_j})\|_{l_2 \rightarrow l_2} &\leq \|A\|_{l_2 \rightarrow l_2}, \quad \|A\|_\infty \leq \|A\|_{l_2 \rightarrow l_2}. \end{aligned}$$

Then, one can obtain Theorem 1 by Lemma 3.1 below.

REMARK 1.3 (Comparison of Theorem 2 with Hanson-Wright inequalities in [15, 21]). (i). When $p_1 = \dots = p_2 = 1$, Theorem 2 recovers the Hanson-Wright inequality (1.4) due to Kolesko and Latała in [21].

(ii). Note that for $r \geq 1$,

$$\begin{aligned} r^{\frac{1}{2} + \frac{1}{\alpha}} \max_i \left(\sum_j a_{ij}^2 p_j \right)^{1/2} &\leq (r^{1/\alpha} \|A\|_\infty^{1/2}) \cdot (r^{1/2} \max_i \left(\sum_j |a_{ij}| p_j \right)^{1/2}) \\ &\leq \frac{1}{2} r^{2/\alpha} \|A\|_\infty + \frac{1}{2} r \max_i \sum_j |a_{ij}| p_j \end{aligned}$$

and

$$\|(a_{ij} \sqrt{p_i p_j})\|_{l_2 \rightarrow l_2} \leq \max_i \sum_j |a_{ij}| \sqrt{p_i p_j} \leq K \max_i \sum_j |a_{ij}| p_j,$$

where $K = K(p_1, \dots, p_n) = \max p_i / \min p_j$. Hence, Theorem 2 yields that

$$\begin{aligned} \left\| \xi^\top A \xi - \mathbb{E} \xi^\top A \xi \right\|_{L_r} &\lesssim_\alpha r^{2/\alpha} \|A\|_\infty + r^{1/2+1/\alpha} \max_i \left(\sum_j a_{ij}^2 p_j \right)^{1/2} \\ &\quad + r \|(a_{ij} \sqrt{p_i p_j})\|_{l_2 \rightarrow l_2} + \sqrt{r} \left(\sum_k a_{kk}^2 p_k + \sum_{i \neq j} a_{ij}^2 p_i p_j \right)^{1/2} \\ &\lesssim_\alpha r^{2/\alpha} \|A\|_\infty + Kr \max_i \sum_j |a_{ij}| p_j + \sqrt{r} \left(\sum_k a_{kk}^2 p_k + \sum_{i \neq j} a_{ij}^2 p_i p_j \right)^{1/2}. \end{aligned}$$

Then, by Lemma 3.1 below, we deduce (1.8) from Theorem 2 but with an extra factor K .

Note that, when A is a diagonal-free matrix, the decay rate of the Hanson-Wright inequality (1.4) is optimal in the sense that, there exist independent centered α -sub-exponential random variables η_1, \dots, η_n such that a matching lower bound holds. See Proposition 1.1 for details and its proof is given in Appendix A.

PROPOSITION 1.1 (Optimality of Theorem 2 when $p_1 = \dots = p_n = 1$). *Let $\eta_1, \dots, \eta_n \stackrel{i.i.d.}{\sim} \mathcal{W}_s(\alpha)$ and A be a symmetric diagonal-free matrix. Then for $0 < \alpha \leq 1$ and $t \geq 0$*

$$\mathbb{P}\{|\eta^\top A \eta - \mathbb{E} \eta^\top A \eta| \geq t\} \geq C(\alpha) \exp(-c(\alpha) f_2(t)),$$

where $\eta = (\eta_1, \dots, \eta_n)^\top$, $C(\alpha), c(\alpha)$ are positive constants depending only on α , and $f_2(t)$ is defined in (1.5).

REMARK 1.4 (Generalization to general A). *Theorems 1 and 2 assume that the matrix A is symmetric. This assumption was made primarily for the convenience of presentation. Indeed, the above results can be extended to general square matrices. The only modification required is that in many places, A should be replaced by $\frac{1}{2}(A + A^\top)$.*

1.2 Discussion

In this section, we assume that $\{\delta_i, 1 \leq i \leq n\}$ is a sequence of independent random variables with $\delta_i \sim \text{Bernoulli}(p_i)$ and $\{\zeta_i, 1 \leq i \leq n\}$ is a sequence of independent centered α -sub-exponential random variables. Assume δ_i 's are independent of ζ_i 's. Denote by $\xi = (\xi_1, \dots, \xi_n)$ the random vector with $\xi_i = \delta_i \cdot \zeta_i$. Let $A = (a_{ij})_{n \times n}$ be a symmetric fixed matrix and $a = (a_1, \dots, a_n)^\top$ be a fixed vector. Set $L(\alpha) = \max_i \|\zeta_i\|_{\Psi_\alpha}$. We mention the following two open questions:

1. For $1 \leq \alpha \leq 2$ and $t > 0$, it holds (see Theorem 6.1 in [3])

$$\mathbb{P}\left\{ \left| \sum_{i=1}^n a_i \zeta_i \right| \geq t \right\} \leq 2 \exp \left(-c(\alpha) \min \left\{ \frac{t^2}{L(\alpha)^2 \|a\|_2^2}, \left(\frac{t}{L(\alpha) \|a\|_{\alpha^*}} \right)^\alpha \right\} \right)$$

and

$$\mathbb{P}\{|\zeta^\top A \zeta - \mathbb{E} \zeta^\top A \zeta| \geq L(\alpha)^2 t\} \leq 2 \exp(-c(\alpha) f_2(t)),$$

where $\alpha^* = \alpha/(\alpha - 1)$ and $f_1(t)$ is defined in (1.3). How to obtain upper bounds for

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_i \xi_i\right| \geq t\right\}, \quad \mathbb{P}\{|\xi^\top A \xi - \mathbb{E} \xi^\top A \xi| \geq t\}$$

which recover the above results when $\delta_1 = \cdots \delta_n = 1$ is an interesting question.

2. In this paper, we give upper bounds for the linear and quadratic forms of sparse α -sub-exponential random vectors:

$$\left\|\sum_{i=1}^n a_i \xi_i\right\|_{L_r}, \quad \|\xi^\top A \xi - \mathbb{E} \xi^\top A \xi\|_{L_r}.$$

Investigating their lower bounds is an interesting question, which may help us establish the optimal bounds for the above quantities.

2 Applications

In this section, we present several applications in covariance estimation, randomized low-rank approximation, and concentration of random vectors. Sparsity can naturally come from the fact that a high-dimensional random vector is sparse, for example, when the elements of the vector are missing at random, or when we intentionally sparsify the vector to speed up computation.

2.1 Covariance matrix estimation under missing observations

Concentration properties of the sample covariance matrix play a central role in numerous high-dimensional inference tasks, including sparse PCA, graphical model selection, discriminant analysis, and high-dimensional regression [22].

We consider the problem of estimating the covariance matrix of a random vector from i.i.d. samples, where each vector is observed with entries missing independently. In the sub-Gaussian setting, this problem was first analyzed in [26], motivated by applications in climate science, gene expression, and cosmology, where an unbiased estimator derived from the sample covariance matrix was proposed. The corresponding error bounds were later sharpened in [19]. More recently, under the L_4 - L_2 moment equivalence assumption, Abdalla [1] introduced an estimator that achieves both optimal dependence on the missingness parameter p and the minimax optimal convergence rate under the spectral norm. However, this estimator is computationally intractable in practice.

Beyond these contributions, a number of related results on covariance estimation with missing data have appeared in the literature, primarily for sub-Gaussian random vectors; see, for example, [6, 29, 27, 20, 36, 37]. In this work, we go beyond the sub-Gaussian framework and develop concentration results for covariance estimation under heavy-tailed distributions with missing observations.

We state our assumptions as follows:

ASSUMPTION 1 (Data assumption). Assume that the entries of $\xi \in \mathbb{R}^m$ are independent, centered, α -sub-exponential random variables with $\alpha \in (0, 2]$ with variance 1. Assume

$$Y = B\xi$$

where B is a fixed $d \times m$ matrix and $BB^\top = \Sigma$.

This means the data vector Y satisfies a multivariate model [10, 6] as a linear transformation of $\xi^{(s)}$, which is a common assumption in high-dimensional statistics. We assume the data vector is observed partially with missing entries, defined as follows:

ASSUMPTION 2 (Missing at random). We observe n independent samples $X^{(1)}, \dots, X^{(n)} \in \mathbb{R}^d$, where each

$$X^{(s)} = \delta^{(s)} \circ Y^{(s)},$$

where

- $\delta^{(s)} \in \{0, 1\}^d$ are i.i.d. Bernoulli random vectors with independent entries $\delta_i^{(s)} \sim \text{Bernoulli}(p_i)$, $i = 1, \dots, d$;
- $Y^{(s)} \in \mathbb{R}^m$ are i.i.d. copies of a random vector Y .

The goal is to estimate $BB^\top = \Sigma$. A natural and unbiased estimator of Σ is defined by

$$\hat{\Sigma}_{jk} = \begin{cases} \frac{1}{np_j p_k} \sum_{s=1}^n X_j^{(s)} X_k^{(s)}, & j \neq k, \\ \frac{1}{np_j} \sum_{s=1}^n (X_j^{(s)})^2, & j = k. \end{cases}$$

This estimator is called the *inverse probability weighting* (IPW) estimator [20, 11, 6, 27]. When Y is a sub-Gaussian random vector, the element-wise deviation of the IPW estimator has been considered in [20, 26, 27].

As soon as the dimension d exceeds the sample size n , however, the sample covariance matrix ceases to be consistent in the operator norm. A standard remedy is to restrict attention to the “significant” coordinates, i.e. to work with a sparse sub-vector. Motivated by this practice we analyse the maximum k -sparse sub-matrix operator norm of the covariance matrix [22]. In particular, we would like to bound the following quantity:

$$\text{RIP}_n(k) := \sup_{\substack{\theta \in \mathbb{R}^d, \|\theta\|_2 \leq 1 \\ \|\theta\|_0 \leq k}} \left| \theta^\top (\hat{\Sigma} - \Sigma) \theta \right|.$$

This norm is also of importance in high-dimensional linear regression due to its connections to the restricted isometry property (RIP) [7] and the restricted eigenvalue (RE) condition [4]. When $k = 1$, $\text{RIP}_n(1)$ is $\|\hat{\Sigma} - \Sigma\|_\infty$ and when $k = d$, $\text{RIP}_n(1)$ is the operator norm $\|\hat{\Sigma} - \Sigma\|_{l_2 \rightarrow l_2}$.

Before giving the main result of this subsection, we introduce some notations. For the vectors $p = (p_1, \dots, p_d)^\top$ and $\theta = (\theta_1, \dots, \theta_d)^\top$, the vector $\theta \circ \frac{1}{p}$ is defined as follows:

$$\theta \circ \frac{1}{p} = (\theta_1/p_1, \dots, \theta_d/p_d)^\top.$$

We define the matrix $A_{\theta,p} \in \mathbb{R}^{d \times d}$ as follows:

$$(A_{\theta,p})_{jk} = \begin{cases} \theta_j^2/p_j, & j = k, \\ \frac{\theta_j \theta_k}{p_j p_k}, & j \neq k. \end{cases}$$

THEOREM 3. Assume that $\max_i \|\xi_i\|_{\Psi_\alpha} \leq 1$. For any $t \geq 0$, with probability at least $1 - 2e^{-t}$,

$$\begin{aligned} \text{RIP}_n(k) &\lesssim_\alpha \frac{(t + k \log(48ed/k))^{1/2}}{\sqrt{n}} \sup_{\theta \in \Omega} \left(\mathbb{E} \|B^\top \text{Diag}(\delta^{(1)}) A_{\theta,p} \text{Diag}(\delta^{(1)}) B\|_F^2 \right)^{1/2} \\ &\quad + \left(\frac{(t + k \log(48ed/k))^{3/4}}{n^{3/4}} + \frac{(t + k \log(48ed/k))^{2/\alpha}}{n} \right) \times \\ &\quad \|B\|_{l_2 \rightarrow l_2} \cdot \|\text{Diag}(\sqrt{p_1}, \dots, \sqrt{p_d}) \cdot B\|_F + \|B\|_{l_2 \rightarrow l_2} \cdot \sup_{\theta \in \Omega} \|\theta \circ \frac{1}{p}\|_2^2, \end{aligned}$$

where $\Omega = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1, \|\theta\|_0 \leq k\}$.

REMARK 2.1 (Comparison with [23]). When all $p_i = 1$, [23] studied covariance estimation for marginally sub-Weibull random vectors. Compared with their result, Theorem 3 removes the extra $\log n$ factor of the $t^{2/\alpha}$ term, and extends the analysis to the missing-observation setting; however, our assumption on the data Y is stronger, whereas [23] does not require X to follow a multivariate model.

REMARK 2.2. (i). By Propositions 5.1 and 5.2 below, we have

$$\begin{aligned} &\left(\mathbb{E} \|B^\top \text{Diag}(\delta^{(1)}) A_{\theta,p} \text{Diag}(\delta^{(1)}) B\|_F^2 \right)^{1/2} \\ &\lesssim \|B\|_{l_2 \rightarrow l_2} \cdot (\|\text{Diag}(\sqrt{p_1}, \dots, \sqrt{p_d}) \cdot B\|_F + \|B\|_{l_2 \rightarrow l_2}) \cdot \|\theta \circ \frac{1}{p}\|_2^2. \end{aligned}$$

(ii). An exact calculation of $\mathbb{E} \|B^\top \text{Diag}(\delta^{(1)}) A_{\theta,p} \text{Diag}(\delta^{(1)}) B\|_F^2$ is given in the Appendix E.

2.2 Low-rank approximation with sparsified sketches

Low-rank approximation is fundamental in data analysis, signal processing, and machine learning. Many large data matrices, such as user–item ratings, gene expression profiles, or kernel matrices, are approximately low-rank due to latent factors or redundancy [34]. Exploiting this structure enables efficient algorithms and reduces storage and computation in downstream tasks.

Recent work [34, 5] shows that large matrices with small spectral norm admit nearby low-rank approximations in the *maximum norm*, providing uniform entrywise guarantees. Building on this perspective, we propose a randomized construction using *sparsified random vectors*. Instead of dense Gaussian sketches, we use probes of the form $\delta \circ \xi$, where δ is a Bernoulli mask and ξ is sub-Gaussian. This reduces storage and matrix–vector costs while introducing dependence on the missingness probability p . Our sparse Hanson–Wright inequalities yield sharp analysis, showing that sparsification preserves robust max-norm guarantees with explicit dependence on matrix coherence.

Sparsified sketches have also appeared in stochastic trace estimation [18], analyzed via sparse Hanson–Wright inequalities [36]. Our results extend this principle to low-rank approximation, demonstrating both computational savings and theoretical control.

Before stating the main result in this subsection we fix notation. For a rank- k matrix $X \in \mathbb{R}^{m \times n}$ with thin SVD $X = U\Sigma V^\top$, the column and row-space coherences are

$$\mu_{\text{col}}(X) = \frac{m}{k} \max_{i \in [m]} \|U^\top e_i\|_2^2, \quad \mu_{\text{row}}(X) = \frac{n}{k} \max_{j \in [n]} \|V^\top \tilde{e}_j\|_2^2,$$

where $e_i \in \mathbb{R}^m$ and $\tilde{e}_j \in \mathbb{R}^n$ are the standard basis vectors.

We propose a randomized algorithm to construct a low-rank approximation of a rank- k matrix X given in Algorithm 1. The analysis of Algorithm 1 uses the sparse Hanson-Wright inequality for a sparse sub-Gaussian random vector.

Algorithm 1 Low rank approximation with sparsified sketches

Input: A rank- k matrix $X \in \mathbb{R}^{m \times n}$, target rank r , sparsity parameter p .

- 1: Compute singular value decomposition $X = U\Sigma V^\top$ where $U \in \mathbb{R}^{m \times k}$, $\Sigma \in \mathbb{R}^{k \times k}$ and $V \in \mathbb{R}^{n \times k}$. Define

$$\tilde{U} = U\Sigma^{1/2} = [\tilde{u}_1^\top, \dots, \tilde{u}_m^\top]^\top, \quad \tilde{V} = V\Sigma^{1/2} = [\tilde{v}_1^\top, \dots, \tilde{v}_n^\top]^\top.$$

- 2: Let $Q = [Q_1, \dots, Q_r] \in \mathbb{R}^{k \times r}$ be a random matrix with independent copies of $r^{-\frac{1}{2}}\delta \circ \xi$, where $\delta \sim \text{Bernoulli}(p)$ and ξ is a centered sub-gaussian random variable satisfying $\|\xi\|_{\Psi_2} \leq 1$ and $\mathbb{E}\xi^2 = 1$. The random variable δ is independent of ξ .

Output: Matrix $Y \in \mathbb{R}^{m \times n}$ such that $Y_{ij} = \frac{1}{p} \tilde{u}_i Q Q^\top \tilde{v}_j^\top$.

THEOREM 4. Assume that $m, n, k \in \mathbb{N}$ satisfies $k \leq \min\{m, n\}$. Let $0 < p \leq 1$ and $\varepsilon > 0$. Then for every matrix $X \in \mathbb{R}^{m \times n}$ of rank k and any constant r such that

$$k \geq r \geq C_1 \log \frac{mn}{\eta} \cdot \max\left\{\frac{p}{\varepsilon^2}, \frac{1}{\varepsilon}\right\},$$

then Algorithm 1 outputs a matrix $Y \in \mathbb{R}^{m \times n}$ of rank no more than r such that with probability at least $1 - \eta$,

$$\|X - Y\|_\infty \leq \frac{k\varepsilon}{p\sqrt{mn}} \sqrt{\mu_{\text{col}}\mu_{\text{row}}} \cdot \|X\|_{l_2 \rightarrow l_2},$$

where μ_{col} and μ_{row} are the coherence of the column and row spaces of X .

REMARK 2.3. Recently, Budzinskiy [5, Theorem 3.4] proved that under the condition $k \geq r \geq \frac{C_1 \log mn}{\varepsilon^2}$, there exists a matrix $Y \in \mathbb{R}^{m \times n}$ of rank no more than r such that

$$\|X - Y\|_\infty \leq \frac{k\varepsilon}{\sqrt{mn}} \sqrt{\mu_{\text{col}}\mu_{\text{row}}} \cdot \|X\|_{l_2 \rightarrow l_2}.$$

Relative to Budzinskiy's theorem, Theorem 4 widens the admissible range of r at the cost of a looser error bound. In practice, we can trade off rank versus accuracy according to the task at hand. Our proof shows that Y can be constructed based on a significantly sparser matrix Q , yielding substantial savings in storage and matrix-vector products compared with the dense approximation proposed in [5].

2.3 Sparse α -sub-exponential concentration

In this subsection we establish a concentration inequality for random vectors with independent α -sub-exponential coordinates; such bounds are widely used [30]. The derivation is standard, so the proof is relegated to the appendix.

THEOREM 5. Assume that $\{\delta_i, 1 \leq i \leq n\}$ and $\{\zeta_i, 1 \leq i \leq n\}$ are two independent sequences of random variables, where $\delta_i \sim \text{Bernoulli}(p)$ are independent and ζ_i are independent α -sub-exponential random variables with mean 0 and variance 1. Consider a random vector $\xi = (\xi_1, \dots, \xi_n)$ with $\xi_i = \delta_i \cdot \zeta_i$. Let A be an $m \times n$ fixed matrix. Then for $0 < \alpha \leq 2$ and $t \geq 0$,

$$\mathbb{P}\left\{\left|\|A\xi\|_2 - \sqrt{p}\|A\|_F\right| > L^2(\alpha)t\right\} \leq 2 \exp\left\{-c(\alpha) \min\left\{\left(\frac{t}{\|A\|_{l_2 \rightarrow l_2}}\right)^2, \left(\frac{t}{\|A\|_{l_2 \rightarrow l_2}}\right)^\alpha\right\}\right\},$$

where $L(\alpha) = \max_i \|\zeta_i\|_{\Psi_\alpha}$ and $c(\alpha) > 0$ is a constant depending only on α .

REMARK 2.4. (i). In the case $0 < \alpha \leq 1$ and $p = 1$, Götze et al. [13, Proposition 2.1] proved that

$$\begin{aligned} & \mathbb{P}\left\{\left|\|A\xi\|_2 - \|A\|_F\right| > L^2(\alpha)t\right\} \\ & \leq 2 \exp\left\{-c(\alpha) \min\left\{\frac{t^2}{\|A\|_F^{2-\alpha}\|A\|_{l_2 \rightarrow l_2}^\alpha}, \left(\frac{t}{\|A\|_{l_2 \rightarrow l_2}}\right)^\alpha\right\}\right\}. \end{aligned}$$

Compared with their result, Theorem 5 gives an improved tail probability due to that

$$\|A\|_{l_2 \rightarrow l_2}^2 \leq \|A\|_F^{2-\alpha}\|A\|_{l_2 \rightarrow l_2}^\alpha.$$

(ii). In the case $\alpha = 2$ and $p = 1$, Rudelson and Vershynin [30, Theorem 2.1] proved that

$$\mathbb{P}\left\{\left|\|A\xi\|_2 - \|A\|_F\right| > L^2(\alpha)t\right\} \leq 2 \exp\left\{-\frac{ct^2}{\|A\|_{l_2 \rightarrow l_2}^2}\right\}. \quad (2.1)$$

Theorem 5 recovers this result in the case $\alpha = 2$ and $p = 1$.

REMARK 2.5. When $\alpha = 2$, we have the following inequality:

$$\|\xi_i\|_{\Psi_2} \leq \|\delta_i\|_{\Psi_2}\|\zeta_i\|_{\Psi_2} \leq L \log^{-1/2}(1 + p^{-1}),$$

where $L = \max_i \|\zeta_i\|_{\Psi_2}$. This implies that ξ_i is a sub-Gaussian random variable. Consequently, we can derive a result of the same type as in Theorem 5 from (2.1).

In particular, Rudelson and Vershynin's result gives us:

$$\mathbb{P}\left\{\left|\|A\xi\|_2^2 - p\|A\|_F^2\right| > L^2t\right\} \leq 2 \exp\left\{-\frac{c \log^2(1 + p^{-1})pt^2}{\|A\|_{l_2 \rightarrow l_2}^2}\right\}.$$

Compared to this result, Theorem 5 provides a better tail probability when p is close to 0. This serves as a good example to illustrate why we study sparse Hanson-Wright type inequalities.

3 Preliminaries

3.1 Tails and Moments

In this subsection, we first introduce the following lemma, which provides the link between L_r estimates and tail probability inequalities. Although such type results are by now common, we give a brief proof for the sake of reading.

LEMMA 3.1. *Let ξ be a random variable such that for $r \geq r_0$*

$$\|\xi\|_{L_r} \leq \sum_{k=1}^m C_k r^{\beta_k} + C_{m+1},$$

where $C_1, \dots, C_{m+1} > 0$ and $\beta_1, \dots, \beta_m > 0$. Then for $t > 0$,

$$\mathbb{P}\{|\xi| > e(mt + C_{m+1})\} \leq e^{r_0} \exp\left(-\min\left\{\left(\frac{t}{C_1}\right)^{1/\beta_1}, \dots, \left(\frac{t}{C_m}\right)^{1/\beta_m}\right\}\right).$$

Proof. Define the following function:

$$f(t) := \min\left\{\left(\frac{t}{C_1}\right)^{1/\beta_1}, \dots, \left(\frac{t}{C_m}\right)^{1/\beta_m}\right\}.$$

If $f(t) \geq r_0$, then

$$\|\xi\|_{L_{f(t)}} \leq \sum_{k=1}^m C_k f(t)^{\beta_k} + C_{m+1} \leq mt + C_{m+1}.$$

Hence, Markov's inequality yields

$$\mathbb{P}\{|\xi| > e(mt + C_{m+1})\} \leq \mathbb{P}\{|\xi| > e\|\xi\|_{L_{f(t)}}\} \leq e^{-f(t)}.$$

As for $f(t) < r_0$, we have the following trivial bound

$$\mathbb{P}\{|\xi| > e(mt + C_{m+1})\} \leq 1 \leq e^{r_0} e^{-f(t)}.$$

Hence, we have for $t \geq 0$

$$\mathbb{P}\{|\xi| > e(mt + C_{m+1})\} \leq e^{r_0} e^{-f(t)}.$$

□

The next lemma estimates the r -th moments of random linear and bilinear sums.

LEMMA 3.2 (Theorem 1.1 in [17]). *Let ξ_1, \dots, ξ_n be independent symmetric random variables with log-convex tails, i.e., $\log \mathbb{P}\{|\xi_i| \geq t\}$ is a convex function for $t \geq 0$. Then for $r \geq 2$*

$$\left\|\sum_i \xi_i\right\|_{L_r} \asymp \left(\sum_i \mathbb{E}|\xi_i|^r\right)^{1/r} + \sqrt{r} \left(\sum_i \mathbb{E}\xi_i^2\right)^{1/2}.$$

In the following lemmas of this subsection, we assume that $\xi_1, \dots, \xi_n \stackrel{i.i.d.}{\sim} \mathcal{W}_s(\alpha)$ and $A = (a_{ij})_{n \times n}$ is a fixed symmetric diagonal-free matrix. Assume further that $\tilde{\xi}_i$ is an independent copy of ξ_i , $1 \leq i \leq n$. For any $1 \leq \alpha \leq 2$, let α^* be its conjugate exponent, i.e., $1/\alpha + 1/\alpha^* = 1$.

LEMMA 3.3 (Theorem 6.1 in [3]). *In the case $1 \leq \alpha \leq 2$, we have for $r \geq 2$*

$$\left\| \sum_{i,j} a_{ij} \xi_i \tilde{\xi}_j \right\|_{L_r} \asymp_{\alpha} r^{1/2} \|A\|_F + r \|A\|_{l_2 \rightarrow l_2} + r^{1/\alpha} \|A\|_{l_{\alpha^*}(l_2)} \\ + r^{(\alpha+2)/2\alpha} \|A\|_{l_2 \rightarrow l_{\alpha^*}} + r^{2/\alpha} \|A\|_{l_{\alpha} \rightarrow l_{\alpha^*}},$$

where $\|A\|_{l_{\alpha^*}(l_2)}$ is defined in (1.9).

LEMMA 3.4 (Example 3 in [21]). *In the case $0 < \alpha \leq 1$, we have for $r \geq 2$*

$$\left\| \sum_{i,j} a_{ij} \xi_i \tilde{\xi}_j \right\|_{L_r} \asymp_{\alpha} r^{1/2} \|A\|_F + r \|A\|_{l_2 \rightarrow l_2} + r^{(\alpha+2)/2\alpha} \|A\|_{l_2 \rightarrow l_{\infty}} + r^{2/\alpha} \|A\|_{\infty}.$$

By employing a similar argument as in [31], we can derive the following simplified result from Lemmas 3.3 and 3.4. For ease of reference, the proof is provided in Appendix C.

COROLLARY 1. *In the case $0 < \alpha \leq 2$, we have for $r \geq 2$,*

$$\left\| \sum_{i,j} a_{ij} \xi_i \tilde{\xi}_j \right\|_{L_r} \lesssim_{\alpha} r^{1/2} \|A\|_F + r^{2/\alpha} \|A\|_{l_2 \rightarrow l_2}.$$

3.2 Decoupling inequality

Decoupling is a technique of replacing quadratic forms of random variables by bilinear forms. The monograph [9] systematically studies decoupling and its applications. In this subsection, we introduce a classic decoupling inequality as follows:

LEMMA 3.5 (Theorem 3.1.1 in [9]). *Let $F : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be a convex function and $A = (a_{ij})_{n \times n}$ be a diagonal-free matrix. If $\{\xi_i, i \leq n\}$ is a sequence of independent random variables and $\tilde{\xi}_i$ is an independent copy of ξ_i , then there exists a universal constant C such that*

$$\mathbb{E}F\left(\left|\sum_{i,j} a_{ij} \xi_i \xi_j\right|\right) \leq \mathbb{E}F\left(C \left|\sum_{i,j} a_{ij} \xi_i \tilde{\xi}_j\right|\right). \quad (3.1)$$

REMARK 3.1. *If, moreover, $A = (a_{ij})$ is a symmetric diagonal-free matrix, then (3.1) can be reversed, that is,*

$$\mathbb{E}F\left(\frac{1}{C} \left|\sum_{i,j} a_{ij} \xi_i \tilde{\xi}_j\right|\right) \leq \mathbb{E}F\left(\left|\sum_{i,j} a_{ij} \xi_i \xi_j\right|\right).$$

3.3 Contraction principle

In this subsection, we present a well-known contraction principle in Banach space. This principle enables the extension of results from Weibull random variables to α -sub-exponential random variables.

LEMMA 3.6 (Lemma 4.6 in [25]). *Let $F : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be a convex non-decreasing function. Assume that $\{\eta_i, i \leq n\}$ and $\{\xi_i, i \leq n\}$ are two sequences of independent symmetric random variables such that for some constant $K \geq 1$*

$$\mathbb{P}\{|\eta_i| > t\} \leq K\mathbb{P}\{|\xi_i| > t\}, \quad i \leq n, \quad t > 0. \quad (3.2)$$

Then, for any sequence $\{a_i, i \leq n\}$ in a Banach space $(\mathcal{A}, \|\cdot\|)$,

$$\mathbb{E}F\left(\left\|\sum_{i=1}^n a_i \eta_i\right\|\right) \leq \mathbb{E}F\left(K\left\|\sum_{i=1}^n a_i \xi_i\right\|\right).$$

REMARK 3.2. *If $\{\eta_i, i \leq n\}$ and $\{\xi_i, i \leq n\}$ are two sequences of independent centered random variables satisfying (3.2), the result in Lemma 3.6 is still valid by a symmetrization argument. Indeed, let $\tilde{\eta}_i$ be independent copy of η_i , $1 \leq i \leq n$. Then, Jensen inequality yields*

$$\begin{aligned} \mathbb{E}F\left(\left\|\sum_{i=1}^n a_i \eta_i\right\|\right) &= \mathbb{E}F\left(\left\|\sum_{i=1}^n a_i (\eta_i - \mathbb{E}_{\tilde{\eta}} \tilde{\eta}_i)\right\|\right) \\ &\leq \mathbb{E}F\left(\left\|\sum_{i=1}^n a_i (\eta_i - \tilde{\eta}_i)\right\|\right) \leq \mathbb{E}F\left(2\left\|\sum_{i=1}^n a_i \varepsilon_i \eta_i\right\|\right), \end{aligned}$$

where $\{\varepsilon_i, i \leq n\}$ are independent Rademacher variables. On the other hand, by the convexity of F

$$\mathbb{E}F\left(2\left\|\sum_{i=1}^n a_i \varepsilon_i \eta_i\right\|\right) \leq \mathbb{E}F\left(2\left\|\sum_{i=1}^n a_i \varepsilon_i (\eta_i - \tilde{\eta}_i)\right\|\right) \leq \mathbb{E}F\left(4\left\|\sum_{i=1}^n a_i \eta_i\right\|\right).$$

Hence, we have

$$\mathbb{E}F\left(\left\|\sum_{i=1}^n a_i \eta_i\right\|\right) \leq \mathbb{E}F\left(2\left\|\sum_{i=1}^n a_i \varepsilon_i \eta_i\right\|\right) \leq \mathbb{E}F\left(4\left\|\sum_{i=1}^n a_i \eta_i\right\|\right).$$

3.4 Concentration inequality

In this subsection, we shall introduce Talagrand's concentration inequality for convex Lipschitz functions.

LEMMA 3.7 (Corollary 4.10 in [24]). *Let ξ_1, \dots, ξ_n be independent random variables such that $|\xi_i| \leq K$ for all $1 \leq i \leq n$. Assume that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a convex and 1-Lipschitz function. Then for $t > 0$*

$$\mathbb{P}\{|f(\xi_1, \dots, \xi_n) - \mathbb{E}f(\xi_1, \dots, \xi_n)| > Kt\} \leq 4e^{-t^2/4}.$$

3.5 The sparse Bernstein inequality

In this subsection, we introduce the following sparse Bernstein inequality in α -sub-exponential random variables, $0 < \alpha \leq 1$. With this property, one can estimate the diagonal sum of the quadratic forms.

LEMMA 3.8. *Assume that $\{\delta_i, 1 \leq i \leq n\}$ and $\{\zeta_i, 1 \leq i \leq n\}$ are two independent sequences of random variables, where $\delta_i \sim \text{Bernoulli}(p_i)$ are independent and ζ_i are independent centered α -sub-exponential random variables. Let $a = (a_1, \dots, a_n)$ be a fixed vector. Then for $0 < \alpha \leq 1$ and $t \geq 0$*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_i \delta_i \zeta_i\right| \geq t\right\} \leq 2 \exp\left(-c(\alpha) \min\left\{\frac{t^2}{L(\alpha)^2(\sum_i a_i^2 p_i)}, \left(\frac{t}{L(\alpha)\|a\|_\infty}\right)^\alpha\right\}\right),$$

where $L(\alpha) = \max_i \|\zeta_i\|_{\Psi_\alpha}$ and $c(\alpha)$ is a positive constant depending only on α .

REMARK 3.3. (i). This property, though likely known in the literature, was explicitly stated and proved using a combinatorial approach in He et al. (see Theorem 3 in [15]). In Appendix D we present an alternative, more concise proof.

(ii). For $0 < \alpha \leq 1$ and $t \geq 0$, it holds (see Corollary 1.4 in [13])

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_i \zeta_i\right| \geq t\right\} \leq 2 \exp\left(-c(\alpha) \min\left\{\frac{t^2}{L(\alpha)^2\|a\|_2^2}, \left(\frac{t}{L(\alpha)\|a\|_\infty}\right)^\alpha\right\}\right).$$

Hence, Lemma 3.8 recovers this classical result when $p_1 = \dots = p_n = 1$.

4 Proofs of main results

4.1 The sparse Hanson-Wright inequality for $0 < \alpha \leq 2$

We begin by proving Theorem 1 for the special case where matrix A has zero diagonal elements.

PROPOSITION 4.1. *Assume that $\{\delta_i, 1 \leq i \leq n\}$ and $\{\zeta_i, 1 \leq i \leq n\}$ are two independent sequences of random variables, where $\delta_i \sim \text{Bernoulli}(p_i)$ are independent and ζ_i are independent centered α -sub-exponential random variables. Consider a random vector $\xi = (\xi_1, \dots, \xi_n)$ with $\xi_i = \delta_i \cdot \zeta_i$. Let $A = (a_{ij})_{n \times n}$ be a symmetric diagonal-free matrix. Then for $0 < \alpha \leq 2$ and $t \geq 0$,*

$$\mathbb{P}\{|\xi^\top A \xi| \geq t\} \leq 2 \exp\left(-c(\alpha) \min\left\{\frac{t^2}{L(\alpha)^4 \sum_{i,j} a_{ij}^2 p_i p_j}, \left(\frac{t}{L(\alpha)^2 \|A\|_{l_2 \rightarrow l_2}}\right)^{\alpha/2}\right\}\right),$$

where $L(\alpha) = \max_i \|\zeta_i\|_{\Psi_\alpha}$ and $c(\alpha)$ is a positive constant depending only on α .

Proof. Let $\tilde{\xi}_i = \tilde{\delta}_i \cdot \tilde{\zeta}_i$ for $1 \leq i \leq n$, where $\{\tilde{\delta}_i, 1 \leq i \leq n\}$ and $\{\tilde{\zeta}_i, 1 \leq i \leq n\}$ are independent copies of $\{\delta_i, 1 \leq i \leq n\}$ and $\{\zeta_i, 1 \leq i \leq n\}$, respectively. Lemma 3.5 yields for $r \geq 1$

$$\|\xi^\top A \xi\|_{L_r} \lesssim \|\xi^\top A \tilde{\xi}\|_{L_r},$$

where $\tilde{\xi} = (\tilde{\xi}_1, \dots, \tilde{\xi}_n)$. Let $\eta_1, \dots, \eta_n \stackrel{i.i.d.}{\sim} \mathcal{W}_s(\alpha), 0 < \alpha \leq 2$, and denote by $\tilde{\eta}_i$ the independent copy of $\eta_i, 1 \leq i \leq n$. Using the conditional probability and Lemma 3.6 twice, we have for $r \geq 1$

$$\begin{aligned} \|\xi^\top A \tilde{\xi} / L(\alpha)^2\|_{L_r} &= \left(\mathbb{E}_\xi \mathbb{E}_{\tilde{\xi}} \left| \frac{1}{L(\alpha)^2} \sum_j \tilde{\xi}_j \sum_i a_{ij} \xi_i \right|^r \right)^{1/r} \\ &\lesssim \left(\mathbb{E}_\xi \mathbb{E}_{\tilde{\eta}} \left| \frac{1}{L(\alpha)} \sum_j \tilde{\delta}_j \tilde{\eta}_j \sum_i a_{ij} \xi_i \right|^r \right)^{1/r} \lesssim \left\| \sum_{i,j} a_{ij} \delta_i \eta_i \tilde{\delta}_j \tilde{\eta}_j \right\|_{L_r}. \end{aligned}$$

Corollary 1 yields for $r \geq 2$

$$\begin{aligned} \mathbb{E}_{\eta, \tilde{\eta}} \left| \sum_{i,j} a_{ij} \delta_i \eta_i \tilde{\delta}_j \tilde{\eta}_j \right|^r &\leq C_1(\alpha)^r r^{r/2} \left(\sum_{i,j} a_{ij}^2 \delta_i \tilde{\delta}_j \right)^{r/2} \\ &\quad + C_2(\alpha)^r r^{2r/\alpha} \|(a_{ij} \delta_i \tilde{\delta}_j)_{n \times n}\|_{l_2 \rightarrow l_2}^r, \end{aligned} \quad (4.1)$$

where $\mathbb{E}_{\eta, \tilde{\eta}}$ means taking expectation with respect to $\eta_i, \tilde{\eta}_i, 1 \leq i \leq n$.

Let $\Lambda = \text{Diag}(\delta_1, \dots, \delta_n)$ be a diagonal matrix with entries $\delta_1, \dots, \delta_n$, and $\tilde{\Lambda} = \text{Diag}(\tilde{\delta}_1, \dots, \tilde{\delta}_n)$. As for the second term on the right side of (4.1), we have

$$\|(a_{ij} \delta_i \tilde{\delta}_j)_{n \times n}\|_{l_2 \rightarrow l_2} = \|\Lambda A \tilde{\Lambda}\|_{l_2 \rightarrow l_2} \leq \|\Lambda\|_{l_2 \rightarrow l_2} \|A\|_{l_2 \rightarrow l_2} \|\tilde{\Lambda}\|_{l_2 \rightarrow l_2} \leq \|A\|_{l_2 \rightarrow l_2}.$$

We next turn to bounding the first term of (4.1). Note that, $f(x_1, \dots, x_n) = \sqrt{\sum_i a_i^2 x_i^2}$ is a convex and Lipschitz function with $\|f\|_{\text{Lip}} = \max_i |a_i|$. Hence, Lemma 3.7 yields for $t > 0$

$$\mathbb{P}_\delta \left\{ \left| \left(\sum_i \delta_i \sum_j a_{ij}^2 \tilde{\delta}_j \right)^{1/2} - \mathbb{E}_\delta \left(\sum_i \delta_i \sum_j a_{ij}^2 \tilde{\delta}_j \right)^{1/2} \right| > \max_i \left(\sum_j a_{ij}^2 \tilde{\delta}_j \right)^{1/2} t \right\} \leq 4e^{-t^2/4},$$

where \mathbb{P}_δ and \mathbb{E}_δ mean taking probability and expectation with respect to $\delta_1, \dots, \delta_n$. Then, a direct integration yields for $r \geq 1$

$$\begin{aligned} &\left(\mathbb{E}_\delta \left| \left(\sum_i \delta_i \sum_j a_{ij}^2 \tilde{\delta}_j \right)^{1/2} \right|^r \right)^{1/r} \\ &\lesssim \mathbb{E}_\delta \left(\sum_i \delta_i \sum_j a_{ij}^2 \tilde{\delta}_j \right)^{1/2} + \sqrt{r} \max_i \left(\sum_j a_{ij}^2 \tilde{\delta}_j \right)^{1/2} \\ &\leq \left(\sum_{i,j} a_{ij}^2 p_i \tilde{\delta}_j \right)^{1/2} + \sqrt{r} \max_i \left(\sum_j a_{ij}^2 \right)^{1/2}. \end{aligned} \quad (4.2)$$

It is obvious to see that

$$\max_i \left(\sum_j a_{ij}^2 \right)^{1/2} = \|A\|_{l_2 \rightarrow l_\infty} \leq \|A\|_{l_2 \rightarrow l_2}.$$

Using Lemma 3.7 again, we have for $t > 0$

$$\mathbb{P} \left\{ \left| \left(\sum_{i,j} a_{ij}^2 p_i \tilde{\delta}_j \right)^{1/2} - \mathbb{E} \left(\sum_{i,j} a_{ij}^2 p_i \tilde{\delta}_j \right)^{1/2} \right| > \max_j \left(\sum_i a_{ij}^2 p_i \right)^{1/2} t \right\} \leq 4e^{-t^2/4}.$$

Then for $r \geq 1$

$$\begin{aligned} \left\| \left(\sum_{i,j} a_{ij}^2 p_i \tilde{\delta}_j \right)^{1/2} \right\|_{L_r} &\lesssim \mathbb{E} \left(\sum_{i,j} a_{ij}^2 p_i \tilde{\delta}_j \right)^{1/2} + \sqrt{r} \max_j \left(\sum_i a_{ij}^2 p_i \right)^{1/2} \\ &\leq \left(\sum_{i,j} a_{ij}^2 p_i p_j \right)^{1/2} + \sqrt{r} \|A\|_{l_2 \rightarrow l_2}. \end{aligned} \quad (4.3)$$

By virtue of (4.2) and (4.3), we have for $r \geq 1$

$$\sqrt{r} \left\| \left(\sum_{i,j} a_{ij}^2 \delta_i \tilde{\delta}_j \right)^{1/2} \right\|_{L_r} \lesssim \sqrt{r} \left(\sum_{i,j} a_{ij}^2 p_i p_j \right)^{1/2} + r \|A\|_{l_2 \rightarrow l_2}. \quad (4.4)$$

As $0 < \alpha \leq 2$, we can absorb the last term above with $r^{2/\alpha} \|A\|_{l_2 \rightarrow l_2}$. Combining (4.1) with (4.4), we have for $r \geq 2$

$$\left\| \sum_{ij} a_{ij} \delta_i \eta_i \tilde{\delta}_j \tilde{\eta}_j \right\|_{L_r} \lesssim_\alpha \sqrt{r} \left(\sum_{ij} a_{ij}^2 p_i p_j \right)^{1/2} + r^{2/\alpha} \|A\|_{l_2 \rightarrow l_2}.$$

Hence, we finish the proof by Lemma 3.1 and adjusting the universal constant. \square

Now, we are prepared to prove Theorem 1.

Proof of Theorem 1. By the triangle inequality, we have for $r \geq 1$

$$\left\| \xi^\top A \xi - \mathbb{E} \xi^\top A \xi \right\|_{L_r} \leq \left\| \sum_{i \neq j} a_{ij} \xi_i \xi_j \right\|_{L_r} + \left\| \sum_i a_{ii} (\xi_i^2 - \mathbb{E} \xi_i^2) \right\|_{L_r}.$$

Proposition 4.1 yields that

$$\left\| \sum_{i \neq j} a_{ij} \xi_i \xi_j / L(\alpha)^2 \right\|_{L_r} \lesssim_\alpha \sqrt{r} \left(\sum_{i \neq j} a_{ij}^2 p_i p_j \right)^{1/2} + r^{2/\alpha} \|A\|_{l_2 \rightarrow l_2}. \quad (4.5)$$

Let $\tilde{\xi}_i$ be an independent copy of ξ_i , $1 \leq i \leq n$. Note that

$$\left\| \sum_i a_{ii} (\xi_i^2 - \mathbb{E} \xi_i^2) \right\|_{L_r} = \left\| \sum_i a_{ii} (\xi_i^2 - \mathbb{E}_{\tilde{\xi}_i} \tilde{\xi}_i^2) \right\|_{L_r} \leq 2 \left\| \sum_i a_{ii} \varepsilon_i \zeta_i^2 \right\|_{L_r},$$

where $\varepsilon_1, \dots, \varepsilon_n$ is a sequence of i.i.d. Rademacher random variables. Due to that

$$\mathbb{P}\{|\varepsilon_i \zeta_i^2| > L(\alpha)^2 t\} \leq \mathbb{P}\{|\zeta_i| > L(\alpha) \sqrt{t}\} \leq 2e^{-ct^{\alpha/2}},$$

Lemma 3.8 yields that

$$\left\| \sum_i a_{ii} (\xi_i^2 - \mathbb{E} \xi_i^2) / L(\alpha)^2 \right\|_{L_r} \lesssim_\alpha \sqrt{r} \left(\sum_i a_{ii}^2 p_i \right)^{1/2} + r^{2/\alpha} \max_i |a_{ii}|. \quad (4.6)$$

By virtue of (4.5) and (4.6), we have

$$\begin{aligned} &\left\| (\xi^\top A \xi - \mathbb{E} \xi^\top A \xi) / L(\alpha)^2 \right\|_{L_r} \\ &\lesssim_\alpha \sqrt{r} \left(\sum_{i \neq j} a_{ij}^2 p_i p_j + \sum_i a_{ii}^2 p_i \right)^{1/2} + r^{2/\alpha} \|A\|_{l_2 \rightarrow l_2}, \end{aligned}$$

where we use the fact $\max_i |a_{ii}| \leq \|A\|_{l_2 \rightarrow l_2}$. Hence, the desired result follows from Lemma 3.1. \square

4.2 An improved sparse Hanson-Wright inequality for $0 < \alpha \leq 1$

We complete the proof of Theorem 2 in this subsection.

Proof of Theorem 2. We first assume $a_{ii} = 0, 1 \leq i \leq n$. By Lemmas 3.5 and 3.6, we only need to bound the quantity $\|\sum_{i,j} a_{ij} \delta_i \eta_i \tilde{\delta}_j \tilde{\eta}_j\|_{L_r}$. Here, $\eta_1, \dots, \eta_n \stackrel{i.i.d.}{\sim} \mathcal{W}_s(\alpha)$, and for $1 \leq i \leq n$, $\tilde{\delta}_i, \tilde{\eta}_i$ are independent copies of δ_i, η_i , respectively.

Note that, for $1 \leq i \leq n$,

$$-\log \mathbb{P}\{|\delta_i \eta_i| \geq t\} = \begin{cases} 0, & t = 0 \\ t^\alpha - \log p_i, & t > 0 \end{cases}$$

is a concave function for $t \geq 0$. Hence, by Lemma 3.2, we have for $r \geq 2$

$$\begin{aligned} \left(\mathbb{E}_{\delta, \eta} \left| \sum_i \delta_i \eta_i \sum_j a_{ij} \tilde{\delta}_j \tilde{\eta}_j \right|^r \right)^{1/r} &\lesssim_\alpha r^{1/\alpha} \left(\sum_i p_i \left| \sum_j a_{ij} \tilde{\delta}_j \tilde{\eta}_j \right|^r \right)^{1/r} \\ &\quad + \sqrt{r} \left(\sum_i p_i \left(\sum_j a_{ij} \tilde{\delta}_j \tilde{\eta}_j \right)^2 \right)^{1/2}, \end{aligned} \quad (4.7)$$

where $\mathbb{E}_{\delta, \eta}$ means taking expectation with respect to $\{\delta_i, \eta_i, 1 \leq i \leq n\}$.

Combining Fubini's theorem with (4.7), we have for $r \geq 2$

$$\begin{aligned} \left\| \sum_{i,j} a_{ij} \delta_i \eta_i \tilde{\delta}_j \tilde{\eta}_j \right\|_{L_r} &= \left(\mathbb{E}_{\tilde{\delta}, \tilde{\eta}} \mathbb{E}_{\delta, \eta} \left| \sum_i \delta_i \eta_i \sum_j a_{ij} \tilde{\delta}_j \tilde{\eta}_j \right|^r \right)^{1/r} \\ &\lesssim_\alpha r^{1/\alpha} \left(\mathbb{E} \sum_i p_i \left| \sum_j a_{ij} \tilde{\delta}_j \tilde{\eta}_j \right|^r \right)^{1/r} \\ &\quad + \sqrt{r} \left\| \left(\sum_i p_i \left(\sum_j a_{ij} \tilde{\delta}_j \tilde{\eta}_j \right)^2 \right)^{1/2} \right\|_{L_r}, \end{aligned} \quad (4.8)$$

where the last inequality we use the fact $(a + b)^r \leq 2^r (|a|^r + |b|^r)$.

Next, we turn to the terms on the right side of (4.8). For the first term, we have

$$\begin{aligned} \left(\mathbb{E} \sum_i p_i \left| \sum_j a_{ij} \tilde{\delta}_j \tilde{\eta}_j \right|^r \right)^{1/r} &= \left(\sum_i p_i \left\| \sum_j a_{ij} \tilde{\delta}_j \tilde{\eta}_j \right\|_{L_r}^r \right)^{1/r} \\ &\lesssim \left(\sum_i p_i \left(\left(\sum_j \mathbb{E} |a_{ij} \tilde{\delta}_j \tilde{\eta}_j|^r \right)^{1/r} + \sqrt{r} \left(\sum_j \mathbb{E} |a_{ij} \tilde{\delta}_j \tilde{\eta}_j|^2 \right)^{1/2} \right)^r \right)^{1/r} \\ &\lesssim_\alpha r^{1/\alpha} \left(\sum_{i,j} |a_{ij}|^r p_i p_j \right)^{1/r} + \sqrt{r} \left(\sum_i p_i \left(\sum_j a_{ij}^2 p_j \right)^{r/2} \right)^{1/r}, \end{aligned}$$

where we use Lemma 3.2 in the second step and use the triangle inequality in the third step.

For the second term on the right side of (4.8), note that for $g_1, \dots, g_n \stackrel{i.i.d.}{\sim} N(0, 1)$ and $a_1, \dots, a_n \in \mathbb{R}$

$$\left\| \sum_i a_i g_i \right\|_{L_r} \asymp \sqrt{r} \left(\sum_i a_i^2 \right)^{1/2}.$$

Hence, we have by Lemma 3.2

$$\begin{aligned}
& \mathbb{E} \left(r \sum_i p_i \left(\sum_j a_{ij} \tilde{\delta}_j \tilde{\eta}_j \right)^2 \right)^{r/2} \\
& \leq C^r \mathbb{E} \mathbb{E}_g \left(\sum_i \sqrt{p_i} g_i \sum_j a_{ij} \tilde{\delta}_j \tilde{\eta}_j \right)^r \\
& = C^r \mathbb{E} \mathbb{E}_{\tilde{\delta}, \tilde{\eta}} \left(\sum_j \tilde{\delta}_j \tilde{\eta}_j \sum_i a_{ij} \sqrt{p_i} g_i \right)^r \\
& \leq C_1(\alpha)^r \mathbb{E} \left(r^{r/\alpha} \sum_j p_j \left| \sum_i a_{ij} \sqrt{p_i} g_i \right|^r + r^{r/2} \left(\sum_j p_j \left(\sum_i a_{ij} \sqrt{p_i} g_i \right)^2 \right)^{r/2} \right) \\
& \leq C_2(\alpha)^r r^{r/\alpha + r/2} \sum_j p_j \left(\sum_i a_{ij}^2 p_i \right)^{r/2} + C_3(\alpha)^r \mathbb{E} \left(\sum_{i,j} a_{ij} \sqrt{p_i p_j} g_i \tilde{g}_j \right)^r,
\end{aligned}$$

where \tilde{g}_j is an independent copy of g_j , $1 \leq j \leq n$. Note that g_i are also sub-gaussian random variables. Hence, we have by Corollary 1 and Lemma 3.6

$$\left\| \sum_{i,j} a_{ij} \sqrt{p_i p_j} g_i \tilde{g}_j \right\|_{L_r} \lesssim \sqrt{r} \left(\sum_{i,j} a_{ij}^2 p_i p_j \right)^{1/2} + r \|(a_{ij} \sqrt{p_i p_j})_{n \times n}\|_{l_2 \rightarrow l_2}.$$

Up to now, we have proved for diagonal free symmetric matrix A

$$\begin{aligned}
\left\| \sum_{i,j} a_{ij} \delta_i \eta_i \tilde{\delta}_j \tilde{\eta}_j \right\|_{L_r} & \lesssim_\alpha r^{2/\alpha} \left(\sum_{i,j} |a_{ij}|^r p_i p_j \right)^{1/r} + r^{1/2+1/\alpha} \left(\sum_i p_i \left(\sum_j a_{ij}^2 p_j \right)^{r/2} \right)^{1/r} \\
& \quad + r \|(a_{ij} \sqrt{p_i p_j})_{n \times n}\|_{l_2 \rightarrow l_2} + \sqrt{r} \left(\sum_{i,j} a_{ij}^2 p_i p_j \right)^{1/2}.
\end{aligned}$$

Note that for $r \geq 3$, by Young's inequality, we have

$$\begin{aligned}
\left(\sum_{i,j} |a_{ij}|^r p_i p_j \right)^{1/r} & \leq \max_{i,j} |a_{ij}|^{(r-2)/r} \cdot \left(\sum_{i,j} a_{ij}^2 p_i p_j \right)^{1/r} \\
& = \left(e^{2r/(r-2)} \max_{i,j} |a_{ij}| \right)^{(r-2)/r} \cdot \left(\sum_{i,j} a_{ij}^2 p_i p_j e^{-2r} \right)^{1/r} \\
& \leq \frac{r-2}{r} \left(e^{2r/(r-2)} \max_{i,j} |a_{ij}| \right) + \frac{2}{r} \left(\sum_{i,j} a_{ij}^2 p_i p_j e^{-2r} \right)^{1/2} \\
& \leq e^6 \|A\|_\infty + e^{-r} \left(\sum_{i,j} a_{ij}^2 p_i p_j \right)^{1/2}.
\end{aligned}$$

Similarly

$$\left(\sum_i p_i \left(\sum_j a_{ij}^2 p_j \right)^{r/2} \right)^{1/r} \leq e^6 \max_i \left(\sum_j a_{ij}^2 p_j \right)^{1/2} + e^{-r} \left(\sum_{i,j} a_{ij}^2 p_i p_j \right)^{1/2}.$$

Note that the e^{-r} factors we introduced will eliminate the $r^{2/\alpha}, r^{1/2+1/\alpha}$ factors in the corresponding terms. Hence, we have for $r \geq 3$

$$\begin{aligned} \left\| \sum_{i,j} a_{ij} \delta_i \eta_i \tilde{\delta}_j \tilde{\eta}_j \right\|_{L_r} &\lesssim_\alpha r^{2/\alpha} \|A\|_\infty + r^{1/2+1/\alpha} \max_i \left(\sum_j a_{ij}^2 p_j \right)^{1/2} \\ &\quad + r \|(a_{ij} \sqrt{p_i p_j})_{n \times n}\|_{l_2 \rightarrow l_2} + \sqrt{r} \left(\sum_{i,j} a_{ij}^2 p_i p_j \right)^{1/2}. \end{aligned} \quad (4.9)$$

For a general symmetric matrix A , we only need to estimate the diagonal case. Following the same line as in the proof of Theorem 1, we have

$$\left\| \sum_i a_{ii} (\xi_i^2 - \mathbb{E} \xi_i^2) / L(\alpha)^2 \right\|_{L_r} \lesssim_\alpha \sqrt{r} \left(\sum_i a_{ii}^2 p_i \right)^{1/2} + r^{2/\alpha} \max_i |a_{ii}|. \quad (4.10)$$

Combining (4.9) with (4.10), the desired result follows from Lemma 3.1. \square

5 Proofs of applications

5.1 Proof of Theorem 3

Although Theorem 3 is not a literal corollary of our main theorem, its proof follows the same roadmap we developed for Theorem 1; thus it can be viewed as an application of our techniques. Before prove Theorem 3, we first show the following properties.

PROPOSITION 5.1. *Let $\delta \in \mathbb{R}^d$ and $B \in \mathbb{R}^{d \times m}$ as in Theorem 3. Let $A = \text{Diag}(a_1, \dots, a_d)$ be a diagonal matrix. For any $r \geq 1$, we have*

$$\begin{aligned} &(\mathbb{E} \|B^\top \cdot \text{Diag}(\delta) \cdot A \cdot \text{Diag}(\delta) \cdot B\|_F^r)^{1/r} \\ &\lesssim \|B\|_{l_2 \rightarrow l_2} \cdot \|A\|_{l_2 \rightarrow l_2} (\|\text{Diag}(\sqrt{p_1}, \dots, \sqrt{p_d}) \cdot B\|_F + r^{1/2} \|B\|_{l_2 \rightarrow l_2}). \end{aligned}$$

Proof. Note that

$$\begin{aligned} &\|B^\top \cdot \text{Diag}(\delta) \cdot A \cdot \text{Diag}(\delta) \cdot B\|_F \\ &\leq \|B^\top \cdot \text{Diag}(\delta) \cdot \sqrt{A}\|_{l_2 \rightarrow l_2} \cdot \|\sqrt{A} \cdot \text{Diag}(\delta) \cdot B\|_F \\ &\leq \|B\|_{l_2 \rightarrow l_2} \cdot \|\sqrt{A}\|_{l_2 \rightarrow l_2} \cdot \|\sqrt{A} \cdot \text{Diag}(\delta) \cdot B\|_F, \end{aligned}$$

where \sqrt{A} is a (complex) diagonal matrix such that $\sqrt{A} \cdot \sqrt{A} = A$. Talagrand's concentration inequality (Lemma 3.7) yields that

$$\begin{aligned} &\mathbb{P} \left\{ \left| \|\sqrt{A} \cdot \text{Diag}(\delta) \cdot B\|_F - \mathbb{E} \|\sqrt{A} \cdot \text{Diag}(\delta) \cdot B\|_F \right| > t \right\} \\ &\leq 2 \exp \left\{ - \frac{t^2}{\left(\|\sqrt{A}\|_{l_2 \rightarrow l_2} \cdot \|B\|_{l_2 \rightarrow \infty} \right)^2} \right\}. \end{aligned}$$

We finish the proof due to that

$$\begin{aligned} \mathbb{E} \|\sqrt{A} \cdot \text{Diag}(\delta) \cdot B\|_F &\leq (\mathbb{E} \|\sqrt{A} \cdot \text{Diag}(\delta) \cdot B\|_F^2)^{1/2} \\ &\leq \|\sqrt{A}\|_{l_2 \rightarrow l_2} \cdot \|\text{Diag}(\sqrt{p_1}, \dots, \sqrt{p_d}) \cdot B\|_F. \end{aligned}$$

\square

PROPOSITION 5.2. Let $\delta \in \mathbb{R}^d$ and $B \in \mathbb{R}^{d \times m}$ as in Theorem 3. For a fixed vector $x \in \mathbb{R}^d$, denote $A = xx^\top$. For any $r \geq 1$, we have

$$(\mathbb{E} \|B^\top \cdot \text{Diag}(\delta) \cdot A \cdot \text{Diag}(\delta) \cdot B\|_F^r)^{1/r} \lesssim \|B\|_{l_2 \rightarrow l_2}^2 \|x\|_2^2.$$

Proof.

$$\begin{aligned} & \|B^\top \cdot \text{Diag}(\delta) \cdot A \cdot \text{Diag}(\delta) \cdot B\|_F \\ &= \|B^\top \cdot \text{Diag}(x) \cdot \delta \delta^\top \cdot \text{Diag}(x) \cdot B\|_F \\ &= \|B^\top \cdot \text{Diag}(x) \cdot \delta\|_2^2 \leq \|B\|_{l_2 \rightarrow l_2}^2 \|\text{Diag}(x) \cdot \delta\|_2^2 \leq \|B\|_{l_2 \rightarrow l_2}^2 \|x\|_2^2. \end{aligned}$$

□

Proof of Theorem 3. For each fixed θ , we can rewrite

$$\begin{aligned} \theta^\top (\widehat{\Sigma} - \Sigma) \theta &= \frac{1}{n} \sum_{s=1}^n \left[X^{(s)\top} A_{\theta,p} X^{(s)} - \mathbb{E}(X^{(s)\top} A_{\theta,p} X^{(s)}) \right] \\ &= \frac{1}{n} \sum_{s=1}^n \left[\xi^{(s)\top} B^\top \text{Diag}(\delta^{(s)}) A_{\theta,p} \text{Diag}(\delta^{(s)}) B \xi^{(s)} - \mathbb{E}(X^{(s)\top} A_{\theta,p} X^{(s)}) \right] \\ &= \tilde{\xi}^\top \tilde{B}^\top \text{Diag}(\tilde{\delta}) \tilde{A}_{\theta,p} \text{Diag}(\tilde{\delta}) \tilde{B} \tilde{\xi} - \mathbb{E} \tilde{\xi}^\top \tilde{B}^\top \text{Diag}(\tilde{\delta}) \tilde{A}_{\theta,p} \text{Diag}(\tilde{\delta}) \tilde{B} \tilde{\xi}. \end{aligned}$$

Here

$$\begin{aligned} \tilde{\xi} &= (\xi^{(1)\top}, \dots, \xi^{(n)\top})^\top \in \mathbb{R}^{mn}, \quad \tilde{B} = \text{Diag}(B, B, \dots, B) \in \mathbb{R}^{dn \times mn}, \\ \tilde{\delta} &= (\delta^{(1)\top}, \dots, \delta^{(n)\top})^\top \in \mathbb{R}^{dn \times dn}, \quad \tilde{A}_{\theta,p} = \text{Diag}(A_{\theta,p}, \dots, A_{\theta,p})/n \in \mathbb{R}^{dn \times dn}. \end{aligned}$$

Hence, by virtue of the decoupling inequality and Corollary 1, we have for $r \geq 1$

$$\begin{aligned} \|\theta^\top (\widehat{\Sigma} - \Sigma) \theta\|_{L_r} &\lesssim_\alpha \|\tilde{\xi}^\top \tilde{B}^\top \text{Diag}(\tilde{\delta}) \tilde{A}_{\theta,p} \text{Diag}(\tilde{\delta}) \tilde{B} \tilde{\eta}\|_{L_r} \\ &\lesssim_\alpha r^{1/2} (\mathbb{E} \|\tilde{B}^\top \text{Diag}(\tilde{\delta}) \tilde{A}_{\theta,p} \text{Diag}(\tilde{\delta}) \tilde{B}\|_F^r)^{1/r} \\ &\quad + r^{2/\alpha} (\mathbb{E} \|\tilde{B}^\top \text{Diag}(\tilde{\delta}) \tilde{A}_{\theta,p} \text{Diag}(\tilde{\delta}) \tilde{B}\|_{l_2 \rightarrow l_2}^r)^{1/r}, \end{aligned} \quad (5.1)$$

where $\tilde{\eta}$ is an independent copy of $\tilde{\xi}$.

For the first term on the right-side of (5.1), we have

$$\|\tilde{B}^\top \text{Diag}(\tilde{\delta}) \tilde{A}_{\theta,p} \text{Diag}(\tilde{\delta}) \tilde{B}\|_F = \left(\frac{1}{n^2} \sum_{s=1}^n \|B^\top \text{Diag}(\delta^{(s)}) A_{\theta,p} \text{Diag}(\delta^{(s)}) B\|_F^2 \right)^{1/2}$$

Note that

$$A_{\theta,p} = (\theta \circ \frac{1}{p})(\theta \circ \frac{1}{p})^\top - \text{Diag}((\theta \circ \frac{1}{p})(\theta \circ \frac{1}{p})^\top) + \text{Diag}(A_{\theta,p}). \quad (5.2)$$

Hence, we can split the above boundary into two parts according to the division of $A_{\theta,p}$, the boundary formed by $(\theta \circ \frac{1}{p})(\theta \circ \frac{1}{p})^\top$ and the boundary formed by $\text{Diag}(A_{\theta,p}) - \text{Diag}((\theta \circ \frac{1}{p})(\theta \circ \frac{1}{p})^\top)$.

$\frac{1}{p})^\top$). Propositions 5.1 and 5.2 yield that, $\|B^\top \text{Diag}(\delta) A_{\theta,p} \text{Diag}(\delta) B\|_F$ is a sub-gaussian random variable with sub-gaussian norm

$$\begin{aligned} K_1(\theta) &= \|B\|_{l_2 \rightarrow l_2} \cdot (\text{Diag}(\sqrt{p_1}, \dots, \sqrt{p_d}) \cdot B\|_F + \|B\|_{l_2 \rightarrow l_2} \cdot \|(\theta \circ \frac{1}{p})(\theta \circ \frac{1}{p})^\top\|_{l_2 \rightarrow l_2}) \\ &= \|B\|_{l_2 \rightarrow l_2} \cdot (\text{Diag}(\sqrt{p_1}, \dots, \sqrt{p_d}) \cdot B\|_F + \|B\|_{l_2 \rightarrow l_2} \cdot \|\theta \circ \frac{1}{p}\|_2^2) \end{aligned}$$

Hence, Bernstein's inequality yields that

$$\begin{aligned} \mathbb{P}\left\{\left|\frac{1}{n^2} \sum_{s=1}^n \left(\|B^\top \text{Diag}(\delta^{(s)}) A_{\theta,p} \text{Diag}(\delta^{(s)}) B\|_F^2 - \mathbb{E}\|B^\top \text{Diag}(\delta^{(s)}) A_{\theta,p} \text{Diag}(\delta^{(s)}) B\|_F^2\right)\right|\right. \\ \left. > K_1(\theta)^2 t\right\} \leq 2 \exp\{-c \min\{n^3 t^2, n^2 t\}\}, \end{aligned}$$

implying that

$$\begin{aligned} \left(\mathbb{E}\left(\frac{1}{n^2} \sum_{s=1}^n \|B^\top \text{Diag}(\delta^{(s)}) A_{\theta,p} \text{Diag}(\delta^{(s)}) B\|_F^2\right)^{r/2}\right)^{2/r} &\lesssim \frac{1}{n} \mathbb{E}\|B^\top \text{Diag}(\delta^{(s)}) A_{\theta,p} \text{Diag}(\delta^{(s)}) B\|_F^2 \\ &\quad + \left(\frac{r}{n^3}\right)^{1/2} K_1^2(\theta) + \frac{r}{n^2} K_1^2(\theta) \end{aligned}$$

Hence, by virtue of (5.2), the first term on the right hand side of (5.1) can be further bounded by

$$\left(\frac{r}{n}\right)^{1/2} K_2(\theta) + \left(\frac{r}{n}\right)^{3/4} K_1(\theta) + \frac{r}{n} K_1(\theta),$$

where

$$K_2(\theta) := \left(\mathbb{E}\|B^\top \text{Diag}(\delta^{(1)}) A_{\theta,p} \text{Diag}(\delta^{(1)}) B\|_F^2\right)^{1/2}.$$

For the second term on the right-side of (5.1), we have

$$\|\tilde{B}^\top \text{Diag}(\tilde{\delta}) \tilde{A}_{\theta,p} \text{Diag}(\tilde{\delta}) \tilde{B}\|_{l_2 \rightarrow l_2} \leq \frac{1}{n} \|B\|_{l_2 \rightarrow l_2}^2 \cdot \|A_{\theta,p}\|_{l_2 \rightarrow l_2}.$$

Hence, we have for $r \geq 1$

$$\|\theta^\top (\hat{\Sigma} - \Sigma) \theta\|_{L_r} \lesssim_\alpha \frac{K_2(\theta) r^{1/2}}{\sqrt{n}} + \frac{K_1(\theta) r^{3/4}}{n^{3/4}} + \frac{K_1(\theta) r^{2/\alpha}}{n},$$

implying that

$$\begin{aligned} &\mathbb{P}\left\{|\theta^\top (\hat{\Sigma} - \Sigma) \theta| > \frac{t^{1/2}}{\sqrt{n}} \sup_{\theta \in \Omega} K_2(\theta) + \frac{t^{3/4}}{n^{3/4}} \sup_{\theta \in \Omega} K_1(\theta) + \frac{t^{2/\alpha}}{n} \sup_{\theta \in \Omega} K_1(\theta)\right\} \\ &\leq \mathbb{P}\left\{|\theta^\top (\hat{\Sigma} - \Sigma) \theta| > \frac{K_2(\theta) t^{1/2}}{\sqrt{n}} + \frac{K_1(\theta) t^{3/4}}{n^{3/4}} + \frac{K_1(\theta) t^{2/\alpha}}{n}\right\} \leq e^{-c(\alpha)t}, \end{aligned}$$

where $c(\alpha)$ is a constant only depending on α .

For the set Ω , there exists an ε -net \mathcal{N} of Ω such that, for $\varepsilon \in (0, 1)$,

$$|\mathcal{N}| \leq \binom{d}{k} (1 + 2/\varepsilon)^k \leq \left(\frac{ed}{k}\right)^k \left(\frac{3}{\varepsilon}\right)^k \leq \left(\frac{3ed}{\varepsilon k}\right)^k.$$

Hence, for any $\theta \in \Omega$, there exists a $\theta' \in \mathcal{N}$ satisfying

$$\begin{aligned} |\theta^\top (\widehat{\Sigma} - \Sigma) \theta| &\leq |(\theta - \theta')^\top (\widehat{\Sigma} - \Sigma) \theta + \theta'^\top (\widehat{\Sigma} - \Sigma) (\theta - \theta')| + |\theta'^\top (\widehat{\Sigma} - \Sigma) \theta'| \\ &\leq 2\varepsilon \sup_{\tilde{\theta} \in \Omega} |\tilde{\theta}^\top (\widehat{\Sigma} - \Sigma) \theta| + |\theta'^\top (\widehat{\Sigma} - \Sigma) \theta'|. \end{aligned}$$

Due to the symmetry of $\widehat{\Sigma} - \Sigma$, we have

$$\tilde{\theta}^\top (\widehat{\Sigma} - \Sigma) \theta = \frac{1}{4} \left((\tilde{\theta} + \theta)^\top (\widehat{\Sigma} - \Sigma) (\tilde{\theta} + \theta) - (\tilde{\theta} - \theta)^\top (\widehat{\Sigma} - \Sigma) (\tilde{\theta} - \theta) \right).$$

Note that $\|\tilde{\theta} \pm \theta\|_2 \leq 2$. Hence, we have

$$\sup_{\theta \in \Omega} |\theta^\top (\widehat{\Sigma} - \Sigma) \theta| \leq 4\varepsilon \sup_{\theta \in \Omega} |\theta^\top (\widehat{\Sigma} - \Sigma) \theta| + \sup_{\theta' \in \mathcal{N}} |\theta'^\top (\widehat{\Sigma} - \Sigma) \theta'|.$$

Taking $\varepsilon = 1/16$, we have

$$\begin{aligned} &\mathbb{P} \left\{ \frac{3}{4} \sup_{\theta \in \Omega} |\theta^\top (\widehat{\Sigma} - \Sigma) \theta| > \frac{t^{1/2}}{\sqrt{n}} \sup_{\theta \in \Omega} K_2(\theta) + \frac{t^{3/4}}{n^{3/4}} \sup_{\theta \in \Omega} K_1(\theta) + \frac{t^{2/\alpha}}{n} \sup_{\theta \in \Omega} K_1(\theta) \right\} \\ &\leq \mathbb{P} \left\{ \sup_{\theta' \in \mathcal{N}} |\theta'^\top (\widehat{\Sigma} - \Sigma) \theta'| > \frac{t^{1/2}}{\sqrt{n}} \sup_{\theta \in \Omega} K_2(\theta) + \frac{t^{3/4}}{n^{3/4}} \sup_{\theta \in \Omega} K_1(\theta) + \frac{t^{2/\alpha}}{n} \sup_{\theta \in \Omega} K_1(\theta) \right\} \\ &\leq |\mathcal{N}| 2e^{-c(\alpha)t} \leq 2 \exp(-c(\alpha)t + k \log(48ed/k)) \end{aligned}$$

We conclude the proof by taking $t = (u + k \log(48ed/k))/c(\alpha)$. □

5.2 Proof of Theorem 4

Proof of Theorem 4. Let $Q = [Q_1, \dots, Q_r] \in \mathbb{R}^{k \times r}$ be a random matrix with independent copies of $r^{-\frac{1}{2}} \delta \circ \xi$, where $\delta \sim \text{Bernoulli}(p)$ and ξ is a centered sub-gaussian random variable satisfying $\|\xi\|_{\Psi_2} \leq 1$ and $\mathbb{E}\xi^2 = 1$. The random variable δ is independent of ξ and $0 \leq p \leq 1$.

Let $X = U\Sigma V^\top$ be the thin singular value decomposition of X , where $U \in \mathbb{R}^{m \times k}$, $\Sigma \in \mathbb{R}^{k \times k}$ and $V \in \mathbb{R}^{n \times k}$. Define the following matrices

$$U\Sigma^{1/2} = \tilde{U} = [\tilde{u}_1^\top, \dots, \tilde{u}_m^\top]^\top, \quad V\Sigma^{1/2} = \tilde{V} = [\tilde{v}_1^\top, \dots, \tilde{v}_n^\top]^\top.$$

Then we have

$$\tilde{u}_i \tilde{v}_j^\top = X_{ij}, \quad \mathbb{E} \tilde{u}_i Q Q^\top \tilde{v}_j^\top = p \cdot X_{ij},$$

where $1 \leq i \leq m$ and $1 \leq j \leq n$. We rewrite $\tilde{u}_i Q Q^\top \tilde{v}_j^\top$ as a quadratic form

$$\tilde{u}_i Q Q^\top \tilde{v}_j^\top = \text{vec}(Q)^\top \cdot (I_r \otimes \tilde{u}_i^\top \tilde{v}_j) \cdot \text{vec}(Q),$$

where $\text{vec}(Q) = [Q_1^\top, \dots, Q_r^\top]^\top \in \mathbb{R}^{kr}$ and \otimes is the Kronecker product. Then, we have by Theorem 1

$$\begin{aligned} \mathbb{P}\left\{|\tilde{u}_i Q Q^\top \tilde{v}_j^\top - p \cdot X_{ij}| > t\right\} &\leq 2 \exp\left(-c \min\left\{\frac{r^2 t^2}{p \|I_r \otimes \tilde{u}_i^\top \tilde{v}_j\|_F^2}, \frac{rt}{\|I_r \otimes \tilde{u}_i^\top \tilde{v}_j\|_{l_2 \rightarrow l_2}}\right\}\right) \\ &\leq 2 \exp\left(-c \min\left\{\frac{rt^2}{p \|\tilde{u}_i\|_2^2 \|\tilde{v}_j\|_2^2}, \frac{rt}{\|\tilde{u}_i\|_2 \|\tilde{v}_j\|_2}\right\}\right). \end{aligned}$$

To obtain the first inequality, we use the following fact:

$$p^2 \sum_{i \neq j} a_{ij}^2 + p \sum_i a_{ii}^2 \leq p \sum_{i,j} a_{ij}^2.$$

Hence, we have for $t > 0$

$$\mathbb{P}\left\{|\tilde{u}_i Q Q^\top \tilde{v}_j^\top - p \cdot X_{ij}| > \|\tilde{u}_i\|_2 \|\tilde{v}_j\|_2 t\right\} \leq \exp(-cr \min\{\frac{t^2}{p}, t\}).$$

Then by the union bound, we have

$$\begin{aligned} \mathbb{P}\left\{\max_{i,j} |\tilde{u}_i Q Q^\top \tilde{v}_j^\top - p \cdot X_{ij}| > \frac{kt}{\sqrt{mn}} \sqrt{\mu_{\text{col}} \mu_{\text{row}}} \cdot \|X\|_{l_2 \rightarrow l_2}\right\} \\ \leq mn \exp(-cr \min\{\frac{t^2}{p}, t\}). \end{aligned}$$

Assume that

$$r > \log \frac{mn}{\eta} \cdot \max\left\{\frac{p}{ct^2}, \frac{1}{ct}\right\},$$

then with probability at least $1 - \eta$, for $1 \leq i \leq m$ and $1 \leq j \leq n$

$$\left|X_{ij} - \frac{1}{p} \tilde{u}_i Q Q^\top \tilde{v}_j^\top\right| \leq \frac{kt}{p \sqrt{mn}} \sqrt{\mu_{\text{col}} \mu_{\text{row}}} \cdot \|X\|_{l_2 \rightarrow l_2}.$$

The desired result follows by setting $Y = (Y_{ij})$ such that $Y_{ij} = \frac{1}{p} \tilde{u}_i Q Q^\top \tilde{v}_j^\top$. \square

Acknowledgment K. Wang was partially supported by Hong Kong RGC grant GRF 16304222. Y. Zhu was partially supported by an AMS-Simons Travel Grant and the Simons Grant MPS-TSM-00013944. This work was partially completed during the first author's visit to Professor Hanchao Wang at Shandong University. The first author would like to express sincere gratitude to Professor Wang for his valuable discussions and support during the visit, which significantly contributed to the development of the ideas presented in this paper.

References

- [1] Pedro Abdalla. Covariance estimation under missing observations and 14-12 moment equivalence. *Electronic Journal of Statistics*, 18(1):2665–2686, 2024.
- [2] Radosław Adamczak and Rafał Łatała. Tail and moment estimates for chaoses generated by symmetric random variables with logarithmically concave tails. In *Annales de l’IHP Probabilités et statistiques*, volume 48, pages 1103–1136, 2012.
- [3] Radosław Adamczak and Paweł Wolff. Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, 162:531–586, 2015.
- [4] Peter J Bickel, Ya’acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. 2009.
- [5] Stanislav Budzinskiy. On the distance to low-rank matrices in the maximum norm. *Linear Algebra and its Applications*, 688:44–58, 2024.
- [6] T Tony Cai and Anru Zhang. Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *Journal of multivariate analysis*, 150:55–74, 2016.
- [7] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n . 2007.
- [8] Abhishek Chakraborty and Arun K Kuchibhotla. Tail bounds for canonical U-statistics and U-processes with unbounded kernels. *arXiv preprint arXiv:2504.01318*, 2025.
- [9] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.
- [10] Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.
- [11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- [12] Evarist Giné, Rafał Łatała, and Joel Zinn. Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II*, pages 13–38. Springer, 2000.
- [13] Friedrich Götze, Holger Sambale, and Arthur Sinulis. Concentration inequalities for polynomials in α -sub-exponential random variables. *Electronic Journal of Probability*, 26, 2021.
- [14] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.
- [15] Yiyun He, Ke Wang, and Yizhe Zhu. Sparse Hanson-Wright inequalities with applications. *arXiv preprint arXiv:2410.15652*, 2024.

- [16] Yukun He, Antti Knowles, and Matteo Marcozzi. Local law and complete eigenvector delocalization for supercritical Erdős–Rényi graphs. *The Annals of Probability*, 47(5):3278–3302, 2019.
- [17] Paweł Hitczenko, Stephen J Montgomery-Smith, and Krzysztof Oleszkiewicz. Moment inequalities for sums of certain independent symmetric random variables. *Studia Math*, 123(1):15–42, 1997.
- [18] Vasileios Kalantzis, Shashanka Ubaru, Chai Wah Wu, Georgios Kollias, and Lior Horesh. Asynchronous randomized trace estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 4294–4302. PMLR, 2024.
- [19] Yegor Klochkov and Nikita Zhivotovskiy. Uniform hanson-wright type concentration inequalities for unbounded entries via the entropy method. *Electron. J. Probab*, 25(22):1–30, 2020.
- [20] Mladen Kolar and Eric P Xing. Estimating sparse precision matrices from data with missing values. In *International Conference on Machine Learning*, pages 635–642, 2012.
- [21] Konrad Kolesko and Rafał Łatała. Moment estimates for chaoses generated by symmetric random variables with logarithmically convex tails. *Statistics & Probability Letters*, 107:210–214, 2015.
- [22] Arun K Kuchibhotla and Abhishek Chakraborty. Moving beyond sub-gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11:1389–1456, 2022.
- [23] Arun Kumar Kuchibhotla and Abhishek Chakraborty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, 2022.
- [24] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.
- [25] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.
- [26] Karim Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, pages 1029–1058, 2014.
- [27] Seongoh Park, Xinlei Wang, and Johan Lim. Estimating high-dimensional covariance and precision matrices under general missing dependence. *Electronic Journal of Statistics*, 15(2):4868–4915, 2021.
- [28] Seongoh Park, Xinlei Wang, and Johan Lim. Sparse Hanson–Wright inequality for a bilinear form of sub-Gaussian variables. *Stat*, 12(1):e539, 2023.
- [29] Eduardo Pavez and Antonio Ortega. Covariance matrix estimation with non uniform and data dependent missing observations. *IEEE Transactions on Information Theory*, 67(2):1201–1215, 2020.

- [30] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.
- [31] Holger Sambale. Some notes on concentration for α -subexponential random variables. In *High Dimensional Probability IX: The Ethereal Volume*, pages 167–192. Springer, 2023.
- [32] Warren Schudy and Maxim Sviridenko. Bernstein-like concentration and moment inequalities for polynomials of independent random variables: multilinear case. *arXiv preprint arXiv:1109.5193*, 2011.
- [33] Warren Schudy and Maxim Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 437–446. SIAM, 2012.
- [34] Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.
- [35] David Wu and Anant Sahai. Precise asymptotic generalization for multiclass classification with overparameterized linear models. *Advances in Neural Information Processing Systems*, 36:42576–42622, 2023.
- [36] Shuheng Zhou. Sparse Hanson–Wright inequalities for subgaussian quadratic forms. *Bernoulli*, 25(3):1603–1639, 2019.
- [37] Shuheng Zhou. Concentration of measure bounds for matrix-variate data with missing values. *Bernoulli*, 30(1):198–226, 2024.

A Proof of Proposition 1.1

Proof. Let $\tilde{\eta} = (\tilde{\eta}_1, \dots, \tilde{\eta}_n)^\top$ be an independent copy of η . Lemma 3.6 and Remark 3.2 yield for $r \geq 1$

$$\|\eta^\top A \eta - \mathbb{E} \eta^\top A \eta\|_{L_r} \asymp \|\eta^\top A \tilde{\eta}\|_{L_r}.$$

By virtue of Lemma 3.4, we have for $r \geq 2$

$$\begin{aligned} \|\eta^\top A \tilde{\eta}\|_{L_r} &\asymp_\alpha \sqrt{r} \|A\|_F + r \|A\|_{l_2 \rightarrow l_2} \\ &\quad + r^{1/\alpha+1/2} \max_i \left(\sum_j a_{ij}^2 \right)^{1/2} + r^{2/\alpha} \|A\|_\infty := f_1(A, r). \end{aligned}$$

Hence, for $r \geq 2$, we have by the Paley-Zygmund inequality

$$\begin{aligned} \mathbb{P} \left\{ |\eta^\top A \eta - \mathbb{E} \eta^\top A \eta| \geq \frac{C_1(\alpha)}{2} f_1(A, r) \right\} &\geq \mathbb{P} \left\{ |\eta^\top A \tilde{\eta}| \geq \frac{1}{2} \|\eta^\top A \tilde{\eta}\|_{L_r} \right\} \\ &\geq (1 - 2^{-r})^2 \left(\frac{\|\eta^\top A \tilde{\eta}\|_{L_r}}{\|\eta^\top A \tilde{\eta}\|_{L_{2r}}} \right)^{2r} \geq \frac{1}{2} e^{-c_1(\alpha)r}, \end{aligned}$$

where in the last inequality, we use Lemma 3.4. If we lower bound $e^{-c_1(\alpha)r}$ by $e^{-c_1(\alpha)(r+2)}$, this inequality is valid for all $r \geq 0$. Hence, we have for $t \geq 0$

$$\mathbb{P}\left\{\left|\eta^\top A\eta - \mathbb{E}\eta^\top A\eta\right| \geq \frac{C_1(\alpha)}{2}f_1(A, t)\right\} \geq \frac{1}{2}e^{-2c_1(\alpha)}e^{-c_1(\alpha)t},$$

which immediately yields the desired result. \square

B Proof of Theorem 5

Proof of Theorem 5. Without loss of generality, we assume $L(\alpha) = 1$. Note that $\mathbb{E}\|A\xi\|_2^2 = p\|A\|_F^2$. Hence, Theorem 1 yields for $u > 0$

$$\begin{aligned} & \mathbb{P}\left\{\left|\|A\xi\|_2^2 - p\|A\|_F^2\right| > t\right\} \\ & \leq 2 \exp\left\{-c(\alpha) \min\left\{\frac{t^2}{p\|A^\top A\|_F^2}, \left(\frac{t}{\|A^\top A\|_{l_2 \rightarrow l_2}}\right)^{\alpha/2}\right\}\right\}. \end{aligned} \quad (\text{B.1})$$

Take $t = up\|A\|_F^2$, $u \geq 0$, and note that

$$\|A^\top A\|_F \leq \|A\|_{l_2 \rightarrow l_2} \|A\|_F, \quad \|A^\top A\|_{l_2 \rightarrow l_2} = \|A\|_{l_2 \rightarrow l_2}^2.$$

Then, (B.1) yields that

$$\begin{aligned} & \mathbb{P}\left\{\left|\|A\xi\|_2^2 - p\|A\|_F^2\right| > up\|A\|_F^2\right\} \\ & \leq 2 \exp\left\{-c(\alpha) \min\left\{\frac{u^2 p\|A\|_F^2}{\|A\|_{l_2 \rightarrow l_2}^2}, \left(\frac{up\|A\|_F^2}{\|A\|_{l_2 \rightarrow l_2}^2}\right)^{\alpha/2}\right\}\right\}. \end{aligned} \quad (\text{B.2})$$

Let $u = \max(\delta, \delta^2)$, $\delta \geq 0$. Then, the event

$$\left\{\left|\|A\xi\|_2^2 - p\|A\|_F^2\right| \leq up\|A\|_F^2\right\}$$

implies the event

$$\left\{\left|\|A\xi\|_2 - \sqrt{p}\|A\|_F\right| \leq \delta\sqrt{p}\|A\|_F\right\},$$

due to the numeric bound $\max(|z-1|, |z-1|^2) \leq |z^2-1|$, $z \geq 0$. Note that $\delta^2 = \min(u, u^2)$, then we have by (B.2)

$$\begin{aligned} & \mathbb{P}\left\{\left|\|A\xi\|_2 - \sqrt{p}\|A\|_F\right| > \delta\sqrt{p}\|A\|_F\right\} \\ & \leq 2 \exp\left\{-c(\alpha) \min\left\{\frac{\delta^2 p\|A\|_F^2}{\|A\|_{l_2 \rightarrow l_2}^2}, \left(\frac{\delta^2 p\|A\|_F^2}{\|A\|_{l_2 \rightarrow l_2}^2}\right)^{\alpha/2}\right\}\right\}. \end{aligned}$$

We finish the proof by taking $t = \delta\sqrt{p}\|A\|_F$. \square

C Proof of Corollary 1

Proof. We first consider the case $0 < \alpha \leq 1$. Note that

$$\|A\|_\infty \leq \|A\|_{l_2 \rightarrow l_\infty} \leq \|A\|_{l_2 \rightarrow l_2}.$$

Therefore, in this case, Corollary 1 can be immediately deduced from Lemma 3.4.

We next turn to the case $1 \leq \alpha \leq 2$. In this case, note that

$$\|A\|_{l_\alpha \rightarrow l_{\alpha^*}} \leq \|A\|_{l_2 \rightarrow l_{\alpha^*}} \leq \|A\|_{l_2 \rightarrow l_2}.$$

Hence, to finish the proof, it is enough to prove for $1 \leq \alpha \leq 2$

$$r^{1/\alpha} \|A\|_{l_{\alpha^*}(l_2)} \leq r^{1/2} \|A\|_F + r \|A\|_{l_2 \rightarrow l_\infty}.$$

Define the following set for $1 \leq \alpha \leq 2$

$$I(r) := \{(x_{ij}) = (z_i y_{ij}) \in \mathbb{R}^{n \times n} : \sum_{i=1}^n |z_i|^\alpha \leq r, \max_{i=1, \dots, n} \sum_{j=1}^n y_{ij}^2 \leq 1\}.$$

We have the following relation

$$r^{1/\alpha} \|A\|_{l_{\alpha^*}(l_2)} = \sup \left\{ \sum_{i,j} a_{ij} x_{ij} : (x_{ij}) \in I(r) \right\}.$$

Indeed, on the one hand, we have

$$\sup_{(x_{ij}) \in I(r)} \sum_{i,j} a_{ij} z_i y_{ij} \leq \sup_{(z_i)} \sum_i z_i \sup_{(y_{ij})} \sum_j a_{ij} y_{ij} \leq r^{1/\alpha} \|A\|_{l_{\alpha^*}(l_2)}.$$

On the other hand, letting

$$y_{ij} = \frac{a_{ij}}{(\sum_{j=1}^n a_{ij}^2)^{1/2}}, \quad z_i = \frac{r^{1/\alpha} (\sum_{j=1}^n a_{ij}^2)^{(\alpha^*-1)/2}}{(\sum_{i=1}^n (\sum_{j=1}^n a_{ij}^2)^{\alpha^*/2})^{1/\alpha}}$$

yields the inverse inequality. Define another subset of $\mathbb{R}^{n \times n}$

$$I_1(r) := \{(x_{ij}) = (z_i y_{ij}) \in \mathbb{R}^{n \times n} : \sum_{i=1}^n \min\{|z_i|^\alpha, z_i^2\} \leq r, \max_{i=1, \dots, n} \sum_{j=1}^n y_{ij}^2 \leq 1\}.$$

Obviously, $I(r) \subset I_1(r)$. Given any (z_i) and (y_{ij}) satisfying the conditions of $I_1(r)$, we have

$$\begin{aligned} \left| \sum_{i,j} a_{ij} z_i y_{ij} \right| &\leq \sum_{i=1}^n |z_i| \left(\sum_{j=1}^n a_{ij}^2 \right)^{1/2} \left(\sum_{j=1}^n y_{ij}^2 \right)^{1/2} \leq \sum_{i=1}^n |z_i| \left(\sum_{j=1}^n a_{ij}^2 \right)^{1/2} \\ &\leq \sum_{i=1}^n |z_i| \mathbb{I}_{\{|z_i| \leq 1\}} \left(\sum_{j=1}^n a_{ij}^2 \right)^{1/2} + \sum_{i=1}^n |z_i| \mathbb{I}_{\{|z_i| > 1\}} \left(\sum_{j=1}^n a_{ij}^2 \right)^{1/2} \\ &\leq \|A\|_F \left(\sum_{i=1}^n z_i^2 \mathbb{I}_{\{|z_i| \leq 1\}} \right)^{1/2} + \|A\|_{l_2 \rightarrow l_\infty} \sum_{i=1}^n |z_i| \mathbb{I}_{\{|z_i| > 1\}} \\ &\leq r^{1/2} \|A\|_F + r \|A\|_{l_2 \rightarrow l_\infty}. \end{aligned}$$

This completes the proof. \square

D Proof of Lemma 3.8

Proof. Let $\eta_1, \dots, \eta_n \stackrel{i.i.d.}{\sim} \mathcal{W}_s(\alpha)$, $0 < \alpha \leq 1$. Note that, for $1 \leq i \leq n$,

$$-\log \mathbb{P}\{|\delta_i \eta_i| \geq t\} = \begin{cases} 0, & t = 0 \\ t^\alpha - \log p_i, & t > 0 \end{cases}$$

is a concave function for $t \geq 0$. Hence, Lemma 3.2 yields for $r \geq 2$

$$\left\| \sum_i a_i \delta_i \eta_i \right\|_{L_r} \asymp_\alpha r^{1/\alpha} \left(\sum_i p_i |a_i|^r \right)^{1/r} + \sqrt{r} \left(\sum_i p_i a_i^2 \right)^{1/2}. \quad (\text{D.1})$$

Here, we use the fact $\|\eta_1\|_{L_r} \asymp r^{1/\alpha}$. As for the first term on the right side of (D.1), we have for $r \geq 3$

$$\begin{aligned} \left(\sum_i p_i |a_i|^r \right)^{1/r} &\leq \max_i |a_i|^{(r-2)/r} \cdot \left(\sum_i p_i a_i^2 \right)^{1/r} \\ &= \left(e^{2r/(r-2)} \max_i |a_i| \right)^{(r-2)/r} \cdot \left(\sum_i p_i a_i^2 e^{-2r} \right)^{1/r} \\ &\leq \frac{r-2}{r} \left(e^{2r/(r-2)} \max_i |a_i| \right) + \frac{2}{r} \left(\sum_i p_i a_i^2 e^{-2r} \right)^{1/2} \\ &\leq e^6 \|a\|_\infty + e^{-r} \left(\sum_i p_i a_i^2 \right)^{1/2}. \end{aligned}$$

This result with (D.1) yields for $r \geq 3$,

$$\left\| \sum_i a_i \delta_i \eta_i \right\|_{L_r} \lesssim_\alpha \sqrt{r} \left(\sum_i p_i a_i^2 \right)^{1/2} + r^{1/\alpha} \|a\|_\infty.$$

For $t \geq 0$ and $1 \leq i \leq n$, we have

$$\mathbb{P}\{|\delta_i \zeta_i| \geq L(\alpha)t\} \leq 2\mathbb{P}\{|\tilde{\delta}_i \eta_i| \geq t\},$$

where $\tilde{\delta}_i$ is an independent copy of δ_i , $1 \leq i \leq n$. Hence, Lemma 3.6 yields for $r \geq 3$

$$\left\| \sum_i a_i \delta_i \zeta_i / L(\alpha) \right\|_{L_r} \lesssim \left\| \sum_i a_i \tilde{\delta}_i \eta_i \right\|_{L_r} \lesssim_\alpha \sqrt{r} \left(\sum_i p_i a_i^2 \right)^{1/2} + r^{1/\alpha} \|a\|_\infty.$$

Then, using Lemma 3.1 and adjusting the universal constant yield the desired result. \square

E Supplement to Theorem 3

In this section, we shall give a detailed expansion of $\mathbb{E} \|B^\top \text{Diag}(\delta^{(1)}) A_{\theta,p} \text{Diag}(\delta^{(1)}) B\|_F^2$ appearing in Theorem 3. For convenience, assume that $B = (b_{ij})_{d \times m}$ and $A_{\theta,p} = (a_{ij})_{d \times d}$.

Note that

$$\begin{aligned}\|B^\top \text{Diag}(\delta^{(1)})A_{\theta,p}\text{Diag}(\delta^{(1)})B\|_F^2 &= \sum_{i=1}^m \sum_{j=1}^m \left(\sum_{l=1}^d \sum_{k=1}^d \delta_l b_{li} a_{lk} \delta_k b_{kj} \right)^2 \\ &= \sum_{i=1}^m \sum_{j=1}^m \sum_{l=1}^d \sum_{p=1}^d \sum_{k=1}^d \sum_{q=1}^d \delta_l b_{li} a_{lk} \delta_k b_{kj} \delta_p b_{pj} a_{pq} \delta_q b_{qj}.\end{aligned}$$

In order to calculate the expectation of the above random variables, we need to make the following classifications based on whether the indicators l, p, k , and q are equal.

- All indicators are different.

$$I_1 = \{(l, p, k, q) : 1 \leq l \neq p \neq k \neq q \leq d\}.$$

- Only two indicators are equal.

$$I_2 = \{(l, p, k, q) : 1 \leq l = p \neq k \neq q \leq d\}, \quad I_3 = \{(l, p, k, q) : 1 \leq l \neq p = k \neq q \leq d\},$$

$$I_4 = \{(l, p, k, q) : 1 \leq l \neq p \neq k = q \leq d\}, \quad I_5 = \{(l, p, k, q) : 1 \leq l = k \neq p \neq q \leq d\},$$

$$I_6 = \{(l, p, k, q) : 1 \leq l = q \neq p \neq k \leq d\}, \quad I_7 = \{(l, p, k, q) : 1 \leq l \neq p = q \neq k \leq d\}.$$

- Only three indicators are equal.

$$I_8 = \{(l, p, k, q) : 1 \leq l = p = k \neq q \leq d\}, \quad I_9 = \{(l, p, k, q) : 1 \leq q = p = k \neq l \leq d\},$$

$$I_{10} = \{(l, p, k, q) : 1 \leq l = q = k \neq p \leq d\}, \quad I_{11} = \{(l, p, k, q) : 1 \leq l = p = q \neq k \leq d\}.$$

- All indicators are equal.

$$I_{12} = \{(l, p, k, q) : 1 \leq l = p = q = k \leq d\}.$$

Hence, we have

$$\mathbb{E}\|B^\top \text{Diag}(\delta^{(1)})A_{\theta,p}\text{Diag}(\delta^{(1)})B\|_F^2 = \sum_{i=1}^m \sum_{j=1}^m \sum_{w=1}^{12} Q_w(i, j).$$

Here,

$$\begin{aligned}
Q_1(i, j) &= \sum_{(l,k,p,q) \in I_1} b_{li} \theta_l \theta_k b_{kj} b_{pj} \theta_p \theta_q b_{qj}, & Q_2(i, j) &= \sum_{(l,k,p,q) \in I_2} \frac{1}{p_l} b_{li} \theta_l \theta_k b_{kj} b_{pj} \theta_p \theta_q b_{qj}, \\
Q_3(i, j) &= \sum_{(l,k,p,q) \in I_3} \frac{1}{p_k} b_{li} \theta_l \theta_k b_{kj} b_{pj} \theta_p \theta_q b_{qj}, & Q_4(i, j) &= \sum_{(l,k,p,q) \in I_4} \frac{1}{p_k} b_{li} \theta_l \theta_k b_{kj} b_{pj} \theta_p \theta_q b_{qj}, \\
Q_5(i, j) &= \sum_{(l,k,p,q) \in I_5} b_{li} \theta_l \theta_k b_{kj} b_{pj} \theta_p \theta_q b_{qj}, & Q_6(i, j) &= \sum_{(l,k,p,q) \in I_6} \frac{1}{p_l} b_{li} \theta_l \theta_k b_{kj} b_{pj} \theta_p \theta_q b_{qj}, \\
Q_7(i, j) &= \sum_{(l,k,p,q) \in I_7} b_{li} \theta_l \theta_k b_{kj} b_{pj} \theta_p \theta_q b_{qj}, & Q_8(i, j) &= \sum_{(l,k,p,q) \in I_8} \frac{1}{p_l} b_{li} \theta_l \theta_k b_{kj} b_{pj} \theta_p \theta_q b_{qj}, \\
Q_9(i, j) &= \sum_{(l,k,p,q) \in I_9} \frac{1}{p_k} b_{li} \theta_l \theta_k b_{kj} b_{pj} \theta_p \theta_q b_{qj}, & Q_{10}(i, j) &= \sum_{(l,k,p,q) \in I_{10}} \frac{1}{p_k} b_{li} \theta_l \theta_k b_{kj} b_{pj} \theta_p \theta_q b_{qj}, \\
Q_{11}(i, j) &= \sum_{(l,k,p,q) \in I_{11}} \frac{1}{p_l} b_{li} \theta_l \theta_k b_{kj} b_{pj} \theta_p \theta_q b_{qj}, & Q_{12}(i, j) &= \sum_{(l,k,p,q) \in I_1} \frac{1}{p_l} b_{li} \theta_l \theta_k b_{kj} b_{pj} \theta_p \theta_q b_{qj}.
\end{aligned}$$