Submitted to Bernoulli

Noise covariance estimation in multi-task high-dimensional linear models

KAI TAN [*1,a], GABRIEL ROMON[2,c] and PIERRE C BELLEC[1,b]

[1]*Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA* , [a]*kai.tan@rutgers.edu*, [b]*pierre.bellec@rutgers.edu*
[2]*CREST, ENSAE, IP Paris, Palaiseau 91120 Cedex, France* , [c]*gabriel.romon@ensae.fr*

This paper studies the multi-task high-dimensional linear regression models where the noise among different tasks is correlated, in the moderately high dimensional regime where sample size n and dimension p are of the same order. Our goal is to estimate the covariance matrix of the noise random vectors, or equivalently the correlation of the noise variables on any pair of two tasks. Treating the regression coefficients as a nuisance parameter, we leverage the multi-task elastic-net and multi-task lasso estimators to estimate the nuisance. By precisely understanding the bias of the squared residual matrix and by correcting this bias, we develop a novel estimator of the noise covariance that converges in Frobenius norm at the rate $n^{-1/2}$ when the covariates are Gaussian distributed with a known covariance matrix. This novel estimator is efficiently computable. Under suitable conditions, the proposed estimator of the noise covariance attains the same rate of convergence as the "oracle" estimator that knows in advance the regression coefficients of the multi-task model. The Frobenius error bounds obtained in this paper also illustrate the advantage of this new estimator compared to a method-of-moments estimator that does not attempt to estimate the nuisance. As byproducts of our techniques, we obtain estimates of the generalization error and out-of-sample error of the multi-task elastic-net and multi-task lasso estimators. Extensive simulation studies are carried out to illustrate the numerical performance of the proposed method.

Keywords: elastic-net; high dimensional analysis; lasso; multi-task model; noise covariance

1. Introduction

1.1. Model and estimation target

Consider a multi-task (also known as multi-response) linear model with T tasks and n i.i.d. observations $(\boldsymbol{x}_i, Y_{i1}, Y_{i2}, \ldots, Y_{iT})$, $\forall i = 1, ..., n$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is a random feature vector and $Y_{i1}, \ldots, Y_{iT}$ are responses in the model

$$\begin{aligned} Y_{it} &= \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(t)} + E_{it} \quad \text{for each } t = 1, ..., T; i = 1, ..., n \quad \text{(scalar form)}, \\ \boldsymbol{y}^{(t)} &= \boldsymbol{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{\varepsilon}^{(t)} \quad \text{for each } t = 1, ..., T \quad \text{(vector form)}, \\ \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E} \quad \text{(matrix form)}, \end{aligned} \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the design matrix with rows $(\boldsymbol{x}_i^\top)_{i=1,...,n}$, $\boldsymbol{y}^{(t)} = (Y_{1t}, ..., Y_{nt})^\top$ is the response vector for task t, $\boldsymbol{\varepsilon}^{(t)} = (E_{1t}, ..., E_{nt})^\top$ is the noise vector for task t, $\boldsymbol{\beta}^{(t)} \in \mathbb{R}^p$ is an unknown fixed coefficient vector for task t. In matrix form, $\boldsymbol{Y} \in \mathbb{R}^{n\times T}$ is the response matrix with columns $\boldsymbol{y}^{(1)}, ..., \boldsymbol{y}^{(T)}$, $\boldsymbol{E} \in \mathbb{R}^{n\times T}$ has columns $\boldsymbol{\varepsilon}^{(1)}, ..., \boldsymbol{\varepsilon}^{(T)}$, and $\boldsymbol{B}^* \in \mathbb{R}^{p\times T}$ is an unknown coefficient matrix with columns $\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(T)}$. The three forms in (1) are equivalent.

While the n vectors $(\boldsymbol{x}_i^\top, y_i^{(1)}, \ldots, y_i^{(T)})_{i=1,...,n}$ of dimension $p + T$ are i.i.d., we assume that for each observation $i = 1, ..., n$, the noise random variables $E_{i1}, ..., E_{iT}$ are centered and correlated. The

focus of the present paper is on estimation of the noise covariance matrix $\boldsymbol{S} \in \mathbb{R}^{T\times T}$, which has entries $\boldsymbol{S}_{tt'} = \mathbb{E}[\varepsilon_1^{(t)}\varepsilon_1^{(t')}]$ for any pair $t, t' = 1, \ldots, T$, or equivalently $\boldsymbol{S} = \frac{1}{n}\mathbb{E}[\boldsymbol{E}^\top \boldsymbol{E}]$.

The noise covariance plays a crucial role in multi-task linear models because it characterizes the noise level and correlation between different tasks: if tasks $t = 1, ..., T$ represent time this captures temporal correlation; if tasks $t = 1, ..., T$ represent different activation areas in the brain (e.g., Bertrand et al. (2019)) this captures spatial correlation.

Since $\boldsymbol{S}$ is the estimation target, we view $\boldsymbol{B}^*$ as an unknown nuisance parameter. If $\boldsymbol{B}^* = \boldsymbol{0}$, then $\boldsymbol{Y} = \boldsymbol{E}$, hence $\boldsymbol{E}$ is directly observed and a natural estimator is the sample covariance $\frac{1}{n}\boldsymbol{E}^\top \boldsymbol{E}$. There are other possible choices for the sample covariance; ours coincides with the maximum likelihood estimator of the centered Gaussian model where the n samples are i.i.d. from $\mathcal{N}_T(\boldsymbol{0}, \boldsymbol{S})$. In the presence of a nuisance parameter $\boldsymbol{B}^* \neq \boldsymbol{0}$, the above sample covariance is not computable since we only observe $(\boldsymbol{X}, \boldsymbol{Y})$ and do not have access to $\boldsymbol{E}$. Thus we will refer to $\frac{1}{n}\boldsymbol{E}^\top \boldsymbol{E} \in \mathbb{R}^{T\times T}$ as the *oracle estimator* for $\boldsymbol{S}$, and its error $\frac{1}{n}\boldsymbol{E}^\top \boldsymbol{E} - \boldsymbol{S}$ will serve as a benchmark.

The nuisance parameter $\boldsymbol{B}^*$ is not of interest by itself, but if an estimator $\widehat{\boldsymbol{B}}$ is available that provides good estimation of $\boldsymbol{B}^*$, we would hope to leverage $\widehat{\boldsymbol{B}}$ to estimate the nuisance and improve estimation of $\boldsymbol{S}$. For instance given an estimate $\widehat{\boldsymbol{B}}$ such that $\|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2/n \to 0$, one may use the estimator

$$\widehat{\boldsymbol{S}}_{(\text{naive})} = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}) \tag{2}$$

to consistently estimate $\boldsymbol{S}$ in Frobenius norm. We refer to this estimator as the *naive estimator* since it is obtained by simply replacing the noise $\boldsymbol{E}$ in the oracle estimator $\frac{1}{n}\boldsymbol{E}^\top \boldsymbol{E}$ with the residual matrix $\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}$. However, in the regime $p/n \to \gamma$ of interest in the present paper, the convergence $\|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2/n \to 0$ does not hold even in the case where $T = 1$ and where $\widehat{\boldsymbol{B}}$ is chosen as the Ridge (Dobriban and Wager, 2018) or the Lasso (Bayati and Montanari, 2012, Miolane and Montanari, 2021): The theory developed in these papers shows that $\|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2/n$ has a non-vanishing limit as p/n converges to a constant. Simulations in Section 4 will show that the naive estimator (2) presents a major bias for estimation of $\boldsymbol{S}$. One goal of this paper is to develop a estimator $\widehat{\boldsymbol{S}}$ of $\boldsymbol{S}$ by exploiting a commonly used estimator $\widehat{\boldsymbol{B}}$ of the nuisance so that in the regime $p/n \to \gamma$ the error $\widehat{\boldsymbol{S}} - \boldsymbol{S}$ is comparable to the benchmark $\frac{1}{n}\boldsymbol{E}^\top \boldsymbol{E} - \boldsymbol{S}$.

1.2. Related literature

If $T = 1$, the above model (1) reduces to the standard linear model with $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and response vector $\boldsymbol{y}^{(1)} \in \mathbb{R}^n$. We will refer to the $T = 1$ case as the single-task linear model and drop the superscript $^{(1)}$ for brevity, i.e., $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \varepsilon_i$, where ε_i are i.i.d. with mean 0, and unknown variance σ^2. The coefficient vector $\boldsymbol{\beta}^*$ is typically assumed to be s-sparse, i.e., $\boldsymbol{\beta}^*$ has at most s nonzero entries. In this single-task linear model, estimation of noise covariance $\boldsymbol{S}$ reduces to estimation of the noise variance $\sigma^2 = \mathbb{E}[\varepsilon_i^2]$, which has been studied in the literature. Fan, Guo and Hao (2012) proposed a consistent estimator for σ^2 based on a refitted cross validation method, which assumes the support of $\boldsymbol{\beta}^*$ is correctly recovered; Belloni, Chernozhukov and Wang (2011) and Sun and Zhang (2012) introduced square-root Lasso (scaled Lasso) to jointly estimate the coefficient $\boldsymbol{\beta}^*$ and noise variance σ^2 by

$$(\hat{\boldsymbol{\beta}}, \widehat{\sigma}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \sigma>0}{\arg\min} \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0\|\boldsymbol{\beta}\|_1. \tag{3}$$

This estimator $\widehat{\sigma}$ is consistent only when the prediction error $\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|^2/n$ goes to 0, which requires $s\log(p)/n \to 0$. Estimation of σ^2 without assumption on $\boldsymbol{X}$ was proposed in Yu and Bien (2019) by

utilizing natural parameterization of the penalized likelihood of the linear model. Their estimator can be expressed as the minimizer of the Lasso problem: $\widehat{\sigma}^2_\lambda = \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + 2\lambda\|\boldsymbol{\beta}\|_1$. Consistency of these estimators Belloni, Chernozhukov and Wang (2011, 2014), Sun and Zhang (2012), Yu and Bien (2019) requires $s\log(p)/n \to 0$ and does not hold in the high-dimensional proportional regime $p/n \to \gamma \in (0,\infty)$. For this proportional regime $p/n \to \gamma \in (0,\infty)$, under the assumption that $\boldsymbol{x}_i$ are i.i.d. $\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})$, Dicker (2014) introduced a method-of-moments estimator $\widehat{\sigma}^2$ of σ^2,

$$\widehat{\sigma}^2 = \frac{n+p+1}{n(n+1)}\|\boldsymbol{y}\|^2 - \frac{1}{n(n+1)}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}^\top\boldsymbol{y}\|^2, \tag{4}$$

which is unbiased, consistent, and asymptotically normal in high-dimensional linear models with Gaussian predictors and errors. Moreover, Janson, Foygel Barber and Candès (2017) developed an EigenPrism procedure for the same task as well as confidence intervals for σ^2. The estimation procedures in these two papers don't attempt to estimate the nuisance parameter $\boldsymbol{\beta}^*$, and require no sparsity on $\boldsymbol{\beta}^*$ and isometry structure, but assume $\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}^*\|^2$ is bounded. Maximum Likelihood Estimators (MLEs) were studied in Dicker and Erdogdu (2016) for joint estimation of noise level and signal strength in high-dimensional linear models with fixed effects; they showed that a classical MLE for random-effects models may also be used effectively in fixed-effects models.

In the proportional regime, Bayati, Erdogdu and Montanari (2013), Miolane and Montanari (2021) used the Lasso to estimate the nuisance $\boldsymbol{\beta}^*$ and produce estimator for σ^2. Their approach requires an uncorrelated Gaussian design assumption with $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Bellec (2020) provided consistent estimators of a similar nature for σ^2 using more general M-estimators with convex penalty without requiring $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. In the special case of the squared loss, this estimator has the form Bayati, Erdogdu and Montanari (2013), Bellec (2020), Miolane and Montanari (2021)

$$\widehat{\sigma}^2 = (n - \hat{\mathsf{df}})^{-2}\big\{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2(n+p-2\hat{\mathsf{df}}) - \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\|\big\}, \tag{5}$$

where $\hat{\mathsf{df}} = \mathrm{Tr}[(\partial/\partial\boldsymbol{y})\boldsymbol{X}\hat{\boldsymbol{\beta}}]$ denotes the degrees of freedom. This estimator coincides with the method-of-moments estimator in Dicker (2014) when $\hat{\boldsymbol{\beta}} = \boldsymbol{0}$.

For multi-task high-dimensional linear model (1) with $T \geq 2$, the estimation of $\boldsymbol{B}^*$ is studied in Lounici et al. (2011), Obozinski, Wainwright and Jordan (2011), Simon, Friedman and Hastie (2013). These works suggest to use a joint convex optimization problem over the tasks to estimate $\boldsymbol{B}^*$. A popular choice is the multi-task elastic-net, which solves the convex optimization problem

$$\widehat{\boldsymbol{B}} = \operatorname*{arg\,min}_{\boldsymbol{B}\in\mathbb{R}^{p\times T}}\Big(\frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{B}\|_{2,1} + \frac{\tau}{2}\|\boldsymbol{B}\|_{\mathrm{F}}^2\Big), \tag{6}$$

where $\|\boldsymbol{B}\|_{2,1} = \sum_{j=1}^p \|\boldsymbol{B}^\top\boldsymbol{e}_j\|_2$, and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix. This optimization problem can be efficiently solved by existing statistical packages, for instance, scikit-learn (Pedregosa et al., 2011), and glmnet (Friedman, Hastie and Tibshirani, 2010). Note that (6) is also referred to as multi-task (group) Lasso and multi-task Ridge if $\tau = 0$ and $\lambda = 0$, respectively. Geer and Stucky (2016) extended square-root Lasso (Belloni, Chernozhukov and Wang, 2011) and scaled Lasso (Sun and Zhang, 2012) to multi-task setting by solving the following problem

$$(\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{S}}) = \operatorname*{arg\,min}_{\boldsymbol{B},\boldsymbol{S}\succ 0}\Big\{\frac{1}{n}\mathrm{Tr}\left((\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})\boldsymbol{S}^{-\frac{1}{2}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^\top\right) + \mathrm{Tr}(\boldsymbol{S}^{\frac{1}{2}}) + 2\lambda_0\|\boldsymbol{B}\|_1\Big\}, \tag{7}$$

where $\|\boldsymbol{B}\|_1 = \sum_{j,t}|B_{jt}|$. Note that the covariance estimator in (7) is constrained to be positive definite. Molstad (2022) studied the same problem and proposed to estimate $\boldsymbol{S}$ by (2) with $\widehat{\boldsymbol{B}}$ in (7), which is

consistent under Frobenius norm loss when $\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_{\mathrm{F}}^2/n \to 0$. Chen and Banerjee (2017) considered another multi-response model with same coefficient for different tasks but different covariates. This literature investigated a similar joint estimation of coefficient and noise covariance. In a recent paper, Bellec and Romon (2021) studied the multi-task Lasso problem and proposed confidence intervals for single entries of $\boldsymbol{B}^*$ and confidence ellipsoids for single rows of $\boldsymbol{B}^*$ under the assumption that $\boldsymbol{S}$ is proportional to the identity, which may be restrictive in practice. This literature generalizes degrees of freedom adjustments from single-task to multi-task models, which we will illustrate in Section 2.

Noise covariance estimation in the high dimensional multi-task linear model is a difficult problem. If the estimand $\boldsymbol{S}$ is known to be diagonal, estimating $\boldsymbol{S}$ reduces to the estimation of noise variance for each task, in which the existing methods for single-task high-dimensional linear models can be applied. Nonetheless, for general positive semi-definite matrix $\boldsymbol{S}$, the noise among different tasks may be correlated, hence the existing methods are not readily applicable, and a more careful analysis is called for to incorporate the correlation between different tasks. Fourdrinier, Haddouche and Mezoued (2021) considered estimating $\boldsymbol{S}$ for the multi-task model (1) where rows of $\boldsymbol{E}$ have elliptically symmetric distribution and in the classical regime $p \le n$. However, their estimator has no statistical guarantee under Frobenius norm loss.

Recently, for the proportional regime $p/n \to \gamma \in (0,\infty)$, Celentano and Montanari (2021) generalized the estimator $\widehat{\sigma}^2$ in Bayati, Erdogdu and Montanari (2013) to the multi-task setting with $T = 2$. Their work covers correlated Gaussian designs, where a Lasso or Ridge regression is used to estimate $\boldsymbol{\beta}^{(1)}$ for the first task, and another Lasso or Ridge regression is used to estimate $\boldsymbol{\beta}^{(2)}$ for the second task. In other words, they estimate the coefficient vector for each task separately instead of using a multi-task estimator like (6). It is not trivial to adapt their estimator from the setting $T = 2$ to larger T, and allow T to increase with n. This present paper takes a different route and aims to fill this gap by proposing a novel noise covariance estimator with theoretical guarantees. Of course, our method applies directly to the 2-task linear model considered in Celentano and Montanari (2021).

1.3. Main contributions

The present paper introduces a novel estimator $\widehat{\boldsymbol{S}}$ in (11) of the noise covariance $\boldsymbol{S}$ when the predictor covariance $\boldsymbol{\Sigma}$ is known. The proposed estimator $\widehat{\boldsymbol{S}}$ is shown to be a consistent estimator of $\boldsymbol{S}$ under Frobenius norm, in the regime where p and n are of the same order. The estimator $\widehat{\boldsymbol{S}}$ is based on the multi-task elastic-net estimator $\widehat{\boldsymbol{B}}$ in (6) of the nuisance, and can be seen as a de-biased version of the naive estimator (2). The naive estimator (2) suffers from a strong bias in the regime where p and n are of the same order, and the estimator $\widehat{\boldsymbol{S}}$ is constructed by precisely understanding this bias and correcting it.

After introducing this novel estimator $\widehat{\boldsymbol{S}}$ in Definition 2.2 below, we prove several rates of convergence for the Frobenius error $\|\widehat{\boldsymbol{S}}-\boldsymbol{S}\|_{\mathrm{F}}$, which is comparable, in terms of rate of convergence, to the benchmark $\|\frac{1}{n}\boldsymbol{E}^\top\boldsymbol{E}-\boldsymbol{S}\|_{\mathrm{F}}$ under suitable assumptions.

As a by-product of the techniques developed for the construction of $\widehat{\boldsymbol{S}}$, we obtain estimates of the generalization error and out-of-sample error of $\widehat{\boldsymbol{B}}$, which are of independent interest and can be used for parameter tuning.

1.4. Notation

Basic notation and definitions that will be used in the rest of the paper are given here. Let $[n] = \{1,2,\dots,n\}$ for all $n \in \mathbb{N}$. The vectors $\boldsymbol{e}_i \in \mathbb{R}^n, \boldsymbol{e}_j \in \mathbb{R}^p, \boldsymbol{e}_t \in \mathbb{R}^T$ denote the canonical basis vector of the corresponding index. We consider restrictions of vectors (resp., of matrices) by zeroing the

corresponding entries (resp., columns). More precisely, for $\boldsymbol{v} \in \mathbb{R}^p$ and index set $B \subset [p]$, $\boldsymbol{v}_B \in \mathbb{R}^p$ is the vector with $(\boldsymbol{v}_B)_j = 0$ if $j \notin B$ and $(\boldsymbol{v}_B)_j = v_j$ if $j \in B$. If $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $B \subset [p]$, $\boldsymbol{X}_B \in \mathbb{R}^{n\times p}$ is such that $(\boldsymbol{X}_B)\boldsymbol{e}_j = \boldsymbol{0}$ if $j \notin B$ and $(\boldsymbol{X}_B)\boldsymbol{e}_j = \boldsymbol{X}\boldsymbol{e}_j$ if $j \in B$. For a real vector $\boldsymbol{a} \in \mathbb{R}^p$, $\|\boldsymbol{a}\|$ denotes its Euclidean norm. For any matrix $\boldsymbol{A}$, $\boldsymbol{A}^\dagger$ is its Moore–Penrose inverse; $\|\boldsymbol{A}\|_{\mathrm{F}}$,$\|\boldsymbol{A}\|_{\mathrm{op}}$, $\|\boldsymbol{A}\|_*$ denote its Frobenius, operator and nuclear norm, respectively. Let $\|\boldsymbol{A}\|_0$ be the number of non-zero rows of $\boldsymbol{A}$. Let $\boldsymbol{A} \otimes \boldsymbol{B}$ be the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$, and $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{B})$ is the Frobenius inner product for matrices of identical size. For $\boldsymbol{A}$ symmetric, $\phi_{\min}(\boldsymbol{A})$ and $\phi_{\max}(\boldsymbol{A})$ denote its smallest and largest eigenvalues, respectively. Let $\boldsymbol{I}_n$ denote the identity matrix of size n for all $n \in \mathbb{N}$. For a random sequence ξ_n, we write $\xi_n = O_P(a_n)$ if ξ_n/a_n is stochastically bounded. C denotes an absolute constant and $C(\tau, \gamma)$ stands for a generic positive constant depending on τ, γ; their expression may vary from place to place.

1.5. Organization

The rest of the paper is organized as follows. Section 2 introduces our proposed estimator for noise covariance. Section 3 presents our main theoretical results on proposed estimator and some relevant estimators. Section 4 demonstrates through numerical experiments that our estimator outperforms several existing methods in the literature, which corroborates our theoretical findings in Section 3. Section 5 provides discussion and points out some future research directions. The appendix provides some technical results and proofs of all theoretical results in the paper. These technical results are futher proved in the supplementary material Tan, Romon and Bellec (2023).

2. Estimating noise covariance, with possibly diverging number of tasks T

Before we can define our noise covariance estimator, we need to introduce the following building blocks. Let $\hat{\mathcal{S}} = \{k \in [p] : \widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k \neq 0\}$ denote the set of nonzero rows of $\widehat{\boldsymbol{B}}$ in (6), and let $|\hat{\mathcal{S}}|$ denote the cardinality of $\hat{\mathcal{S}}$. For each $k \in \hat{\mathcal{S}}$, define $\boldsymbol{H}^{(k)} = \lambda\|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k\|^{-1}(\boldsymbol{I}_T - \widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k \boldsymbol{e}_k^\top \widehat{\boldsymbol{B}}\, \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k\|^{-2})$, which is the Hessian of the map $\boldsymbol{u} \mapsto \lambda\|\boldsymbol{u}\|$ at $\boldsymbol{u} = \widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k$ when $\boldsymbol{u} \neq \boldsymbol{0}$. Define $\boldsymbol{M} \in \mathbb{R}^{pT\times pT}$ by

$$\boldsymbol{M} = \boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + \tau n \boldsymbol{P}_{\hat{\mathcal{S}}}) + n \sum_{k\in\hat{\mathcal{S}}} (\boldsymbol{H}^{(k)} \otimes \boldsymbol{e}_k \boldsymbol{e}_k^\top), \tag{8}$$

where $\boldsymbol{P}_{\hat{\mathcal{S}}} = \sum_{k\in\hat{\mathcal{S}}} \boldsymbol{e}_k \boldsymbol{e}_k^\top \in \mathbb{R}^{p\times p}$. While the expression of $\boldsymbol{M}$ looks a bit long, it is obtained by differentiating the multi-task estimate $\widehat{\boldsymbol{B}}$ in (6) w.r.t. noise E_{it}, in the sense that $\boldsymbol{M}\frac{\partial \mathbf{vec}(\widehat{\boldsymbol{B}})}{\partial E_{it}} = (\boldsymbol{e}_t \otimes \boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{e}_i)$. The detailed derivation is presented in Proof of Lemma E.6. in supplementary file Tan, Romon and Bellec (2023). Define the residual matrix $\boldsymbol{F}$, the error matrix $\boldsymbol{H}$ by

$$\boldsymbol{F} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}, \qquad \boldsymbol{H} = \boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}). \tag{9}$$

To construct our estimator we also make use of the so-called interaction matrix $\widehat{\mathbf{A}} \in \mathbb{R}^{T\times T}$.

Definition 2.1 (Bellec and Romon (2021)). The *interaction matrix* $\widehat{\mathbf{A}} \in \mathbb{R}^{T\times T}$ of the estimator $\widehat{\boldsymbol{B}}$ in (6) is defined by

$$\widehat{\mathbf{A}} = \sum_{i=1}^n (\boldsymbol{I}_T \otimes \boldsymbol{e}_i^\top \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{X}^\top \boldsymbol{e}_i). \tag{10}$$

The matrix $\widehat{\mathbf{A}}$ was introduced in Bellec and Romon (2021), where it is used alongside the multi-task Lasso estimator ($\tau = 0$ in (6)). The interaction matrix $\widehat{\mathbf{A}}$ is essentially the derivatives of the mapping $\boldsymbol{E} \mapsto \boldsymbol{X}\widehat{\boldsymbol{B}}$, in the sense that $\widehat{\mathbf{A}}_{tt'} = \mathrm{Tr}\big(\frac{\partial \boldsymbol{X}\widehat{\boldsymbol{B}}\boldsymbol{e}_t}{\partial \boldsymbol{E}\boldsymbol{e}_{t'}}\big)$. Lemma E.6 in the supplementary material Tan, Romon and Bellec (2023) gives a formal statement of this explanation. It generalizes the degrees of freedom from Stein (1981) to the multi-task case. Intuitively, it captures the correlation between the residuals on different tasks (Bellec and Romon, 2021, Lemma F.1). Our definition of the noise covariance estimator involves $\widehat{\mathbf{A}}$, although our statistical purposes differ greatly from the confidence intervals developed in Bellec and Romon (2021).

With the above definitions, we are now ready to introduce our estimator $\widehat{\boldsymbol{S}}$ of the noise covariance $\boldsymbol{S}$.

Definition 2.2 (Noise covariance estimator). Let $\boldsymbol{F} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}$ and $\widehat{\mathbf{A}}$ be defined as before, we define our new noise covariance estimator as

$$\widehat{\boldsymbol{S}} = (n\boldsymbol{I}_T - \widehat{\mathbf{A}})^{-1}\Big[\boldsymbol{F}^\top\big((p+n)\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top\big)\boldsymbol{F} - \widehat{\mathbf{A}}\boldsymbol{F}^\top\boldsymbol{F} - \boldsymbol{F}^\top\boldsymbol{F}\widehat{\mathbf{A}}\Big](n\boldsymbol{I}_T - \widehat{\mathbf{A}})^{-1}. \tag{11}$$

The construction of above $\widehat{\boldsymbol{S}}$ follows from novel applications of variants of Stein's formulae in order to remove the bias in the naive estimator $\frac{1}{n}\boldsymbol{F}^\top\boldsymbol{F}$. The detailed derivation is presented in the Proof of Theorem 3.3. in Appendix B.2. The estimator $\widehat{\boldsymbol{S}}$ generalizes the scalar estimator (5) to the multi-task setting in the sense that for $T = 1$, $\widehat{\boldsymbol{S}}$ is exactly equal to (5). Note that unlike in (5), here $\boldsymbol{F}^\top\boldsymbol{F}$, $\widehat{\mathbf{A}}$ and $(n\boldsymbol{I}_T - \widehat{\mathbf{A}})$ are matrices of size $T \times T$: the order of matrix multiplication in $\widehat{\boldsymbol{S}}$ matters and should not be switched. This non-commutativity is not present for $T = 1$ in (5) where matrices in $\mathbb{R}^{T\times T}$ are reduced to scalars. Another special case of $\widehat{\boldsymbol{S}}$ can be seen in Celentano and Montanari (2021) for $T = 2$ where the matrix $\widehat{\mathbf{A}} \in \mathbb{R}^{2\times 2}$ is diagonal and the two columns of $\widehat{\boldsymbol{B}} \in \mathbb{R}^{p\times 2}$ are two Lasso or Ridge estimators computed independently of each other, one for each task. Except in these two special cases — (5) for $T = 1$, Celentano and Montanari (2021) for $T = 2$ and two Lasso/Ridge — we are not aware of previously proposed estimators of the same form as $\widehat{\boldsymbol{S}}$.

The definition of our estimator $\widehat{\boldsymbol{S}}$ involves simple algebraic operations between the matrices $\boldsymbol{X}, \boldsymbol{\Sigma}$, the residual $\boldsymbol{F}$ and the interaction matrix $\widehat{\mathbf{A}}$. The multi-task estimte $\widehat{\boldsymbol{B}}$ in (6) can be efficiently solved by existing solver (e.g., `sklearn.linear_model.MultiTaskElasticNet` in scikit-learn library (Pedregosa et al., 2011)), computation of $\boldsymbol{F}$ is then straightforward, and computing the matrix $\widehat{\mathbf{A}}$ only requires inverting a matrix of size $|\hat{\mathcal{S}}|$ thanks to the Sherman-Morrison-Woodbury formula (Bellec and Romon, 2021, Section 5).

3. Theoretical analysis

3.1. Oracle and method-of-moments estimator

Before moving on to the theoretical analysis of $\widehat{\boldsymbol{S}}$, we state our randomness assumptions for $\boldsymbol{E}, \boldsymbol{X}$ and we study two preliminary estimators: the oracle $\frac{1}{n}\boldsymbol{E}^\top\boldsymbol{E}$ and another estimator obtained by the method of moments.

Assumption 1 (Gaussian noise). $\boldsymbol{E} \in \mathbb{R}^{n\times T}$ is a Gaussian noise matrix with i.i.d. $\mathcal{N}_T(\mathbf{0}, \boldsymbol{S})$ rows, where $\boldsymbol{S} \in \mathbb{R}^{T\times T}$ is an unknown positive semi-definite matrix.

An oracle with access to the noise matrix $\boldsymbol{E}$ may compute the oracle estimator $\widehat{\boldsymbol{S}}_{(\text{oracle})} \overset{\text{def}}{=} \frac{1}{n}\boldsymbol{E}^\top\boldsymbol{E}$, with convergence rate given by the following theorem, which will serve as a benchmark.

Proposition 3.1 (Convergence rate of $\widehat{S}_{(\text{oracle})}$). *Under Assumption 1,*

$$\mathbb{E}\big[\|\widehat{S}_{(\text{oracle})} - S\|_{\text{F}}^2\big] = \frac{1}{n}[(\text{Tr}(S))^2 + \text{Tr}(S^2)]. \tag{12}$$

Consequently, $n^{-1}(\text{Tr}(S))^2 \le \mathbb{E}\big[\|\widehat{S}_{(\text{oracle})} - S\|_{\text{F}}^2\big] \le 2n^{-1}(\text{Tr}(S))^2$.

The next assumption concerns the design matrix X with rows $x_1^\top, \dots, x_n^\top$.

Assumption 2 (Gaussian design). $X \in \mathbb{R}^{n\times p}$ is a Gaussian design matrix with i.i.d. $\mathcal{N}_p(0, \Sigma)$ rows, where Σ is a known positive definite matrix. The matrices E and X are independent.

Under the preceding assumptions, we obtain the following method-of-moments estimator, which extends the estimator for noise variance in Dicker (2014) to the multi-task setting. Its error will also serve as a benchmark.

Proposition 3.2. *Under Assumptions 1 and 2, the method-of-moments estimator defined as*

$$\widehat{S}_{(\text{mm})} = \frac{(n+1+p)}{n(n+1)} Y^\top Y - \frac{1}{n(n+1)} Y^\top X \Sigma^{-1} X^\top Y \tag{13}$$

is unbiased for S, i.e., $\mathbb{E}[\widehat{S}_{(\text{mm})}] = S$. Furthermore, the Frobenius error is bounded from below as

$$\mathbb{E}[\|\widehat{S}_{(\text{mm})} - S\|_{\text{F}}^2] \ge \frac{p-2}{(n+1)^2}\big[\,\text{Tr}(S) + \|\Sigma^{\frac{1}{2}} B^*\|_{\text{F}}^2\big]^2. \tag{14}$$

By (14), a larger norm $\|\Sigma^{1/2} B^*\|_{\text{F}}$ induces a larger variance for $\widehat{S}_{(\text{mm})}$. Our goal with an estimate $\widehat{S}$, when a good estimator $\widehat{B}$ of the nuisance is available, is to improve upon the right-hand side of (14) when the estimation error $\|\Sigma^{1/2}(\widehat{B} - B^*)\|_{\text{F}}$ is smaller than $\|\Sigma^{1/2} B^*\|_{\text{F}}$.

A high-probability upper bound of the form $\|\widehat{S}_{(\text{mm})} - S\|_{\text{F}}^2 \le C\frac{n+p}{n^2}[\text{Tr}(S) + \|\Sigma^{\frac{1}{2}} B^*\|_{\text{F}}^2]^2$, that matches the lower bound (14) when $p > n$, is a consequence of our main result Theorem 3.4 in the next subsection. Indeed, when $\widehat{B} = 0$ then $\widehat{A} = 0$ and our estimator $\widehat{S}$ from Definition 2.2 coincides with $\widehat{S}_{(\text{mm})}$ up to the minor modification of replacing $n+1$ by n in (13). This replacement is immaterial compared to the right-hand side in (14). Furthermore, such $\widehat{S}$ corresponds to one of τ or λ being $+\infty$ in (6) and the aforementioned upper bound follows by taking $\tau = +\infty$ in the proof of Theorem 3.4 below. The empirical results in Section 4 confirm that $\widehat{S}$ has smaller variance compared to $\widehat{S}_{(\text{mm})}$ in simulations.

3.2. Theoretical results for proposed estimator $\widehat{S}$

We have established lower bounds for the oracle estimator and the method-of-moments estimator that will serve as benchmarks. We turn to the analysis of the proposed estimator $\widehat{S}$ from Definition 2.2 under the following additional assumptions.

Assumption 3 (High-dimensional regime). n, p satisfy $p/n \le \gamma$ for a constant $\gamma \in (0, \infty)$.

For asymptotic statements such as those involving the stochastically bounded notation $O_p(\cdot)$ or the convergence in probability in (23) below, we implicitly consider a sequence of multi-task problems

indexed by n where $p, T, \boldsymbol{B}^*, \widehat{\boldsymbol{B}}, \boldsymbol{S}$ all implicitly depend on n. The assumptions, such as $p/n \le \gamma$ above, are required to hold at all points of the sequence. In particular, $p/n \to \gamma'$ is allowed for any limit $\gamma' \le \gamma$ under Assumption 3, although our results do not require a specific value for the limit.

Assumption 4. Recall τ is a regularization parameter in problem (6), we assume either one of the following:

i) $\tau > 0$, and let $\tau' = \tau / \|\boldsymbol{\Sigma}\|_{\mathrm{op}}$.

ii) $\tau = 0$ and for $c > 0$, $\mathbb{P}(U_1) \ge 1 - \frac{1}{T}$ and $\mathbb{P}(U_1) \to 1$ as $n \to \infty$, where $U_1 = \{\|\widehat{\boldsymbol{B}}\|_0 \le n(1-c)/2\}$ is the event that $\widehat{\boldsymbol{B}}$ has at most $n(1-c)/2$ nonzero rows. Finally, $T \le e^{\sqrt{n}}$.

iii) $\tau = 0$, $\boldsymbol{\Sigma}_{jj} = 1$ for all $j \in [p]$, $0 < \phi_* \le \phi_{\min}(\boldsymbol{\Sigma}) \le \phi_{\max}(\boldsymbol{\Sigma}) \le \phi^*$, $T \le e^{\sqrt{n}}$, $\|\boldsymbol{B}^*\|_0 \le \min\{c^*(\gamma, \boldsymbol{\Sigma}), c^{**}(\gamma, \boldsymbol{\Sigma})\} n$, $\lambda \ge \mu^*(\gamma, \boldsymbol{\Sigma}) \sqrt{\mathrm{Tr}(\boldsymbol{S})/n}$, where $\mu^*(\gamma, \boldsymbol{\Sigma}) = (30\|\boldsymbol{\Sigma}\|_{\mathrm{op}})^{1/2}(2 + \sqrt{\gamma})$, $c^*(\gamma) = \sup_{c \in [0, \frac{1}{16} \wedge \gamma]} \{c \log(e\gamma/c) \le 1/64\}$, $c^*(\gamma, \boldsymbol{\Sigma}) = \frac{c^*(\gamma)\, \phi_{\min}(\boldsymbol{\Sigma})}{64}$, $c^{**}(\gamma, \boldsymbol{\Sigma}) = \frac{\phi_{\min}(\boldsymbol{\Sigma})}{192 \|\boldsymbol{\Sigma}\|_{\mathrm{op}} (2 + \sqrt{\gamma})^2}$.

Assumption 4(i) requires that the Ridge penalty in (6) be enforced, so that the objective function is strongly convex. Assumption 4(ii), on the other hand, does not require strong convexity but that the number of nonzero rows of $\widehat{\boldsymbol{B}}$ is small enough with high-probability, which is a reasonable assumption when the tuning parameter λ in (6) is large enough and $\boldsymbol{B}^*$ is sparse enough. While we do not prove in the present paper that $\mathbb{P}(U_1) \to 1$ under assumptions on the tuning parameter λ and the sparsity of $\boldsymbol{B}^*$, results of a similar nature have been obtained previously in several group-Lasso settings (Lounici et al., 2011, Theorem 3.1), (Liu and Zhang, 2009, Lemma 6), (Bellec and Romon, 2021, Lemma C.3), (Bellec and Kuchibhotla, 2019, Proposition 3.7). However, the proofs in those papers are not valid for the porportional regime. Under Assumption 4(iii), the following Proposition 3.3 provides a bound for the support size of $\widehat{\boldsymbol{B}}$ and shows that Assumption 4(iii) implies Assumption 4(ii).

Proposition 3.3. *Under Assumption 4(iii), we have* $\mathbb{P}(\|\widehat{\boldsymbol{B}}\|_0 \le n/3) \ge 1 - \exp(-c(\gamma, \boldsymbol{\Sigma})n)$ *for some positive constant* $c(\gamma, \boldsymbol{\Sigma})$ *depending on* $\gamma, \boldsymbol{\Sigma}$ *only.*

Theorem 3.4. *Suppose that Assumptions 1 to 4 hold for all* n, p *as* $n \to \infty$*, then almost surely*

$$\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)(\widehat{\boldsymbol{S}} - \boldsymbol{S})(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)\|_{\mathrm{F}} \le \Theta_1 n^{-\frac{1}{2}} \big(\|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \|\boldsymbol{H}\|_{\mathrm{F}}^2 + \mathrm{Tr}(\boldsymbol{S}) \big) \tag{15}$$

for some non-negative random variable Θ_1 *of constant order, in the sense that* $\mathbb{E}[\Theta_1^2] \le C(\tau')(T \wedge (1 + \frac{p}{n}))(1 + \frac{p}{n})) \le C(\gamma, \tau')$ *under Assumption 4(i), and* $\mathbb{E}[I(\Omega)\Theta_1^2] \le C(\gamma, c)$ *under Assumption 4(ii), where* $I(\Omega)$ *is the indicator function of an event* Ω *with* $\mathbb{P}(\Omega) \to 1$.

Above, $\Theta_1 \ge 0$ is said to be of constant order because $\Theta_1 = O_P(1)$ follows from $\mathbb{E}[\Theta_1^2] \le C(\gamma, \tau')$ or from $\mathbb{E}[I(\Omega)\Theta_1^2] \le C(\gamma, c)$ if the stochastically bounded notation $O_P(1)$ is allowed to hide constants depending on (γ, τ') or (γ, c) only. In the left-hand side of (15), multiplication by $\boldsymbol{I}_T - \widehat{\mathbf{A}}/n$ on both sides of the error $\widehat{\boldsymbol{S}} - \boldsymbol{S}$ can be further removed, as

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}} \le \|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)(\widehat{\boldsymbol{S}} - \boldsymbol{S})(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)\|_{\mathrm{F}} \|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\|_{\mathrm{op}}^2 \tag{16}$$

and the fact that $\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\|_{\mathrm{op}}$ is bounded from above with high probability by a constant depending on γ, τ', c only. Upper bounds on $\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\|_{\mathrm{op}}$ are formally stated in Appendix B.1.

We are now ready to present our main result on the error bounds for $\widehat{\boldsymbol{S}}$.

Theorem 3.5. *Let Assumptions 1 to 4 be fulfilled and $T = o(n)$. Then*

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}} \le O_P(n^{-\frac{1}{2}})(\|\boldsymbol{F}\|_{\mathrm{F}}^2/n), \tag{17}$$

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}} \le O_P(n^{-\frac{1}{2}})[\mathrm{Tr}(\boldsymbol{S}) + \|\boldsymbol{H}\|_{\mathrm{F}}^2]. \tag{18}$$

Here the $O_P(n^{-\frac{1}{2}})$ notation involves constants depending on γ, τ', c.

It is instructive at this point to compare (18) with the lower bound (14) on the Frobenius error of the method-of-moments estimator. When $p \ge n$ then $\mathbb{E}[\|\widehat{\boldsymbol{S}}_{(\mathrm{mm})} - \boldsymbol{S}\|_{\mathrm{F}}^2] \ge \frac{c}{n}[\mathrm{Tr}[\boldsymbol{S}] + \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{B}^*\|_{\mathrm{F}}^2]^2$; this is the situation where the Statistician does not attempt to estimate $\boldsymbol{B}^*$, and pays a price of $[\mathrm{Tr}[\boldsymbol{S}] + \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{B}^*\|_{\mathrm{F}}^2]^2/n$. On the other hand, by definition of $\boldsymbol{H}$ in (9), the right-hand side of (18), when squared, is of order $n^{-1}[\mathrm{Tr}[\boldsymbol{S}] + \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2]^2$. Here the error bound only depends on $\boldsymbol{B}^*$ through the estimation error for the nuisance $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2$. This explains that when $\widehat{\boldsymbol{B}}$ is a good estimator of $\boldsymbol{B}^*$ and $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2$ is smaller compared to $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{B}^*\|_{\mathrm{F}}^2$, the estimator $\widehat{\boldsymbol{S}}$ that leverages $\widehat{\boldsymbol{B}}$ will outperform the method-of-moments estimator $\widehat{\boldsymbol{S}}_{(\mathrm{mm})}$ which does not attempt to estimate the nuisance $\boldsymbol{B}^*$.

Finally, the next results show that under additional assumptions, the estimator $\widehat{\boldsymbol{S}}$ enjoys Frobenius error bounds similar to the oracle estimator $\frac{1}{n}\boldsymbol{E}^\top\boldsymbol{E}$.

Assumption 5. SNR $\le \mathfrak{snr}$ for some positive constant $\mathfrak{snr}$ independent of n, p, T, where $\mathrm{SNR} = \|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{B}^*\|_{\mathrm{F}}^2/\mathrm{Tr}(\boldsymbol{S})$ denotes the signal-to-noise ratio of the multi-task linear model (1).

Corollary 3.6. *Suppose that Assumptions 1, 2, 3, 4(i), 5 and $T = o(n)$ hold, then*

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}} \le O_P(n^{-\frac{1}{2}})\,\mathrm{Tr}(\boldsymbol{S}), \tag{19}$$

where $O_P(\cdot)$ hides constants depending on $\gamma, \tau', \mathfrak{snr}$. Furthermore,

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}}^2 \le O_P(T/n)\|\boldsymbol{S}\|_{\mathrm{F}}^2 = o_P(1)\|\boldsymbol{S}\|_{\mathrm{F}}^2,$$

$$\left|\|\widehat{\boldsymbol{S}}\|_* - \mathrm{Tr}(\boldsymbol{S})\right| \le O_P(\sqrt{T/n})\,\mathrm{Tr}(\boldsymbol{S}) = o_P(1)\,\mathrm{Tr}(\boldsymbol{S}).$$

Corollary 3.7. *Suppose that Assumptions 1, 2, 3, 4(ii) and $T = o(n)$ hold. If $\|\boldsymbol{B}^*\|_0 \le (1-c)n/2$ and the tuning parameter λ is of the form $\lambda = \mu\sqrt{\mathrm{Tr}(\boldsymbol{S})/n}$ for some positive constant μ, then*

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}} \le O_P(n^{-\frac{1}{2}})(1+\mu^2)\,\mathrm{Tr}(\boldsymbol{S}), \tag{20}$$

where $O_P(\cdot)$ hides constants depending on $c, \gamma, \phi_{\min}(\boldsymbol{\Sigma})$.

The following corollary is a direct consequence of Corollary 3.7 and Proposition 3.3.

Corollary 3.8. *Suppose that Assumptions 1, 2, 3, 4(iii) and $T = o(n)$ hold. We have*

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}} \le O_P(n^{-\frac{1}{2}})\,\mathrm{Tr}(\boldsymbol{S}), \tag{21}$$

where $O_P(\cdot)$ hides constants depending on $c, \gamma, \phi_{\min}(\boldsymbol{\Sigma}), \phi_{\max}(\boldsymbol{\Sigma})$.

Comparing Corollaries 3.6 to 3.8 with Proposition 3.1, we conclude that $\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}}^2$ is of the same order as the Frobenius error of the oracle estimator in (12) up to constants depending on the signal-to-noise ratio, γ, and τ' under Assumption 4(i), and up to constants depending on $\mu, c, \gamma, \phi_{\min}(\boldsymbol{\Sigma})$ under Assumption 4(ii).

The error bounds in (18)–(20) are measured in Frobenius norm, similarly to existing works on noise covariance estimation Molstad (2022). Outside the context of linear regression models, much work has been devoted to covariance estimation in the operator norm. By the loose bound $\|\boldsymbol{M}\|_{\mathrm{op}} \le \|\boldsymbol{M}\|_{\mathrm{F}}$, our upper bounds carry over to the operator norm. The same cannot be said for lower bounds, since for instance $\mathbb{E}\big[\|\widehat{\boldsymbol{S}}_{(\mathrm{oracle})} - \boldsymbol{S}\|_{\mathrm{op}}^2\big] \asymp n^{-1}\|\boldsymbol{S}\|_{\mathrm{op}}\operatorname{Tr}(\boldsymbol{S})$ (see, e.g., (Koltchinskii and Lounici, 2017, Corollary 2)).

3.3. Generalization error estimation

By analogy with single task models, we define the *generalization error* in multi-task models as the matrix $\boldsymbol{H}^\top\boldsymbol{H}+\boldsymbol{S}$ of size $T\times T$, whose (t,t')-th entry is $\mathbb{E}[(Y_t^{new} - \boldsymbol{x}_{new}^\top\widehat{\boldsymbol{B}}\boldsymbol{e}_t)(Y_{t'}^{new} - \boldsymbol{x}_{new}^\top\widehat{\boldsymbol{B}}\boldsymbol{e}_{t'})|(\boldsymbol{X},\boldsymbol{Y})]$ where $(Y_t^{new}, Y_{t'}^{new}, \boldsymbol{x}_{new})$ is independent of $(\boldsymbol{X},\boldsymbol{Y})$ and has the same distribution as $(Y_{it}, Y_{it'}, \boldsymbol{x}_i)$ for some $i = 1,\dots,n$. Estimating the generalization error is useful for parameter tuning: since

$$\operatorname{Tr}[\boldsymbol{H}^\top\boldsymbol{H}+\boldsymbol{S}] = \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2 + \operatorname{Tr}[\boldsymbol{S}], \tag{22}$$

minimizing an estimator of $\operatorname{Tr}[\boldsymbol{H}^\top\boldsymbol{H}+\boldsymbol{S}]$ is a useful proxy to minimize the Frobenius error $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2$ of $\widehat{\boldsymbol{B}}$.

The following theorem suggests an estimate for the generalization error matrix as well as a consistent estimator for its trace (22).

Theorem 3.9 (Generalization error). *Let Assumptions 1 to 4 be fulfilled. Then*

$$\|\boldsymbol{F}^\top\boldsymbol{F}/n - (\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)(\boldsymbol{H}^\top\boldsymbol{H}+\boldsymbol{S})(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)\|_{\mathrm{F}} \le \Theta_2 n^{-\frac{1}{2}}\big(\|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \|\boldsymbol{H}\|_{\mathrm{F}}^2 + \operatorname{Tr}(\boldsymbol{S})\big),$$

for some non-negative random variable Θ_2 of constant order, in the sense that $\mathbb{E}[\Theta_2] \le C(\gamma,\tau')$ under Assumption 4(i), and with $\mathbb{E}[I(\Omega)\Theta_2] \le C(\gamma,c)$ under Assumption 4(ii) or (iii), where $I(\Omega)$ is the indicator function of an event Ω with $\mathbb{P}(\Omega)\to 1$.

Furthermore, if $T = o(n)$ as $n, p \to \infty$ while τ', γ, c stay constant, then

$$\frac{\operatorname{Tr}(\boldsymbol{S}) + \|\boldsymbol{H}\|_{\mathrm{F}}^2}{\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{F}^\top\|_{\mathrm{F}}^2/n} \xrightarrow{p} 1. \tag{23}$$

In the above theorem, $\boldsymbol{S}$ and $\boldsymbol{H}$ are unknown, while $\widehat{\mathbf{A}}$ and $\boldsymbol{F}$ can be computed from the observed data $(\boldsymbol{X},\boldsymbol{Y})$. Thus (23) shows that $\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{F}^\top\|_{\mathrm{F}}^2/n$ is a consistent estimate for the unobservable quantity $\operatorname{Tr}(\boldsymbol{S}) + \|\boldsymbol{H}\|_{\mathrm{F}}^2$ without requiring the knowledge of $\boldsymbol{\Sigma}$. Theorem 3.9 also suggests an useful method for selecting tuning parameter, i.e., to choose the paramter that minimizes the estimated generalization error.

3.4. Out-of-sample error estimation

In this section, we present a by-product of our techniques for estimating the noise covariance. For evaluating the performance of a regression method on a new data, we define the out-of-sample error for

the multi-task linear model (1) as

$$\mathbb{E}\big[(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top \boldsymbol{x}_{\text{new}}\boldsymbol{x}_{\text{new}}^\top(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)|(\boldsymbol{X},\boldsymbol{Y})\big]=\boldsymbol{H}^\top\boldsymbol{H},$$

where $\boldsymbol{x}_{\text{new}}$ is independent of the data $(\boldsymbol{X};\boldsymbol{Y})$ with the same distribution as any row of $\boldsymbol{X}$. The following theorem on estimation of out-of-sample error is a by-product of our technique for constructing $\widehat{\boldsymbol{S}}$.

Theorem 3.10 (Out-of-sample error). *Under the same conditions of Theorem 3.4, with* $\boldsymbol{Z}=\boldsymbol{X}\boldsymbol{\Sigma}^{-\frac{1}{2}}$, *we have*

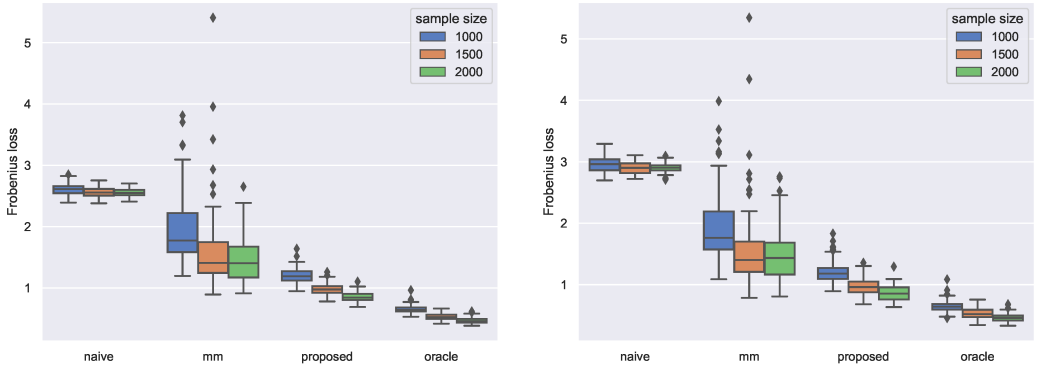
$$\begin{aligned}&\|(\boldsymbol{I}_T-\widehat{\mathbf{A}}/n)\boldsymbol{H}^\top\boldsymbol{H}(\boldsymbol{I}_T-\widehat{\mathbf{A}}/n)-\frac{1}{n^2}\big(\boldsymbol{F}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{F}+\widehat{\mathbf{A}}\boldsymbol{F}^\top\boldsymbol{F}+\boldsymbol{F}^\top\boldsymbol{F}\widehat{\mathbf{A}}-p\boldsymbol{F}^\top\boldsymbol{F}\big)\|_{\text{F}}\\ \le&\Theta_3 n^{-\frac{1}{2}}\big(\|\boldsymbol{H}\|_{\text{F}}^2+\|\boldsymbol{F}\|_{\text{F}}^2/n\big)\end{aligned}$$

for some non-negative random variable Θ_3 *of constant order, in the sense that* $\mathbb{E}[\Theta_3]\le C(\gamma,\tau')$ *under Assumption 4(i), and with* $\mathbb{E}[I(\Omega)\Theta_3]\le C(\gamma,c)$ *under Assumption 4(ii) or (iii), where* $I(\Omega)$ *is the indicator function of an event* Ω *with* $\mathbb{P}(\Omega)\to 1$.

Theorem 3.10 generalizes the result in Bellec (2020) to multi-task setting. While the out-of-sample error $\boldsymbol{H}^\top\boldsymbol{H}$ is unknown, the quantities $\boldsymbol{Z}$, $\boldsymbol{F}$, $\widehat{\mathbf{A}}$ are observable when $\boldsymbol{\Sigma}$ is known. Since typically the quantity $(\|\boldsymbol{H}\|_{\text{F}}^2+\|\boldsymbol{F}\|_{\text{F}}^2/n)$ is of a constant order, Theorem 3.10 suggests the following estimate of $\boldsymbol{H}^\top\boldsymbol{H}$:

$$\frac{1}{n^2}(\boldsymbol{I}_T-\widehat{\mathbf{A}}/n)^{-1}\big(\boldsymbol{F}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{F}+\widehat{\mathbf{A}}\boldsymbol{F}^\top\boldsymbol{F}+\boldsymbol{F}^\top\boldsymbol{F}\widehat{\mathbf{A}}-p\boldsymbol{F}^\top\boldsymbol{F}\big)(\boldsymbol{I}_T-\widehat{\mathbf{A}}/n)^{-1},$$

which can further be used for parameter tuning in multi-task linear models.



(a) Boxplot for estimating a *full rank* $\boldsymbol{S}$

(b) Boxplot for estimating a *low rank* $\boldsymbol{S}$

Figure 1: Boxplots for Frobenius norm loss over 100 repetitions.

Table 1. Frobenius norm loss for different methods

		$n = 1000$		$n = 1500$		$n = 2000$	
$\boldsymbol{S}$	method	mean	sd	mean	sd	mean	sd
full rank	naive	2.610	0.086	2.560	0.075	2.555	0.062
	mm	1.970	0.549	1.586	0.621	1.475	0.381
	proposed	1.206	0.114	0.982	0.098	0.853	0.081
	oracle	0.652	0.061	0.534	0.052	0.469	0.045
low rank	naive	2.960	0.121	2.899	0.100	2.901	0.077
	mm	1.916	0.559	1.560	0.643	1.480	0.428
	proposed	1.208	0.178	0.971	0.130	0.861	0.120
	oracle	0.654	0.096	0.531	0.081	0.464	0.065

4. Numerical experiments

In this section, we evaluate the empirical performance of our proposed method and compare it with relevant methods for noise covariance estimation.

Regarding parameters for our simulations, we set $T = 20$, $p = 1.5n$ and n equals successively $1000, 1500, 2000$. We consider two types of noise covariance matrix: (i) $\boldsymbol{S}$ is full-rank with (t,t')-th entry $\boldsymbol{S}_{t,t'} = \frac{\cos(t-t')}{1+\sqrt{|t-t'|}}$; (ii) $\boldsymbol{S}$ is low-rank with $\boldsymbol{S} = \boldsymbol{u}\boldsymbol{u}^\top$, where $\boldsymbol{u} \in \mathbb{R}^{T\times 10}$ has i.i.d. entries from $\mathcal{N}(0, 1/T)$. The design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is constructed by independently sampling its rows from $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}_{jk} = 0.5^{|j-k|}$. To build the coefficient matrix $\boldsymbol{B}^*$, we first set its sparsity pattern, i.e., we define the support $\mathcal{S}$ with cardinality $|\mathcal{S}| = 0.1p$, then we generate an intermediate matrix $\boldsymbol{B} \in \mathbb{R}^{p\times T}$. The j-th row of $\boldsymbol{B}$ is sampled from $\mathcal{N}_T(\boldsymbol{0}, p^{-1}\boldsymbol{I}_T)$ if $j \in \mathcal{S}$, otherwise we set it to be the zero vector. Finally we let $\boldsymbol{B}^* = \boldsymbol{B}[\mathrm{Tr}(\boldsymbol{S})/\mathrm{Tr}(\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B})]^{\frac{1}{2}}$, which forces a signal-to-noise ratio of exactly 1.

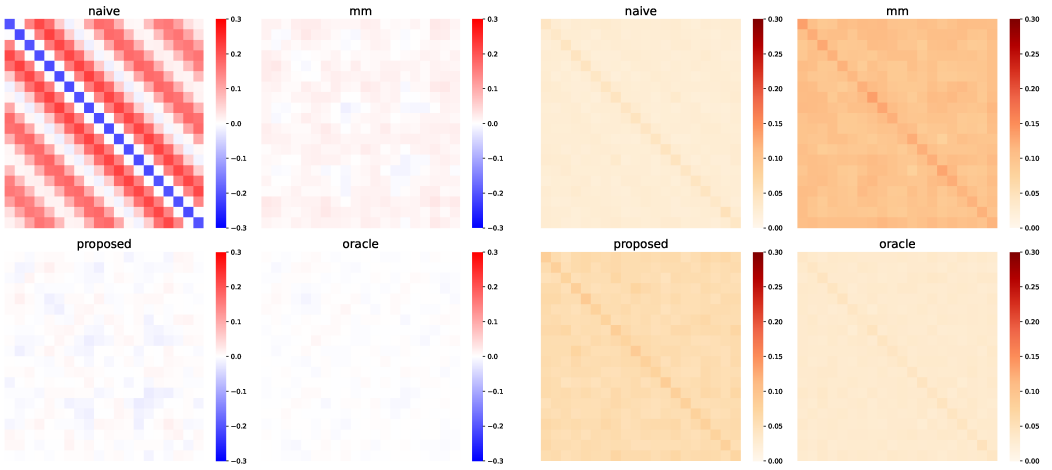
For calculation of multi-task estimates $\widehat{\boldsymbol{B}}$ in (6), we use Python library Scikit-learn (Pedregosa et al., 2011). More precisely we invoke `MultiTaskElasticNetCV` to obtain $\widehat{\boldsymbol{B}}$ by 5-fold cross-validation with parameters `l1_ratio=[0.5, 0.7, 0.9, 1]`, `n_alpha=100`. To compute the interaction matrix $\widehat{\mathbf{A}}$ we use the efficient implementation described in (Bellec and Romon, 2021, Section 5). The full code needed to reproduce our experiments is provided in the supplementary material Tan, Romon and Bellec (2023).

We compare our proposed estimator $\widehat{\boldsymbol{S}}$ (11) with relevant estimators including (1) the naive estimate $\widehat{\boldsymbol{S}}_{(\mathrm{naive})} = n^{-1}\boldsymbol{F}^\top\boldsymbol{F}$, (2) the method-of-moments estimate $\widehat{\boldsymbol{S}}_{(\mathrm{mm})}$ defined in Proposition 3.2, and (3) the oracle estimate $\widehat{\boldsymbol{S}}_{(\mathrm{oracle})} = n^{-1}\boldsymbol{E}^\top\boldsymbol{E}$. The performance of each estimator is measured in Frobenius norm error: for instance, $\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}}$ is the error for proposed estimator $\widehat{\boldsymbol{S}}$. For each configurations of (n, p) and the aforementioned two types of noise covariance $\boldsymbol{S}$, we run 100 repetitions and report in Figure 1 the boxplots of the Frobenius error from different methods, and report in Table 1 the corresponding mean and standard deviation of the Frobenius error.

Figure 1 and Table 1 show that, besides the oracle estimator, our proposed estimator has the best performance with significantly smaller loss compared to the naive and method-of-moments estimators. As expected, oracle estimator outperforms all three other methods. However, since it is typically not available in practice, we recommend using our proposed estimator because it has the nearest performance to oracle estimator.

Since the estimation target $\boldsymbol{S}$ is a $T \times T$ matrix, we also want to compare different estimators in terms of the bias and standard deviation for each entry of $\boldsymbol{S}$. Figure 2 and Figure 3 present the heatmaps of bias and standard deviation from different estimators for full-rank and low-rank $\boldsymbol{S}$ when $n = 1000$. We display

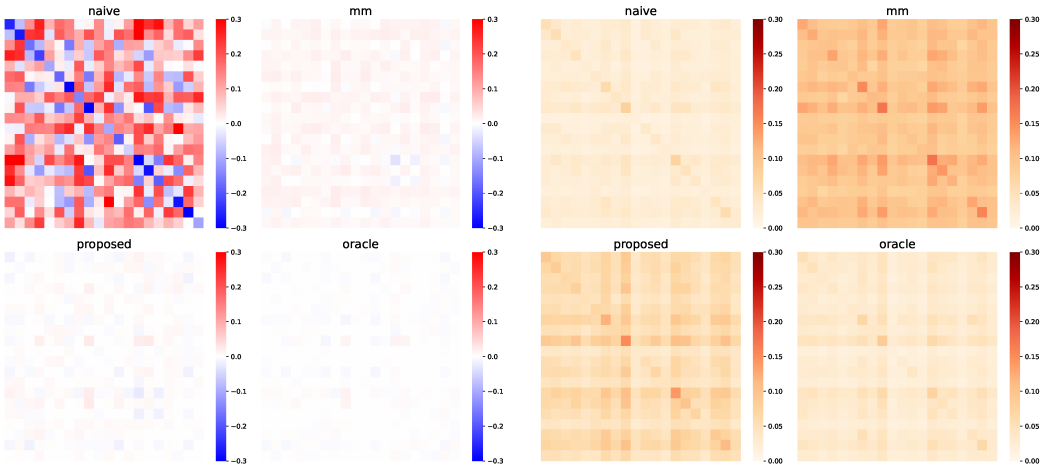
the remaining heatmaps from other n (1500 and 2000) in the supplementary material Tan, Romon and Bellec (2023) as they exhibit similar pattern to Figures 2 and 3.



(a) Heatmap of *bias* for each entry

(b) Heatmap of *standard deviation* for each entry

Figure 2: Heatmaps for estimation of full rank $\boldsymbol{S}$ with $n = 1000$ over 100 repetitions.



(a) Heatmap of *bias* for each entry

(b) Heatmap of *standard deviation* for each entry

Figure 3: Heatmaps for estimation of low rank $\boldsymbol{S}$ with $n = 1000$ over 100 repetitions.

The comparison pattern of the four estimation methods in Figures 2 and 3 are consistent with Figure 1. It is not surprising that the oracle estimator has smallest bias and standard deviation. Figures 2 and 3 also

convince us that the naive estimator has large bias, though it has small standard deviation. The method-of-moments estimator is unbiased but its variance is relatively large, which means its performance is not stable, as was reflected in Figure 1 and Table 1. Our proposed estimator improves on both the naive and method-of-moments estimators because it has much smaller bias than the former, while having smaller standard deviation than the latter. Therefore, our proposed estimator is the clear winner besides the impractical oracle estimator.

5. Limitations and future work

One limitation of the proposed estimator $\widehat{\boldsymbol{S}}$ is that its construction necessitates the knowledge of $\boldsymbol{\Sigma}$. Let us first mention that the estimator $n^{-1}\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{F}^\top\|_{\mathrm{F}}$ of $\mathrm{Tr}(\boldsymbol{S}) + \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2$ in Theorem 3.9 does not require knowing $\boldsymbol{\Sigma}$. Thus, this estimator can further be used as a proxy of the error $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2$, say for parameter tuning, without the knowledge of $\boldsymbol{\Sigma}$. The problem of estimating $\boldsymbol{S}$ with known $\boldsymbol{\Sigma}$ was studied in Celentano and Montanari (2021) for $T = 2$: in this inaccurate covariate model and for $p/n \le \gamma$, our results yield the convergence rate $n^{-1/2}$ for $\boldsymbol{S}$ which improves upon the rate n^{-c_0} for a non-explicit constant $c_0 > 0$ in (Celentano and Montanari, 2021, Theorem 2.1).

In order to use $\widehat{\boldsymbol{S}}$ when $\boldsymbol{\Sigma}$ is unknown, one may plug-in an estimator $\widehat{\boldsymbol{\Sigma}}$ in Equation (11), resulting in an extra term of order $\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_{\mathrm{op}}\|\boldsymbol{F}\|_{\mathrm{F}}$ for the Frobenius error. See (Dicker, 2014, §4) for related discussions in the $T = 1$ (single-task) case. While, under the proportional regime $p/n \to \gamma$, no estimator is consistent for all covariance matrices $\boldsymbol{\Sigma}$ in operator norm, consistent estimators do exist under additional structural assumptions Bickel and Levina (2008), Cai, Zhang and Zhou (2010), El Karoui (2008). If available, additional unlabeled samples $(\boldsymbol{x}_i)_{i\ge n+1}$ can also be used to construct norm-consistent estimator of $\boldsymbol{\Sigma}$.

Future directions include extending estimator $\widehat{\boldsymbol{S}}$ to utilize other estimators of the nuisance $\boldsymbol{B}^*$ than the multi-task elastic-net (6); for instance (7) or the estimators studied in Bertrand et al. (2019), Geer and Stucky (2016), Molstad (2022). In the simpler case where columns of $\boldsymbol{B}^*$ are estimated independently on each task, e.g., if the T columns of $\widehat{\boldsymbol{B}}$ are Lasso estimators $(\hat{\boldsymbol{\beta}}^{(t)})_{t\in[T]}$ each computed from $\boldsymbol{y}^{(t)}$, then minor modifications of our proof yield that the estimator (11) with $\widehat{\mathbf{A}} = \mathbf{diag}(\|\hat{\boldsymbol{\beta}}^{(1)}\|_0, ..., \|\hat{\boldsymbol{\beta}}^{(T)}\|_0)$ enjoys similar Frobenius norm bounds of order $n^{-1/2}$.

Appendix A: Proof of Propositions 3.1 to 3.3

Notation. We first introduce a few notations that will be used throughout the proofs. We use indexes i and l only to loop or sum over $[n] = \{1, 2, \dots, n\}$, use j and k only to loop or sum over $[p] = \{1, 2, \dots, p\}$, use t and t' only to loop or sum over $[T] = \{1, 2, \dots, T\}$, so that $\boldsymbol{e}_i$ (and $\boldsymbol{e}_l$) refer to the i-th (and l-th) canonical basis vector in $\mathbb{R}^n$, $\boldsymbol{e}_j$ (and $\boldsymbol{e}_k$) refer to the j-th (and k-th) canonical basis vector in $\mathbb{R}^p$, $\boldsymbol{e}_t$ (and $\boldsymbol{e}_{t'}$) refer to the t-th (and t'-th) canonical basis vector in $\mathbb{R}^T$. For any two real numbers a and b, let $a \vee b = \max(a, b)$, and $a \wedge b = \min(a, b)$. Positive constants that depend on γ, τ' only are denoted by $C(\gamma, \tau')$, and positive constants that depend on γ, c only are denoted by $C(\gamma, c)$. The values of these constants may vary from place to place.

Proof of Proposition 3.1. Let $\boldsymbol{S} = \sum_{t=1}^T \sigma_t^2 \boldsymbol{u}_t \boldsymbol{u}_t^\top$ be the spectral decomposition of $\boldsymbol{S}$, then $\|\boldsymbol{E}^\top\boldsymbol{E} - n\boldsymbol{S}\|_{\mathrm{F}}^2 = \sum_{t\in[T]}\sum_{t'\in[T]}[\boldsymbol{u}_{t'}^\top((\boldsymbol{E}^\top\boldsymbol{E} - n\boldsymbol{S})\boldsymbol{u}_t]^2$. We now compute the expectation of one term indexed by (t, t'). The random variable $\boldsymbol{u}_{t'}^\top(\boldsymbol{E}^\top\boldsymbol{E} - n\boldsymbol{S})\boldsymbol{u}_t$ is the sum of n i.i.d. mean zero random variables with the same distribution as $z_{t'}z_t - \boldsymbol{u}_{t'}^\top\boldsymbol{S}\boldsymbol{u}_t$ where $(z_t, z_{t'}) \sim \mathcal{N}_2(\mathbf{0}, \mathbf{diag}(\sigma_t^2, \sigma_{t'}^2))$. Thus

$$\mathbb{E}[(\boldsymbol{u}_{t'}^\top(\boldsymbol{E}^\top\boldsymbol{E} - n\boldsymbol{S})\boldsymbol{u}_t)^2] = n\mathrm{Var}[z_{t'}z_t - \boldsymbol{u}_{t'}^\top\boldsymbol{S}\boldsymbol{u}_t] = n2\sigma_t^4 I_{t=t'} + n\sigma_t^2\sigma_{t'}^2 I_{t\ne t'}$$

due to $\mathrm{Var}[\chi_1^2] = 2$ if $t = t'$ and independence if $t \neq t'$. Summing over all $(t, t') \in [T] \times [T]$ yields $2n\sum_{t=1}^T \sigma_t^4 + n\sum_{t\neq t'} \sigma_t^2\sigma_{t'}^2 = n\sum_{t=1}^T \sigma_t^4 + n(\sum_{t=1}^T \sigma_t^2)^2 = n\|\boldsymbol{S}\|_{\mathrm{F}}^2 + n[\mathrm{Tr}(\boldsymbol{S})]^2$ as desired.

The inequality simply follows from $\|\boldsymbol{S}\|_{\mathrm{F}}^2 \leq [\mathrm{Tr}(\boldsymbol{S})]^2$ since $\boldsymbol{S}$ is positive semi-definite. ■

Proof of Proposition 3.2. Without loss of generality, we assume $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. For general positive definite $\boldsymbol{\Sigma}$, the proof follows by replacing $(\boldsymbol{X}, \boldsymbol{B}^*)$ with $(\boldsymbol{X}\boldsymbol{\Sigma}^{-\frac{1}{2}}, \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{B}^*)$.

We first derive the method-of-moments estimator $\widehat{\boldsymbol{S}}_{(\mathrm{mm})}$. Under Assumptions 1 and 2 with $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\boldsymbol{X}$ has i.i.d. rows from $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$, $\boldsymbol{E}$ has i.i.d. rows from $\mathcal{N}_T(\boldsymbol{0}, \boldsymbol{S})$, and $\boldsymbol{X}$ and $\boldsymbol{E}$ are independent. Then, the expectations of $\boldsymbol{Y}^\top\boldsymbol{Y}$ and $\boldsymbol{Y}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}$ are given by

$$\mathbb{E}(\boldsymbol{Y}^\top\boldsymbol{Y}) = \mathbb{E}\big[(\boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E})^\top(\boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E})\big] = n(\boldsymbol{B}^{*\top}\boldsymbol{B}^* + \boldsymbol{S}), \tag{24}$$

and

$$\begin{aligned}
\mathbb{E}(\boldsymbol{Y}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}) &= \mathbb{E}\big[(\boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E})^\top\boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E})\big] \\
&= \mathbb{E}(\boldsymbol{B}^{*\top}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}^*) + \mathbb{E}(\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{E}) \\
&= \boldsymbol{B}^{*\top}\mathbb{E}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{B}^* + \mathbb{E}(\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{E}) \\
&= n(n+p+1)\boldsymbol{B}^{*\top}\boldsymbol{B}^* + np\boldsymbol{S},
\end{aligned} \tag{25}$$

where the last line uses

$$\begin{aligned}
&\mathbb{E}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}) \\
&= \mathbb{E}\Big[\sum_{i=1}^n(\boldsymbol{x}_i\boldsymbol{x}_i^\top)\sum_{l=1}^n(\boldsymbol{x}_l\boldsymbol{x}_l^\top)\Big] \\
&= \sum_{i\neq l}\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{x}_l\boldsymbol{x}_l^\top) + \sum_{i=l}\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{x}_l\boldsymbol{x}_l^\top) \\
&= n(n-1)\boldsymbol{I}_p^2 + n\mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1^\top\boldsymbol{x}_1\boldsymbol{x}_1^\top) \\
&= n(n-1)\boldsymbol{I}_p + n[2\boldsymbol{I}_p^2 + \mathrm{Tr}(\boldsymbol{I}_p)\boldsymbol{I}_p] \\
&= n(n+p+1)\boldsymbol{I}_p,
\end{aligned}$$

and

$$\mathbb{E}(\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{E}) = \mathbb{E}\big[\mathbb{E}(\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{E}|\boldsymbol{E})\big] = \mathbb{E}\big[\boldsymbol{E}^\top\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{E}\big] = np\boldsymbol{S}.$$

Solving for $\boldsymbol{S}$ from the system of equations (24) and (25), we obtain the method-of-moments estimator

$$\widehat{\boldsymbol{S}}_{(\mathrm{mm})} = \frac{(n+p+1)}{n(n+1)}\boldsymbol{Y}^\top\boldsymbol{Y} - \frac{1}{n(n+1)}\boldsymbol{Y}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y},$$

and $\mathbb{E}[\widehat{\boldsymbol{S}}_{(\mathrm{mm})}] = \boldsymbol{S}$.

Now we derive the variance lower bound for $\widehat{\boldsymbol{S}}_{(\mathrm{mm})}$. Since $\mathbb{E}[\widehat{\boldsymbol{S}}_{(\mathrm{mm})}] = \boldsymbol{S}$, $\mathbb{E}\big[\|\widehat{\boldsymbol{S}}_{(\mathrm{mm})} - \boldsymbol{S}\|_{\mathrm{F}}^2\big] = \sum_{t,t'}\mathrm{Var}\big\{[\widehat{\boldsymbol{S}}_{(\mathrm{mm})}]_{t,t'}\big\}$. By definition of $\widehat{\boldsymbol{S}}_{(\mathrm{mm})}$,

$$[\widehat{\boldsymbol{S}}_{(\mathrm{mm})}]_{t,t'} = \frac{n+p+1}{n(n+1)}[\boldsymbol{y}^{(t)}]^\top\boldsymbol{y}^{(t')} - \frac{1}{n(n+1)}[\boldsymbol{y}^{(t)}]^\top\boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top\boldsymbol{y}^{(t')}.$$

Since $\boldsymbol{y}^{(t)} = \boldsymbol{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{\varepsilon}^{(t)}$, $\boldsymbol{y}^{(t')} = \boldsymbol{X}\boldsymbol{\beta}^{(t')} + \boldsymbol{\varepsilon}^{(t')}$, for $t \neq t'$, without loss of generality, we assume $\boldsymbol{\beta}^{(t)} = a_0\boldsymbol{e}_1$ and $\boldsymbol{\beta}^{(t')} = a_1\boldsymbol{e}_1 + a_2\boldsymbol{e}_2$ for some constants a_0, a_1, a_2. If necessary, we could let $\boldsymbol{u}_1 = \boldsymbol{\beta}^{(t)}/\|\boldsymbol{\beta}^{(t)}\|$, and $\boldsymbol{u}_2 = \widetilde{\boldsymbol{u}}_2/\|\widetilde{\boldsymbol{u}}_2\|$ where $\widetilde{\boldsymbol{u}}_2 = \boldsymbol{\beta}^{(t')} - \boldsymbol{P}_{\boldsymbol{u}_1}\boldsymbol{\beta}^{(t')}$, and completing the basis to obtain an orthonormal basis $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p\}$ for $\mathbb{R}^p$. Let $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p]$, then $\boldsymbol{U}$ is an orthogonal matrix, hence $\boldsymbol{X}\boldsymbol{U}$ and $\boldsymbol{X}$ have the same distribution, only the first coordinate of $\boldsymbol{U}^\top\boldsymbol{\beta}^{(t)}$ is nonzero, and only the first two coordinates of $\boldsymbol{U}^\top\boldsymbol{\beta}^{(t')}$ are be nonzero. That is, we could perform change of variables by replacing $(\boldsymbol{X}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t')})$ with $(\boldsymbol{X}\boldsymbol{U}, \boldsymbol{U}^\top\boldsymbol{\beta}^{(t)}, \boldsymbol{U}^\top\boldsymbol{\beta}^{(t')})$.

Therefore, $\boldsymbol{y}^{(t)}$ and $\boldsymbol{y}^{(t')}$ are independent of $\{\boldsymbol{X}\boldsymbol{e}_j : 3 \le j \le p\}$. Let $\mathcal{F} = \sigma(\boldsymbol{y}^{(t)}, \boldsymbol{y}^{(t')}, \boldsymbol{X}\boldsymbol{e}_1, \boldsymbol{X}\boldsymbol{e}_2)$ be the $\sigma-$field generated by $(\boldsymbol{y}^{(t)}, \boldsymbol{y}^{(t')}, \boldsymbol{X}\boldsymbol{e}_1, \boldsymbol{X}\boldsymbol{e}_2)$, then

$$\mathrm{Var}\big\{[\widehat{\boldsymbol{S}}_{(\mathrm{mm})}]_{t,t'}\big\} \ge \mathbb{E}\big[\mathrm{Var}\big\{[\widehat{\boldsymbol{S}}_{(\mathrm{mm})}]_{t,t'}|\mathcal{F}\big\}\big] = \frac{1}{n^2(n+1)^2}\mathbb{E}\big[\mathrm{Var}\big\{[\boldsymbol{y}^{(t)}]^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{y}^{(t')}|\mathcal{F}\big\}\big].$$

Note that in the above display,

$$[\boldsymbol{y}^{(t)}]^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{y}^{(t')} = \sum_{j=1}^{2}[\boldsymbol{y}^{(t)}]^\top\boldsymbol{X}\boldsymbol{e}_j\boldsymbol{e}_j^\top\boldsymbol{X}^\top\boldsymbol{y}^{(t')} + \sum_{j=3}^{p}[\boldsymbol{y}^{(t)}]^\top\boldsymbol{X}\boldsymbol{e}_j\boldsymbol{e}_j^\top\boldsymbol{X}^\top\boldsymbol{y}^{(t')},$$

where the first term is measurable with respect to $\mathcal{F}$, and the second term is a quadratic form

$$\sum_{j=3}^{p}[\boldsymbol{y}^{(t)}]^\top\boldsymbol{X}\boldsymbol{e}_j\boldsymbol{e}_j^\top\boldsymbol{X}^\top\boldsymbol{y}^{(t')} = \sum_{j=3}^{p}\boldsymbol{e}_j^\top\boldsymbol{X}^\top\boldsymbol{y}^{(t')}[\boldsymbol{y}^{(t)}]^\top\boldsymbol{X}\boldsymbol{e}_j = \boldsymbol{\xi}^\top\boldsymbol{\Lambda}\boldsymbol{\xi},$$

here $\boldsymbol{\xi} = [\boldsymbol{e}_3^\top\boldsymbol{X}^\top, \ldots, \boldsymbol{e}_p^\top\boldsymbol{X}^\top]^\top \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{n(p-2)})$, and $\boldsymbol{\Lambda} = \boldsymbol{I}_{p-2} \otimes \boldsymbol{y}^{(t')}[\boldsymbol{y}^{(t)}]^\top$. Thus, for $t \neq t'$,

$$\begin{aligned}
\mathrm{Var}\big\{[\widehat{\boldsymbol{S}}_{(\mathrm{mm})}]_{t,t'}\big\} &\ge \frac{1}{n^2(n+1)^2}\mathbb{E}\Big\{\mathrm{Var}\big\{\boldsymbol{\xi}^\top\boldsymbol{\Lambda}\boldsymbol{\xi}|\mathcal{F}\big\}\Big\} \\
&= \frac{1}{n^2(n+1)^2}\mathbb{E}\Big\{\|\boldsymbol{\Lambda}\|_{\mathrm{F}}^2 + \mathrm{Tr}(\boldsymbol{\Lambda}^2)\Big\} \\
&\ge \frac{1}{n^2(n+1)^2}\mathbb{E}[\|\boldsymbol{\Lambda}\|_{\mathrm{F}}^2] \\
&= \frac{p-2}{n^2(n+1)^2}\mathbb{E}[\|\boldsymbol{y}^{(t)}\|^2\|\boldsymbol{y}^{(t')}\|^2].
\end{aligned}$$

For $t = t'$, using a similar argument we obtain

$$\mathrm{Var}\big\{[\widehat{\boldsymbol{S}}_{(\mathrm{mm})}]_{t,t'}\big\} \ge \frac{p-1}{n^2(n+1)^2}\mathbb{E}[\|\boldsymbol{y}^{(t)}\|^2\|\boldsymbol{y}^{(t')}\|^2].$$

Summing over all $(t, t') \in [T] \times [T]$ yields

$$\begin{aligned}
\mathbb{E}\big[\|\widehat{\boldsymbol{S}}_{(\mathrm{mm})} - \boldsymbol{S}\|_{\mathrm{F}}^2\big] &\ge \frac{p-2}{n^2(n+1)^2}\sum_{t,t'}\mathbb{E}[\|\boldsymbol{y}^{(t)}\|^2\|\boldsymbol{y}^{(t')}\|^2] \\
&= \frac{p-2}{n^2(n+1)^2}\mathbb{E}[\|\boldsymbol{Y}\|_{\mathrm{F}}^4] \\
&\ge \frac{p-2}{n^2(n+1)^2}(\mathbb{E}[\|\boldsymbol{Y}\|_{\mathrm{F}}^2])^2
\end{aligned}$$

$$= \frac{p-2}{(n+1)^2}[\mathrm{Tr}(\boldsymbol{S}) + \|\boldsymbol{B}^*\|^2]^2.$$

■

Proof of Proposition 3.3. Let us first introduce the following two lemmas and defer their proofs to the end of this section.

Lemma A.1 (deterministic bound for $\|\widehat{\boldsymbol{B}}\|_0$). *For the multi-task Lasso estimator*

$$\widehat{\boldsymbol{B}} = \operatorname*{arg\,min}_{\boldsymbol{B}\in\mathbb{R}^{p\times T}} \Big(\frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{B}\|_{2,1}\Big),$$

we have

$$\|\widehat{\boldsymbol{B}}\|_0 \le \|\boldsymbol{X}\|_{\mathrm{op}}^2 \max\left\{\frac{2\|\boldsymbol{E}\|_{\mathrm{F}}^2}{n^2\lambda^2}, \frac{4\|\boldsymbol{B}^*\|_0}{n\kappa^2}\right\},$$

where $\kappa^2 = \inf_{\boldsymbol{B}\in\mathbb{R}^{p\times T}:\|\boldsymbol{B}\|_{2,1}\le\|\boldsymbol{B}^*\|_{2,1}} \frac{\|\boldsymbol{X}(\boldsymbol{B}-\boldsymbol{B}^*)\|_{\mathrm{F}}^2}{n\|\boldsymbol{B}-\boldsymbol{B}^*\|_{\mathrm{F}}^2}$.

Lemma A.1 is a short deterministic argument that provides an upper bound for $\|\widehat{\boldsymbol{B}}\|_0$ of the multi-task Lasso. It follows the same lines as that for the Lasso in (Bellec, 2020, Scetion 10). This argument is sufficient in the $n \asymp p$ regime of the present paper; in the regime $s \lll n \lll p$ regime, sharper bounds are available ((Lounici et al., 2011, Theorem 3.1), (Liu and Zhang, 2009, Lemma 6), (Bellec and Romon, 2021, Lemma C.3), (Bellec and Kuchibhotla, 2019, Proposition 3.7)).

Lemma A.2 is a generalization of (Lecué and Mendelson, 2017, Lemma 2.7) to the multi-task setting and will be useful to provide a lower bound for κ in Lemma A.1.

Lemma A.2. *Let $\boldsymbol{X}\in\mathbb{R}^{n\times p}$, $k\in[p], \delta>0$ be such that*

$$\inf_{\boldsymbol{b}\in\mathbb{R}^p:\|\boldsymbol{b}\|_0\le k} \frac{\|\boldsymbol{X}\boldsymbol{b}\|}{\sqrt{n}\|\boldsymbol{b}\|} \ge \delta, \tag{26}$$

then for any matrix $\boldsymbol{A}\in\mathbb{R}^{p\times T}$ we have

$$(1-\frac{1}{k})\frac{\|\boldsymbol{X}\boldsymbol{A}\|_{\mathrm{F}}^2}{n} \ge \delta^2(1-\frac{1}{k})\|\boldsymbol{A}\|_{\mathrm{F}}^2 - \frac{(\sum_{j=1}^p \|\boldsymbol{A}^\top\boldsymbol{e}_j\|)^2}{k}\Big[\sum_j\Big(\mu_j\frac{\|\boldsymbol{X}\boldsymbol{e}_j\|^2}{n} - \delta^2\Big)\Big],$$

where $\mu_j = \frac{\|\boldsymbol{A}^\top\boldsymbol{e}_j\|}{\sum_j\|\boldsymbol{A}^\top\boldsymbol{e}_j\|}$.

Our first step is to bound κ from below using Lemma A.2. To this end, we first verify condition (26) by finding suitable expressions for k and δ. Note that

$$\inf_{\boldsymbol{b}\in\mathbb{R}^p:\|\boldsymbol{b}\|_0\le k} \frac{\|\boldsymbol{X}\boldsymbol{b}\|}{\sqrt{n}\|\boldsymbol{b}\|} \ge \phi_{\min}^{1/2}(\boldsymbol{\Sigma}) \inf_{\boldsymbol{b}\in\mathbb{R}^p:\|\boldsymbol{b}\|_0\le k} \frac{\|\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}\boldsymbol{b}\|}{\sqrt{n}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{b}\|} = \phi_{\min}^{1/2}(\boldsymbol{\Sigma}) \inf_{\substack{C\subseteq[p]\\|C|=k}} \inf_{\boldsymbol{u}\in V_C} \frac{\|\boldsymbol{Z}\boldsymbol{u}\|}{\sqrt{n}\|\boldsymbol{u}\|},$$

where $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\in\mathbb{R}^{n\times p}$ is a matrix with iid $\mathcal{N}(0,1)$ entries, and $V_C = \{\boldsymbol{\Sigma}^{1/2}\boldsymbol{b} : \boldsymbol{b}\in\mathbb{R}^p, \mathrm{supp}(\boldsymbol{b})\subseteq C\subseteq[p], |C|=k\}$. The linear subspace V_C has dimension k and we consider a matrix $Q\in\mathbb{R}^{p\times k}$ such

that its columns form an orthonormal basis of V_C. By construction, $QQ^\top$ is the orthogonal projection onto V_C and $Q^\top Q = I_k$. For any $\boldsymbol{u} \in V_C$, let $\boldsymbol{v} = Q^\top \boldsymbol{u}$, so that

$$\frac{\|\boldsymbol{Z}\boldsymbol{u}\|}{\sqrt{n}\|\boldsymbol{u}\|} = \frac{\|\boldsymbol{Z}Q\boldsymbol{v}\|}{\sqrt{n}\|\boldsymbol{v}\|} \geq \frac{\phi_{\min}(\boldsymbol{Z}Q)}{\sqrt{n}},$$

where the matrix $\boldsymbol{Z}Q \in \mathbb{R}^{n\times k}$ has iid $\mathcal{N}(0,1)$ entries since $Q^\top Q = I_k$. Using (Davidson and Szarek, 2001, Theorem II.13) on the smallest singular value of a Gaussian matrix, we have for every $t \geq 0$

$$\mathbb{P}(\phi_{\min}(\boldsymbol{Z}Q) \geq \sqrt{n} - \sqrt{k} - t) \geq 1 - e^{-t^2/2}. \tag{27}$$

We take $t = \frac{\sqrt{n}}{4}$, and choose k such that $\sqrt{k} \leq \frac{\sqrt{n}}{4}$, so that $\sqrt{n} - \sqrt{k} - t \geq \frac{\sqrt{n}}{2}$. Taking all the $\binom{p}{k}$ combinations of support set C such that $C \subseteq [p]$ and $|C| = k$, we obtain

$$\mathbb{P}\Big(\inf_{\boldsymbol{b}\in\mathbb{R}^p:\|\boldsymbol{b}\|_0\leq k} \frac{\|\boldsymbol{X}\boldsymbol{b}\|}{\sqrt{n}\|\boldsymbol{b}\|} \geq \phi_{\min}^{1/2}(\boldsymbol{\Sigma})/2\Big) \geq 1 - \binom{p}{k} e^{-n/32} \geq 1 - e^{k\log(ep/k) - n/32}.$$

We take $k = c^*(\gamma)n$ with $c^*(\gamma) = \sup_{c\in[0,\frac{1}{16}\wedge\gamma]}\{c\log(e\gamma/c) \leq 1/64\}$, then

$$\mathbb{P}\Big(\inf_{\boldsymbol{b}\in\mathbb{R}^p:\|\boldsymbol{b}\|_0\leq k} \frac{\|\boldsymbol{X}\boldsymbol{b}\|}{\sqrt{n}\|\boldsymbol{b}\|} \geq \phi_{\min}^{1/2}(\boldsymbol{\Sigma})/2\Big) \geq 1 - e^{-n/64}.$$

With $\delta = \phi_{\min}^{1/2}(\boldsymbol{\Sigma})/2$, and $k = c^*(\gamma)n$, then condition (26) holds with probability at least $1 - e^{-n/64}$. In other words, letting

$$A_0 = \{\inf_{\boldsymbol{b}\in\mathbb{R}^p:\|\boldsymbol{b}\|_0\leq c^*(\gamma)n} \frac{\|\boldsymbol{X}\boldsymbol{b}\|}{\sqrt{n}\|\boldsymbol{b}\|} \geq \phi_{\min}^{1/2}(\boldsymbol{\Sigma})/2\},$$

then $\mathbb{P}(A_0) \geq 1 - e^{-n/64}$. Now we are ready to bound κ^2 in Lemma A.1 using Lemma A.2. To this end, we first derive a cone condition for $\boldsymbol{B} - \boldsymbol{B}^*$ using $\|\boldsymbol{B}\|_{2,1} \leq \|\boldsymbol{B}^*\|_{2,1}$. By rearranging the terms and Cauchy-Schwarz inequality, we have

$$\sum_{j\in\mathcal{S}^c} \|\boldsymbol{B}^\top \boldsymbol{e}_j\| \leq \sum_{j\in\mathcal{S}} (\|\boldsymbol{B}^{*\top}\boldsymbol{e}_j\| - \|\boldsymbol{B}^\top \boldsymbol{e}_j\|) \leq \sum_{j\in\mathcal{S}} \|(\boldsymbol{B} - \boldsymbol{B}^*)^\top \boldsymbol{e}_j\| \leq \sqrt{|\mathcal{S}|}\|\boldsymbol{B} - \boldsymbol{B}^*\|_{\mathrm{F}}.$$

Since $\sum_{j\in\mathcal{S}^c}\|(\boldsymbol{B} - \boldsymbol{B}^*)^\top \boldsymbol{e}_j\| = \sum_{j\in\mathcal{S}^c}\|\boldsymbol{B}^\top \boldsymbol{e}_j\| \leq \sqrt{|\mathcal{S}|}\|\boldsymbol{B} - \boldsymbol{B}^*\|_{\mathrm{F}}$, we obtain the cone condition

$$\sum_{j=1}^{p} \|(\boldsymbol{B} - \boldsymbol{B}^*)^\top \boldsymbol{e}_j\| \leq 2\sqrt{|\mathcal{S}|}\|\boldsymbol{B} - \boldsymbol{B}^*\|_{\mathrm{F}}. \tag{28}$$

Let $A_1 = \big\{\max_{j\in[p]} \frac{\|\boldsymbol{X}\boldsymbol{e}_j\|^2}{n} \leq 1.1\big\}$, then under the assumption that $\boldsymbol{\Sigma}_{jj} = 1$ for all $j \in [p]$, we have $\mathbb{P}(A_1) \geq 1 - e^{-c(\gamma)n}$ by (Laurent and Massart, 2000, Lemma 1). Now we apply Lemma A.2 to $\boldsymbol{A} = \boldsymbol{B} - \boldsymbol{B}^*$, using the cone (28) and $\sum_j \mu_j = 1$, we have on the event A_1,

$$\frac{\|\boldsymbol{X}(\boldsymbol{B} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2}{n} \geq \delta^2(1 - \frac{1}{k})\|\boldsymbol{B} - \boldsymbol{B}^*\|_{\mathrm{F}}^2 - \frac{4.4|\mathcal{S}|}{k}\|\boldsymbol{B} - \boldsymbol{B}^*\|_{\mathrm{F}}^2.$$

Since $1-\frac{1}{k}=1-\frac{1}{c^*(\gamma)n}\geq 1/2$ for n large enough, with $|\mathscr{S}|\leq \frac{k\delta^2}{17.6}\leq \frac{k\delta^2}{16}$, we have

$$\frac{\|\boldsymbol{X}(\boldsymbol{B}-\boldsymbol{B}^*)\|_{\mathrm{F}}^2}{n}\geq \frac{\delta^2}{4}\|\boldsymbol{B}-\boldsymbol{B}^*\|_{\mathrm{F}}^2, \tag{29}$$

and thus $\kappa\geq\frac{\delta}{2}$ with $\delta=\phi_{\min}^{1/2}(\boldsymbol{\Sigma})/2$.

Now we are ready to bound $\|\widehat{\boldsymbol{B}}\|_0$ using Lemma A.1. Let $A_2=\{\|\boldsymbol{X}^\top\boldsymbol{X}\|_{\mathrm{op}}\leq\|\boldsymbol{\Sigma}\|_{\mathrm{op}}(2+\sqrt{\gamma})^2 n\}$, $A_3=\{\|\boldsymbol{E}\|_{\mathrm{F}}^2\leq 5\,\mathrm{Tr}(\boldsymbol{S})n\}$, then $\mathbb{P}(A_2)\geq 1-\exp(-n/2)$ by (Davidson and Szarek, 2001, Theorem II.13), and $\mathbb{P}(A_3)\geq 1-\exp(-n)$ by (Laurent and Massart, 2000, Lemma 1). On the event $A_0\cap A_1\cap A_2\cap A_3$, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{B}}\|_0 &\leq \|\boldsymbol{X}\|_{\mathrm{op}}^2\max\left\{\frac{2\|\boldsymbol{E}\|_{\mathrm{F}}^2}{n^2\lambda^2},\frac{4\|\boldsymbol{B}^*\|_0}{n\kappa^2}\right\}\\
&\leq \|\boldsymbol{\Sigma}\|_{\mathrm{op}}(2+\sqrt{\gamma})^2\max\left\{\frac{10\,\mathrm{Tr}(\boldsymbol{S})}{\lambda^2},\frac{4\|\boldsymbol{B}^*\|_0}{\kappa^2}\right\}\\
&\leq \|\boldsymbol{\Sigma}\|_{\mathrm{op}}(2+\sqrt{\gamma})^2\max\left\{\frac{10n}{\mu^2},\frac{64\|\boldsymbol{B}^*\|_0}{\phi_{\min}(\boldsymbol{\Sigma})}\right\},
\end{aligned}$$

where the last inequality uses $\lambda=\mu\sqrt{\mathrm{Tr}(\boldsymbol{S})/n}$, and $\kappa\geq\delta/2=\phi_{\min}^{1/2}(\boldsymbol{\Sigma})/4$.

To guarantee $\|\widehat{\boldsymbol{B}}\|_0\leq n/3$, we need (1) $\|\boldsymbol{\Sigma}\|_{\mathrm{op}}(2+\sqrt{\gamma})^2\frac{10}{\mu^2}\leq 1/3$, that is, $\mu\geq\mu^*(\gamma,\boldsymbol{\Sigma}):=(30\|\boldsymbol{\Sigma}\|_{\mathrm{op}})^{1/2}(2+\sqrt{\gamma})$, and (2) $\|\boldsymbol{\Sigma}\|_{\mathrm{op}}(2+\sqrt{\gamma})^2\frac{64\|\boldsymbol{B}^*\|_0}{\phi_{\min}(\boldsymbol{\Sigma})}\leq n/3$, that is, $\|\boldsymbol{B}^*\|_0\leq\frac{\phi_{\min}(\boldsymbol{\Sigma})}{192\|\boldsymbol{\Sigma}\|_{\mathrm{op}}(2+\sqrt{\gamma})^2}n$. Recall that we also need $\|\boldsymbol{B}^*\|_0\leq\frac{k\delta^2}{16}\leq\frac{c^*(\gamma)\phi_{\min(\boldsymbol{\Sigma})}}{64}n:=c^*(\gamma,\boldsymbol{\Sigma})n$ to ensure (29). So under the Assumption 4(iii), on the event $A_0\cap A_1\cap A_2\cap A_3$, we have $\|\widehat{\boldsymbol{B}}\|_0\leq n/3$. Using the union bound for the events A_0,A_1,A_2,A_3, we have

$$\mathbb{P}\Big(\|\widehat{\boldsymbol{B}}\|_0\leq n/3\Big)\geq 1-\exp(-c(\gamma)n)>1-\frac{1}{T}$$

if $T\leq e^{\sqrt{n}}$. ■

To complete the proof of Proposition 3.3, we now provide proofs for Lemmas A.1 and A.2.

Proof of Lemma A.1. The KKT conditions for $\widehat{\boldsymbol{B}}$ read $\boldsymbol{X}^\top(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})=n\lambda\partial\|\widehat{\boldsymbol{B}}\|_{2,1}$. Let $\hat{\mathscr{S}}$ be the support of $\widehat{\boldsymbol{B}}$, then for each $j\in\hat{\mathscr{S}}$, we have $\boldsymbol{e}_j^\top\boldsymbol{X}^\top(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})=n\lambda\frac{\boldsymbol{e}_j^\top\widehat{\boldsymbol{B}}}{\|\widehat{\boldsymbol{B}}^\top\boldsymbol{e}_j\|}$. Taking squared norm on both sides and summing over $j\in\hat{\mathscr{S}}$ gives

$$(n\lambda)^2|\hat{\mathscr{S}}|=\sum_{j\in\hat{\mathscr{S}}}\|\boldsymbol{e}_j^\top\boldsymbol{X}^\top(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})\|^2\leq\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}}\|_{\mathrm{F}}^2\|\boldsymbol{X}^\top\boldsymbol{X}\|_{\mathrm{op}}. \tag{30}$$

Now we bound $\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}}\|_{\mathrm{F}}^2$. Multplying the KKT conditions by $\widehat{\boldsymbol{B}}-\boldsymbol{B}^*$ gives

$$\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_{\mathrm{F}}^2+\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}}\|_{\mathrm{F}}^2\leq\|\boldsymbol{E}\|_{\mathrm{F}}^2+2n\lambda(\|\boldsymbol{B}^*\|_{2,1}-\|\widehat{\boldsymbol{B}}\|_{2,1}).$$

We bound the RHS by comparing the two terms. If $\|\boldsymbol{E}\|_{\mathrm{F}}^2\geq 2n\lambda(\|\boldsymbol{B}^*\|_{2,1}-\|\widehat{\boldsymbol{B}}\|_{2,1})$, then

$$\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}}\|_{\mathrm{F}}^2\leq(RHS)\leq 2\|\boldsymbol{E}\|_{\mathrm{F}}^2.$$

Otherwise $\|\boldsymbol{E}\|_{\mathrm{F}}^2 \leq 2n\lambda(\|\boldsymbol{B}^*\|_{2,1} - \|\widehat{\boldsymbol{B}}\|_{2,1})$, then we have

$$\begin{aligned}\|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2 + \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}\|_{\mathrm{F}}^2 &\leq 4n\lambda(\|\boldsymbol{B}^*\|_{2,1} - \|\widehat{\boldsymbol{B}}\|_{2,1}) \\ &\leq 4n\lambda\sqrt{|\mathscr{S}|}\|\boldsymbol{B}^* - \widehat{\boldsymbol{B}}\|_{\mathrm{F}} \\ &\leq 4\sqrt{n}\lambda\sqrt{|\mathscr{S}|}\kappa^{-1}\|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}.\end{aligned}$$

where $\mathscr{S}$ is the support of $\boldsymbol{B}^*$. Using $4ab \leq 4a^2 + b^2$ with $a = \sqrt{n}\lambda\sqrt{|\mathscr{S}|}\kappa^{-1}$ and $b = \|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}$, then the term b^2 cancels out in both sides. We have

$$\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}\|_{\mathrm{F}}^2 \leq 4n\lambda^2|\mathscr{S}|\kappa^{-2}.$$

Plugging the maximum of the previous two bounds for $\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}\|_{\mathrm{F}}^2$ into (30) gives the desired inequality. ■

Proof of Lemma A.2. Replacing $\boldsymbol{A}$ by $\boldsymbol{A}/\sum_{j=1}^p \|\boldsymbol{A}^\top \boldsymbol{e}_j\|$ if necessary, we assume with out of generality that $\sum_{j=1}^p \|\boldsymbol{A}^\top \boldsymbol{e}_j\| = 1$ and we also assume for simplicity that $\forall j \in [p]$, $\boldsymbol{A}^\top \boldsymbol{e}_j \neq 0$. Define a discrete random matrix $\widetilde{\mathbf{A}}$ by $\mathbb{P}(\widetilde{\mathbf{A}} = \boldsymbol{e}_j \boldsymbol{e}_j^\top \boldsymbol{A}/\mu_j) = \mu_j$, where $\mu_j = \|\boldsymbol{A}^\top \boldsymbol{e}_j\|$. Then we have $\mathbb{E}\widetilde{\mathbf{A}} = \boldsymbol{A}$. Let $\widetilde{\mathbf{A}}_1, \ldots \widetilde{\mathbf{A}}_k$ be iid copies of $\widetilde{\mathbf{A}}$ and set $\bar{\boldsymbol{A}} = \frac{1}{k}\sum_{m=1}^k \widetilde{\mathbf{A}}_m$, then $\bar{\boldsymbol{A}}$ is $k-$row sparse by construction and condition (26) gives $\|\boldsymbol{X}\bar{\boldsymbol{A}}\|_{\mathrm{F}}^2/n \geq \delta^2\|\bar{\boldsymbol{A}}\|_{\mathrm{F}}^2$. Hence

$$\mathbb{E}[\|\boldsymbol{X}\bar{\boldsymbol{A}}\|_{\mathrm{F}}^2/n] \geq \delta^2\mathbb{E}[\|\bar{\boldsymbol{A}}\|_{\mathrm{F}}^2]. \tag{31}$$

Note that all expectations are conditionally on $\boldsymbol{X}$ and the randomness stems from $\widetilde{\mathbf{A}}_1, \ldots, \widetilde{\mathbf{A}}_k$. Now we calculate the expectations (conditionally on $\boldsymbol{X}$) by expanding the squares. For $u \neq v$, we have $\mathbb{E}\langle \boldsymbol{X}\widetilde{\mathbf{A}}_u, \boldsymbol{X}\widetilde{\mathbf{A}}_v\rangle = \|\boldsymbol{X}\boldsymbol{A}\|_{\mathrm{F}}^2$ by independence. For $u = v$, using $\mu_j = \|\boldsymbol{A}^\top \boldsymbol{e}_j\|$, we have

$$\mathbb{E}[\|\boldsymbol{X}\widetilde{\mathbf{A}}_u\|_{\mathrm{F}}^2/n] = \sum_j \mu_j\|\boldsymbol{X}\boldsymbol{e}_j\boldsymbol{e}_j^\top\boldsymbol{A}\|_{\mathrm{F}}^2/(\mu_j^2 n) = \sum_j \mu_j\|\boldsymbol{X}\boldsymbol{e}_j\|^2/n.$$

Therefore,

$$\mathbb{E}[\|\boldsymbol{X}\bar{\boldsymbol{A}}\|_{\mathrm{F}}^2/n] = \Big(1 - \frac{1}{k}\Big)\|\boldsymbol{X}\boldsymbol{A}\|_{\mathrm{F}}^2/n + \frac{1}{k}\sum_j \mu_j\|\boldsymbol{X}\boldsymbol{e}_j\|^2/n.$$

The same argument gives

$$\mathbb{E}[\|\bar{\boldsymbol{A}}\|_{\mathrm{F}}^2/n] = \Big(1 - \frac{1}{k}\Big)\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \frac{1}{k}\sum_j \mu_j = \Big(1 - \frac{1}{k}\Big)\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \frac{1}{k}.$$

Plugging the above two equalities into (31) gives the desired bound. ■

Appendix B: Proof of main results

We first present two lemmas and three propositions in Appendix B.1. Based on these technical results, we prove the main results of the paper in Appendix B.2. The proofs of technical results in Appendix B.1 are deferred to the supplementary material Tan, Romon and Bellec (2023).

B.1. Preliminary results

The following Lemmas B.1 and B.2 provide operator norm bounds for $\boldsymbol{I}_T - \widehat{\mathbf{A}}/n$ and $(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}$ under suitable conditions.

Lemma B.1. *Suppose that Assumption 2 holds. If $\tau > 0$ in (6) with $\tau' = \tau/\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$, then*

(i) $\|\boldsymbol{I}_T - \widehat{\mathbf{A}}/n\|_{\mathrm{op}} \le 1$.

(ii) *In the event $\big\{\|\boldsymbol{X}\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_{\mathrm{op}} < 2\sqrt{n} + \sqrt{p}\big\}$, we have $\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\|_{\mathrm{op}} \le 1 + (\tau')^{-1}(2 + \sqrt{p/n})^2$. Furthermore, $\mathbb{E}[\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\|_{\mathrm{op}}] \le 1 + (\tau')^{-1}[(1 + \sqrt{p/n})^2 + n^{-1}]$.*

Lemma B.2. *Suppose that Assumption 2 holds. If $\tau = 0$ in (6), then*

(i) *In the event U_1, we have $\|\boldsymbol{I}_T - \widehat{\mathbf{A}}/n\|_{\mathrm{op}} \le 1$.*

(ii) *In the event U_1, $\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\|_{\mathrm{op}} \le C(c)$. Hence, $\mathbb{E}[I(U_1)\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\|_{\mathrm{op}}] \le C(c)$.*

The proof of main results rely on the following three propositions, which are derived from non-trivial applications of second-order Stein's identity (Bellec and Zhang, 2021).

Proposition B.3. *Suppose that Assumption 1 holds.*
Let $\boldsymbol{Q}_1 = \frac{\frac{1}{n}\left(\boldsymbol{F}^\top\boldsymbol{F} + \boldsymbol{H}^\top\boldsymbol{Z}^\top\boldsymbol{F} - \boldsymbol{S}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})\right)}{\|\boldsymbol{S}^{\frac{1}{2}}\|_{\mathrm{F}}(\|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \mathrm{Tr}(\boldsymbol{S}))^{\frac{1}{2}} n^{-\frac{1}{2}}}$, then $\mathbb{E}[\|\boldsymbol{Q}_1\|_{\mathrm{F}}^2] \le 4$.

Proposition B.4. *Suppose that Assumptions 2, 3 and 4 hold. Let*

$$\boldsymbol{Q}_2 = \frac{\frac{1}{n^2}\left(\boldsymbol{F}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{F} - \boldsymbol{F}^\top\boldsymbol{F}(p\boldsymbol{I}_T - \widehat{\mathbf{A}}) + (n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{H}^\top\boldsymbol{Z}^\top\boldsymbol{F}\right)}{(\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{F}\|_{\mathrm{F}}^2/n)n^{-\frac{1}{2}}},$$

then $\mathbb{E}[\|\boldsymbol{Q}_2\|_{\mathrm{F}}^2] \le C(\tau')(T \wedge (1 + \frac{p}{n}))(1 + \frac{p}{n})$ under Assumption 4(i), and $\mathbb{E}[I(\Omega)\|\boldsymbol{Q}_2\|_{\mathrm{F}}^2] \le C(\gamma, c)$ under Assumption 4(ii) for some set Ω with $\mathbb{P}(\Omega) \to 1$.

Proposition B.5. *Suppose that Assumptions 2, 3 and 4 hold. Let $\Xi = (n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{H}^\top\boldsymbol{Z}^\top\boldsymbol{F}$, and*

$$\boldsymbol{Q}_3 = \frac{\frac{1}{n^2}\left(p\boldsymbol{F}^\top\boldsymbol{F} - \boldsymbol{F}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{F} - (n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{H}^\top\boldsymbol{H}(n\boldsymbol{I}_T - \widehat{\mathbf{A}}) - \Xi - \Xi^\top\right)}{(\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{F}\|_{\mathrm{F}}^2/n)n^{-\frac{1}{2}}},$$

then $\mathbb{E}[\|\boldsymbol{Q}_3\|_{\mathrm{F}}] \le C(\gamma, \tau')$ under Assumption 4(i), and $\mathbb{E}[I(\Omega)\|\boldsymbol{Q}_3\|_{\mathrm{F}}] \le C(\gamma, c)$ under Assumption 4(ii) for some set Ω with $\mathbb{P}(\Omega) \to 1$.

B.2. Proofs of main results

With the preliminary results in previous section, we are ready to present the proofs of main results.

Proof of Theorem 3.4. Recall definition of $\widehat{\boldsymbol{S}}$ in Definition 2.2, and let $\boldsymbol{Q}_1, \boldsymbol{Q}_2$ be defined as in Propositions B.3 and B.4. With $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$, we obtain

$$n^2\Big[\boldsymbol{Q}_2(\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{F}\|_{\mathrm{F}}^2/n)n^{-\frac{1}{2}} - n^{-1}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{Q}_1\big(\|\boldsymbol{S}^{\frac{1}{2}}\|_{\mathrm{F}}\big(\|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \mathrm{Tr}(\boldsymbol{S})\big)^{\frac{1}{2}} n^{-\frac{1}{2}}\big)\Big]$$

$$
\begin{aligned}
&= \big(\boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{Z}^\top \boldsymbol{F} - \boldsymbol{F}^\top \boldsymbol{F}(p\boldsymbol{I}_T - \widehat{\mathbf{A}}) + (n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{H}^\top \boldsymbol{Z}^\top \boldsymbol{F}\big) \\
&\quad - \big[(n\boldsymbol{I}_T - \widehat{\mathbf{A}})(\boldsymbol{F}^\top \boldsymbol{F} + \boldsymbol{H}^\top \boldsymbol{Z}^\top \boldsymbol{F} - \boldsymbol{S}(n\boldsymbol{I}_T - \widehat{\mathbf{A}}))\big] \\
&= \big(\boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{Z}^\top \boldsymbol{F} - \boldsymbol{F}^\top \boldsymbol{F}(p\boldsymbol{I}_T - \widehat{\mathbf{A}})\big) - (n\boldsymbol{I}_T - \widehat{\mathbf{A}})(\boldsymbol{F}^\top \boldsymbol{F} - \boldsymbol{S}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})) \\
&= \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{Z}^\top \boldsymbol{F} + \boldsymbol{F}^\top \boldsymbol{F}\widehat{\mathbf{A}} + \widehat{\mathbf{A}}\boldsymbol{F}^\top \boldsymbol{F} - (n+p)\boldsymbol{F}^\top \boldsymbol{F} + (n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{S}(n\boldsymbol{I}_T - \widehat{\mathbf{A}}) \\
&= (n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{S}(n\boldsymbol{I}_T - \widehat{\mathbf{A}}) + \boldsymbol{F}^\top \boldsymbol{F}\widehat{\mathbf{A}} + \widehat{\mathbf{A}}\boldsymbol{F}^\top \boldsymbol{F} - \boldsymbol{F}^\top((n+p)\boldsymbol{I}_T - \boldsymbol{Z}\boldsymbol{Z}^\top)\boldsymbol{F} \\
&= (n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{S}(n\boldsymbol{I}_T - \widehat{\mathbf{A}}) - (n\boldsymbol{I}_T - \widehat{\mathbf{A}})\widehat{\boldsymbol{S}}(n\boldsymbol{I}_T - \widehat{\mathbf{A}}) \\
&= (n\boldsymbol{I}_T - \widehat{\mathbf{A}})(\boldsymbol{S} - \widehat{\boldsymbol{S}})(n\boldsymbol{I}_T - \widehat{\mathbf{A}}).
\end{aligned}
$$

Therefore, by the triangle inequality and $\|\boldsymbol{I}_T - \widehat{\mathbf{A}}/n\|_{\mathrm{op}} \le 1$ in Lemmas B.1 and B.2,

$$
\begin{aligned}
&\big\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)(\boldsymbol{S} - \widehat{\boldsymbol{S}})(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)\big\|_{\mathrm{F}} \\
&\le \|\boldsymbol{Q}_2\|_{\mathrm{F}} n^{-\frac{1}{2}}(\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{F}\|_{\mathrm{F}}^2/n) + \|\boldsymbol{I}_T - \widehat{\mathbf{A}}/n\|_{\mathrm{op}}\|\boldsymbol{Q}_1\|_{\mathrm{F}} n^{-\frac{1}{2}}\|\boldsymbol{S}^{\frac{1}{2}}\|_{\mathrm{F}}\big(\|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \mathrm{Tr}(\boldsymbol{S})\big)^{\frac{1}{2}} \\
&\le \|\boldsymbol{Q}_2\|_{\mathrm{F}} n^{-\frac{1}{2}}(\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{F}\|_{\mathrm{F}}^2/n) + \|\boldsymbol{Q}_1\|_{\mathrm{F}} n^{-\frac{1}{2}}\|\boldsymbol{S}^{\frac{1}{2}}\|_{\mathrm{F}}\big(\|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \mathrm{Tr}(\boldsymbol{S})\big)^{\frac{1}{2}} \\
&\le \|\boldsymbol{Q}_2\|_{\mathrm{F}} n^{-\frac{1}{2}}(\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{F}\|_{\mathrm{F}}^2/n) + \|\boldsymbol{Q}_1\|_{\mathrm{F}} n^{-\frac{1}{2}}\frac{1}{2}\big[\mathrm{Tr}(\boldsymbol{S}) + \big(\|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \mathrm{Tr}(\boldsymbol{S})\big)\big] \\
&\le (\|\boldsymbol{Q}_2\|_{\mathrm{F}} + \|\boldsymbol{Q}_1\|_{\mathrm{F}}) n^{-\frac{1}{2}}(\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \mathrm{Tr}(\boldsymbol{S})).
\end{aligned}
$$

Therefore,

$$
\big\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)(\boldsymbol{S} - \widehat{\boldsymbol{S}})(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)\big\|_{\mathrm{F}} \le \Theta_1 n^{-\frac{1}{2}}(\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \mathrm{Tr}(\boldsymbol{S})),
$$

where $\Theta_1 = \|\boldsymbol{Q}_1\|_{\mathrm{F}} + \|\boldsymbol{Q}_2\|_{\mathrm{F}}$. Note that we have $\mathbb{E}[\|\boldsymbol{Q}_1\|_{\mathrm{F}}^2] \le 4$ from Proposition B.3. By Proposition B.4, we have

(1) under Assumption 4(i), $\mathbb{E}[\|\boldsymbol{Q}_2\|_{\mathrm{F}}^2] \le C(\tau')(T \wedge (1 + \frac{p}{n}))(1 + \frac{p}{n})$. Hence

$$
\begin{aligned}
\mathbb{E}[\Theta_1^2] \le 2\mathbb{E}[\|\boldsymbol{Q}_1\|_{\mathrm{F}}^2 + \|\boldsymbol{Q}_2\|_{\mathrm{F}}^2] &\le 2[4 + C(\tau')(T \wedge (1 + \frac{p}{n}))(1 + \frac{p}{n})] \\
&\le C(\tau')(T \wedge (1 + \frac{p}{n}))(1 + \frac{p}{n}).
\end{aligned}
$$

(2) under Assumption 4(ii), $\mathbb{E}[I(\Omega)\|\boldsymbol{Q}_2\|_{\mathrm{F}}^2] \le C(\gamma, c)$ with $\mathbb{P}(\Omega) \to 1$. Thus,

$$
\mathbb{E}[I(\Omega)\Theta_1^2] \le 2\mathbb{E}[\|\boldsymbol{Q}_1\|_{\mathrm{F}}^2 + I(\Omega)\|\boldsymbol{Q}_2\|_{\mathrm{F}}^2] \le C(\gamma, c).
$$

■

Proof of Theorem 3.9. From the definitions of $\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3$ in Propositions B.3, B.4 and B.5, we have

$$
\begin{aligned}
&(\boldsymbol{Q}_2^\top + \boldsymbol{Q}_3) n^{-\frac{1}{2}}(\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{F}\|_{\mathrm{F}}^2/n) + (\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)\boldsymbol{Q}_1 n^{-\frac{1}{2}}\|\boldsymbol{S}^{\frac{1}{2}}\|_{\mathrm{F}}\big(\|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \mathrm{Tr}(\boldsymbol{S})\big)^{\frac{1}{2}} \\
&= \frac{1}{n}\boldsymbol{F}^\top \boldsymbol{F} - (\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)(\boldsymbol{H}^\top \boldsymbol{H} + \boldsymbol{S})(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n) \\
&= (\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)\big[n^{-1}(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{F}^\top \boldsymbol{F}(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1} - (\boldsymbol{H}^\top \boldsymbol{H} + \boldsymbol{S})\big](\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)
\end{aligned}
$$

$$= (\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)(\widehat{\boldsymbol{R}} - \boldsymbol{R})(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n),$$

where $\widehat{\boldsymbol{R}} \overset{\text{def}}{=} n^{-1}(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{F}^\top \boldsymbol{F}(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}$, and $\boldsymbol{R} \overset{\text{def}}{=} \boldsymbol{H}^\top \boldsymbol{H} + \boldsymbol{S}$.

Therefore, by the triangle inequality and $\|\boldsymbol{I}_T - \widehat{\mathbf{A}}/n\|_{\text{op}} \le 1$ from Lemmas B.1 and B.2,

$$\begin{aligned}
&\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)(\widehat{\boldsymbol{R}} - \boldsymbol{R})(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)\|_{\text{F}} \\
&\le (\|\boldsymbol{Q}_2\|_{\text{F}} + \|\boldsymbol{Q}_3\|_{\text{F}})n^{-\frac{1}{2}}(\|\boldsymbol{H}\|_{\text{F}}^2 + \|\boldsymbol{F}\|_{\text{F}}^2/n) + \|\boldsymbol{Q}_1\|_{\text{F}} n^{-\frac{1}{2}} \|\boldsymbol{S}^{\frac{1}{2}}\|_{\text{F}} \big(\|\boldsymbol{F}\|_{\text{F}}^2/n + \text{Tr}(\boldsymbol{S})\big)^{\frac{1}{2}} \\
&\le (\|\boldsymbol{Q}_2\|_{\text{F}} + \|\boldsymbol{Q}_3\|_{\text{F}} + \|\boldsymbol{Q}_1\|_{\text{F}})n^{-\frac{1}{2}}(\|\boldsymbol{F}\|_{\text{F}}^2/n + \|\boldsymbol{H}\|_{\text{F}}^2 + \text{Tr}(\boldsymbol{S})) \\
&= \Theta_2 n^{-\frac{1}{2}}(\|\boldsymbol{F}\|_{\text{F}}^2/n + \|\boldsymbol{H}\|_{\text{F}}^2 + \text{Tr}(\boldsymbol{S})),
\end{aligned}$$

where $\Theta_2 = \|\boldsymbol{Q}_1\|_{\text{F}} + \|\boldsymbol{Q}_2\|_{\text{F}} + \|\boldsymbol{Q}_3\|_{\text{F}}$. By Propositions B.3, B.4 and B.5, we obtain $\mathbb{E}[\Theta_2] \le C(\gamma, \tau')$ under Assumption 4(i) and $\mathbb{E}[I(\Omega)\Theta_2] \le C(\gamma, c)$ with $P(\Omega) \to 1$ under Assumption 4(ii).

Furthermore, since $\Theta_2 = O_P(1)$, and $\big\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\big\|_{\text{op}} = O_P(1)$ from Lemmas B.1 and B.2,

$$\begin{aligned}
\|\widehat{\boldsymbol{R}} - \boldsymbol{R}\|_{\text{F}} &\le \|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\|_{\text{op}}^2 \Theta_2 n^{-\frac{1}{2}}(\|\boldsymbol{F}\|_{\text{F}}^2/n + \|\boldsymbol{H}\|_{\text{F}}^2 + \text{Tr}(\boldsymbol{S})) \\
&= O_P(n^{-\frac{1}{2}})(\|\boldsymbol{F}\|_{\text{F}}^2/n + \|\boldsymbol{H}\|_{\text{F}}^2 + \text{Tr}(\boldsymbol{S})).
\end{aligned}$$

Since $\frac{1}{n}\boldsymbol{F}^\top \boldsymbol{F} = (\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)\widehat{\boldsymbol{R}}(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)$, taking trace of both sides gives $\frac{1}{n}\|\boldsymbol{F}\|_{\text{F}}^2 \le \|\widehat{\boldsymbol{R}}\|_*$ thanks to $\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)\|_{\text{op}} \le 1$. Note that $\|\boldsymbol{R}\|_* = \|\boldsymbol{H}\|_{\text{F}}^2 + \text{Tr}(\boldsymbol{S})$ by definition of $\boldsymbol{R}$, we obtain

$$\|\widehat{\boldsymbol{R}} - \boldsymbol{R}\|_{\text{F}} \le O_P(n^{-\frac{1}{2}})(\|\widehat{\boldsymbol{R}}\|_* + \|\boldsymbol{R}\|_*). \tag{32}$$

Since $\widehat{\boldsymbol{R}}$ and $\boldsymbol{R}$ are both $T \times T$ positive semi-definite matrices, whose ranks are at most T,

$$\begin{aligned}
&\big|\|\widehat{\boldsymbol{R}}\|_* - \|\boldsymbol{R}\|_*\big| \le \|\widehat{\boldsymbol{R}} - \boldsymbol{R}\|_* \le \sqrt{T}\|\widehat{\boldsymbol{R}} - \boldsymbol{R}\|_{\text{F}} \\
&\le O_P((T/n)^{\frac{1}{2}})(\|\widehat{\boldsymbol{R}}\|_* + \|\boldsymbol{R}\|_*) = o_P(1)(\|\widehat{\boldsymbol{R}}\|_* + \|\boldsymbol{R}\|_*),
\end{aligned}$$

thanks to $T = o(n)$. That is,

$$\frac{\big|\|\widehat{\boldsymbol{R}}\|_* - \|\boldsymbol{R}\|_*\big|}{\|\widehat{\boldsymbol{R}}\|_* + \|\boldsymbol{R}\|_*} \le O_P((T/n)^{\frac{1}{2}}),$$

which implies $\frac{\|\boldsymbol{R}\|_*}{\|\widehat{\boldsymbol{R}}\|_*} - 1 = O_P((T/n)^{\frac{1}{2}})$, i.e.,

$$\frac{\text{Tr}(\boldsymbol{S}) + \|\boldsymbol{H}\|_{\text{F}}^2}{\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{F}^\top\|_{\text{F}}^2/n} - 1 = O_P((T/n)^{\frac{1}{2}}) = o_P(1).$$

■

Proof of Theorem 3.5. This proof is based on results of Theorems 3.4 and 3.9. We begin with the result of Theorem 3.9,

$$\frac{\text{Tr}(\boldsymbol{S}) + \|\boldsymbol{H}\|_{\text{F}}^2}{\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{F}^\top\|_{\text{F}}^2/n} \xrightarrow{p} 1.$$

In other words,

$$\mathrm{Tr}(\boldsymbol{S}) + \|\boldsymbol{H}\|_{\mathrm{F}}^2 = (1 + o_P(1))\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{F}^\top\|_{\mathrm{F}}^2/n.$$

Thus, the upper bound in Theorem 3.4 can be bounded from above as follows

$$\begin{aligned}
&\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)(\widehat{\boldsymbol{S}} - \boldsymbol{S})(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)\|_{\mathrm{F}} \\
\le\ & \Theta_1 n^{-\frac{1}{2}}(\|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{S}^{\frac{1}{2}}\|_{\mathrm{F}}^2) \\
\le\ & \Theta_1 n^{-\frac{1}{2}}(\|\boldsymbol{F}\|_{\mathrm{F}}^2/n + (1 + o_P(1))\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{F}^\top\|_{\mathrm{F}}^2/n) \\
\le\ & \Theta_1 n^{-\frac{1}{2}}\big(1 + (1 + o_P(1))\|(\boldsymbol{I}_T - \widehat{\mathbf{A}})^{-1}\|_{\mathrm{op}}^2\big)\|\boldsymbol{F}\|_{\mathrm{F}}^2/n \\
=\ & O_P(n^{-\frac{1}{2}})\|\boldsymbol{F}\|_{\mathrm{F}}^2/n.
\end{aligned}$$

Using $\|(\boldsymbol{I}_T - \widehat{\mathbf{A}})^{-1}\|_{\mathrm{op}} = O_P(1)$ again, it follows

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}} \le O_P(n^{-\frac{1}{2}})\|\boldsymbol{F}\|_{\mathrm{F}}^2/n. \tag{33}$$

A similar argument leads to

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}} \le O_P(n^{-\frac{1}{2}})(\mathrm{Tr}(\boldsymbol{S}) + \|\boldsymbol{H}\|_{\mathrm{F}}^2). \tag{34}$$

■

Proof of Corollary 3.6. Under Assumption 4(i) and 5, we proceed to bound $\|\boldsymbol{F}\|_{\mathrm{F}}^2/n$ in terms of $\mathrm{Tr}(\boldsymbol{S})$. Let $L(\boldsymbol{B}) = \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{B}\|_{2,1} + \frac{\tau}{2}\|\boldsymbol{B}\|_{\mathrm{F}}^2$ be the objective function in (6), then $L(\widehat{\boldsymbol{B}}) \le L(\mathbf{0})$ by definition of $\widehat{\boldsymbol{B}}$. Thus,

$$\frac{1}{2n}\|\boldsymbol{F}\|_{\mathrm{F}}^2 \le \frac{1}{2n}\|\boldsymbol{F}\|_{\mathrm{F}}^2 + \lambda\|\widehat{\boldsymbol{B}}\|_{2,1} + \frac{\tau}{2}\|\widehat{\boldsymbol{B}}\|_{\mathrm{F}}^2 \le \frac{1}{2n}\|\boldsymbol{Y}\|_{\mathrm{F}}^2.$$

Now we bound $\frac{1}{n}\|\boldsymbol{Y}\|_{\mathrm{F}}^2$ by Hanson-Wright inequality. Since $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E}$, the rows of $\boldsymbol{Y}$ are i.i.d. $\mathcal{N}_T(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{y}})$ with $\boldsymbol{\Sigma}_{\boldsymbol{y}} = (\boldsymbol{B}^*)^\top\boldsymbol{\Sigma}\boldsymbol{B}^* + \boldsymbol{S}$, then $\mathbf{vec}(\boldsymbol{Y}^\top) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{\mathbf{n}} \otimes \boldsymbol{\Sigma}_{\boldsymbol{y}})$, and $\boldsymbol{\xi} \overset{\text{def}}{=} [\boldsymbol{I}_n \otimes \boldsymbol{\Sigma}_{\boldsymbol{y}}]^{-\frac{1}{2}}\mathbf{vec}(\boldsymbol{Y}^\top) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{nT})$. Since $\|\boldsymbol{Y}\|_{\mathrm{F}}^2 = [\mathbf{vec}(\boldsymbol{Y}^\top)]^\top\mathbf{vec}(\boldsymbol{Y}^\top) = \boldsymbol{\xi}^\top(\boldsymbol{I}_n \otimes \boldsymbol{\Sigma}_{\boldsymbol{y}})\boldsymbol{\xi}$, we apply the following variant of Hanson-Wright inequality.

Lemma B.6 (Lemma 1 in Laurent and Massart (2000)). *For $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$ and matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, the following inequality holds for any $x > 0$,*

$$\mathbb{P}(\boldsymbol{\xi}^\top\boldsymbol{A}\boldsymbol{\xi} - \mathrm{Tr}(\boldsymbol{A}) \le 2\sqrt{x}\|\boldsymbol{A}\|_{\mathrm{F}} + 2x\|\boldsymbol{A}\|_{\mathrm{op}}) \ge 1 - \exp(-x).$$

In our case, we take $\boldsymbol{A} = (\boldsymbol{I}_n \otimes \boldsymbol{\Sigma}_{\boldsymbol{y}})$, then $\mathrm{Tr}(\boldsymbol{A}) = n\,\mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{y}})$, $\|\boldsymbol{A}\|_{\mathrm{F}} = \sqrt{n}\|\boldsymbol{\Sigma}_{\boldsymbol{y}}\|_{\mathrm{F}} \le \sqrt{n}\,\mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{y}})$, $\|\boldsymbol{A}\|_{\mathrm{op}} = \|\boldsymbol{\Sigma}_{\boldsymbol{y}}\|_{\mathrm{op}} \le \mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{y}})$, thus with probability at least $1 - \exp(-x)$,

$$\|\boldsymbol{Y}\|_{\mathrm{F}}^2 - n\,\mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{y}}) \le 2\sqrt{nx}\,\mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{y}}) + 2x\,\mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{y}}).$$

Take $x = n$, then with probability at least $1 - \exp(-n)$,

$$\|\boldsymbol{F}\|_{\mathrm{F}}^2/n \le \|\boldsymbol{Y}\|_{\mathrm{F}}^2/n \le 5\,\mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{y}}).$$

Thus, $\|\boldsymbol{F}\|_{\mathrm{F}}^2/n = O_P(1)\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{y}})$. Together with (33), we obtain

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}} \le O_P(n^{-\frac{1}{2}})\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{y}}).$$

Note that by Assumption 5, $\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{y}}) = \|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{B}^*\|_{\mathrm{F}}^2 + \operatorname{Tr}(\boldsymbol{S}) \le (1 + \mathfrak{snr})\operatorname{Tr}(\boldsymbol{S})$. Therefore, we obtain

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}} \le O_P(n^{-\frac{1}{2}})\operatorname{Tr}(\boldsymbol{S}).$$

Furthermore, since $\operatorname{Tr}(\boldsymbol{S}) \le \sqrt{T}\|\boldsymbol{S}\|_{\mathrm{F}}$ and $T = o(n)$, we have

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}} \le O_P(\sqrt{T/n})\|\boldsymbol{S}\|_{\mathrm{F}} = o_P(1)\|\boldsymbol{S}\|_{\mathrm{F}}.$$

Finally, since $\|\boldsymbol{S}\|_* = \operatorname{Tr}(\boldsymbol{S})$, by the triangular inequality

$$\left|\|\widehat{\boldsymbol{S}}\|_* - \operatorname{Tr}(\boldsymbol{S})\right| \le \|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_* \le \sqrt{T}\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}} \le O_P(\sqrt{T/n})\operatorname{Tr}(\boldsymbol{S}) = o_P(1)\operatorname{Tr}(\boldsymbol{S}).$$

■

Proof of Corollary 3.7. For $\tau = 0$, by the optimality of $\widehat{\boldsymbol{B}}$ in (6),

$$\frac{1}{2n}\|\boldsymbol{F}\|_{\mathrm{F}}^2 + \lambda\|\widehat{\boldsymbol{B}}\|_{2,1} \le \frac{1}{2n}\|\boldsymbol{E}\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{B}^*\|_{2,1}.$$

Note that $\boldsymbol{F} = \boldsymbol{E} - \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) = \boldsymbol{E} - \boldsymbol{Z}\boldsymbol{H}$, expanding the squares and rearranging terms yields

$$\|\boldsymbol{Z}\boldsymbol{H}\|_{\mathrm{F}}^2 \le 2\langle \boldsymbol{E}, \boldsymbol{Z}\boldsymbol{H}\rangle + 2n\lambda(\|\boldsymbol{B}^*\|_{2,1} - \|\widehat{\boldsymbol{B}}\|_{2,1}) \le 2\langle \boldsymbol{E}, \boldsymbol{Z}\boldsymbol{H}\rangle + 2n\lambda\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{2,1}. \tag{35}$$

From assumptions in this corollary, $\widehat{\boldsymbol{B}} - \boldsymbol{B}^*$ has at most $(1-c)n$ rows. Thus, in the event U_2, we have

$$n\eta\|\boldsymbol{H}\|_{\mathrm{F}}^2 = n\eta\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2 \le \|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2 = \|\boldsymbol{Z}\boldsymbol{H}\|_{\mathrm{F}}^2.$$

We bound the right-hand side two terms in (35) by Cauchy-Schwarz inequality,

$$\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{2,1} \le \sqrt{(1-c)n}\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{\mathrm{F}} \le \frac{\sqrt{(1-c)n}}{\sqrt{\phi_{\min}(\boldsymbol{\Sigma})}}\|\boldsymbol{H}\|_{\mathrm{F}} \le \frac{\sqrt{1-c}}{\sqrt{\eta\phi_{\min}(\boldsymbol{\Sigma})}}\|\boldsymbol{Z}\boldsymbol{H}\|_{\mathrm{F}},$$

and $\langle \boldsymbol{E}, \boldsymbol{Z}\boldsymbol{H}\rangle \le \|\boldsymbol{E}\|_{\mathrm{F}}\|\boldsymbol{Z}\boldsymbol{H}\|_{\mathrm{F}} \le \|\boldsymbol{S}^{\frac{1}{2}}\|_{\mathrm{F}}\|\boldsymbol{E}\boldsymbol{S}^{-\frac{1}{2}}\|_{\mathrm{op}}\|\boldsymbol{Z}\boldsymbol{H}\|_{\mathrm{F}}$.

Therefore, by canceling a factor $\|\boldsymbol{Z}\boldsymbol{H}\|_{\mathrm{F}}$ from both sides of (35), we have

$$\sqrt{n\eta}\|\boldsymbol{H}\|_{\mathrm{F}} \le \|\boldsymbol{Z}\boldsymbol{H}\|_{\mathrm{F}} \le 2\|\boldsymbol{S}^{\frac{1}{2}}\|_{\mathrm{F}}\|\boldsymbol{E}\boldsymbol{S}^{-\frac{1}{2}}\|_{\mathrm{op}} + \frac{2\sqrt{(1-c)}n\lambda}{\sqrt{\eta\phi_{\min}(\boldsymbol{\Sigma})}}.$$

Using $(a+b)^2 \le 2a^2 + 2b^2$,

$$\|\boldsymbol{H}\|_{\mathrm{F}}^2 \le \frac{4}{n\eta}\operatorname{Tr}(\boldsymbol{S})\|\boldsymbol{E}\boldsymbol{S}^{-\frac{1}{2}}\|_{\mathrm{op}}^2 + \frac{4(1-c)n\lambda^2}{\eta^2\phi_{\min}(\boldsymbol{\Sigma})}.$$

Hence, using λ is of the form $\mu\sqrt{\mathrm{Tr}(\boldsymbol{S})/n}$, we have

$$\begin{aligned}
&\mathrm{Tr}(\boldsymbol{S}) + \|\boldsymbol{H}\|_{\mathrm{F}}^2 \\
&\leq (1 + 4\eta^{-1}n^{-1}\|\boldsymbol{E}\boldsymbol{S}^{-\frac{1}{2}}\|_{\mathrm{op}}^2)\,\mathrm{Tr}(\boldsymbol{S}) + \frac{4(1-c)\mu^2}{\eta^2\phi_{\min}(\boldsymbol{\Sigma})}\,\mathrm{Tr}(\boldsymbol{S}) \\
&\leq O_P(1)(1+\mu^2)\,\mathrm{Tr}(\boldsymbol{S}),
\end{aligned}$$

where we used that $n^{-1}\|\boldsymbol{E}\boldsymbol{S}^{-\frac{1}{2}}\|_{\mathrm{op}} = O_P(1)$ by (Davidson and Szarek, 2001, Theorem II.13) and $T = o(n)$. Now, by Theorem 3.5,

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}} \leq O_P(n^{-\frac{1}{2}})[\mathrm{Tr}(\boldsymbol{S}) + \|\boldsymbol{H}\|_{\mathrm{F}}^2] \leq O_P(n^{-\frac{1}{2}})(1+\mu^2)\,\mathrm{Tr}(\boldsymbol{S}),$$

where the $O_P(\cdot)$ hides constants depending on $\gamma, c, \phi_{\min}(\boldsymbol{\Sigma})$ since η is a constant that only depends on γ, c. ■

Proof of Theorem 3.10. From the definitions of $\boldsymbol{Q}_2, \boldsymbol{Q}_3$ in Propositions B.4 and B.5, we have

$$\begin{aligned}
&\boldsymbol{Q}_2 + \boldsymbol{Q}_2^\top + \boldsymbol{Q}_3 \\
&= \frac{n^{-2}\big(\boldsymbol{F}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{F} + \widehat{\mathbf{A}}\boldsymbol{F}^\top\boldsymbol{F} + \boldsymbol{F}^\top\boldsymbol{F}\widehat{\mathbf{A}} - p\boldsymbol{F}^\top\boldsymbol{F} - (n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{H}^\top\boldsymbol{H}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})}{(\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{F}\|_{\mathrm{F}}^2/n)n^{-\frac{1}{2}}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\big\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)\boldsymbol{H}^\top\boldsymbol{H}(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n) - n^{-2}\big(\boldsymbol{F}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{F} + \widehat{\mathbf{A}}\boldsymbol{F}^\top\boldsymbol{F} + \boldsymbol{F}^\top\boldsymbol{F}\widehat{\mathbf{A}} - p\boldsymbol{F}^\top\boldsymbol{F}\big)\big\|_{\mathrm{F}} \\
&= \|\boldsymbol{Q}_2 + \boldsymbol{Q}_2^\top + \boldsymbol{Q}_3\|_{\mathrm{F}}(\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{F}\|_{\mathrm{F}}^2/n)n^{-\frac{1}{2}} \\
&\leq \Theta_3(\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{F}\|_{\mathrm{F}}^2/n)n^{-\frac{1}{2}},
\end{aligned}$$

where $\Theta_3 = 2\|\boldsymbol{Q}_2\|_{\mathrm{F}} + \|\boldsymbol{Q}_3\|_{\mathrm{F}}$. The conclusion thus follows by Propositions B.4 and B.5. ■

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper. The authors acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here. URL: https://oarc.rutgers.edu.

Funding

Pierre C. Bellec is partially supported by the NSF Grants DMS-1811976 and DMS-1945428. Gabriel Romon is supported by a PhD scholarship from CREST.

Supplementary Material

Supplement I
Additional simulation results and proofs of results in Appendix B.1.

Supplement II
Code for reproducing the numerical experiments in the main paper and the Supplemnt I.

References

BAYATI, M., ERDOGDU, M. A. and MONTANARI, A. (2013). Estimating lasso risk and noise level. *Adv. Neural Inf. Process. Syst.* **26**.

BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inform. Theory* **58** 1997–2017.

BELLEC, P. C. (2020). Out-of-sample error estimate for robust M-estimators with convex penalty. *arXiv preprint arXiv:2008.11840*.

BELLEC, P. and KUCHIBHOTLA, A. (2019). First order expansion of convex regularized estimators. *Adv. Neural Inf. Process. Syst.* **32**.

BELLEC, P. C. and ROMON, G. (2021). Chi-square and normal inference in high-dimensional multi-task regression. *arXiv preprint arXiv:2107.07828*.

BELLEC, P. and TSYBAKOV, A. (2016). Bounds on the prediction error of penalized least squares estimators with convex penalty. In *International Conference on Modern Problems of Stochastic Analysis and Statistics* 315–333. Springer.

BELLEC, P. C. and ZHANG, C.-H. (2019). De-biasing convex regularized estimators and interval estimation in linear models. *arXiv preprint arXiv:1912.11943*.

BELLEC, P. C. and ZHANG, C.-H. (2021). Second-order Stein: SURE for SURE and other applications in high-dimensional inference. *Ann. Statist.* **49** 1864–1903.

BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806.

BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2014). Pivotal estimation via square-root Lasso in nonparametric regression. *Ann. Statist.* **42** 757–788.

BERTRAND, Q., MASSIAS, M., GRAMFORT, A. and SALMON, J. (2019). Handling correlated and repeated measurements with the smoothed multivariate square-root Lasso. *Adv. Neural Inf. Process. Syst.* **32**.

BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604.

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144.

CELENTANO, M. and MONTANARI, A. (2021). CAD: Debiasing the Lasso with inaccurate covariate model. *arXiv preprint arXiv:2107.14172*.

CHEN, S. and BANERJEE, A. (2017). Alternating estimation for structured high-dimensional multi-response models. *Adv. Neural Inf. Process. Syst.* **30**.

DAVIDSON, K. R. and SZAREK, S. J. (2001). Local operator theory, random matrices and Banach spaces. *Handbook of the geometry of Banach spaces* **1** 131.

DICKER, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika* **101** 269–284.

DICKER, L. H. and ERDOGDU, M. A. (2016). Maximum likelihood for variance estimation in high-dimensional linear models. In *Artificial Intelligence and Statistics* 159–167. PMLR.

DOBRIBAN, E. and WAGER, S. (2018). High-dimensional asymptotics of prediction: ridge regression and classification. *Ann. Statist.* **46** 247–279.

EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756.

FAN, J., GUO, S. and HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 37–65.

FOURDRINIER, D., HADDOUCHE, A. M. and MEZOUED, F. (2021). Covariance matrix estimation under data-based loss. *Statist. Probab. Lett.* **177** Paper No. 109160, 7.

FRIEDMAN, J. H., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33** 1–22.

GEER, S. V. D. and STUCKY, B. (2016). χ 2-confidence sets in high-dimensional regression. In *Statistical analysis for high-dimensional data* 279–306. Springer.

JANSON, L., FOYGEL BARBER, R. and CANDÈS, E. (2017). EigenPrism: inference for high dimensional signal-to-noise ratios. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1037–1065.

KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133.

LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338.

LECUÉ, G. and MENDELSON, S. (2017). Sparse recovery under weak moment assumptions. *J. Eur. Math. Soc. (JEMS)* **19** 881–904.

LIU, H. and ZHANG, J. (2009). Estimation consistency of the group lasso and its applications. In *Artificial Intelligence and Statistics* 376–383. PMLR.

LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204.

MIOLANE, L. and MONTANARI, A. (2021). The distribution of the Lasso: uniform control over sparse balls and adaptive parameter tuning. *Ann. Statist.* **49** 2313–2335.

MOLSTAD, A. J. (2022). New Insights for the Multivariate Square-Root Lasso. *J. Mach. Learn. Res.* **23** 1–52.

OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Statist.* **39** 1–47.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M. and DUCHESNAY, E. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12** 2825–2830.

SIMON, N., FRIEDMAN, J. and HASTIE, T. (2013). A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *arXiv preprint arXiv:1311.6529*.

STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* 1135–1151.

SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898.

TAN, K., ROMON, G. and BELLEC, P. C. (2023). Supplement to "Noise covariance estimation in multi-task high-dimensional linear models".

YU, G. and BIEN, J. (2019). Estimating the error variance in a high-dimensional linear model. *Biometrika* **106** 533–546.

Noise Covariance Estimation in Multi-Task High-dimensional Linear Models Supplementary Material

Kai Tan, Gabriel Romon, Pierre C. Bellec

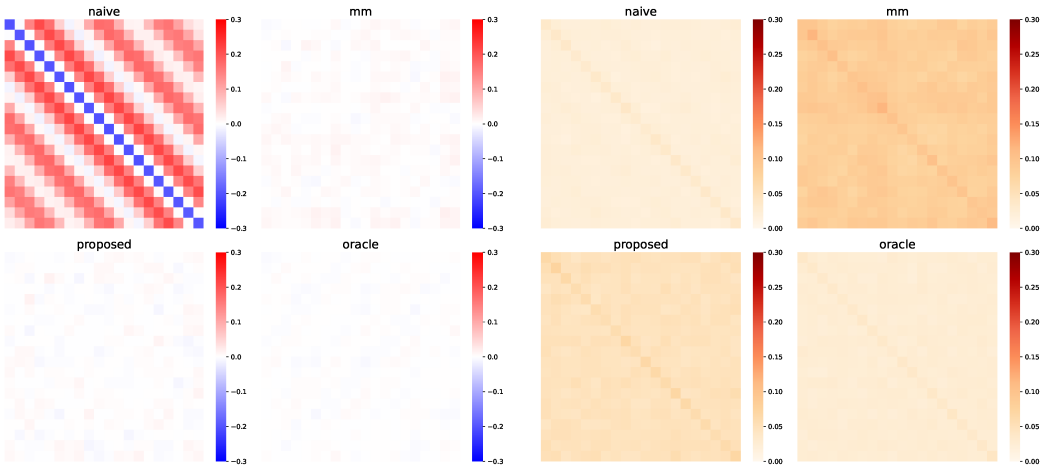
In this supplement, we first present additional simulation results for estimating two types of noise covariance with different (n, p) in Section C. To complete the proofs of all theoretical results in the paper, it remains to prove Lemma B.1, B.2 and Propositions B.3 to B.5 in Appendix B of the paper. We prove Lemmas B.1 and B.2 in Section D, and present in Section E a few useful lemmas before proving Propositions B.3 to B.5 in Section F. The proofs of lemmas in Section E are deferred to Section G.

C. Additional simulation results

While Figures 2 and 3 in the main paper present heatmaps for estimating full and low-rank noise covariance $\boldsymbol{S}$ with $n = 1000$, this section provides additional estimating results from $n = 1500$ and $n = 2000$ (both with $p = 1.5n$).

C.1. Heatmaps for estimating full rank S with $n = 1500, 2000$

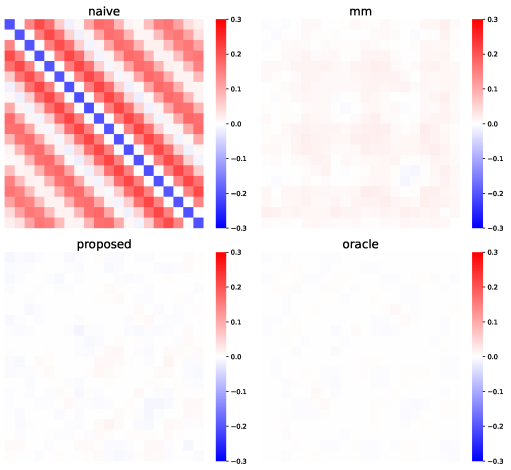
When estimating the full rank noise covariance matrix $\boldsymbol{S}$ with (t, t')-th entry $\boldsymbol{S}_{t,t'} = \frac{\cos(t-t')}{1+\sqrt{|t-t'|}}$, the heatmaps for different estimators from $n = 1500$ and $n = 2000$ are presented in Figures 4 and 5, respectively. The comparisons between four methods in Figures 4 and 5 are similar to the case $n = 1000$ in Figure 2 of the main paper; our proposed estimator outperforms the naive estimator and method-of-moments estimator.



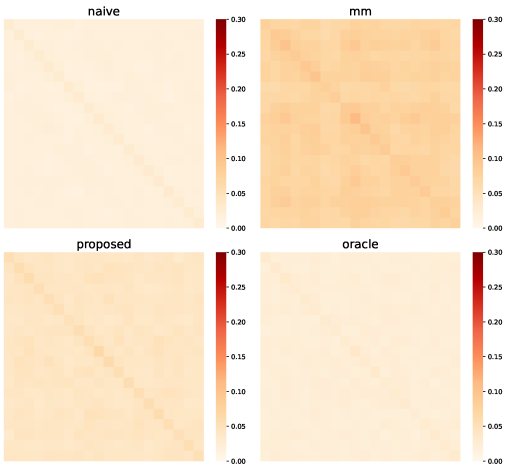
(a) Heatmap of *bias* for each entry

(b) Heatmap of *standard deviation* in each entry

Figure 4: Heatmaps for estimation of full rank $\boldsymbol{S}$ with $n = 1500$ over 100 repetitions.



(a) Heatmap of *bias* for each entry

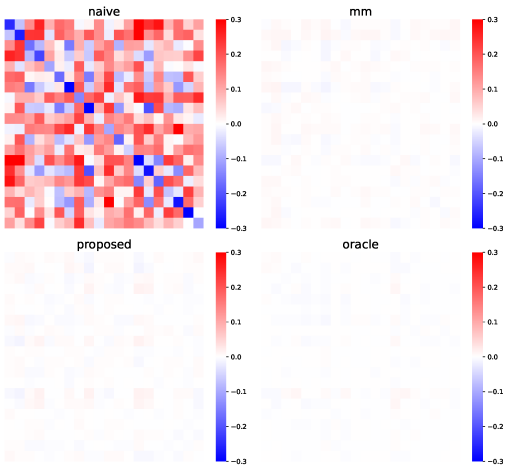


(b) Heatmap of *standard deviation* in each entry

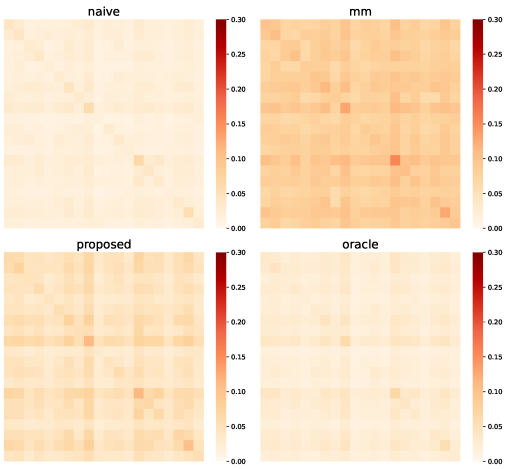
Figure 5: Heatmaps for estimation of full rank $\boldsymbol{S}$ with $n = 2000$ over 100 repetitions.

C.2. Heatmaps for estimating low rank S with $n = 1500, 2000$

When estimating the low rank noise covariance matrix with $\boldsymbol{S} = \boldsymbol{u}\boldsymbol{u}^\top$, and $\boldsymbol{u} \in \mathbb{R}^{T\times 10}$ with entries are i.i.d. from $\mathcal{N}(0, 1/T)$. We present the heatmaps for different estimators with $1500, 2000$ in Figures 6 and 7 below. The comparisons between four methods in Figures 6 and 7 are similar to the case $n = 1000$ in Figure 3 of the main paper. All of these figures convince us that besides the oracle estimator, the proposed estimator has the best performance.

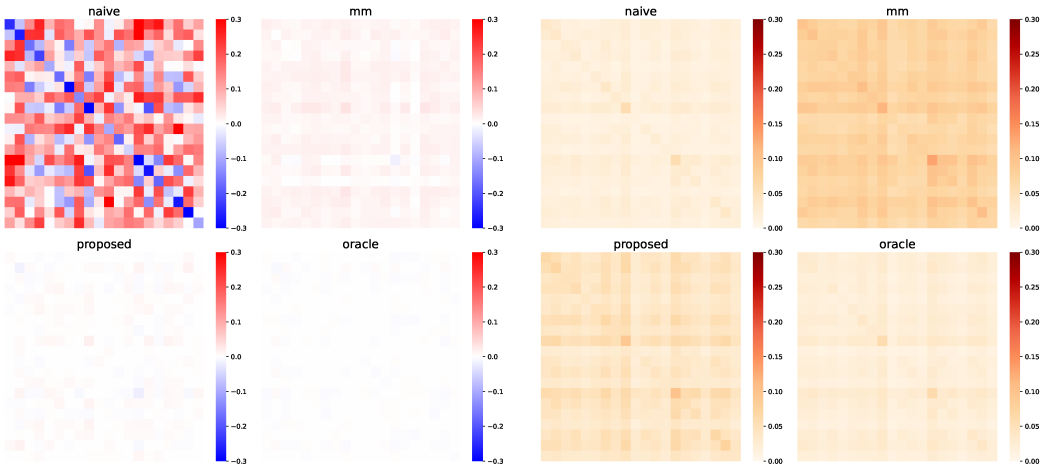


(a) Heatmap of *bias* for each entry



(b) Heatmap of *standard deviation* in each entry

Figure 6: Heatmaps for estimation of low rank $\boldsymbol{S}$ with $n = 1500$ over 100 repetitions.



(a) Heatmap of *bias* for each entry

(b) Heatmap of *standard deviation* in each entry

Figure 7: Heatmaps for estimation of low rank $\boldsymbol{S}$ with $n = 2000$ over 100 repetitions.

D. Proofs of Lemmas B.1 and B.2

Proof of Lemma B.1. (i) For any $\boldsymbol{u} \in \mathbb{R}^T$, by definition (10),

$$\begin{aligned}
\boldsymbol{u}^\top \widehat{\mathbf{A}} \boldsymbol{u} &= \operatorname{Tr}\left[(\boldsymbol{u}^\top \otimes \boldsymbol{X}_{\hat{\mathcal{S}}}) \boldsymbol{M}^\dagger (\boldsymbol{u} \otimes \boldsymbol{X}_{\hat{\mathcal{S}}}^\top)\right] \\
&\le \operatorname{Tr}\left[(\boldsymbol{u}^\top \otimes \boldsymbol{X}_{\hat{\mathcal{S}}})[\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger](\boldsymbol{u} \otimes \boldsymbol{X}_{\hat{\mathcal{S}}}^\top)\right] \\
&= \operatorname{Tr}\left[(\boldsymbol{u}^\top \boldsymbol{I}_T \boldsymbol{u}) \otimes [\boldsymbol{X}_{\hat{\mathcal{S}}} (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger \boldsymbol{X}_{\hat{\mathcal{S}}}^\top]\right] \\
&= \|\boldsymbol{u}\|^2 \operatorname{Tr}[\boldsymbol{X}_{\hat{\mathcal{S}}} (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger \boldsymbol{X}_{\hat{\mathcal{S}}}^\top] \\
&= \|\boldsymbol{u}\|^2 \operatorname{Tr}[\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger].
\end{aligned}$$

Let $r = \operatorname{rank}(\boldsymbol{X}_{\hat{\mathcal{S}}}) \le \min(n, |\hat{\mathcal{S}}|)$ be the rank of $\boldsymbol{X}_{\hat{\mathcal{S}}}$, and $\widehat{\phi}_1 \ge \cdots \ge \widehat{\phi}_r > 0$ be the nonzero eigenvalues of $\frac{1}{n} \boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}}$. We have

$$\begin{aligned}
\|\widehat{\mathbf{A}}/n\|_{\mathrm{op}} &\le \frac{1}{n} \operatorname{Tr}[\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger] \\
&= \frac{1}{n} \operatorname{Tr}[\frac{1}{n} \boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} (\frac{1}{n} \boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + \tau \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger] \\
&\le \frac{r}{n} \left\| \frac{1}{n} \boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} (\frac{1}{n} \boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + \tau \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger \right\|_{\mathrm{op}} \\
&\le \frac{\widehat{\phi}_1}{\widehat{\phi}_1 + \tau} \le 1.
\end{aligned}$$

Thus, $\|\boldsymbol{I} - \widehat{\mathbf{A}}/n\|_{\mathrm{op}} \le 1$ as $\widehat{\mathbf{A}}$ is positive semi-definite.

(ii) Note that $\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\|_{\text{op}} = (1 - \|\widehat{\mathbf{A}}/n\|_{\text{op}})^{-1} \le 1 + \frac{\widehat{\phi}_1}{\tau}$, where $\widehat{\phi}_1 = \|\frac{1}{n}\boldsymbol{X}_{\widehat{\mathcal{S}}}^\top \boldsymbol{X}_{\widehat{\mathcal{S}}}\|_{\text{op}} \le \|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\|_{\text{op}} \le \frac{1}{n}\|\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_{\text{op}}^2\|\boldsymbol{\Sigma}\|_{\text{op}}$. Therefore,

(1) in the event $\{\|\boldsymbol{X}\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_{\text{op}} < 2\sqrt{n} + \sqrt{p}\}$, we have

$$\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\|_{\text{op}} \le 1 + \tau^{-1}(2 + \sqrt{p/n})^2\|\boldsymbol{\Sigma}\|_{\text{op}} = 1 + (\tau')^{-1}(2 + \sqrt{p/n})^2.$$

(2) $\mathbb{E}[\widehat{\phi}_1] \le \mathbb{E}[n^{-1}\|\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_{\text{op}}^2\|\boldsymbol{\Sigma}\|_{\text{op}}] \le [(1 + \sqrt{p/n})^2 + n^{-1}]\|\boldsymbol{\Sigma}\|_{\text{op}}$ by (36). Hence,

$$\mathbb{E}\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\|_{\text{op}} \le 1 + \tau^{-1}\mathbb{E}[\widehat{\phi}_1] \le 1 + (\tau')^{-1}[(1 + \sqrt{p/n})^2 + n^{-1}].$$

■

Proof of Lemma B.2. (i) For $\tau = 0$, using the same argument as proof of Lemma B.1, we obtain

$$\boldsymbol{u}^\top\widehat{\mathbf{A}}\boldsymbol{u} \le \|\boldsymbol{u}\|^2 \operatorname{Tr}[\boldsymbol{X}_{\hat{\mathcal{S}}}^\top\boldsymbol{X}_{\hat{\mathcal{S}}}(\boldsymbol{X}_{\hat{\mathcal{S}}}^\top\boldsymbol{X}_{\hat{\mathcal{S}}})^\dagger] \le \|\boldsymbol{u}\|^2|\hat{\mathcal{S}}|.$$

Thus, in the event U_1, we have $\|\widehat{\mathbf{A}}\|_{\text{op}}/n \le |\hat{\mathcal{S}}|/n \le (1-c)/2 < 1$, hence

$$\|\boldsymbol{I}_T - \widehat{\mathbf{A}}/n\|_{\text{op}} \le 1.$$

(ii) In the event U_1, we have $\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\|_{\text{op}} = (1 - \|\widehat{\mathbf{A}}/n\|_{\text{op}})^{-1} \le (1 - (1-c)/2)^{-1}$. Furthermore, $\mathbb{E}[I(U_1)\|(\boldsymbol{I}_T - \widehat{\mathbf{A}}/n)^{-1}\|_{\text{op}}] \le (1 - (1-c)/2)^{-1}$.

■

E. Preliminary results needed for proving Propositions B.3 to B.5

Let us first introduce two events besides the event $U_1 = \{\|\widehat{\boldsymbol{B}}\|_0 \le n(1-c)/2\}$ in Assumption 4(ii), we define events U_2 and U_3 as below,

$$U_2 = \Big\{\inf_{\boldsymbol{b}\in\mathbb{R}^p:\|\boldsymbol{b}\|_0\le(1-c)n} \|\boldsymbol{X}\boldsymbol{b}\|^2/(n\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{b}\|^2) > \eta\Big\},$$

$$U_3 = \big\{\|\boldsymbol{X}\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_{\text{op}} < 2\sqrt{n} + \sqrt{p}\big\}.$$

Under Assumptions 2 and 3, (Bellec and Zhang, 2019, Lemma B.1) guarantees $\mathbb{P}(U_2) \ge 1 - C(\gamma, c)e^{-C(\gamma,c)n}$ for some constant η that only depends on constants γ, c. Under Assumption 2, (Davidson and Szarek, 2001, Theorem II.13) guarantees $\mathbb{P}(U_3) > 1 - e^{-n/2}$ and there exists a random variable $z \sim \mathcal{N}(0,1)$ s.t. $\|\boldsymbol{X}\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_{\text{op}} \le \sqrt{n} + \sqrt{p} + z$ almost surely. Therefore, under Assumptions 2 and 3, we have

$$\mathbb{E}[\|n^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\Sigma}^{-\frac{1}{2}}\|_{\text{op}}^2] \le (1 + \sqrt{p/n})^2 + n^{-1} \le C(\gamma). \tag{36}$$

Furthermore, under Assumptions 2, 3 and 4(ii), $\mathbb{P}(U_1 \cap U_2 \cap U_3) \to 1$ by a union bound, and for large enough n,

$$\begin{aligned}\mathbb{P}\big\{(U_1 \cap U_2 \cap U_3)^c\big\} &< 1/T + C(\gamma, c)e^{-n/C(\gamma,c)} \\ &= \frac{1}{T}(1 + TC(\gamma, c)e^{-n/C(\gamma,c)})\end{aligned}$$

$$< \frac{1}{T}(1 + C(\gamma, c)e^{\sqrt{n} - n/C(\gamma,c)})$$
$$< \frac{1}{T}C(\gamma, c). \tag{37}$$

E.1. Lipschitz and differential properties for a given, fixed noise matrix $\boldsymbol{E}$

We need to study Lipschitz and differential properties of certain mappings when the noise matrix $\boldsymbol{E}$ is fixed. Let $g : \mathbb{R}^{p\times T} \to \mathbb{R}$ defined by $g(\boldsymbol{B}) = \tau\|\boldsymbol{B}\|_{\mathrm{F}}^2/2 + \lambda\|\boldsymbol{B}\|_{2,1}$ be the penalty in (6). For a fixed value of $\boldsymbol{E}$, define the mappings

$$\boldsymbol{Z} \longmapsto \boldsymbol{H}(\boldsymbol{Z}) = \operatorname*{arg\,min}_{\bar{\boldsymbol{H}}\in\mathbb{R}^{p\times T}} \frac{1}{2n}\|\boldsymbol{E} - \boldsymbol{Z}\bar{\boldsymbol{H}}\|_{\mathrm{F}}^2 + g(\boldsymbol{\Sigma}^{-1/2}\bar{\boldsymbol{H}}) \qquad (\mathbb{R}^{n\times p} \to \mathbb{R}^{p\times T}) \tag{38}$$

$$\boldsymbol{Z} \longmapsto \boldsymbol{F}(\boldsymbol{Z}) = \boldsymbol{E} - \boldsymbol{Z}\boldsymbol{H}(\boldsymbol{Z}) \qquad (\mathbb{R}^{n\times p} \to \mathbb{R}^{n\times T}) \tag{39}$$

$$\boldsymbol{Z} \longmapsto D(\boldsymbol{Z}) = (\|\boldsymbol{H}(\boldsymbol{Z})\|_{\mathrm{F}}^2 + \|\boldsymbol{F}(\boldsymbol{Z})\|_{\mathrm{F}}^2/n)^{1/2} \qquad (\mathbb{R}^{n\times p} \to \mathbb{R}). \tag{40}$$

Next, define the random variable $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Sigma}^{-\frac{1}{2}} \in \mathbb{R}^{n\times p}$, and let us use the convention that if arguments of $\boldsymbol{H}, \boldsymbol{F}$ or D are omitted then these mappings are implicitly taken at the realized value of the random variable $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Sigma}^{-\frac{1}{2}} \in \mathbb{R}^{n\times p}$ where $\boldsymbol{X}$ is the observed design matrix. With this convention and by definition of the above mappings, we then have $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{Z}) = \boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)$ as well as $\boldsymbol{F} = \boldsymbol{F}(\boldsymbol{Z}) = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}$ and $D = [\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{F}\|_{\mathrm{F}}^2/n]^{1/2}$ so that the notation is consistent with the rest of the paper (in particular, with (9) in the main paper).

Finally, denote the (i, j)-th entry of $\boldsymbol{Z}$ by z_{ij} throughout this supplement, and the corresponding partial derivatives of the above mappings by $\frac{\partial}{\partial z_{ij}}$.

Lemma E.1. *For multi-task elastic-net (i.e., $\tau > 0$ in (6)), the mapping $\boldsymbol{Z} \mapsto D^{-1}\boldsymbol{F}/\sqrt{n}$ is $n^{-\frac{1}{2}}L$-Lipschitz with $L = 8\max(1, (2\tau')^{-1})$, where $\tau' = \tau/\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$.*

Lemma E.2. *For multi-task group Lasso (i.e., $\tau = 0$ in (6)). we have*

(1) *In $U_1 \cap U_2$, the map $\boldsymbol{Z} \mapsto D^{-1}\boldsymbol{F}/\sqrt{n}$ is $n^{-\frac{1}{2}}L$-Lipschitz with $L = 8\max(1, (2\eta)^{-1})$.*
(2) *In $U_1 \cap U_2 \cap U_3$, the map $\boldsymbol{Z} \mapsto D^{-1}\boldsymbol{Z}^\top\boldsymbol{F}/n$ is $n^{-1/2}(1 + (2 + \sqrt{p/n})L)$-Lipschitz, where $L = 8\max(1, (2\eta)^{-1})$ as in (1).*

Corollary E.3. *Suppose that Assumption 4 holds, then*

(1) *Under Assumption 4(i) that $\tau > 0$ and $\tau' = \tau/\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$, we have*

$$\sum_{ij}\left(\frac{\partial D}{\partial z_{ij}}\right)^2 \le n^{-1}D^2[4\max(1, (2\tau')^{-1})]^2.$$

This implies that $nD^{-2}\sum_{ij}(\frac{\partial D}{\partial z_{ij}})^2 \le C(\tau')$.

(2) *Under Assumption 4(ii) that $\tau = 0$ and $\mathbb{P}(U_1) \to 1$, in the event $U_1 \cap U_2$, we have*

$$\sum_{ij}\left(\frac{\partial D}{\partial z_{ij}}\right)^2 \le n^{-1}D^2[4\max(1, (2\eta)^{-1})]^2.$$

This implies that $nD^{-2}\sum_{ij}(\frac{\partial D}{\partial z_{ij}})^2 \le C(\eta) = C(\gamma, c)$ *since* η *is a constant that only depends on* γ, c.

Note that with a fixed noise $\boldsymbol{E}$, Lemmas E.1 and E.2 guarantee that the map $\boldsymbol{Z} \mapsto \boldsymbol{F}$ is Lipschitz, hence the derivative exists almost everywhere by Rademacher's theorem. We present the formula for derivative of this map in Lemma E.4.

Lemma E.4. *Recall* $\boldsymbol{F} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}$ *with* $\widehat{\boldsymbol{B}}$ *defined in* (6). *Under Assumption 4(i)* $\tau > 0$, *or under Assumption 4(ii)* $\tau = 0$ *and in the event* $U_1 \cap U_2$, *for each* $i, l \in [n], j \in [p], t \in [T]$, *the following derivative exists almost everywhere and has the expression*

$$\frac{\partial F_{lt}}{\partial z_{ij}} = D_{ij}^{lt} + \Delta_{ij}^{lt},$$

where $D_{ij}^{lt} = -(\boldsymbol{e}_j^\top \boldsymbol{H} \otimes \boldsymbol{e}_i^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{e}_t \otimes \boldsymbol{e}_l)$, *and* $\Delta_{ij}^{lt} = -(\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\big(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p\big)(\boldsymbol{e}_i \otimes \boldsymbol{e}_j)$. *Furthermore, a straightforward calculation yields*

$$\sum_{i=1}^{n} D_{ij}^{it} = -\boldsymbol{e}_j^\top \boldsymbol{H}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{e}_t.$$

Lemma E.5. *Suppose that Assumption 4 holds.*

(1) *Under Assumption 4(i) that* $\tau > 0$ *and* $\tau' = \tau/\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$, *we have*

$$\frac{1}{n}\sum_{ij}\left\|\frac{\partial(\boldsymbol{F}/D)}{\partial z_{ij}}\right\|_{\mathrm{F}}^2 \le \underbrace{4\max(1, (\tau')^{-1}(T \wedge \frac{p}{n})) + 2n^{-1}[4\max(1, (2\tau')^{-1})]^2}_{f(\tau', T, n, p)}.$$

(2) *Under Assumption 4(ii) that* $\tau = 0$ *and* $\mathbb{P}(U_1) \to 1$, *in the event* $U_1 \cap U_2$, *we have*

$$\frac{1}{n}\sum_{ij}\left\|\frac{\partial(\boldsymbol{F}/D)}{\partial z_{ij}}\right\|_{\mathrm{F}}^2 \le \underbrace{4\max(1, (\eta)^{-1}(T \wedge \frac{p}{n})) + 2n^{-1}[4\max(1, (2\eta)^{-1})]^2}_{f(\eta, T, n, p)}.$$

Furthermore, the right-hand side in (1) can be bounded from above by $C(\tau')(T \wedge \frac{p}{n})$, *and the right-hand side in (2) can be bounded from above by* $C(\gamma, c)$ *in the regime* $p/n \le \gamma$.

E.2. Lipschitz and differential properties for a given, fixed design matrix

We also need to study Lipschitz and derivative properties of functions of the noise $\boldsymbol{E}$ when the design $\boldsymbol{X}$ is fixed. Formally, for a given and fixed design matrix $\boldsymbol{X}$, define the function $\boldsymbol{E} \mapsto \boldsymbol{F}(\boldsymbol{E})$ by the value $\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}$ of the residual matrix when the observed data $(\boldsymbol{X}, \boldsymbol{Y})$ is $(\boldsymbol{X}, \boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E})$ and with $\widehat{\boldsymbol{B}}$ the estimator (6). Note that this map is 1-Lipschitz by (Bellec and Tsybakov, 2016, Proposition 3), Rademacher's theorem thus guarantees this map is differentiable almost everywhere. We denote its partial derivative by $\frac{\partial}{\partial E_{it}}$ for each entry $(E_{it})_{i\in[n], t\in[T]}$ of the noise matrix $\boldsymbol{E}$. We present its derivative formula in Lemma E.6 below.

Lemma E.6. *For each $i, l \in [n], t, t' \in [T]$, the following derivative exists almost everywhere and has the expression*

$$\frac{\partial F_{lt}}{\partial E_{it'}} = \boldsymbol{e}_l^\top \boldsymbol{e}_i \boldsymbol{e}_t^\top \boldsymbol{e}_{t'} - \boldsymbol{e}_l^\top (\boldsymbol{e}_t^\top \otimes \boldsymbol{X}) \boldsymbol{M}^\dagger (\boldsymbol{e}_{t'} \otimes \boldsymbol{X}^\top) \boldsymbol{e}_i.$$

As a consequence, we further have

$$\sum_{i=1}^n \frac{\partial F_{it}}{\partial E_{it'}} = \boldsymbol{e}_t^\top (n\boldsymbol{I}_T - \widehat{\mathbf{A}}) \boldsymbol{e}_{t'}, \quad \sum_{i=1}^n \frac{\partial \boldsymbol{e}_i^\top \boldsymbol{Z}\boldsymbol{H}\boldsymbol{e}_t}{\partial E_{it'}} = \boldsymbol{e}_t^\top \widehat{\mathbf{A}} \boldsymbol{e}_{t'}.$$

E.3. Probabilistic tools

We first list some useful variants of Stein's formulae and Gaussian-Poincaré inequalities. Let f' denote the derivative of a differentiable univariate function. For a differentiable vector-valued function $\boldsymbol{f}(z) : \mathbb{R}^n \to \mathbb{R}^n$, denote its Jacobian (derivative) and divergence respectively by $\nabla \boldsymbol{f}(z)$ and $\operatorname{div} \boldsymbol{f}(z)$, i.e., $[\nabla \boldsymbol{f}(z)]_{i,l} = \frac{\partial f_i(z)}{\partial z_l}$ for $i, l \in [n]$, and $\operatorname{div} \boldsymbol{f}(z) = \operatorname{Tr}(\nabla \boldsymbol{f}(z))$.

Lemma E.7 (Second-order Stein's formula Bellec and Zhang (2021)). *The following identities hold provided the involved derivatives exist a.e. and the expectations are finite.*

i) *$z \sim \mathcal{N}(0,1)$, $f : \mathbb{R} \to \mathbb{R}$, then*

$$\mathbb{E}[(zf(z) - f'(z))^2] = \mathbb{E}[f(z)^2] + \mathbb{E}[(f'(z))^2].$$

ii) *$\boldsymbol{z} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$, $f : \mathbb{R}^n \to \mathbb{R}^n$, then*

$$\mathbb{E}[(\boldsymbol{z}^\top \boldsymbol{f}(\boldsymbol{z}) - \operatorname{div} \boldsymbol{f}(\boldsymbol{z}))^2] = \mathbb{E}\big[\|\boldsymbol{f}(\boldsymbol{z})\|^2 + \operatorname{Tr}[(\nabla \boldsymbol{f}(\boldsymbol{z}))^2]\big] \le \mathbb{E}\big[\|\boldsymbol{f}(\boldsymbol{z})\|^2 + \|\nabla \boldsymbol{f}(\boldsymbol{z})\|_{\mathrm{F}}^2\big],$$

where the inequality uses Cauchy-Schwarz inequality.

iii) *More generally, for $\boldsymbol{z} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{\Sigma})$, $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$, then*

$$\begin{aligned}\mathbb{E}[(\boldsymbol{z}^\top \boldsymbol{f}(\boldsymbol{z}) - \operatorname{Tr}(\boldsymbol{\Sigma}\nabla \boldsymbol{f}(\boldsymbol{z}))^2] &= \mathbb{E}\big[\|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{f}(\boldsymbol{z})\|^2 + \operatorname{Tr}[(\boldsymbol{\Sigma}\nabla \boldsymbol{f}(\boldsymbol{z}))^2]\big] \\ &\le \mathbb{E}\big[\|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{f}(\boldsymbol{z})\|^2 + \|(\boldsymbol{\Sigma}\nabla \boldsymbol{f}(\boldsymbol{z})\|_{\mathrm{F}}^2]\big],\end{aligned}$$

where the inequality uses Cauchy-Schwarz inequality.

Lemma E.8 (Gaussian-Poincaré inequality Boucheron, Lugosi and Massart (2013)). *The following inequalities hold provided the right-hand side derivatives exist a.e. and the expectations are finite.*

i) *$z \sim \mathcal{N}(0,1)$, $f : \mathbb{R} \to \mathbb{R}$, then*

$$Var[f(z)] \le \mathbb{E}[(f'(z))^2].$$

ii) *$\boldsymbol{z} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$, $f : \mathbb{R}^n \to \mathbb{R}$, then*

$$Var[f(\boldsymbol{z})] \le \mathbb{E}[\|\nabla f(\boldsymbol{z})\|^2].$$

iii) *$\boldsymbol{z} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$, $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, then*

$$\mathbb{E}[\|\boldsymbol{f}(\boldsymbol{z}) - \mathbb{E}[\boldsymbol{f}(\boldsymbol{z})]\|^2] \le \mathbb{E}[\|\nabla \boldsymbol{f}(\boldsymbol{z})\|_{\mathrm{F}}^2].$$

iv) *More generally, for* $\boldsymbol{z} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{\Sigma})$, $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, *then*

$$\mathbb{E}[\|\boldsymbol{f}(\boldsymbol{z}) - \mathbb{E}[\boldsymbol{f}(\boldsymbol{z})]\|^2] \le \mathbb{E}[\|\boldsymbol{\Sigma}^{\frac{1}{2}} \nabla \boldsymbol{f}(\boldsymbol{z})\|_{\mathrm{F}}^2].$$

Now we present a few important lemmas, whose proofs are based on Lemma E.7 and Lemma E.8.

Lemma E.9. *Assume that Assumption 1 holds. For fixed* $\boldsymbol{X}$, *we have*

$$\mathbb{E}\Big[\|\boldsymbol{E}^\top \boldsymbol{F}/\tilde{D} - \boldsymbol{S}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})/\tilde{D}\|_{\mathrm{F}}^2\Big] \le 4\,\mathrm{Tr}(\boldsymbol{S}),$$

where $\tilde{D} = \big(\|\boldsymbol{F}\|_{\mathrm{F}}^2 + n\,\mathrm{Tr}(\boldsymbol{S})\big)^{\frac{1}{2}}$.

Lemma E.10. *Let* $\boldsymbol{U}, \boldsymbol{V} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times T}$ *be two locally Lipschitz functions of* $\boldsymbol{Z}$ *with i.i.d.* $\mathcal{N}(0,1)$ *entries, then*

$$\mathbb{E}\Big[\Big\|\boldsymbol{U}^\top \boldsymbol{Z}\boldsymbol{V} - \sum_{j=1}^{p}\sum_{i=1}^{n} \frac{\partial}{\partial z_{ij}}\Big(\boldsymbol{U}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{V}\Big)\Big\|_{\mathrm{F}}^2\Big]$$

$$\le \mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}}^2\|\boldsymbol{V}\|_{\mathrm{F}}^2 + \mathbb{E}\sum_{ij}\Big[2\|\boldsymbol{V}\|_{\mathrm{F}}^2\Big\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2 + 2\|\boldsymbol{U}\|_{\mathrm{F}}^2\Big\|\frac{\partial \boldsymbol{V}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big].$$

Corollary E.11. *Assume the same setting as Lemma E.10. If on some open set* $\Omega \subset \mathbb{R}^{n\times p}$ *with* $\mathbb{P}(\Omega^c) \le C/T$ *for some constant* C, *we have (i)* $\boldsymbol{U}$ *is* $n^{-1/2}L_1$ *-Lipschitz and* $\|\boldsymbol{U}\|_{\mathrm{F}} \le 1$, *(ii)* $\boldsymbol{V}$ *is* $n^{-1/2}L_2$ *-Lipschitz and* $\|\boldsymbol{V}\|_{\mathrm{F}} \le K$. *Then*

$$\mathbb{E}\Big[I(\Omega)\Big\|\boldsymbol{U}^\top \boldsymbol{Z}\boldsymbol{V} - \sum_{j=1}^{p}\sum_{i=1}^{n} \frac{\partial}{\partial z_{ij}}\Big(\boldsymbol{U}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{V}\Big)\Big\|_{\mathrm{F}}^2\Big]$$

$$\le K^2 + 2C(K^2L_1^2 + L_2^2) + 2\mathbb{E}\Big[I(\Omega)\sum_{ij}\Big(K^2\Big\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2 + \Big\|\frac{\partial \boldsymbol{V}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big)\Big].$$

Lemma E.12. *Let* $\boldsymbol{U}, \boldsymbol{V} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times T}$ *be two locally Lipschitz functions of* $\boldsymbol{Z}$ *with i.i.d.* $\mathcal{N}(0,1)$ *entries. Assume also that* $\|\boldsymbol{U}\|_{\mathrm{F}} \vee \|\boldsymbol{V}\|_{\mathrm{F}} \le 1$ *almost surely. Then*

$$\mathbb{E}\Big[\Big\|p\boldsymbol{U}^\top \boldsymbol{V} - \sum_{j=1}^{p}\Big(\sum_{i=1}^{n}\partial_{ij}\boldsymbol{U}^\top \boldsymbol{e}_i - \boldsymbol{U}^\top \boldsymbol{Z}\boldsymbol{e}_j\Big)\Big(\sum_{i=1}^{n}\partial_{ij}\boldsymbol{e}_i^\top \boldsymbol{V} - \boldsymbol{e}_j^\top \boldsymbol{Z}^\top \boldsymbol{V}\Big)\Big\|_{\mathrm{F}}\Big]$$

$$\le 2\|\boldsymbol{U}\|_\partial\|\boldsymbol{V}\|_\partial + \sqrt{p}\big(\sqrt{2} + (3+\sqrt{2})(\|\boldsymbol{U}\|_\partial + \|\boldsymbol{V}\|_\partial)\big),$$

where $\partial_{ij}\boldsymbol{U} \overset{def}{=} \partial \boldsymbol{U}/\partial z_{ij}$, *and* $\|\boldsymbol{U}\|_\partial \overset{def}{=} \mathbb{E}[\sum_{i=1}^{n}\sum_{j=1}^{p}\|\partial_{ij}\boldsymbol{U}\|_{\mathrm{F}}^2]^{\frac{1}{2}}$.

Lemma E.13. *Let* $\boldsymbol{N} = (\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{X}^\top)$, *we have* $\|\boldsymbol{N}\|_{\mathrm{op}} \le 1$.

Lemma E.14. *We have*

$$\sum_{ij} \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j\boldsymbol{e}_i^\top \frac{\partial \boldsymbol{F}}{\partial z_{ij}} = \boldsymbol{J}_1 - \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{H}(n\boldsymbol{I}_T - \widehat{\mathbf{A}}),$$
$$\sum_{ij} \Big(\frac{\partial \boldsymbol{F}}{\partial z_{ij}}\Big)^\top \boldsymbol{Z}\boldsymbol{e}_j\boldsymbol{e}_i^\top \boldsymbol{F} = \boldsymbol{J}_2 - \widehat{\mathbf{A}}\boldsymbol{F}^\top \boldsymbol{F}.$$

where $\boldsymbol{J}_1 = \sum_{ij} \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j \sum_t \Delta_{ij}^{it}\boldsymbol{e}_t^\top$ *with* $\|\boldsymbol{J}_1\|_{\mathrm{F}} \le n^{\frac{1}{2}}\|\boldsymbol{F}\|_{\mathrm{F}}^2$, *and* $\boldsymbol{J}_2 = \sum_{ijlt} \boldsymbol{e}_t D_{ij}^{lt}\boldsymbol{e}_l^\top \boldsymbol{Z}\boldsymbol{e}_j\boldsymbol{e}_i^\top \boldsymbol{F}$ *with* $\|\boldsymbol{J}_2\|_{\mathrm{F}} \le n^{\frac{1}{2}}\|\boldsymbol{Z}\|_{\mathrm{op}}\|\boldsymbol{H}\|_{\mathrm{F}}\|\boldsymbol{F}\|_{\mathrm{F}}$.

F. Proofs of Propositions B.3 to B.5

F.1. Proof of Proposition B.3

Proof of Proposition B.3. Note that $\boldsymbol{Q}_1 = \frac{\boldsymbol{E}^\top \boldsymbol{F}/\tilde{D} - \boldsymbol{S}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})/\tilde{D}}{\|\boldsymbol{S}^{\frac{1}{2}}\|_{\mathrm{F}}}$, where $\tilde{D} = \big(\|\boldsymbol{F}\|_{\mathrm{F}}^2 + n\,\mathrm{Tr}(\boldsymbol{S})\big)^{\frac{1}{2}}$ is defined in Lemma E.9. Now, apply Lemma E.9, we obtain

$$\mathbb{E}\|\boldsymbol{Q}_1\|_{\mathrm{F}}^2 = \mathbb{E}\big[\|\boldsymbol{E}^\top \boldsymbol{F}/\tilde{D} - \boldsymbol{S}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})/\tilde{D}\|_{\mathrm{F}}^2\big]\frac{1}{\mathrm{Tr}(\boldsymbol{S})} \le 4.$$

■

F.2. Proof of Proposition B.4

Proof of Proposition B.4. We first apply Lemma E.10. To be more specific, let $\boldsymbol{U} = n^{-\frac{1}{2}}\boldsymbol{F}/D$ and $\boldsymbol{V} = n^{-\frac{1}{2}}\boldsymbol{Z}^\top \boldsymbol{U}$ with $D = (\|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \|\boldsymbol{H}\|_{\mathrm{F}}^2)^{\frac{1}{2}}$, then $\|\boldsymbol{U}\|_{\mathrm{F}} \le 1$, $\|\boldsymbol{V}\|_{\mathrm{F}} \le n^{-\frac{1}{2}}\|\boldsymbol{Z}\|_{\mathrm{op}}$. Lemma E.10 yields

$$\mathbb{E}\Big(\Big\|\boldsymbol{U}^\top \boldsymbol{Z}\boldsymbol{V} - \sum_{j=1}^{p}\sum_{i=1}^{n} \frac{\partial}{\partial z_{ij}}\Big(\boldsymbol{U}^\top \boldsymbol{e}_i\boldsymbol{e}_j^\top \boldsymbol{V}\Big)\Big\|_{\mathrm{F}}^2\Big) \tag{41}$$
$$\le \mathbb{E}[\|\boldsymbol{U}\|_{\mathrm{F}}^2\|\boldsymbol{V}\|_{\mathrm{F}}^2] + \mathbb{E}\sum_{ij}\Big[2\|\boldsymbol{V}\|_{\mathrm{F}}^2\Big\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2 + 2\|\boldsymbol{U}\|_{\mathrm{F}}^2\Big\|\frac{\partial \boldsymbol{V}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big]$$
$$\le \mathbb{E}\big(n^{-1}\|\boldsymbol{Z}\|_{\mathrm{op}}^2\big) + 2\mathbb{E}\Big(n^{-1}\|\boldsymbol{Z}\|_{\mathrm{op}}^2\sum_{ij}\Big\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big) + 2\mathbb{E}\Big(\sum_{ij}\Big\|\frac{\partial \boldsymbol{V}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big) \tag{42}$$
$$\le \frac{4p}{n} + \mathbb{E}\big(\frac{1}{n}\|\boldsymbol{Z}\|_{\mathrm{op}}^2\big) + 6\mathbb{E}\Big(n^{-1}\|\boldsymbol{Z}\|_{\mathrm{op}}^2\sum_{ij}\Big\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big),$$

where the last inequality uses the following bound derived using $\boldsymbol{V} = n^{-\frac{1}{2}}\boldsymbol{Z}^\top \boldsymbol{U}$, and $\|\boldsymbol{U}\|_{\mathrm{F}} \le 1$,

$$\sum_{ij}\Big\|\frac{\partial \boldsymbol{V}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2 = \sum_{ij} n^{-1}\Big\|\frac{\partial \boldsymbol{Z}^\top}{\partial z_{ij}}\boldsymbol{U} + \boldsymbol{Z}^\top \frac{\partial \boldsymbol{U}^\top}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2$$

$$\leq 2n^{-1}\Big(p\|\boldsymbol{U}\|_{\mathrm{F}}^2 + \|\boldsymbol{Z}\|_{\mathrm{op}}^2 \sum_{ij}\Big\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big)$$

$$\leq \frac{2p}{n} + 2n^{-1}\|\boldsymbol{Z}\|_{\mathrm{op}}^2 \sum_{ij}\Big\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2. \tag{43}$$

Note that $\mathbb{E}[(n^{-\frac{1}{2}}\|\boldsymbol{Z}\|_{\mathrm{op}})^2] \leq (1+\sqrt{p/n})^2 + 1/n$ by (36). Now we establish the connection between $\boldsymbol{Q}_2$ and the term inside Frobenius norm in (41). By definitions of $\boldsymbol{U}$ and $\boldsymbol{V}$,

$$\boldsymbol{U}^\top \boldsymbol{Z}\boldsymbol{V} = n^{-\frac{3}{2}} D^{-2}\boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{Z}^\top \boldsymbol{F}. \tag{44}$$

Next, by product rule,

$$\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\partial}{\partial z_{ij}}\Big(\boldsymbol{U}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{V}\Big)$$

$$= n^{-\frac{3}{2}}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\partial}{\partial z_{ij}}\Big(\boldsymbol{F}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{Z}^\top \boldsymbol{F} D^{-2}\Big)$$

$$= n^{-\frac{3}{2}}\sum_{j=1}^{p}\sum_{i=1}^{n}\Big(\underbrace{\frac{\partial \boldsymbol{F}^\top}{\partial z_{ij}}\boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{Z}^\top \boldsymbol{F} D^{-2}}_{(i)} + \underbrace{\boldsymbol{F}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \frac{\partial \boldsymbol{Z}^\top \boldsymbol{F}}{\partial z_{ij}} D^{-2}}_{(ii)} + \underbrace{\boldsymbol{F}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{Z}^\top \boldsymbol{F}\frac{\partial D^{-2}}{\partial z_{ij}}}_{(iii)}\Big).$$

We now rewrite the above three terms (i), (ii) and (iii).

(i) For term (i), by Lemma E.14,

$$n^{-\frac{3}{2}} D^{-2}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\partial \boldsymbol{F}^\top}{\partial z_{ij}}\boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{Z}^\top \boldsymbol{F}$$

$$= n^{-\frac{3}{2}} D^{-2}\big[\boldsymbol{J}_1 - \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{H}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})\big]^\top$$

$$= n^{-\frac{3}{2}} D^{-2}\big[\boldsymbol{J}_1^\top - (n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{H}^\top \boldsymbol{Z}^\top \boldsymbol{F}\big].$$

(ii) For term (ii), by product rule and Lemma E.14,

$$n^{-\frac{3}{2}} D^{-2}\sum_{j=1}^{p}\sum_{i=1}^{n}\boldsymbol{F}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \frac{\partial \boldsymbol{Z}^\top \boldsymbol{F}}{\partial z_{ij}}$$

$$= n^{-\frac{3}{2}} D^{-2}\Big(p\boldsymbol{F}^\top \boldsymbol{F} + \sum_{j=1}^{p}\sum_{i=1}^{n}\boldsymbol{F}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{Z}^\top \frac{\partial \boldsymbol{F}}{\partial z_{ij}}\Big)$$

$$= n^{-\frac{3}{2}} D^{-2}\Big(p\boldsymbol{F}^\top \boldsymbol{F} + (\boldsymbol{J}_2 - \widehat{\mathbf{A}}\boldsymbol{F}^\top \boldsymbol{F})^\top\Big)$$

$$= n^{-\frac{3}{2}} D^{-2}\big(p\boldsymbol{F}^\top \boldsymbol{F} - \boldsymbol{F}^\top \boldsymbol{F}\widehat{\mathbf{A}} + \boldsymbol{J}_2^\top\big).$$

(iii) For term (*iii*), by chain rule,

$$
\begin{aligned}
& n^{-\frac{3}{2}} \sum_{j=1}^{p} \sum_{i=1}^{n} \boldsymbol{F}^{\top} \boldsymbol{e}_i \boldsymbol{e}_j^{\top} \boldsymbol{Z}^{\top} \boldsymbol{F} \frac{\partial D^{-2}}{\partial z_{ij}} \\
= & -2 n^{-\frac{3}{2}} D^{-3} \sum_{j=1}^{p} \sum_{i=1}^{n} \boldsymbol{F}^{\top} \boldsymbol{e}_i \boldsymbol{e}_j^{\top} \boldsymbol{Z}^{\top} \boldsymbol{F} \frac{\partial D}{\partial z_{ij}} \\
= & -n^{-\frac{3}{2}} D^{-2} \underbrace{\Big(2 D^{-1} \sum_{j=1}^{p} \sum_{i=1}^{n} \boldsymbol{F}^{\top} \boldsymbol{e}_i \boldsymbol{e}_j^{\top} \boldsymbol{Z}^{\top} \boldsymbol{F} \frac{\partial D}{\partial z_{ij}} \Big)}_{\boldsymbol{J}_3}
\end{aligned}
$$

Combining (44) and the above three expressions for (i)-(ii)-(iii),

$$
\begin{aligned}
& \boldsymbol{U}^{\top} \boldsymbol{Z} \boldsymbol{V} - \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{\partial}{\partial z_{ij}} \Big(\boldsymbol{U}^{\top} \boldsymbol{e}_i \boldsymbol{e}_j^{\top} \boldsymbol{V} \Big) \\
= & \, n^{-\frac{3}{2}} D^{-2} \Big[\boldsymbol{F}^{\top} \boldsymbol{Z} \boldsymbol{Z}^{\top} \boldsymbol{F} + (n \boldsymbol{I}_T - \widehat{\mathbf{A}}) \boldsymbol{H}^{\top} \boldsymbol{Z}^{\top} \boldsymbol{F} - p \boldsymbol{F}^{\top} \boldsymbol{F} + \boldsymbol{F}^{\top} \boldsymbol{F} \widehat{\mathbf{A}} \\
& - \boldsymbol{J}_1^{\top} - \boldsymbol{J}_2^{\top} + \boldsymbol{J}_3 \Big] \\
= & \, \boldsymbol{Q}_2 - n^{-\frac{3}{2}} D^{-2} (\boldsymbol{J}_1^{\top} + \boldsymbol{J}_2^{\top} - \boldsymbol{J}_3).
\end{aligned}
$$

That is,

$$
\boldsymbol{Q}_2 = \boldsymbol{U}^{\top} \boldsymbol{Z} \boldsymbol{V} - \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{\partial}{\partial z_{ij}} \Big(\boldsymbol{U}^{\top} \boldsymbol{e}_i \boldsymbol{e}_j^{\top} \boldsymbol{V} \Big) + n^{-\frac{3}{2}} D^{-2} (\boldsymbol{J}_1^{\top} + \boldsymbol{J}_2^{\top} - \boldsymbol{J}_3). \tag{45}
$$

Note that Lemma E.14 implies that

$$
n^{-\frac{3}{2}} D^{-2} \|\boldsymbol{J}_1\|_{\mathrm{F}} \leq \frac{\|\boldsymbol{F}\|_{\mathrm{F}}^2 / n}{\|\boldsymbol{F}\|_{\mathrm{F}}^2 / n + \|\boldsymbol{H}\|_{\mathrm{F}}^2} \leq 1, \tag{46}
$$

and

$$
n^{-\frac{3}{2}} D^{-2} \|\boldsymbol{J}_2\|_{\mathrm{F}} \leq \Big[n^{-\frac{1}{2}} \|\boldsymbol{Z}\|_{\mathrm{op}} \frac{\|\boldsymbol{F}\|_{\mathrm{F}} n^{-\frac{1}{2}} \|\boldsymbol{H}\|_{\mathrm{F}}}{\|\boldsymbol{F}\|_{\mathrm{F}}^2 / n + \|\boldsymbol{H}\|_{\mathrm{F}}^2} \Big] \leq \frac{1}{2} (n^{-\frac{1}{2}} \|\boldsymbol{Z}\|_{\mathrm{op}}). \tag{47}
$$

Since $\boldsymbol{J}_3 = 2 D^{-1} \sum_{j=1}^{p} \sum_{i=1}^{n} \boldsymbol{F}^{\top} \boldsymbol{e}_i \boldsymbol{e}_j^{\top} \boldsymbol{Z}^{\top} \boldsymbol{F} \frac{\partial D}{\partial z_{ij}}$, by Cauchy-Schwarz inequality

$$
\begin{aligned}
n^{-3} D^{-4} \|\boldsymbol{J}_3\|_{\mathrm{F}}^2 = & \sum_{t,t'} ([\boldsymbol{J}_3]_{t,t'})^2 \\
= & \, 4 n^{-3} D^{-6} \sum_{t,t'} \Big[\sum_{i,j} \boldsymbol{e}_t^{\top} \boldsymbol{F}^{\top} \boldsymbol{e}_i \boldsymbol{e}_j^{\top} \boldsymbol{Z}^{\top} \boldsymbol{F} \boldsymbol{e}_{t'} \frac{\partial D}{\partial z_{ij}} \Big]^2 \\
\leq & \, 4 n^{-3} D^{-6} \sum_{t,t'} \Big[\sum_{i,j} [\boldsymbol{e}_t^{\top} \boldsymbol{F}^{\top} \boldsymbol{e}_i \boldsymbol{e}_j^{\top} \boldsymbol{Z}^{\top} \boldsymbol{F} \boldsymbol{e}_{t'}]^2 \sum_{i,j} \Big(\frac{\partial D}{\partial z_{ij}} \Big)^2 \Big]
\end{aligned}
$$

$$= 4n^{-3}D^{-6}\|\boldsymbol{F}\|_{\mathrm{F}}^2\|\boldsymbol{Z}^\top\boldsymbol{F}\|_{\mathrm{F}}^2\sum_{i,j}\big(\frac{\partial D}{\partial z_{ij}}\big)^2$$

$$= 4n^{-3}D^{-6}\|\boldsymbol{F}\|_{\mathrm{F}}^4\|\boldsymbol{Z}\|_{\mathrm{op}}^2\sum_{i,j}\big(\frac{\partial D}{\partial z_{ij}}\big)^2.$$

(1) Under Assumption 4(i) that $\tau > 0$. By Corollary E.3, we have $\sum_{i,j}\big(\frac{\partial D}{\partial z_{ij}}\big)^2 \le C(\tau')n^{-1}D^2$. Then,

$$n^{-3}D^{-4}\|\boldsymbol{J}_3\|_{\mathrm{F}}^2 \le C(\tau')n^{-4}D^2\|\boldsymbol{F}\|_{\mathrm{F}}^4\|\boldsymbol{Z}\|_{\mathrm{op}}^2 \le C(\tau')n^{-2}\|\boldsymbol{Z}\|_{\mathrm{op}}^2, \tag{48}$$

by $\|\boldsymbol{F}\|_{\mathrm{F}}^2/(nD^2) \le 1$.

By (45) and triangular inequality, we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{Q}_2\|_{\mathrm{F}}^2] \le 2\mathbb{E}\Big[\|\boldsymbol{U}^\top\boldsymbol{Z}\boldsymbol{V} - \sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\partial}{\partial z_{ij}}\Big(\boldsymbol{U}^\top\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{V}\Big)\|_{\mathrm{F}}^2\Big] \\ + 2\mathbb{E}[\|n^{-\frac{3}{2}}D^{-2}(\boldsymbol{J}_1^\top + \boldsymbol{J}_2^\top - \boldsymbol{J}_3)\|_{\mathrm{F}}^2].\end{aligned} \tag{49}$$

By (46), (47), (48), the second term in (49) can be upper bounded by

$$6\mathbb{E}[(1 + \frac{1}{4}n^{-1}\|\boldsymbol{Z}\|_{\mathrm{op}}^2) + C(\tau')n^{-2}\|\boldsymbol{Z}\|_{\mathrm{op}}^2)] \le C(\tau')(1 + \frac{p}{n}), \tag{50}$$

where the last inequality uses (36).

For the first term in (49), since $\sum_{ij}\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\|_{\mathrm{F}}^2 \le C(\tau')(T \wedge \frac{p}{n})$ by Lemma E.5, we have

$$\begin{aligned}&\mathbb{E}\Big(\Big\|\boldsymbol{U}^\top\boldsymbol{Z}\boldsymbol{V} - \sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\partial}{\partial z_{ij}}\Big(\boldsymbol{U}^\top\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{V}\Big)\Big\|_{\mathrm{F}}^2\Big) \\ &\le \frac{4p}{n} + \mathbb{E}\big(n^{-1}\|\boldsymbol{Z}\|_{\mathrm{op}}^2\big) + 6\mathbb{E}\Big(n^{-1}\|\boldsymbol{Z}\|_{\mathrm{op}}^2\sum_{ij}\Big\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big) \\ &\le \frac{4p}{n} + [1 + C(\tau')(T \wedge \frac{p}{n})]\mathbb{E}\big(n^{-1}\|\boldsymbol{Z}\|_{\mathrm{op}}^2\big) \\ &\le \frac{4p}{n} + [1 + C(\tau')(T \wedge \frac{p}{n})][(1 + \sqrt{p/n})^2 + 1/n] \\ &\le C(\tau')(T \wedge (1 + \frac{p}{n}))(1 + \frac{p}{n}).\end{aligned}$$

Therefore, under Assumption 4(i), we obtain

$$\mathbb{E}[\|\boldsymbol{Q}_2\|_{\mathrm{F}}^2] \le C(\tau')(T \wedge (1 + \frac{p}{n}))(1 + \frac{p}{n}).$$

(2) Under Assumption 4(ii), let $\Omega = U_1 \cap U_2 \cap U_3$, then we have $\mathbb{P}(\Omega^c) \le C(\gamma, c)\frac{1}{T}$ by (37). By Lemma E.2, on Ω, we have (i) The map $\boldsymbol{Z} \mapsto \boldsymbol{U}$ is $n^{-1/2}L_1$-Lipschitz, where $L_1 = 8\max(1, (2\eta)^{-1})$, and $\|\boldsymbol{U}\|_{\mathrm{F}} \le 1$, (ii) The map $\boldsymbol{Z} \mapsto \boldsymbol{V}$ is $n^{-1/2}L_2$-Lipschitz, where $L_2 = (1 + (2 + \sqrt{p/n})L_1)$, and

$\|\boldsymbol{V}\|_{\mathrm{F}} \le (2+\sqrt{p/n})$. Applying Corollary E.11 with $K = 2+\sqrt{p/n}$ yields

$$\begin{aligned}
&\mathbb{E}\Big[I(\Omega)\|\boldsymbol{U}^\top \boldsymbol{Z}\boldsymbol{V} - \sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\partial}{\partial z_{ij}}\Big(\boldsymbol{U}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{V}\Big)\|_{\mathrm{F}}^2\Big] \\
&\le K^2 + C(\gamma,c)(K^2L_1^2 + L_2^2) + 2\mathbb{E}\Big[I(\Omega)\sum_{ij}\Big(K^2\Big\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2 + \Big\|\frac{\partial \boldsymbol{V}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big)\Big] \\
&\le C(\gamma,c),
\end{aligned}$$

where the last inequality holds because $K \le C(\gamma), L_1 = C(\gamma,c), L_2 = C(\gamma,c)$, and on Ω, $\sum_{ij}\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\|_{\mathrm{F}}^2 \le C(\gamma,c)$ from Lemma E.5, and $\mathbb{E}[\|\frac{\partial \boldsymbol{V}}{\partial z_{ij}}\|_{\mathrm{F}}^2] \le C(\gamma,c)$ by product rule. Therefore, under Assumption 4(i), we obtain

$$\begin{aligned}
\mathbb{E}[I(\Omega)\|\boldsymbol{Q}_2\|_{\mathrm{F}}^2] &\le 2\mathbb{E}\Big[\|\boldsymbol{U}^\top \boldsymbol{Z}\boldsymbol{V} - \sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\partial}{\partial z_{ij}}\Big(\boldsymbol{U}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{V}\Big)\|_{\mathrm{F}}^2\Big] \\
&\quad + 2\mathbb{E}[\|n^{-\frac{3}{2}}D^{-2}(\boldsymbol{J}_1^\top + \boldsymbol{J}_2^\top - \boldsymbol{J}_3)\|_{\mathrm{F}}^2] \\
&\le C(\gamma,c),
\end{aligned}$$

where the last inequality used (46), (47), and that $n^{-3}\mathbb{E}[D^{-4}\|\boldsymbol{J}_3\|_{\mathrm{F}}^2] \le C(\gamma,c)$ in analogy to (48).

■

F.3. Proof of Proposition B.5

Proof of Proposition B.5. We will apply Lemma E.12. Let $\boldsymbol{U} = \boldsymbol{V} = n^{-\frac{1}{2}}D^{-1}\boldsymbol{F}$ with $D = (\|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \|\boldsymbol{H}\|_{\mathrm{F}}^2)^{\frac{1}{2}}$, then $\|\boldsymbol{U}\|_{\mathrm{F}} = \|\boldsymbol{V}\|_{\mathrm{F}} \le 1$. Let $\boldsymbol{W}_0 = p\boldsymbol{U}^\top\boldsymbol{U} - \sum_{j=1}^{p}\big(\sum_{i=1}^{n}\frac{\partial \boldsymbol{U}^\top \boldsymbol{e}_i}{\partial z_{ij}} - \boldsymbol{U}^\top \boldsymbol{Z}\boldsymbol{e}_j\big)\big(\sum_{i=1}^{n}\frac{\partial \boldsymbol{U}^\top \boldsymbol{e}_i}{\partial z_{ij}} - \boldsymbol{U}^\top \boldsymbol{Z}\boldsymbol{e}_j\big)^\top$, then Lemma E.12 gives

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{W}_0\|_{\mathrm{F}}] &\le 2\|\boldsymbol{U}\|_\partial\|\boldsymbol{V}\|_\partial + \sqrt{p}\big(\sqrt{2} + (3+\sqrt{2})(\|\boldsymbol{U}\|_\partial + \|\boldsymbol{V}\|_\partial)\big) \\
&= 2\|\boldsymbol{U}\|_\partial^2 + \sqrt{p}(\sqrt{2} + 2(3+\sqrt{2})\|\boldsymbol{U}\|_\partial).
\end{aligned}$$

We will prove under Assumption 4(i), and the proof under Assumption 4(ii) on set $\Omega = U_1 \cap U_2 \cap U_3$ follows from almost similar arguments with τ' replaced by η, which is a constant that depends only on γ, c.

By definition of $\|\boldsymbol{U}\|_\partial$ and Lemma E.5, $\|\boldsymbol{U}\|_\partial^2 = \sum_{ij}\mathbb{E}\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\|_{\mathrm{F}}^2 \le C(\gamma,\tau')$. Thus $\mathbb{E}[\|\boldsymbol{W}_0\|_{\mathrm{F}}] \le C(\tau',\gamma)\sqrt{p}$.

Now we establish the connection between $\boldsymbol{W}_0$ and $\boldsymbol{Q}_3$. Since $\boldsymbol{U} = n^{-\frac{1}{2}}D^{-1}\boldsymbol{F}$, by product rule,

$$\begin{aligned}
&\sum_{i=1}^{n}\frac{\partial \boldsymbol{U}^\top \boldsymbol{e}_i}{\partial z_{ij}} - \boldsymbol{U}^\top \boldsymbol{Z}\boldsymbol{e}_j \\
&= n^{-\frac{1}{2}}\Big(\sum_{i=1}^{n}\frac{\partial D^{-1}\boldsymbol{F}^\top \boldsymbol{e}_i}{\partial z_{ij}} - D^{-1}\boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j\Big)
\end{aligned}$$

$$= n^{-\frac{1}{2}}\Big(\sum_{i=1}^{n} D^{-1}\frac{\partial \boldsymbol{F}^\top \boldsymbol{e}_i}{\partial z_{ij}} - D^{-1}\boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j\Big) + n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{F}^\top \boldsymbol{e}_i \frac{\partial D^{-1}}{\partial z_{ij}}$$

$$= n^{-\frac{1}{2}} D^{-1}\Big(\sum_{i=1}^{n}\frac{\partial \boldsymbol{F}^\top \boldsymbol{e}_i}{\partial z_{ij}} - \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j\Big) - n^{-\frac{1}{2}} D^{-2}\sum_{i=1}^{n}\boldsymbol{F}^\top \boldsymbol{e}_i \frac{\partial D}{\partial z_{ij}}.$$

For the first term in the last display, we have by Lemma E.4

$$\sum_{i=1}^{n}\frac{\partial \boldsymbol{e}_i^\top \boldsymbol{F}}{\partial z_{ij}} = \sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\partial \boldsymbol{e}_i^\top \boldsymbol{F}\boldsymbol{e}_t}{\partial z_{ij}}\boldsymbol{e}_t^\top = \sum_{it}(D_{ij}^{it} + \Delta_{ij}^{it})\boldsymbol{e}_t^\top = -\boldsymbol{e}_j^\top \boldsymbol{H}(n\boldsymbol{I}_T - \widehat{\mathbf{A}}) + \sum_{it}\Delta_{ij}^{it}\boldsymbol{e}_t^\top.$$

Hence,

$$\begin{aligned}
&\sum_{i=1}^{n}\frac{\partial \boldsymbol{U}^\top \boldsymbol{e}_i}{\partial z_{ij}} - \boldsymbol{U}^\top \boldsymbol{Z}\boldsymbol{e}_j \\
&= n^{-\frac{1}{2}} D^{-1}\big[-(n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{H}^\top \boldsymbol{e}_j + \sum_{it}\Delta_{ij}^{it}\boldsymbol{e}_t - \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j\big] - n^{-\frac{1}{2}} D^{-2}\sum_{i=1}^{n}\boldsymbol{F}^\top \boldsymbol{e}_i\frac{\partial D}{\partial z_{ij}} \\
&= -n^{-\frac{1}{2}} D^{-1}\big[(n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{H}^\top + \boldsymbol{F}^\top \boldsymbol{Z}\big]\boldsymbol{e}_j + n^{-\frac{1}{2}} D^{-1}\sum_{it}\Delta_{ij}^{it}\boldsymbol{e}_t - n^{-\frac{1}{2}} D^{-2}\sum_{i=1}^{n}\boldsymbol{F}^\top \boldsymbol{e}_i\frac{\partial D}{\partial z_{ij}}. \qquad (51)
\end{aligned}$$

Let $\boldsymbol{W}_1 = -n^{-\frac{1}{2}} D^{-1}\big[(n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{H}^\top + \boldsymbol{F}^\top \boldsymbol{Z}\big]$ be the first term in (51). For the second term in (51), recall $\Delta_{ij}^{lt} = -(\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\big(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p\big)(\boldsymbol{e}_i \otimes \boldsymbol{e}_j)$ in Lemma E.4,

$$\begin{aligned}
&n^{-\frac{1}{2}} D^{-1}\sum_{it}\Delta_{ij}^{it}\boldsymbol{e}_t \\
&= -n^{-\frac{1}{2}} D^{-1}\sum_{it}\boldsymbol{e}_t(\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_i^\top)(\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\big(\boldsymbol{F}^\top \boldsymbol{e}_i \otimes \boldsymbol{e}_j\big) \\
&= -n^{-\frac{1}{2}} D^{-1}\sum_{i}(\boldsymbol{I}_T \otimes \boldsymbol{e}_i^\top \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\big(\boldsymbol{F}^\top \boldsymbol{e}_i \otimes \boldsymbol{I}_p\big)\boldsymbol{e}_j \\
&= -n^{-\frac{1}{2}} D^{-1}\sum_{i=1}^{n}(\boldsymbol{I}_T \otimes \boldsymbol{e}_i^\top \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})(\boldsymbol{F}^\top \boldsymbol{e}_i \otimes \boldsymbol{I}_p)\boldsymbol{e}_j \\
&= \boldsymbol{W}_2\boldsymbol{e}_j,
\end{aligned}$$

where $\boldsymbol{W}_2 = -n^{-\frac{1}{2}} D^{-1}\sum_{i=1}^{n}(\boldsymbol{I}_T \otimes \boldsymbol{e}_i^\top \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})(\boldsymbol{F}^\top \boldsymbol{e}_i \otimes \boldsymbol{I}_p)$. For the third term in (51),

$$-n^{-\frac{1}{2}} D^{-2}\boldsymbol{F}^\top \sum_{i=1}^{n}\boldsymbol{e}_i\frac{\partial D}{\partial z_{ij}} = \boldsymbol{W}_3\boldsymbol{e}_j,$$

where $\boldsymbol{W}_3 = -n^{-\frac{1}{2}} D^{-2} \boldsymbol{F}^\top \frac{\partial D}{\partial \boldsymbol{Z}}$, here we slightly abuse the notation and let $\frac{\partial D}{\partial \boldsymbol{Z}}$ denote the $n \times p$ matrix with (i, j)-th entry being $\frac{\partial D}{\partial z_{ij}}$. Therefore, (51) can be simplified as

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{U}^\top \boldsymbol{e}_i}{\partial z_{ij}} - \boldsymbol{U}^\top \boldsymbol{Z} \boldsymbol{e}_j = \big[\boldsymbol{W}_1 + \boldsymbol{W}_2 + \boldsymbol{W}_3\big] \boldsymbol{e}_j.$$

Furthermore,

$$\begin{aligned}
\boldsymbol{W}_0 &= p\boldsymbol{U}^\top \boldsymbol{U} - \sum_{j=1}^{p} \Big(\sum_{i=1}^{n} \frac{\partial \boldsymbol{U}^\top \boldsymbol{e}_i}{\partial z_{ij}} - \boldsymbol{U}^\top \boldsymbol{Z} \boldsymbol{e}_j \Big) \Big(\sum_{i=1}^{n} \frac{\partial \boldsymbol{U}^\top \boldsymbol{e}_i}{\partial z_{ij}} - \boldsymbol{U}^\top \boldsymbol{Z} \boldsymbol{e}_j \Big)^\top \\
&= p\boldsymbol{U}^\top \boldsymbol{U} - [\boldsymbol{W}_1 + \boldsymbol{W}_2 + \boldsymbol{W}_3][\boldsymbol{W}_1 + \boldsymbol{W}_2 + \boldsymbol{W}_3]^\top \\
&= n^{\frac{1}{2}} \boldsymbol{Q}_3 - \boldsymbol{W}_1 (\boldsymbol{W}_2 + \boldsymbol{W}_3)^\top - (\boldsymbol{W}_2 + \boldsymbol{W}_3) \boldsymbol{W}_1^\top - (\boldsymbol{W}_2 + \boldsymbol{W}_3)(\boldsymbol{W}_2 + \boldsymbol{W}_3)^\top,
\end{aligned}$$

where the last equality is due to

$$\begin{aligned}
& p\boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{W}_1 \boldsymbol{W}_1^\top \\
&= n^{-1} D^{-2} \Big[p\boldsymbol{F}^\top \boldsymbol{F} - \boldsymbol{F}^\top \boldsymbol{Z} \boldsymbol{Z}^\top \boldsymbol{F} - (n\boldsymbol{I}_T - \widehat{\mathbf{A}}) \boldsymbol{H}^\top \boldsymbol{H} (n\boldsymbol{I}_T - \widehat{\mathbf{A}}) \\
&\quad - (n\boldsymbol{I}_T - \widehat{\mathbf{A}}) \boldsymbol{H}^\top \boldsymbol{Z}^\top \boldsymbol{F} - \boldsymbol{F}^\top \boldsymbol{Z} \boldsymbol{H} (n\boldsymbol{I}_T - \widehat{\mathbf{A}}) \Big] \\
&= n^{\frac{1}{2}} \boldsymbol{Q}_3.
\end{aligned}$$

Therefore,

$$\boldsymbol{Q}_3 = n^{-\frac{1}{2}} \Big[\boldsymbol{W}_0 + \boldsymbol{W}_1 (\boldsymbol{W}_2 + \boldsymbol{W}_3)^\top + (\boldsymbol{W}_2 + \boldsymbol{W}_3) \boldsymbol{W}_1^\top + (\boldsymbol{W}_2 + \boldsymbol{W}_3)(\boldsymbol{W}_2 + \boldsymbol{W}_3)^\top \Big]. \tag{52}$$

We then bound the norms of $\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3$. For $\boldsymbol{W}_1$,

$$\begin{aligned}
\|\boldsymbol{W}_1\|_{\mathrm{F}} &= n^{-\frac{1}{2}} D^{-1} \|(n\boldsymbol{I}_T - \widehat{\mathbf{A}}) \boldsymbol{H}^\top + \boldsymbol{F}^\top \boldsymbol{Z}\|_{\mathrm{F}} \\
&\le n^{\frac{1}{2}} \big(D^{-1} \|\boldsymbol{H}\|_{\mathrm{F}} + n^{-1} D^{-1} \|\boldsymbol{F}\|_{\mathrm{F}} \|\boldsymbol{Z}\|_{\mathrm{op}} \big) \\
&\le n^{\frac{1}{2}} + n^{\frac{1}{2}} \big(D^{-1} \|\boldsymbol{F}\|_{\mathrm{F}} / \sqrt{n} \big\| \boldsymbol{Z} / \sqrt{n} \big\|_{\mathrm{op}} \big) \\
&\le n^{\frac{1}{2}} (1 + \|\boldsymbol{Z} / \sqrt{n}\|_{\mathrm{op}}),
\end{aligned}$$

where we used $\|\boldsymbol{I}_T - \widehat{\mathbf{A}}/n\|_{\mathrm{op}} \le 1$ by Lemma B.1 , $D^{-1} \|\boldsymbol{H}\|_{\mathrm{F}} \le 1$, and $D^{-1} \|\boldsymbol{F}\|_{\mathrm{F}} / \sqrt{n} \le 1$.

For $\boldsymbol{W}_2$,

$$\begin{aligned}
\|\boldsymbol{W}_2\|_{\mathrm{op}} &= n^{-\frac{1}{2}} D^{-1} \Big\| \sum_{i=1}^{n} (\boldsymbol{I}_T \otimes \boldsymbol{e}_i^\top \boldsymbol{X}) \boldsymbol{M}^\dagger (\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}}) (\boldsymbol{F}^\top \boldsymbol{e}_i \otimes \boldsymbol{I}_p) \Big\|_{\mathrm{op}} \\
&\le n^{-\frac{1}{2}} D^{-1} \sum_{i=1}^{n} \|(\boldsymbol{I}_T \otimes \boldsymbol{e}_i^\top \boldsymbol{X}) \boldsymbol{M}^\dagger (\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}}) (\boldsymbol{F}^\top \boldsymbol{e}_i \otimes \boldsymbol{I}_p)\|_{\mathrm{op}}
\end{aligned}$$

$$
\begin{aligned}
&\le n^{-\frac{1}{2}} D^{-1} \sum_{i=1}^{n} \|(\boldsymbol{I}_T \otimes \boldsymbol{e}_i^\top \boldsymbol{X}) \boldsymbol{M}^\dagger (\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\|_{\mathrm{op}} \|(\boldsymbol{F}^\top \boldsymbol{e}_i \otimes \boldsymbol{I}_p)\|_{\mathrm{op}} \\
&\le n^{-\frac{1}{2}} \|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{\frac{1}{2}} D^{-1} \sum_{i=1}^{n} \|(\boldsymbol{I}_T \otimes \boldsymbol{e}_i^\top \boldsymbol{X})(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger)\|_{\mathrm{op}} \|(\boldsymbol{F}^\top \boldsymbol{e}_i \otimes \boldsymbol{I}_p)\|_{\mathrm{op}} \\
&= n^{-\frac{1}{2}} \|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{\frac{1}{2}} D^{-1} \sum_{i=1}^{n} \|\boldsymbol{I}_T \otimes [\boldsymbol{e}_i^\top \boldsymbol{X}_{\hat{\mathcal{S}}} (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger]\|_{\mathrm{op}} \|\boldsymbol{F}^\top \boldsymbol{e}_i\| \\
&\le n^{-\frac{1}{2}} \|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{\frac{1}{2}} D^{-1} \|\boldsymbol{X}_{\hat{\mathcal{S}}} (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger\|_{\mathrm{op}} \sum_{i=1}^{n} \|\boldsymbol{F}^\top \boldsymbol{e}_i\| \\
&\le n^{-\frac{1}{2}} D^{-1} n^{-\frac{1}{2}} (\tau / \|\boldsymbol{\Sigma}\|_{\mathrm{op}})^{-\frac{1}{2}} n^{\frac{1}{2}} \|\boldsymbol{F}\|_{\mathrm{F}} \\
&\le (\tau')^{-\frac{1}{2}},
\end{aligned}
$$

where the third inequality uses $\boldsymbol{M}^\dagger \preceq \boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger$, the fourth inequality uses the result that $\|\boldsymbol{e}_i^\top \boldsymbol{A}\|_{\mathrm{op}} \le \|\boldsymbol{A}\|_{\mathrm{op}}$, the penultimate inequality uses $\|\boldsymbol{X}_{\hat{\mathcal{S}}} (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger\|_{\mathrm{op}} \le (n\tau)^{-1/2}$ and the Cauchy-Schwarz inequality, the last inequality follows from $n^{-1/2} D^{-1} \|\boldsymbol{F}\|_{\mathrm{F}} \le 1$. It immediately follows that $\|\boldsymbol{W}_2\|_{\mathrm{F}} \le \sqrt{T} C(\tau')$ since the rank of $\boldsymbol{W}_2$ is at most T.

For $\boldsymbol{W}_3$, using $n^{-\frac{1}{2}} D^{-1} \|\boldsymbol{F}\|_{\mathrm{F}} \le 1$, and $\left\|\frac{\partial D}{\partial \boldsymbol{Z}}\right\|_{\mathrm{F}} \le n^{-\frac{1}{2}} D C(\tau')$ from Corollary E.3, we obtain

$$
\|\boldsymbol{W}_3\|_{\mathrm{F}} = n^{-\frac{1}{2}} D^{-2} \|\boldsymbol{F}^\top \frac{\partial D}{\partial \boldsymbol{Z}}\|_{\mathrm{F}} \le n^{-\frac{1}{2}} D^{-2} \|\boldsymbol{F}\|_{\mathrm{F}} \left\|\frac{\partial D}{\partial \boldsymbol{Z}}\right\|_{\mathrm{F}} \le D^{-1} \left\|\frac{\partial D}{\partial \boldsymbol{Z}}\right\|_{\mathrm{F}} \le n^{-1/2} C(\tau').
$$

The desired inequality follows by combining (52) and the bounds for $\boldsymbol{W}_0, \boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3$. ■

G. Proof of technical results in Section E

G.1. Proofs of results in Section E.1

Proof of Lemma E.1. Fixing $\boldsymbol{E}$, if $\boldsymbol{X}, \bar{\boldsymbol{X}}$ are two design matrices, and $\widehat{\boldsymbol{B}}, \bar{\boldsymbol{B}}$ are the two corresponding multi-task elastic net estimates. Let $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Sigma}^{-\frac{1}{2}}$, $\bar{\boldsymbol{Z}} = \bar{\boldsymbol{X}}\boldsymbol{\Sigma}^{-\frac{1}{2}}$, $\bar{\boldsymbol{H}} = \boldsymbol{\Sigma}^{\frac{1}{2}}(\bar{\boldsymbol{B}} - \boldsymbol{B}^*)$, $\bar{\boldsymbol{F}} = \boldsymbol{Y} - \bar{\boldsymbol{X}}\bar{\boldsymbol{B}}$, and $\bar{D} = [\|\bar{\boldsymbol{H}}\|_{\mathrm{F}}^2 + \|\bar{\boldsymbol{F}}\|_{\mathrm{F}}^2/n]^{\frac{1}{2}}$. Without loss of generality, we assume $\bar{D} \le D$. Recall the multi-task elastic net estimate $\widehat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B} \in \mathbb{R}^{p \times T}} \left(\frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2 + g(\boldsymbol{B})\right)$, where $g(\boldsymbol{B}) = \lambda\|\boldsymbol{B}\|_{2,1} + \frac{\tau}{2}\|\boldsymbol{B}\|_{\mathrm{F}}^2$. Define $\varphi : \boldsymbol{B} \mapsto \frac{1}{2n}\|\boldsymbol{E} + \boldsymbol{X}(\boldsymbol{B}^* - \boldsymbol{B})\|_{\mathrm{F}}^2 + g(\boldsymbol{B})$, $\psi : \boldsymbol{B} \mapsto \frac{1}{2n}\|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B})\|_{\mathrm{F}}^2$ and $\zeta : \boldsymbol{B} \mapsto \varphi(\boldsymbol{B}) - \psi(\boldsymbol{B})$. When expanding the squares, it is clear that ζ is the sum of a linear function and a τ-strong convex penalty, thus ζ is τ-strongly convex of $\boldsymbol{B}$. Additivity of subdifferentials yields $\partial\varphi(\widehat{\boldsymbol{B}}) = \partial\zeta(\widehat{\boldsymbol{B}}) + \partial\psi(\widehat{\boldsymbol{B}}) = \partial\zeta(\widehat{\boldsymbol{B}})$. By optimality of $\widehat{\boldsymbol{B}}$ we have $\boldsymbol{0}_{p \times T} \in \partial\varphi(\widehat{\boldsymbol{B}})$, thus $\boldsymbol{0}_{p \times T} \in \partial\zeta(\widehat{\boldsymbol{B}})$. By strong convexity of ζ, $\zeta(\bar{\boldsymbol{B}}) - \zeta(\widehat{\boldsymbol{B}}) \ge \langle \partial\zeta(\widehat{\boldsymbol{B}}), \bar{\boldsymbol{B}} - \widehat{\boldsymbol{B}} \rangle + \frac{\tau}{2}\|\bar{\boldsymbol{B}} - \widehat{\boldsymbol{B}}\|_{\mathrm{F}}^2 = \frac{\tau}{2}\|\bar{\boldsymbol{B}} - \widehat{\boldsymbol{B}}\|_{\mathrm{F}}^2$, which can further be rewritten as

$$
\|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \bar{\boldsymbol{B}})\|_{\mathrm{F}}^2 + n\tau\|\widehat{\boldsymbol{B}} - \bar{\boldsymbol{B}}\|_{\mathrm{F}}^2 \le \|\boldsymbol{E} - \boldsymbol{X}(\bar{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2 - \|\boldsymbol{E} - \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{\mathrm{F}}^2 + 2n(g(\bar{\boldsymbol{B}}) - g(\widehat{\boldsymbol{B}})),
$$

i.e.,

$$
\|\boldsymbol{Z}(\boldsymbol{H} - \bar{\boldsymbol{H}})\|_{\mathrm{F}}^2 + n\tau\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{H} - \bar{\boldsymbol{H}})\|_{\mathrm{F}}^2 \le \|\boldsymbol{E} - \boldsymbol{Z}\bar{\boldsymbol{H}}\|_{\mathrm{F}}^2 - \|\boldsymbol{E} - \boldsymbol{Z}\boldsymbol{H}\|_{\mathrm{F}}^2 + 2n(g(\bar{\boldsymbol{B}}) - g(\widehat{\boldsymbol{B}})).
$$

Summing the above inequality with its counterpart obtained by replacing $(\boldsymbol{X}, \widehat{\boldsymbol{B}}, \boldsymbol{H})$ with $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{B}}, \bar{\boldsymbol{H}})$, we have

$$
\begin{aligned}
&(LHS)\\
&\stackrel{\text{def}}{=} \|\boldsymbol{Z}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}^2 + \|\bar{\boldsymbol{Z}}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}^2 + 2n\tau'\|\boldsymbol{H}-\bar{\boldsymbol{H}}\|_{\mathrm{F}}^2\\
&\le \|\boldsymbol{Z}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}^2 + \|\bar{\boldsymbol{Z}}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}^2 + 2n\tau\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}^2\\
&\le \|\boldsymbol{E}-\boldsymbol{Z}\bar{\boldsymbol{H}}\|_{\mathrm{F}}^2 - \|\boldsymbol{E}-\boldsymbol{Z}\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{E}-\bar{\boldsymbol{Z}}\boldsymbol{H}\|_{\mathrm{F}}^2 - \|\boldsymbol{E}-\bar{\boldsymbol{Z}}\bar{\boldsymbol{H}}\|_{\mathrm{F}}^2\\
&= \langle \boldsymbol{Z}(\boldsymbol{H}-\bar{\boldsymbol{H}}), \boldsymbol{F}+\bar{\boldsymbol{F}}+(\bar{\boldsymbol{Z}}-\boldsymbol{Z})\bar{\boldsymbol{H}}\rangle + \langle -\bar{\boldsymbol{Z}}(\boldsymbol{H}-\bar{\boldsymbol{H}}), \boldsymbol{F}+\bar{\boldsymbol{F}}+(\boldsymbol{Z}-\bar{\boldsymbol{Z}})\boldsymbol{H}\rangle\\
&= \langle (\boldsymbol{Z}-\bar{\boldsymbol{Z}})(\boldsymbol{H}-\bar{\boldsymbol{H}}), \boldsymbol{F}+\bar{\boldsymbol{F}}\rangle + \langle \boldsymbol{Z}(\boldsymbol{H}-\bar{\boldsymbol{H}}), (\bar{\boldsymbol{Z}}-\boldsymbol{Z})\bar{\boldsymbol{H}}\rangle + \langle \bar{\boldsymbol{Z}}(\bar{\boldsymbol{H}}-\boldsymbol{H}), (\boldsymbol{Z}-\bar{\boldsymbol{Z}})\boldsymbol{H}\rangle\\
&\le \|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}\|\boldsymbol{H}-\bar{\boldsymbol{H}}\|_{\mathrm{F}}(\|\boldsymbol{F}\|_{\mathrm{F}} + \|\bar{\boldsymbol{F}}\|_{\mathrm{F}}) + \|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}\|\boldsymbol{Z}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}\|\bar{\boldsymbol{H}}\|_{\mathrm{F}}\\
&\quad + \|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}\|\bar{\boldsymbol{Z}}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}\|\boldsymbol{H}\|_{\mathrm{F}}\\
&\le \|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}\Big[\sqrt{\frac{(LHS)}{2n\tau'}}(\|\boldsymbol{F}\|_{\mathrm{F}} + \|\bar{\boldsymbol{F}}\|_{\mathrm{F}}) + \sqrt{(LHS)}(\|\bar{\boldsymbol{H}}\|_{\mathrm{F}} + \|\boldsymbol{H}\|_{\mathrm{F}})\Big]\\
&\le \|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}\sqrt{(LHS)}(D+\bar{D})\max(1, (2\tau')^{-\frac{1}{2}})
\end{aligned}
$$

where $\tau' = \tau\phi_{\min}(\boldsymbol{\Sigma}^{-1}) = \tau/\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$. That is,

$$\sqrt{(LHS)} \le \|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}} 2D \max(1, (2\tau')^{-\frac{1}{2}}).$$

Therefore,

$$
\begin{aligned}
n^{-\frac{1}{2}}\|\boldsymbol{F}-\bar{\boldsymbol{F}}\|_{\mathrm{F}} &= n^{-\frac{1}{2}}\|\boldsymbol{Z}\boldsymbol{H}-\bar{\boldsymbol{Z}}\bar{\boldsymbol{H}}\|_{\mathrm{F}}\\
&\le n^{-\frac{1}{2}}[\|\boldsymbol{Z}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}} + \|(\boldsymbol{Z}-\bar{\boldsymbol{Z}})\bar{\boldsymbol{H}}\|_{\mathrm{F}}]\\
&\le n^{-\frac{1}{2}}[\|\boldsymbol{Z}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}} + \|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}\|\boldsymbol{H}\|_{\mathrm{F}}]\\
&\le n^{-\frac{1}{2}}[\sqrt{(LHS)} + \|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}D]\\
&\le n^{-\frac{1}{2}}\|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}D[2\max(1, (2\tau')^{-\frac{1}{2}}) + 1].
\end{aligned}
$$

So far we obtained

$$
\begin{aligned}
\|\boldsymbol{H}-\bar{\boldsymbol{H}}\|_{\mathrm{F}} \le \sqrt{\frac{(LHS)}{2n\tau'}} &\le n^{-\frac{1}{2}}\|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}D(2\tau')^{-\frac{1}{2}}2\max(1, (2\tau')^{-\frac{1}{2}}),\\
n^{-\frac{1}{2}}\|\boldsymbol{F}-\bar{\boldsymbol{F}}\|_{\mathrm{F}} &\le n^{-\frac{1}{2}}\|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}D[2\max(1, (2\tau')^{-\frac{1}{2}}) + 1].
\end{aligned}
$$

Let $\boldsymbol{Q} = [\boldsymbol{H}^\top, \boldsymbol{F}^\top/\sqrt{n}]^\top$ and $\bar{\boldsymbol{Q}} = [\bar{\boldsymbol{H}}^\top, \bar{\boldsymbol{F}}^\top/\sqrt{n}]^\top$, then $D = \|\boldsymbol{Q}\|_{\mathrm{F}}$, $\bar{D} = \|\bar{\boldsymbol{Q}}\|_{\mathrm{F}}$. By triangular inequality,

$$
\begin{aligned}
|D-\bar{D}| \le \|\boldsymbol{Q}-\bar{\boldsymbol{Q}}\|_{\mathrm{F}} &\le \|\boldsymbol{H}-\bar{\boldsymbol{H}}\|_{\mathrm{F}} + \|\boldsymbol{F}-\bar{\boldsymbol{F}}\|_{\mathrm{F}}/\sqrt{n}\\
&\le n^{-\frac{1}{2}}\|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}D[4\max(1, (2\tau')^{-1})],
\end{aligned}
$$

where the last inequality uses the elementary inequality $\max(a,b)(a+b) \le 2[\max(a,b)]^2$ for $a,b>0$ with $a=1, b=(2\tau')^{-\frac{1}{2}}$. Let $\frac{\partial D}{\partial \boldsymbol{Z}} \stackrel{\text{def}}{=} \frac{\partial D}{\partial \mathbf{vec}(\boldsymbol{Z})} \in \mathbb{R}^{1\times np}$, then $\left\|\frac{\partial D}{\partial \boldsymbol{Z}}\right\| \le n^{-\frac{1}{2}} D L_1$ with $L_1 = [4\max(1,(2\tau')^{-1})]$. Hence,

$$\sum_{ij}\left(\frac{\partial D}{\partial z_{ij}}\right)^2 = \left\|\frac{\partial D}{\partial \boldsymbol{Z}}\right\|^2 \le n^{-1} D^2 L_1^2.$$

Furthermore, by triangle inequality

$$\begin{aligned}
\left\|\frac{\boldsymbol{Q}}{D} - \frac{\bar{\boldsymbol{Q}}}{\bar{D}}\right\|_{\mathrm{F}} &\le \frac{1}{D}\|\boldsymbol{Q}-\bar{\boldsymbol{Q}}\|_{\mathrm{F}} + \left|\frac{1}{D}-\frac{1}{\bar{D}}\right|\|\bar{\boldsymbol{Q}}\|_{\mathrm{F}} \\
&= \frac{1}{D}\|\boldsymbol{Q}-\bar{\boldsymbol{Q}}\|_{\mathrm{F}} + \frac{|D-\bar{D}|}{D\bar{D}}\|\bar{\boldsymbol{Q}}\|_{\mathrm{F}} \\
&\le \frac{1}{D}\|\boldsymbol{Q}-\bar{\boldsymbol{Q}}\|_{\mathrm{F}} + \frac{1}{D}\|\boldsymbol{Q}-\bar{\boldsymbol{Q}}\|_{\mathrm{F}} \\
&\le n^{-\frac{1}{2}}\|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}} L,
\end{aligned}$$

where $L = 8\max(1,(2\tau')^{-1})$. Therefore, when $\tau>0$, we obtain the two mappings $\boldsymbol{Z}\mapsto D^{-1}\boldsymbol{F}/\sqrt{n}$, and $\boldsymbol{Z}\mapsto D^{-1}\boldsymbol{H}$ are both $n^{-\frac{1}{2}}L$-lipschitz with $L=8\max(1,(2\tau')^{-1})$, where $\tau'=\tau/\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$. ■

The proof of Lemma E.2 uses a similar argument as proof of Lemma E.1, we present it here for completeness.

Proof of Lemma E.2. For multi-task group Lasso ($\tau=0$), we restrict our analysis in the event $U_1\cap U_2$, where $U_1 = \{\|\widehat{\boldsymbol{B}}\|_0 \le n(1-c)/2\}$, $U_2 = \{\inf_{\boldsymbol{b}\in\mathbb{R}^p:\|\boldsymbol{b}\|_0\le(1-c)n}\|\boldsymbol{X}\boldsymbol{b}\|^2/(n\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{b}\|^2) > \eta\}$.

Since the only randomness of the problem comes from $\boldsymbol{X}$ and $\boldsymbol{E}$, there exists a measurable set $\mathcal{U}$ such that $U_1\cap U_2 = \{(\boldsymbol{X},\boldsymbol{E})\in\mathcal{U}\}$. For some noise matrix $\boldsymbol{E}$, consider $\boldsymbol{X},\bar{\boldsymbol{X}}$ two design matrices such that $(\boldsymbol{X},\boldsymbol{E})\in\mathcal{U}$ and $(\bar{\boldsymbol{X}},\boldsymbol{E})\in\mathcal{U}$. We slightly abuse the notation and let $\widehat{\boldsymbol{B}},\bar{\boldsymbol{B}}$ denote the two corresponding multi-task group-Lasso estimates. Thus, the row sparsity of $\widehat{\boldsymbol{B}}-\bar{\boldsymbol{B}}$ is at most $n(1-c)$. Let $\bar{\boldsymbol{H}} = \bar{\boldsymbol{B}}-\boldsymbol{B}^*$, $\bar{\boldsymbol{F}} = \boldsymbol{Y}-\bar{\boldsymbol{X}}\bar{\boldsymbol{B}}$, and $\bar{D} = [\|\bar{\boldsymbol{H}}\|_{\mathrm{F}}^2 + \|\bar{\boldsymbol{F}}\|_{\mathrm{F}}^2/n]^{\frac{1}{2}}$. Without loss of generality, we assume $\bar{D}\le D$. Since when $\tau=0$, the multi-task group Lasso estimate is $\widehat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}\in\mathbb{R}^{p\times T}}\left(\frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2 + g(\boldsymbol{B})\right)$, where $g(\boldsymbol{B}) = \lambda\|\boldsymbol{B}\|_{2,1}$. Define $\varphi:\boldsymbol{B}\mapsto\frac{1}{2n}\|\boldsymbol{E}+\boldsymbol{X}(\boldsymbol{B}^*-\boldsymbol{B})\|_{\mathrm{F}}^2 + g(\boldsymbol{B})$, $\psi:\boldsymbol{B}\mapsto\frac{1}{2n}\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B})\|_{\mathrm{F}}^2$ and $\zeta:\boldsymbol{B}\mapsto\varphi(\boldsymbol{B})-\psi(\boldsymbol{B})$. Under $\tau=0$, by the same arguments in proof of E.1 with the same functions $\varphi(\cdot),\psi(\cdot),\zeta(\cdot)$, we obtain

$$\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\bar{\boldsymbol{B}})\|_{\mathrm{F}}^2 \le \|\boldsymbol{E}-\boldsymbol{X}(\bar{\boldsymbol{B}}-\boldsymbol{B}^*)\|_{\mathrm{F}}^2 - \|\boldsymbol{E}-\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_{\mathrm{F}}^2 + 2n(g(\bar{\boldsymbol{B}})-g(\widehat{\boldsymbol{B}})).$$

Summing the above inequality with its counterpart obtained by replacing $(\boldsymbol{X},\widehat{\boldsymbol{B}},\boldsymbol{H})$ with $(\bar{\boldsymbol{X}},\bar{\boldsymbol{B}},\bar{\boldsymbol{H}})$, we have

$$\begin{aligned}
&\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\bar{\boldsymbol{B}})\|_{\mathrm{F}}^2 + \|\bar{\boldsymbol{X}}(\widehat{\boldsymbol{B}}-\bar{\boldsymbol{B}})\|_{\mathrm{F}}^2 \\
\le\ &\|\boldsymbol{E}-\boldsymbol{Z}\bar{\boldsymbol{H}}\|_{\mathrm{F}}^2 - \|\boldsymbol{E}-\boldsymbol{Z}\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{E}-\bar{\boldsymbol{Z}}\boldsymbol{H}\|_{\mathrm{F}}^2 - \|\boldsymbol{E}-\bar{\boldsymbol{Z}}\bar{\boldsymbol{H}}\|_{\mathrm{F}}^2.
\end{aligned}$$

Note that in event $U_1\cap U_2$, we have

$$\eta n\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\widehat{\boldsymbol{B}}-\bar{\boldsymbol{B}})\|_{\mathrm{F}}^2 \le \|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\bar{\boldsymbol{B}})\|_{\mathrm{F}}^2, \quad \eta n\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\widehat{\boldsymbol{B}}-\bar{\boldsymbol{B}})\|_{\mathrm{F}}^2 \le \|\bar{\boldsymbol{X}}(\widehat{\boldsymbol{B}}-\bar{\boldsymbol{B}})\|_{\mathrm{F}}^2.$$

Thus, $2\eta n\|(\widehat{\boldsymbol{H}}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}^2 \le \|\boldsymbol{Z}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}^2 + \|\bar{\boldsymbol{Z}}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}^2$, and

$$\begin{aligned}(LHS) &\stackrel{\text{def}}{=} \max(2\eta n\|\boldsymbol{H}-\bar{\boldsymbol{H}}\|_{\mathrm{F}}^2, \|\boldsymbol{Z}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}^2 + \|\bar{\boldsymbol{Z}}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}^2)\\ &= \|\boldsymbol{Z}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}^2 + \|\bar{\boldsymbol{Z}}(\boldsymbol{H}-\bar{\boldsymbol{H}})\|_{\mathrm{F}}^2\\ &\le \|\boldsymbol{E}-\boldsymbol{Z}\bar{\boldsymbol{H}}\|_{\mathrm{F}}^2 - \|\boldsymbol{E}-\boldsymbol{Z}\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{E}-\bar{\boldsymbol{Z}}\boldsymbol{H}\|_{\mathrm{F}}^2 - \|\boldsymbol{E}-\bar{\boldsymbol{Z}}\bar{\boldsymbol{H}}\|_{\mathrm{F}}^2.\end{aligned}$$

Now, in $U_1 \cap U_2$, the Lipschitz property of the map $\boldsymbol{Z} \mapsto D^{-1}\boldsymbol{F}/\sqrt{n}$ follows from the same arguments in proof of Lemma E.1, with τ' in E.1 replaced by η in this proof.

Furthermore, in the event $U_1 \cap U_2 \cap U_3$, the Lipschitz property of $\boldsymbol{Z} \mapsto D^{-1}\boldsymbol{Z}^\top\boldsymbol{F}/n$ follows by triangle inequality. To see this, let $\boldsymbol{U} = D^{-1}\boldsymbol{F}/\sqrt{n}$, and $\boldsymbol{V} = D^{-1}\boldsymbol{Z}^\top\boldsymbol{F}/n = n^{-1/2}\boldsymbol{Z}^\top\boldsymbol{U}$, thus by triangle inequality

$$\begin{aligned}\|\boldsymbol{V}-\bar{\boldsymbol{V}}\|_{\mathrm{op}} &= n^{-1/2}\|\boldsymbol{Z}^\top\boldsymbol{U} - \bar{\boldsymbol{Z}}^\top\bar{\boldsymbol{U}}\|_{\mathrm{op}}\\ &= n^{-1/2}[\|(\boldsymbol{Z}-\bar{\boldsymbol{Z}})^\top\boldsymbol{U}\|_{\mathrm{op}} + \|\bar{\boldsymbol{Z}}^\top(\boldsymbol{U}-\bar{\boldsymbol{U}})\|_{\mathrm{op}}]\\ &\le n^{-1/2}[\|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}} + \|\bar{\boldsymbol{Z}}\|_{\mathrm{op}}\|\boldsymbol{U}-\bar{\boldsymbol{U}}\|_{\mathrm{op}}]\\ &\le n^{-1/2}(1+n^{-1/2}\|\bar{\boldsymbol{Z}}\|_{\mathrm{op}}L)\|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}\\ &\le n^{-1/2}(1+(2+\sqrt{p/n})L).\end{aligned}$$

where the last line uses $\|\bar{\boldsymbol{Z}}\|_{\mathrm{op}} \le 2\sqrt{n}+\sqrt{p}$ in the event U_3. ■

Proof of Corollary E.3. Corollary E.3 (1) is a direct consequence of the intermediate result $|D-\bar{D}| \le n^{-\frac{1}{2}}\|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}D[4\max(1,(2\tau')^{-1})]$ in proof of Lemma E.1, while Corollary E.3 (2) is a direct consequence of the intermediate result $|D-\bar{D}| \le n^{-\frac{1}{2}}\|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_{\mathrm{op}}D[4\max(1,(2\eta)^{-1})]$ in proof of Lemma E.2. ■

Before proving the derivative formula, we restate $\widehat{\boldsymbol{B}}$ (defined in (6) of the full paper) below,

$$\widehat{\boldsymbol{B}} = \underset{\boldsymbol{B}\in\mathbb{R}^{p\times T}}{\arg\min}\left(\frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{B}\|_{2,1} + \frac{\tau}{2}\|\boldsymbol{B}\|_{\mathrm{F}}^2\right), \tag{53}$$

where $\|\boldsymbol{B}\|_{2,1} = \sum_{j=1}^p \|\boldsymbol{B}^\top\boldsymbol{e}_j\|_2$.

For the reader's convenience, we recall some useful notations. $\boldsymbol{P}_{\hat{\mathcal{S}}} = \sum_{k\in\hat{\mathcal{S}}}\boldsymbol{e}_k\boldsymbol{e}_k^\top$. For each $k \in \hat{\mathcal{S}}$, $\boldsymbol{H}^{(k)} = \lambda\|\widehat{\boldsymbol{B}}^\top\boldsymbol{e}_k\|_2^{-1}\left(\boldsymbol{I}_T - \widehat{\boldsymbol{B}}^\top\boldsymbol{e}_k\boldsymbol{e}_k^\top\widehat{\boldsymbol{B}}\,\|\widehat{\boldsymbol{B}}^\top\boldsymbol{e}_k\|_2^{-2}\right)$. $\tilde{\boldsymbol{H}} = \sum_{k\in\hat{\mathcal{S}}}(\boldsymbol{H}^{(k)}\otimes\boldsymbol{e}_k\boldsymbol{e}_k^\top)$. $\boldsymbol{M}_1 = \boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top\boldsymbol{X}_{\hat{\mathcal{S}}} + \tau n\boldsymbol{P}_{\hat{\mathcal{S}}})$, $\boldsymbol{M} = \boldsymbol{M}_1 + n\tilde{\boldsymbol{H}} \in \mathbb{R}^{pT\times pT}$, and let $\boldsymbol{N} = (\boldsymbol{I}_T\otimes\boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T\otimes\boldsymbol{X}^\top)$.

Proof of Lemma E.4. We first derive $\frac{\partial F_{lt}}{\partial x_{ij}}$. Since $\boldsymbol{F} = \boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}} = \boldsymbol{E}-\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)$, by product rule,

$$\frac{\partial F_{lt}}{\partial x_{ij}} = \boldsymbol{e}_l^\top\frac{\partial\boldsymbol{E}-\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)}{\partial x_{ij}}\boldsymbol{e}_t = -\boldsymbol{e}_l^\top(\dot{\boldsymbol{X}}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)+\boldsymbol{X}\dot{\boldsymbol{B}})\boldsymbol{e}_t,$$

where $\dot{\boldsymbol{X}} \stackrel{\text{def}}{=} \frac{\partial\boldsymbol{X}}{\partial x_{ij}} = \boldsymbol{e}_i\boldsymbol{e}_j^\top$, and $\dot{\boldsymbol{B}} \stackrel{\text{def}}{=} \frac{\partial\widehat{\boldsymbol{B}}}{\partial x_{ij}}$.

Now we derive $\mathbf{vec}(\dot{\boldsymbol{B}})$ from KKT conditions for $\widehat{\boldsymbol{B}}$ defined in (53):

1) For $k \in \hat{\mathcal{S}}$, i.e., $\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k \neq \boldsymbol{0}$,

$$\boldsymbol{e}_k^\top \boldsymbol{X}^\top \left[\boldsymbol{E} - \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\right] - n\tau \boldsymbol{e}_k^\top \widehat{\boldsymbol{B}} = \frac{n\lambda}{\|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k\|_2} \boldsymbol{e}_k^\top \widehat{\boldsymbol{B}} \quad \in \mathbb{R}^{1\times T}.$$

2) For $k \notin \hat{\mathcal{S}}$, i.e., $\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k = \boldsymbol{0}$,

$$\left\| \boldsymbol{e}_k^\top \boldsymbol{X}^\top \left[\boldsymbol{E} - \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\right] - n\tau \boldsymbol{e}_k^\top \widehat{\boldsymbol{B}} \right\| < n\lambda.$$

Here the strict inequality is guaranteed by Proposition 2.3 of Bellec (2020).

Keeping $\boldsymbol{E}$ fixed, differentiation of the above display for $k \in \hat{\mathcal{S}}$ w.r.t. x_{ij} yields

$$\boldsymbol{e}_k^\top \left[\dot{\boldsymbol{X}}^\top \boldsymbol{F} - \boldsymbol{X}^\top [\dot{\boldsymbol{X}}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) + \boldsymbol{X}\dot{\boldsymbol{B}}] - n\tau \dot{\boldsymbol{B}} \right] = n \boldsymbol{e}_k^\top \dot{\boldsymbol{B}} \boldsymbol{H}^{(k)},$$

with $\boldsymbol{H}^{(k)} = \lambda \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k\|_2^{-1} \left(\boldsymbol{I}_T - \widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k \boldsymbol{e}_k^\top \widehat{\boldsymbol{B}} \, \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k\|_2^{-2} \right) \in \mathbb{R}^{T\times T}$. Rearranging and using $\dot{\boldsymbol{X}} = \boldsymbol{e}_i \boldsymbol{e}_j^\top$,

$$\boldsymbol{e}_k^\top \left[\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{F} - \boldsymbol{X}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top (\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) \right] = \boldsymbol{e}_k^\top [(\boldsymbol{X}^\top \boldsymbol{X} + n\tau \boldsymbol{I}_p)\dot{\boldsymbol{B}} + n\dot{\boldsymbol{B}} \boldsymbol{H}^{(k)}].$$

Recall $\boldsymbol{P}_{\hat{\mathcal{S}}} = \sum_{k\in\hat{\mathcal{S}}} \boldsymbol{e}_k \boldsymbol{e}_k^\top$. Multiplying by $\boldsymbol{e}_k$ to the left and summing over $k \in \hat{\mathcal{S}}$, we obtain

$$\boldsymbol{P}_{\hat{\mathcal{S}}} \left[\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{F} - \boldsymbol{X}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top (\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) \right] = \boldsymbol{P}_{\hat{\mathcal{S}}} (\boldsymbol{X}^\top \boldsymbol{X} + n\tau \boldsymbol{I}_p)\dot{\boldsymbol{B}} + n \sum_{k\in\hat{\mathcal{S}}} \boldsymbol{e}_k \boldsymbol{e}_k^\top \dot{\boldsymbol{B}} \boldsymbol{H}^{(k)}.$$

Since $\hat{\mathcal{S}}$ is locally constant in a small neighborhood of $\boldsymbol{X}$, $\widehat{\boldsymbol{B}}_{\hat{\mathcal{S}}^c} = 0$, $\mathrm{supp}(\dot{\boldsymbol{B}}) \subseteq \hat{\mathcal{S}}$. Hence, $\boldsymbol{P}_{\hat{\mathcal{S}}} \dot{\boldsymbol{B}} = \dot{\boldsymbol{B}}$, and $\boldsymbol{X}\dot{\boldsymbol{B}} = \boldsymbol{X}_{\hat{\mathcal{S}}} \dot{\boldsymbol{B}}$. The above display can be rewritten as

$$\boldsymbol{P}_{\hat{\mathcal{S}}} \boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{F} - \boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top (\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) = (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})\dot{\boldsymbol{B}} + n \sum_{k\in\hat{\mathcal{S}}} \boldsymbol{e}_k \boldsymbol{e}_k^\top \dot{\boldsymbol{B}} \boldsymbol{H}^{(k)}.$$

Vectorizing the above display using property $\mathbf{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^\top \otimes \boldsymbol{A})\,\mathbf{vec}(\boldsymbol{A})$ yields

$$\begin{aligned}
&(\boldsymbol{F}^\top \otimes \boldsymbol{P}_{\hat{\mathcal{S}}} \boldsymbol{e}_j)\,\mathbf{vec}(\boldsymbol{e}_i^\top) - ((\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \boldsymbol{e}_j \otimes \boldsymbol{X}_{\hat{\mathcal{S}}}^\top)\,\mathbf{vec}(\boldsymbol{e}_i) \\
&= [\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}}) + n \sum_{k\in\hat{\mathcal{S}}} (\boldsymbol{H}^{(k)} \otimes \boldsymbol{e}_k \boldsymbol{e}_k^\top)]\,\mathbf{vec}(\dot{\boldsymbol{B}}) \\
&= (\boldsymbol{M}_1 + n\tilde{\boldsymbol{H}})\,\mathbf{vec}(\dot{\boldsymbol{B}}) \\
&= \boldsymbol{M}\,\mathbf{vec}(\dot{\boldsymbol{B}}),
\end{aligned}$$

where $\boldsymbol{M}_1 = \boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})$, and $\tilde{\boldsymbol{H}} = \sum_{k\in\hat{\mathcal{S}}} (\boldsymbol{H}^{(k)} \otimes \boldsymbol{e}_k \boldsymbol{e}_k^\top)$.

Under Assumption 4(i) that $\tau > 0$, it's obviously that $\mathrm{rank}(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})) = T|\hat{\mathcal{S}}|$. Under Assumption 4(ii) that $\tau = 0$ with $\mathbb{P}(U_1) \to 1$. In the event $U_1 \cap U_2$, we know $\mathrm{rank}(\boldsymbol{X}_{\hat{\mathcal{S}}}) = |\hat{\mathcal{S}}|$ from (Bellec and Romon, 2021, Lemma C.4), hence $\mathrm{rank}(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})) = T|\hat{\mathcal{S}}|$. In either of the above two scenarios, we thus have $\dim(\ker(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}}))) = T(p - |\hat{\mathcal{S}}|)$ by rank-nullity theorem. Since $[\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})](\boldsymbol{e}_t \otimes \boldsymbol{e}_k) = \boldsymbol{0}$ for $t \in [T], k \in \hat{\mathcal{S}}^c$. Let $V = \{(\boldsymbol{e}_t \otimes \boldsymbol{e}_k) : t \in$

$[T], k \in \hat{\mathcal{S}}^c\}$ be a vector space, then the elements of V are linear independent, and $\dim(V) = T(p - |\hat{\mathcal{S}}|)$. Thus, V forms a basis for $\ker(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}}))$. Since for any $(\boldsymbol{e}_t \otimes \boldsymbol{e}_k) \in V$, we also have $\tilde{\boldsymbol{H}}(\boldsymbol{e}_t \otimes \boldsymbol{e}_k) = \boldsymbol{0}$, $\ker(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})) \subseteq \ker(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}}) + n\tilde{\boldsymbol{H}})$. On the other hand, if any vector $\boldsymbol{v}$ s.t. $[\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}}) + n\tilde{\boldsymbol{H}}]\boldsymbol{v} = \boldsymbol{0}$, since these matrices are all positive semi-definite, we have $\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})\boldsymbol{v} = \boldsymbol{0}$, which implies that $\ker(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}}) + n\tilde{\boldsymbol{H}}) \subseteq \ker(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}}))$. Therefore,

$$\begin{aligned}\ker(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}}) + n\tilde{\boldsymbol{H}}) &= \ker(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})) \\ &= \operatorname{span}\{(\boldsymbol{e}_t \otimes \boldsymbol{e}_k) : t \in [T], k \in \hat{\mathcal{S}}^c\},\end{aligned}$$

and

$$\operatorname{range}(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}}) + n\tilde{\boldsymbol{H}}) = \operatorname{span}\{(\boldsymbol{e}_t \otimes \boldsymbol{e}_k) : t \in [T], k \in \hat{\mathcal{S}}\}.$$

Since $\dot{\boldsymbol{B}} = \boldsymbol{P}_{\hat{\mathcal{S}}}\dot{\boldsymbol{B}}$, $\mathbf{vec}(\dot{\boldsymbol{B}}) = (\boldsymbol{I}_T \otimes \boldsymbol{P}_{\hat{\mathcal{S}}})\,\mathbf{vec}(\dot{\boldsymbol{B}})$, then $\mathbf{vec}(\dot{\boldsymbol{B}}) \in \operatorname{col}(\boldsymbol{I}_T \otimes \boldsymbol{P}_{\hat{\mathcal{S}}}) = \operatorname{range}(\boldsymbol{M})$. Since $\boldsymbol{M}$ is symmetric, $\boldsymbol{M}^\dagger \boldsymbol{M}$ is the orthogonal projection on the range of $\boldsymbol{M}$. Therefore,

$$\mathbf{vec}(\dot{\boldsymbol{B}}) = \boldsymbol{M}^\dagger \boldsymbol{M}\,\mathbf{vec}(\dot{\boldsymbol{B}}) = \boldsymbol{M}^\dagger[(\boldsymbol{F}^\top \otimes \boldsymbol{e}_j) - ((\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \boldsymbol{e}_j \otimes \boldsymbol{X}^\top)]\boldsymbol{e}_i. \tag{54}$$

Since $\operatorname{supp}(\dot{\boldsymbol{B}}) \subseteq \hat{\mathcal{S}}$, $\boldsymbol{X}\dot{\boldsymbol{B}} = \boldsymbol{X}_{\hat{\mathcal{S}}}\dot{\boldsymbol{B}}$, we have

$$\begin{aligned}\frac{\partial F_{lt}}{\partial x_{ij}} &= -\boldsymbol{e}_l^\top(\dot{\boldsymbol{X}}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) + \boldsymbol{X}\dot{\boldsymbol{B}})\boldsymbol{e}_t \\ &= -(\boldsymbol{e}_l^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_t + \boldsymbol{e}_l^\top \boldsymbol{X}_{\hat{\mathcal{S}}}\dot{\boldsymbol{B}}\boldsymbol{e}_t) \\ &= -(\boldsymbol{e}_l^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_t + (\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top \boldsymbol{X}_{\hat{\mathcal{S}}})\,\mathbf{vec}(\dot{\boldsymbol{B}})) \\ &= -\boldsymbol{e}_l^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_t - (\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top \boldsymbol{X}_{\hat{\mathcal{S}}})\boldsymbol{M}^\dagger[(\boldsymbol{F}^\top \otimes \boldsymbol{e}_j) - ((\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \boldsymbol{e}_j \otimes \boldsymbol{X}^\top)]\boldsymbol{e}_i \\ &= -(e_j^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) \otimes \boldsymbol{e}_i^\top)(\boldsymbol{e}_t \otimes \boldsymbol{e}_l) + (\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top \boldsymbol{X}_{\hat{\mathcal{S}}})\boldsymbol{M}^\dagger((\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \boldsymbol{e}_j \otimes \boldsymbol{X}^\top \boldsymbol{e}_i) \\ &\quad - (\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top \boldsymbol{X}_{\hat{\mathcal{S}}})\boldsymbol{M}^\dagger(\boldsymbol{F}^\top \otimes \boldsymbol{e}_j)\boldsymbol{e}_i \\ &= -(e_j^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) \otimes \boldsymbol{e}_i^\top)(\boldsymbol{e}_t \otimes \boldsymbol{e}_l) + (e_j^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) \otimes \boldsymbol{e}_i^\top)\boldsymbol{N}(\boldsymbol{e}_t \otimes \boldsymbol{e}_l) \\ &\quad - (\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top \boldsymbol{X}_{\hat{\mathcal{S}}})\boldsymbol{M}^\dagger(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p)(\boldsymbol{e}_i \otimes \boldsymbol{e}_j) \\ &= -(e_j^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) \otimes \boldsymbol{e}_i^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{e}_t \otimes \boldsymbol{e}_l) - (\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p)(\boldsymbol{e}_i \otimes \boldsymbol{e}_j)\end{aligned}$$

Now we calculate $\frac{\partial F_{lt}}{\partial z_{ij}}$. Since $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{\frac{1}{2}}$, $x_{ik} = \sum_{j=1}^p z_{ij}(\boldsymbol{\Sigma}^{\frac{1}{2}})_{jk}$, $\frac{\partial x_{ik}}{\partial z_{ij}} = (\boldsymbol{\Sigma}^{\frac{1}{2}})_{jk}$,

$$\frac{\partial F_{lt}}{\partial z_{ij}} = \sum_{k=1}^p \frac{\partial F_{lt}}{\partial x_{ik}}\frac{\partial x_{ik}}{\partial z_{ij}} = \sum_{k=1}^p \frac{\partial F_{lt}}{\partial x_{ik}}(\boldsymbol{\Sigma}^{\frac{1}{2}})_{jk} = D_{ij}^{lt} + \Delta_{ij}^{lt},$$

where

$$D_{ij}^{lt} = -\sum_{k=1}^p (\boldsymbol{e}_k^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) \otimes \boldsymbol{e}_i^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{e}_t \otimes \boldsymbol{e}_l)(\boldsymbol{\Sigma}^{\frac{1}{2}})_{jk}$$

$$= -(\boldsymbol{e}_j^\top \boldsymbol{\Sigma}^{\frac{1}{2}}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) \otimes \boldsymbol{e}_i^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{e}_t \otimes \boldsymbol{e}_l)$$
$$= -(\boldsymbol{e}_j^\top \boldsymbol{H} \otimes \boldsymbol{e}_i^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{e}_t \otimes \boldsymbol{e}_l),$$

and

$$\Delta_{ij}^{lt} = -\sum_{k=1}^{p}(\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger\big(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p\big)(\boldsymbol{e}_i \otimes \boldsymbol{e}_k)(\boldsymbol{\Sigma}^{\frac{1}{2}})_{jk}$$
$$= -(\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger\big(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p\big)(\boldsymbol{e}_i \otimes \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{e}_j)$$
$$= -(\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\big(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p\big)(\boldsymbol{e}_i \otimes \boldsymbol{e}_j)$$

It follows that

$$\sum_{i=1}^{n} D_{ij}^{it} = -\sum_{i=1}^{n}(\boldsymbol{e}_j^\top \boldsymbol{H} \otimes \boldsymbol{e}_i^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{e}_t \otimes \boldsymbol{e}_i)$$
$$= -\boldsymbol{e}_j^\top \boldsymbol{H}\big[\sum_{i=1}^{n}(\boldsymbol{I}_T \otimes \boldsymbol{e}_i^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{I}_T \otimes \boldsymbol{e}_i)\big]\boldsymbol{e}_t$$
$$= -\boldsymbol{e}_j^\top \boldsymbol{H}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{e}_t,$$

where the last line follows from definition of $\widehat{\mathbf{A}}$ in (10). ■

Proof of Lemma E.5. (1) For $\tau > 0$, by formula of $\frac{\partial F_{lt}}{\partial z_{ij}}$ in Lemma E.4, we have

$$\sum_{ij}\left\|\frac{\partial \boldsymbol{F}}{\partial z_{ij}}\right\|_{\mathrm{F}}^2 = \sum_{ij}\sum_{lt}\left(\frac{\partial F_{lt}}{\partial z_{ij}}\right)^2 = \sum_{ij}\sum_{lt}\left(D_{ij}^{lt} + \Delta_{ij}^{lt}\right)^2$$
$$\leq 2\sum_{ij,lt}(D_{ij}^{lt})^2 + 2\sum_{ij,lt}(\Delta_{ij}^{lt})^2$$
$$= 2\|(\boldsymbol{H} \otimes \boldsymbol{I}_n)(\boldsymbol{I}_{nT} - \boldsymbol{N})\|_{\mathrm{F}}^2 + 2\|(\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p)\|_{\mathrm{F}}^2$$
$$\leq 2n\|\boldsymbol{H}\|_{\mathrm{F}}^2 + 2\|(\boldsymbol{I}_T \otimes \boldsymbol{X}_{\hat{\mathscr{S}}})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p)\|_{\mathrm{F}}^2.$$

Since $0 \preceq \boldsymbol{M}^\dagger \preceq \boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathscr{S}}}^\top \boldsymbol{X}_{\hat{\mathscr{S}}} + \tau n \boldsymbol{P}_{\hat{\mathscr{S}}})^\dagger$,

$$\|(\boldsymbol{I}_T \otimes \boldsymbol{X}_{\hat{\mathscr{S}}})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p)\|_{\mathrm{F}}^2$$
$$\leq \|(\boldsymbol{I}_T \otimes \boldsymbol{X}_{\hat{\mathscr{S}}})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\|_{\mathrm{op}}^2\|(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p)\|_{\mathrm{F}}^2$$
$$\leq p\|\boldsymbol{\Sigma}\|_{\mathrm{op}}\|\boldsymbol{F}\|_{\mathrm{F}}^2\|(\boldsymbol{I}_T \otimes \boldsymbol{X}_{\hat{\mathscr{S}}})\boldsymbol{M}^\dagger\|_{\mathrm{op}}^2$$
$$\leq p\|\boldsymbol{\Sigma}\|_{\mathrm{op}}\|\boldsymbol{F}\|_{\mathrm{F}}^2\|(\boldsymbol{X}_{\hat{\mathscr{S}}}^\top \boldsymbol{X}_{\hat{\mathscr{S}}} + \tau n \boldsymbol{P}_{\hat{\mathscr{S}}})^\dagger \boldsymbol{X}_{\hat{\mathscr{S}}}^\top\|_{\mathrm{op}}^2$$
$$\leq \frac{p}{n\tau}\|\boldsymbol{\Sigma}\|_{\mathrm{op}}\|\boldsymbol{F}\|_{\mathrm{F}}^2$$

$$= \frac{p}{n\tau'}\|\boldsymbol{F}\|_{\mathrm{F}}^2,$$

where the last inequality uses $\|(\boldsymbol{X}_{\hat{\mathscr{S}}}^\top \boldsymbol{X}_{\hat{\mathscr{S}}} + \tau n \boldsymbol{P}_{\hat{\mathscr{S}}})^\dagger \boldsymbol{X}_{\hat{\mathscr{S}}}^\top\|_{\mathrm{op}} \le (n\tau)^{-1}$.

On the other hand, we also have

$$\begin{aligned}
&\|(\boldsymbol{I}_T \otimes \boldsymbol{X}_{\hat{\mathscr{S}}})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p)\|_{\mathrm{F}}^2 \\
\le\ & \|(\boldsymbol{I}_T \otimes \boldsymbol{X}_{\hat{\mathscr{S}}})\boldsymbol{M}^\dagger\|_{\mathrm{F}}^2 \|(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p)\|_{\mathrm{op}}^2 \\
\le\ & \|(\boldsymbol{I}_T \otimes \boldsymbol{X}_{\hat{\mathscr{S}}})(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathscr{S}}}^\top \boldsymbol{X}_{\hat{\mathscr{S}}} + \tau n \boldsymbol{P}_{\hat{\mathscr{S}}})^\dagger)\|_{\mathrm{F}}^2 \|\boldsymbol{F}\|_{\mathrm{F}}^2 \|\boldsymbol{\Sigma}\|_{\mathrm{op}} \\
\le\ & T\|\boldsymbol{X}_{\hat{\mathscr{S}}}(\boldsymbol{X}_{\hat{\mathscr{S}}}^\top \boldsymbol{X}_{\hat{\mathscr{S}}} + \tau n \boldsymbol{P}_{\hat{\mathscr{S}}})^\dagger\|_{\mathrm{F}}^2 \|\boldsymbol{F}\|_{\mathrm{F}}^2 \|\boldsymbol{\Sigma}\|_{\mathrm{op}} \\
\le\ & T\,\mathrm{Tr}\left[(\boldsymbol{X}_{\hat{\mathscr{S}}}^\top \boldsymbol{X}_{\hat{\mathscr{S}}} + \tau n \boldsymbol{P}_{\hat{\mathscr{S}}})^\dagger \boldsymbol{X}_{\hat{\mathscr{S}}}^\top \boldsymbol{X}_{\hat{\mathscr{S}}} (\boldsymbol{X}_{\hat{\mathscr{S}}}^\top \boldsymbol{X}_{\hat{\mathscr{S}}} + \tau n \boldsymbol{P}_{\hat{\mathscr{S}}})^\dagger\right] \|\boldsymbol{F}\|_{\mathrm{F}}^2 \|\boldsymbol{\Sigma}\|_{\mathrm{op}} \\
\le\ & T\,\mathrm{Tr}\left[(\boldsymbol{X}_{\hat{\mathscr{S}}}^\top \boldsymbol{X}_{\hat{\mathscr{S}}} + \tau n \boldsymbol{P}_{\hat{\mathscr{S}}})^\dagger\right] \|\boldsymbol{F}\|_{\mathrm{F}}^2 \|\boldsymbol{\Sigma}\|_{\mathrm{op}} \\
\le\ & T(\tau)^{-1}\|\boldsymbol{F}\|_{\mathrm{F}}^2 \|\boldsymbol{\Sigma}\|_{\mathrm{op}} \\
\le\ & T\,\mathrm{Tr}\left[(\tau n \boldsymbol{P}_{\hat{\mathscr{S}}})^\dagger\right] \|\boldsymbol{F}\|_{\mathrm{F}}^2 \|\boldsymbol{\Sigma}\|_{\mathrm{op}} \\
\le\ & T(\tau)^{-1}\|\boldsymbol{F}\|_{\mathrm{F}}^2 \|\boldsymbol{\Sigma}\|_{\mathrm{op}} \\
=\ & \frac{T}{\tau'}\|\boldsymbol{F}\|_{\mathrm{F}}^2,
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{1}{n}\sum_{ij}\left\|\frac{\partial \boldsymbol{F}}{\partial z_{ij}}\right\|_{\mathrm{F}}^2 &\le 2\|\boldsymbol{H}\|_{\mathrm{F}}^2 + 2(\tau')^{-1}(T \wedge \frac{p}{n})\|\boldsymbol{F}\|_{\mathrm{F}}^2/n \\
&\le 2\max(1, (\tau')^{-1}(T \wedge \frac{p}{n}))(\|\boldsymbol{F}\|_{\mathrm{F}}^2/n + \|\boldsymbol{H}\|_{\mathrm{F}}^2) \\
&= 2\max(1, (\tau')^{-1}(T \wedge \frac{p}{n}))D^2.
\end{aligned}$$

Now by product rule and triangle inequality

$$\begin{aligned}
&\frac{1}{n}\sum_{ij}\left\|\frac{\partial \boldsymbol{F}/D}{\partial z_{ij}}\right\|_{\mathrm{F}}^2 \\
\le\ & 2D^{-2}\frac{1}{n}\sum_{ij}\left\|\frac{\partial \boldsymbol{F}}{\partial z_{ij}}\right\|_{\mathrm{F}}^2 + 2\frac{1}{n}\sum_{ij}\left\|\boldsymbol{F}\frac{\partial D^{-1}}{\partial z_{ij}}\right\|_{\mathrm{F}}^2 \\
=\ & 2D^{-2}\frac{1}{n}\sum_{ij}\left\|\frac{\partial \boldsymbol{F}}{\partial z_{ij}}\right\|_{\mathrm{F}}^2 + 2D^{-4}\frac{1}{n}\|\boldsymbol{F}\|_{\mathrm{F}}^2\sum_{ij}\left(\frac{\partial D}{\partial z_{ij}}\right)^2 \\
\le\ & 2D^{-2}\frac{1}{n}\sum_{ij}\left\|\frac{\partial \boldsymbol{F}}{\partial z_{ij}}\right\|_{\mathrm{F}}^2 + 2D^{-4}\frac{1}{n}\|\boldsymbol{F}\|_{\mathrm{F}}^2 n^{-1}D^2[4\max(1, (2\tau')^{-1})]^2
\end{aligned}$$

$$\leq 2D^{-2}\frac{1}{n}\sum_{ij}\left\|\frac{\partial \boldsymbol{F}}{\partial z_{ij}}\right\|_{\mathrm{F}}^2 + 2n^{-1}[4\max(1,(2\tau')^{-1})]^2$$
$$\leq 4\max(1,(\tau')^{-1}(T\wedge\frac{p}{n})) + 2n^{-1}[4\max(1,(2\tau')^{-1})]^2$$
$$:= f(\tau', T, n, p),$$

where the second inequality is by Corollary E.3.

(2) For $\tau = 0$, by Lemma E.2, in the event $U_1 \cap U_2$, we obtain the same upper bounds as in the first case (1) with τ' replaced by η. To see this,

$$\begin{aligned}
&\|(\boldsymbol{I}_T \otimes \boldsymbol{X}_{\hat{\mathcal{S}}})\boldsymbol{M}^{\dagger}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})(\boldsymbol{F}^{\top} \otimes \boldsymbol{I}_p)\|_{\mathrm{F}}^2 \\
\leq\ &\|(\boldsymbol{I}_T \otimes \boldsymbol{X}_{\hat{\mathcal{S}}})\boldsymbol{M}^{\dagger}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\|_{\mathrm{op}}^2\|(\boldsymbol{F}^{\top} \otimes \boldsymbol{I}_p)\|_{\mathrm{F}}^2 \\
=\ &\|(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\boldsymbol{M}^{\dagger}(\boldsymbol{I}_T \otimes \boldsymbol{X}_{\hat{\mathcal{S}}}^{\top}\boldsymbol{X}_{\hat{\mathcal{S}}})\boldsymbol{M}^{\dagger}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\|_{\mathrm{op}} p\|\boldsymbol{F}\|_{\mathrm{F}}^2 \\
\leq\ &\|(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\boldsymbol{M}^{\dagger}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\|_{\mathrm{op}} p\|\boldsymbol{F}\|_{\mathrm{F}}^2 \\
\leq\ &p\ \|\boldsymbol{F}\|_{\mathrm{F}}^2\frac{1}{n\eta} \\
=\ &\frac{p}{n\eta}\|\boldsymbol{F}\|_{\mathrm{F}}^2,
\end{aligned}$$

where the third inequality is by Lemma B.2. Also, we have

$$\begin{aligned}
&\|(\boldsymbol{I}_T \otimes \boldsymbol{X}_{\hat{\mathcal{S}}})\boldsymbol{M}^{\dagger}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})(\boldsymbol{F}^{\top} \otimes \boldsymbol{I}_p)\|_{\mathrm{F}}^2 \\
\leq\ &\|(\boldsymbol{I}_T \otimes \boldsymbol{X}_{\hat{\mathcal{S}}})\boldsymbol{M}^{\dagger}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\|_{\mathrm{F}}^2\|(\boldsymbol{F}^{\top} \otimes \boldsymbol{I}_p)\|_{\mathrm{op}}^2 \\
\leq\ &\mathrm{Tr}\left[(\boldsymbol{I}_T \otimes \boldsymbol{X}_{\hat{\mathcal{S}}}^{\top}\boldsymbol{X}_{\hat{\mathcal{S}}})\boldsymbol{M}^{\dagger}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_{\hat{\mathcal{S}},\hat{\mathcal{S}}})\boldsymbol{M}^{\dagger}\right]\|\boldsymbol{F}\|_{\mathrm{F}}^2 \\
\leq\ &\mathrm{Tr}\left[(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_{\hat{\mathcal{S}},\hat{\mathcal{S}}})\boldsymbol{M}^{\dagger}\right]\|\boldsymbol{F}\|_{\mathrm{F}}^2 \\
\leq\ &\mathrm{Tr}\left[(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_{\hat{\mathcal{S}},\hat{\mathcal{S}}})(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^{\top}\boldsymbol{X}_{\hat{\mathcal{S}}})^{\dagger})\right]\|\boldsymbol{F}\|_{\mathrm{F}}^2 \\
=\ &T\,\mathrm{Tr}\left[\boldsymbol{\Sigma}_{\hat{\mathcal{S}},\hat{\mathcal{S}}}(\boldsymbol{X}_{\hat{\mathcal{S}}}^{\top}\boldsymbol{X}_{\hat{\mathcal{S}}})^{\dagger}\right]\|\boldsymbol{F}\|_{\mathrm{F}}^2 \\
\leq\ &T\,\mathrm{Tr}\left[(n\eta)^{-1}\boldsymbol{X}_{\hat{\mathcal{S}}}^{\top}\boldsymbol{X}_{\hat{\mathcal{S}}}(\boldsymbol{X}_{\hat{\mathcal{S}}}^{\top}\boldsymbol{X}_{\hat{\mathcal{S}}})^{\dagger}\right]\|\boldsymbol{F}\|_{\mathrm{F}}^2 \\
\leq\ &\frac{T}{\eta}\|\boldsymbol{F}\|_{\mathrm{F}}^2,
\end{aligned}$$

where the penultimate inequality uses $\boldsymbol{\Sigma}_{\hat{\mathcal{S}},\hat{\mathcal{S}}} \preceq (n\eta)^{-1}\boldsymbol{X}_{\hat{\mathcal{S}}}^{\top}\boldsymbol{X}_{\hat{\mathcal{S}}}$ in the event $U_1 \cap U_2$. Therefore, on $U_1 \cap U_2$, we have

$$\frac{1}{n}\sum_{ij}\left\|\frac{\partial \boldsymbol{F}/D}{\partial z_{ij}}\right\|_{\mathrm{F}}^2 \leq 4\max(1,(\eta)^{-1}(T\wedge\frac{p}{n})) + 2n^{-1}[4\max(1,(2\tau')^{-1})]^2$$
$$:= f(\eta, T, n, p),$$

where the function f is the same as in case (1). The only difference is that τ' in the upper bound for case (1) is replaced by η in case (2).

■

G.2. Proofs of results in Section E.2

The following proof of Lemma E.6 relies on a similar argument as proof of Lemma E.4, we present the proof here for completeness.

Proof of Lemma E.6. Recall the KKT conditions for $\widehat{\boldsymbol{B}}$ defined in (6):

1) For $k \in \hat{\mathcal{S}}$, i.e., $\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k \neq \boldsymbol{0}$,

$$\boldsymbol{e}_k^\top \boldsymbol{X}^\top \left[\boldsymbol{E} - \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\right] - n\tau \boldsymbol{e}_k^\top \widehat{\boldsymbol{B}} = \frac{n\lambda}{\|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k\|_2} \boldsymbol{e}_k^\top \widehat{\boldsymbol{B}} \quad \in \mathbb{R}^{1\times T}.$$

2) For $k \notin \hat{\mathcal{S}}$, i.e., $\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k = \boldsymbol{0}$,

$$\left\| \boldsymbol{e}_k^\top \boldsymbol{X}^\top \left[\boldsymbol{E} - \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\right] - n\tau \boldsymbol{e}_k^\top \widehat{\boldsymbol{B}} \right\| < n\lambda.$$

Here the strict inequality is guaranteed by Proposition 2.3 of Bellec (2020).

Let $\ddot{\boldsymbol{B}} = \frac{\partial \widehat{\boldsymbol{B}}}{\partial E_{it'}}$, $\dot{\boldsymbol{E}} = \frac{\partial \boldsymbol{E}}{\partial E_{it'}}$. Differentiation of the above display for $k \in \hat{\mathcal{S}}$ w.r.t. $E_{it'}$ yields

$$\boldsymbol{e}_k^\top \boldsymbol{X}^\top (\dot{\boldsymbol{E}} - \boldsymbol{X}\ddot{\boldsymbol{B}}) - n\tau \boldsymbol{e}_k^\top \ddot{\boldsymbol{B}} = n \boldsymbol{e}_k^\top \ddot{\boldsymbol{B}} \boldsymbol{H}^{(k)}$$

with $\boldsymbol{H}^{(k)} = \lambda \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k\|_2^{-1} \left(\boldsymbol{I}_T - \widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k \boldsymbol{e}_k^\top \widehat{\boldsymbol{B}} \, \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_k\|_2^{-2} \right) \in \mathbb{R}^{T\times T}$. Rearranging and using $\dot{\boldsymbol{E}} = \boldsymbol{e}_i \boldsymbol{e}_{t'}^\top$,

$$\boldsymbol{e}_k^\top \boldsymbol{X}^\top \boldsymbol{e}_i \boldsymbol{e}_{t'}^\top = \boldsymbol{e}_k^\top [n\ddot{\boldsymbol{B}} \boldsymbol{H}^{(k)} + (\boldsymbol{X}^\top \boldsymbol{X} + n\tau \boldsymbol{I}_{p\times p})\ddot{\boldsymbol{B}}].$$

Recall $\boldsymbol{P}_{\hat{\mathcal{S}}} = \sum_{k\in\hat{\mathcal{S}}} \boldsymbol{e}_k \boldsymbol{e}_k^\top \in \mathbb{R}^{p\times p}$. Multiplying by $\boldsymbol{e}_k$ to the left and summing over $k \in \hat{\mathcal{S}}$, we obtain

$$\boldsymbol{P}_{\hat{\mathcal{S}}} \boldsymbol{X}^\top \boldsymbol{e}_i \boldsymbol{e}_{t'}^\top = n \sum_{k\in\hat{\mathcal{S}}} \boldsymbol{e}_k \boldsymbol{e}_k^\top \ddot{\boldsymbol{B}} \boldsymbol{H}^{(k)} + \boldsymbol{P}_{\hat{\mathcal{S}}} (\boldsymbol{X}^\top \boldsymbol{X} + n\tau \boldsymbol{I}_{p\times p})\ddot{\boldsymbol{B}},$$

which reduces to the following by $\operatorname{supp}(\ddot{\boldsymbol{B}}) \subseteq \hat{\mathcal{S}}$ and $\boldsymbol{X}\ddot{\boldsymbol{B}} = \boldsymbol{X}_{\hat{\mathcal{S}}} \ddot{\boldsymbol{B}}$,

$$\begin{aligned}
\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{e}_i \boldsymbol{e}_{t'}^\top &= n \sum_{k\in\hat{\mathcal{S}}} \boldsymbol{e}_k \boldsymbol{e}_k^\top \ddot{\boldsymbol{B}} \boldsymbol{H}^{(k)} + \boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} \ddot{\boldsymbol{B}} \boldsymbol{I}_T + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}} \ddot{\boldsymbol{B}} \boldsymbol{I}_T \\
&= n \sum_{k\in\hat{\mathcal{S}}} \boldsymbol{e}_k \boldsymbol{e}_k^\top \ddot{\boldsymbol{B}} \boldsymbol{H}^{(k)} + (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}}) \ddot{\boldsymbol{B}} \boldsymbol{I}_T.
\end{aligned}$$

Vectorizing the above yields

$$\begin{aligned}
(\boldsymbol{e}_{t'} \otimes \boldsymbol{X}_{\hat{\mathcal{S}}}^\top) \operatorname{vec}(\boldsymbol{e}_i) &= [n \sum_{k\in\hat{\mathcal{S}}} (\boldsymbol{H}^{(k)} \otimes \boldsymbol{e}_k \boldsymbol{e}_k^\top) + \boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})] \operatorname{vec}(\ddot{\boldsymbol{B}}) \\
&= (n\tilde{\boldsymbol{H}} + \boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + n\tau \boldsymbol{P}_{\hat{\mathcal{S}}})) \operatorname{vec}(\ddot{\boldsymbol{B}}) \\
&= \boldsymbol{M} \operatorname{vec}(\ddot{\boldsymbol{B}}).
\end{aligned}$$

A similar argument as in Proof of Lemma E.4 leads to

$$\mathbf{vec}(\ddot{\boldsymbol{B}}) = \boldsymbol{M}^{\dagger}\boldsymbol{M}\,\mathbf{vec}(\ddot{\boldsymbol{B}}) = \boldsymbol{M}^{\dagger}(\boldsymbol{e}_{t'} \otimes \boldsymbol{X}_{\hat{\mathscr{P}}}^{\top})\boldsymbol{e}_i.$$

Therefore, by $\boldsymbol{X}\ddot{\boldsymbol{B}} = \boldsymbol{X}_{\hat{\mathscr{P}}}\ddot{\boldsymbol{B}}$,

$$\begin{aligned}
\frac{\partial F_{lt}}{\partial E_{it'}} &= \boldsymbol{e}_l^{\top}\frac{\partial \boldsymbol{E} - \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)}{\partial E_{it'}}\boldsymbol{e}_t \\
&= \boldsymbol{e}_l^{\top}\big(\boldsymbol{e}_i\boldsymbol{e}_{t'}^{\top} - \boldsymbol{X}\ddot{\boldsymbol{B}}\big)\boldsymbol{e}_t \\
&= \boldsymbol{e}_l^{\top}\boldsymbol{e}_i\boldsymbol{e}_{t'}^{\top}\boldsymbol{e}_t - \boldsymbol{e}_l^{\top}\boldsymbol{X}\ddot{\boldsymbol{B}}\boldsymbol{e}_t \\
&= \boldsymbol{e}_l^{\top}\boldsymbol{e}_i\boldsymbol{e}_{t'}^{\top}\boldsymbol{e}_t - (\boldsymbol{e}_t^{\top} \otimes \boldsymbol{e}_l^{\top}\boldsymbol{X}_{\hat{\mathscr{P}}})\,\mathbf{vec}(\ddot{\boldsymbol{B}}) \\
&= \boldsymbol{e}_l^{\top}\boldsymbol{e}_i\boldsymbol{e}_{t'}^{\top}\boldsymbol{e}_t - (\boldsymbol{e}_t^{\top} \otimes \boldsymbol{e}_l^{\top}\boldsymbol{X}_{\hat{\mathscr{P}}})\boldsymbol{M}^{\dagger}(\boldsymbol{e}_{t'} \otimes \boldsymbol{X}_{\hat{\mathscr{P}}}^{\top})\boldsymbol{e}_i \\
&= \boldsymbol{e}_l^{\top}\boldsymbol{e}_i\boldsymbol{e}_{t'}^{\top}\boldsymbol{e}_t - \boldsymbol{e}_l^{\top}(\boldsymbol{e}_t^{\top} \otimes \boldsymbol{X}_{\hat{\mathscr{P}}})\boldsymbol{M}^{\dagger}(\boldsymbol{e}_{t'} \otimes \boldsymbol{X}_{\hat{\mathscr{P}}}^{\top})\boldsymbol{e}_i \\
&= \boldsymbol{e}_l^{\top}\boldsymbol{e}_i\boldsymbol{e}_{t'}^{\top}\boldsymbol{e}_t - \boldsymbol{e}_l^{\top}(\boldsymbol{e}_t^{\top} \otimes \boldsymbol{X})\boldsymbol{M}^{\dagger}(\boldsymbol{e}_{t'} \otimes \boldsymbol{X}^{\top})\boldsymbol{e}_i,
\end{aligned}$$

where the last equality is due to $\boldsymbol{M}^{\dagger} = (\boldsymbol{I}_T \otimes \boldsymbol{P}_{\hat{\mathscr{P}}})\boldsymbol{M}^{\dagger}(\boldsymbol{I}_T \otimes \boldsymbol{P}_{\hat{\mathscr{P}}})$.

Now the calculation of $\sum_{i=1}^n \frac{\partial F_{it}}{\partial E_{it'}}$ is straightforward,

$$\begin{aligned}
\sum_{i=1}^n \frac{\partial F_{it}}{\partial E_{it'}} &= \sum_{i=1}^n \big[\boldsymbol{e}_i^{\top}\boldsymbol{e}_i\boldsymbol{e}_t^{\top}\boldsymbol{e}_{t'} - \boldsymbol{e}_i^{\top}(\boldsymbol{e}_t^{\top} \otimes \boldsymbol{X})\boldsymbol{M}^{\dagger}(\boldsymbol{e}_{t'} \otimes \boldsymbol{X}^{\top})\boldsymbol{e}_i\big] \\
&= n\boldsymbol{e}_t^{\top}\boldsymbol{e}_{t'} - \mathrm{Tr}[(\boldsymbol{e}_t^{\top} \otimes \boldsymbol{X})\boldsymbol{M}^{\dagger}(\boldsymbol{e}_{t'} \otimes \boldsymbol{X}^{\top})] \\
&= n\boldsymbol{e}_t^{\top}\boldsymbol{e}_{t'} - \boldsymbol{e}_t^{\top}\widehat{\mathbf{A}}\boldsymbol{e}_{t'} \\
&= \boldsymbol{e}_t^{\top}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{e}_{t'},
\end{aligned}$$

where the third equality is due to the formula of $\widehat{\mathbf{A}}$ in (10).

Noting that $\boldsymbol{F} = \boldsymbol{E} - \boldsymbol{Z}\boldsymbol{H}$, it follows that $\sum_{i=1}^n \frac{\partial \boldsymbol{e}_i^{\top}\boldsymbol{Z}\boldsymbol{H}\boldsymbol{e}_t}{\partial E_{it'}} = \boldsymbol{e}_t^{\top}\widehat{\mathbf{A}}\boldsymbol{e}_{t'}$. ■

G.3. Proofs of results in Section E.3

Proof of Lemma E.9. Let $z = \mathbf{vec}(\boldsymbol{E})$, then $z \sim \mathcal{N}(0, \boldsymbol{K})$ with $\boldsymbol{K} = \boldsymbol{S} \otimes \boldsymbol{I}_n$ by Assumption 1. For each $t_0, t_0' \in [T]$, let $\boldsymbol{G}^{(t_0,t_0')} = \boldsymbol{F}\boldsymbol{e}_{t_0'}\boldsymbol{e}_{t_0}^{\top}$, and $\boldsymbol{f}(z)^{(t_0,t_0')} = \mathbf{vec}(\boldsymbol{G})\tilde{D}^{-1}$. For convenience, we will drop the superscript (t_0, t_0') from $\boldsymbol{G}^{(t_0,t_0')}$ and $\boldsymbol{f}(z)^{(t_0,t_0')}$ in this proof. By $\mathrm{Tr}(\boldsymbol{A}^{\top}\boldsymbol{B}) = \mathbf{vec}(\boldsymbol{A})^{\top}\mathbf{vec}(\boldsymbol{B})$, we obtain

$$\boldsymbol{e}_{t_0}^{\top}\boldsymbol{E}^{\top}\boldsymbol{F}\tilde{D}^{-1}\boldsymbol{e}_{t_0'} = \mathrm{Tr}(\boldsymbol{E}^{\top}\boldsymbol{F}\boldsymbol{e}_{t_0'}\boldsymbol{e}_{t_0}^{\top})\tilde{D}^{-1} = \mathrm{Tr}(\boldsymbol{E}^{\top}\boldsymbol{G}\tilde{D}^{-1}) = z^{\top}\boldsymbol{f}(z). \tag{55}$$

By product rule, we have

$$\nabla \boldsymbol{f}(z) = \frac{\partial\,\mathbf{vec}(\boldsymbol{G})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\tilde{D}^{-1} + \underbrace{\mathbf{vec}(\boldsymbol{G})\frac{\partial \tilde{D}^{-1}}{\partial\,\mathbf{vec}(\boldsymbol{E})}}_{\text{Rem}}, \tag{56}$$

where $\mathrm{Rem} = \boldsymbol{u}\boldsymbol{v}^\top$ with $\boldsymbol{u} = \mathbf{vec}(\boldsymbol{G}) \in \mathbb{R}^{nT\times 1}$, $\boldsymbol{v}^\top = \frac{\partial \tilde{D}^{-1}}{\partial \mathbf{vec}(\boldsymbol{E})} \in \mathbb{R}^{1\times nT}$. It follows that

$$\mathrm{Tr}(\boldsymbol{K}\nabla \boldsymbol{f}(z)) = \mathrm{Tr}\left(\boldsymbol{K}\frac{\partial \mathbf{vec}(\boldsymbol{G})}{\partial \mathbf{vec}(\boldsymbol{E})}\right)\tilde{D}^{-1} + \mathrm{Tr}(\boldsymbol{K}\mathrm{Rem}). \tag{57}$$

Since $\boldsymbol{K} = \boldsymbol{S}\otimes \boldsymbol{I}_n$ and $\boldsymbol{G} = \boldsymbol{F}\boldsymbol{e}_{t_0'}\boldsymbol{e}_{t_0}^\top$, $\boldsymbol{K}_{it,lt'} = S_{tt'}I(i=l)$, and $G_{it} = F_{it_0'}I(t=t_0)$. It follows

$$\mathrm{Tr}\left(\boldsymbol{K}\frac{\partial \mathbf{vec}(\boldsymbol{G})}{\partial \mathbf{vec}(\boldsymbol{E})}\right) = \sum_{i,t}\sum_{l,t'}\boldsymbol{K}_{it,lt'}\frac{\partial G_{it}}{\partial E_{lt'}} = \sum_{t'}S_{t_0t'}\sum_{i}\frac{\partial F_{it_0'}}{\partial E_{it'}} = \boldsymbol{e}_{t_0}^\top \boldsymbol{S}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{e}_{t_0'}, \tag{58}$$

where the last equality used Lemma E.6 and that $\widehat{\mathbf{A}}$ is symmetric.

Now we rewrite the quantity we want to bound as

$$\begin{aligned}
&\mathbb{E}\Big[\|\boldsymbol{E}^\top\boldsymbol{F}/\tilde{D} - \boldsymbol{S}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})/\tilde{D}\|_{\mathrm{F}}^2\Big] \\
&= \sum_{t_0,t_0'}\mathbb{E}\Big[\Big(\boldsymbol{e}_{t_0}^\top\boldsymbol{E}^\top\boldsymbol{F}\tilde{D}^{-1}\boldsymbol{e}_{t_0'} - \boldsymbol{e}_{t_0}^\top\boldsymbol{S}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})\boldsymbol{e}_{t_0'}\tilde{D}^{-1}\Big)^2\Big] \\
&= \sum_{t_0,t_0'}\mathbb{E}\Big[\big(z^\top\boldsymbol{f}(z) - \mathrm{Tr}(\boldsymbol{K}\nabla\boldsymbol{f}(z)) + \mathrm{Tr}(\boldsymbol{K}\mathrm{Rem})\big)^2\Big] \\
&\le 2\sum_{t_0,t_0'}\Big\{\mathbb{E}\Big[\big(z^\top\boldsymbol{f}(z) - \mathrm{Tr}(\boldsymbol{K}\nabla\boldsymbol{f}(z))\big)^2\Big] + \mathbb{E}\Big[\big(\mathrm{Tr}(\boldsymbol{K}\mathrm{Rem})\big)^2\Big]\Big\},
\end{aligned} \tag{59}$$

where the second equality follows from (55), (57) and (58), and the last inequality uses elementary inequality $(a+b)^2 \le 2(a^2+b^2)$. We next bound the two terms in (59). **First term in (59).** By second-order Stein formula in Lemma E.7,

$$\sum_{t_0,t_0'}\mathbb{E}\big(z^\top\boldsymbol{f}(z) - \mathrm{Tr}(\boldsymbol{K}\nabla\boldsymbol{f}(z))\big)^2 = \sum_{t_0,t_0'}\mathbb{E}\Big[\|\boldsymbol{K}^{\frac12}\boldsymbol{f}(z)\|_{\mathrm{F}}^2 + \mathrm{Tr}\left[\big(\boldsymbol{K}\nabla\boldsymbol{f}(z)\big)^2\right]\Big]. \tag{60}$$

Now we bound the two terms in the right-hand side of (60). For the first term, recall $\boldsymbol{f}(z) = \mathbf{vec}(\boldsymbol{G})\tilde{D}^{-1}$, and $\boldsymbol{G} = \boldsymbol{F}\boldsymbol{e}_{t_0'}\boldsymbol{e}_{t_0}^\top$, we obtain

$$\|\boldsymbol{K}^{\frac12}\boldsymbol{f}(z)\|_{\mathrm{F}}^2 = \tilde{D}^{-2}\|(\boldsymbol{S}^{\frac12}\otimes\boldsymbol{I}_n)\,\mathbf{vec}(\boldsymbol{G})\|_{\mathrm{F}}^2 = \tilde{D}^{-2}\|\boldsymbol{G}\boldsymbol{S}^{\frac12}\|_{\mathrm{F}}^2 = \tilde{D}^{-2}\|\boldsymbol{S}^{\frac12}\boldsymbol{e}_{t_0}\|_{\mathrm{F}}^2\|\boldsymbol{F}\boldsymbol{e}_{t_0'}\|_{\mathrm{F}}^2.$$

Summing over all $(t_0, t_0') \in [T]\times[T]$, we obtain

$$\sum_{t_0,t_0'}\|\boldsymbol{K}^{\frac12}\boldsymbol{f}(z)\|_{\mathrm{F}}^2 = \tilde{D}^{-2}\|\boldsymbol{F}\|_{\mathrm{F}}^2\,\mathrm{Tr}(\boldsymbol{S}). \tag{61}$$

For the second term in RHS of (60), recall $\nabla\boldsymbol{f}(z) = \frac{\partial \mathbf{vec}(\boldsymbol{G})}{\partial \mathbf{vec}(\boldsymbol{E})}\tilde{D}^{-1} + \mathrm{Rem}$,

$$\begin{aligned}
&\mathrm{Tr}\left[\big(\boldsymbol{K}\nabla\boldsymbol{f}(z)\big)^2\right] \\
&= \tilde{D}^{-2}\,\mathrm{Tr}\left[\left(\boldsymbol{K}\frac{\partial \mathbf{vec}(\boldsymbol{G})}{\partial \mathbf{vec}(\boldsymbol{E})}\right)^2\right] + \mathrm{Tr}[(\boldsymbol{K}\mathrm{Rem})^2] + 2\tilde{D}^{-1}\,\mathrm{Tr}\left[\boldsymbol{K}\frac{\partial \mathbf{vec}(\boldsymbol{G})}{\partial \mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}\mathrm{Rem}\right].
\end{aligned} \tag{62}$$

By property of vectorization operation, $\mathbf{vec}(\boldsymbol{G}) = \mathbf{vec}(\boldsymbol{F}\boldsymbol{e}_{t_0'}\boldsymbol{e}_{t_0}^\top) = (\boldsymbol{e}_{t_0}\boldsymbol{e}_{t_0'}^\top \otimes \boldsymbol{I}_n)\,\mathbf{vec}(\boldsymbol{F})$, hence

$$\frac{\partial\,\mathbf{vec}(\boldsymbol{G})}{\partial\,\mathbf{vec}(\boldsymbol{E})} = (\boldsymbol{e}_{t_0}\boldsymbol{e}_{t_0'}^\top \otimes \boldsymbol{I}_n)\frac{\partial\,\mathbf{vec}(\boldsymbol{F})}{\partial\,\mathbf{vec}(\boldsymbol{E})},$$

where $\|\frac{\partial\,\mathbf{vec}(\boldsymbol{F})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\|_{\mathrm{op}} \le 1$ since the map $\mathbf{vec}(\boldsymbol{E}) \mapsto \mathbf{vec}(\boldsymbol{F})$ is 1-Lipschitz by (Bellec and Tsybakov, 2016, proposition 3).

Now we bound the three terms in (62). For the first term, by Cauchy-Schwarz inequality,

$$\begin{aligned}
&\tilde{D}^{-2}\,\mathrm{Tr}\left[\left(\boldsymbol{K}\frac{\partial\,\mathbf{vec}(\boldsymbol{G})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\right)^2\right] \\
&= \tilde{D}^{-2}\,\mathrm{Tr}\left(\boldsymbol{K}(\boldsymbol{e}_{t_0}\boldsymbol{e}_{t_0'}^\top \otimes \boldsymbol{I}_n)\frac{\partial\,\mathbf{vec}(\boldsymbol{F})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}(\boldsymbol{e}_{t_0}\boldsymbol{e}_{t_0'}^\top \otimes \boldsymbol{I}_n)\frac{\partial\,\mathbf{vec}(\boldsymbol{F})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\right) \\
&\le \tilde{D}^{-2}\|(\boldsymbol{e}_{t_0'}^\top \otimes \boldsymbol{I}_n)\frac{\partial\,\mathbf{vec}(\boldsymbol{F})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}(\boldsymbol{e}_{t_0} \otimes \boldsymbol{I}_n)\|_{\mathrm{F}}^2.
\end{aligned}$$

For the second term in (62), recall $\mathrm{Rem} = \boldsymbol{u}\boldsymbol{v}^\top$, and $\boldsymbol{u} = \mathbf{vec}(\boldsymbol{G})$, $\boldsymbol{v}^\top = \frac{\partial\tilde{D}^{-1}}{\partial\,\mathbf{vec}(\boldsymbol{E})}$ from (56), then $\mathrm{Tr}[(\boldsymbol{K}\mathrm{Rem})^2] = \mathrm{Tr}(\boldsymbol{K}\boldsymbol{u}\boldsymbol{v}^\top\boldsymbol{K}\boldsymbol{u}\boldsymbol{v}^\top) = (\boldsymbol{v}^\top\boldsymbol{K}\boldsymbol{u})^2$, thus,

$$\begin{aligned}
&\mathrm{Tr}[(\boldsymbol{K}\mathrm{Rem})^2] \\
&= \left[\frac{\partial\tilde{D}^{-1}}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}\,\mathbf{vec}(\boldsymbol{G})\right]^2 \\
&= \tilde{D}^{-6}\left[\,\mathbf{vec}(\boldsymbol{F})^\top\frac{\partial\,\mathbf{vec}(\boldsymbol{F})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}(\boldsymbol{e}_{t_0}\boldsymbol{e}_{t_0'}^\top \otimes \boldsymbol{I}_n)\,\mathbf{vec}(\boldsymbol{F})\right]^2 \\
&\le \tilde{D}^{-6}\|\mathbf{vec}(\boldsymbol{F})^\top\frac{\partial\,\mathbf{vec}(\boldsymbol{F})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}(\boldsymbol{e}_{t_0} \otimes \boldsymbol{I}_n)\|^2\|(\boldsymbol{e}_{t_0'}^\top \otimes \boldsymbol{I}_n)\,\mathbf{vec}(\boldsymbol{F})\|^2.
\end{aligned}$$

where the second equality uses $\frac{\partial\tilde{D}^{-1}}{\partial\,\mathbf{vec}(\boldsymbol{E})} = -\frac{1}{2}\tilde{D}^{-3}\frac{\partial\|\boldsymbol{F}\|_{\mathrm{F}}^2}{\partial\,\mathbf{vec}(\boldsymbol{E})} = -\tilde{D}^{-3}\,\mathbf{vec}(\boldsymbol{F})^\top\frac{\partial\,\mathbf{vec}(\boldsymbol{F})}{\partial\,\mathbf{vec}(\boldsymbol{E})}$ by chain rule, and the inequality uses Cauchy-Schwarz inequality.

For the third term in (62), recall $\mathrm{Rem} = \boldsymbol{u}\boldsymbol{v}^\top$, and $\boldsymbol{u} = \mathbf{vec}(\boldsymbol{G})$, $\boldsymbol{v}^\top = \frac{\partial\tilde{D}^{-1}}{\partial\,\mathbf{vec}(\boldsymbol{E})}$, then we have $\mathrm{Tr}\left[\boldsymbol{K}\frac{\partial\,\mathbf{vec}(\boldsymbol{G})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}\mathrm{Rem}\right] = \boldsymbol{v}^\top\boldsymbol{K}\frac{\partial\,\mathbf{vec}(\boldsymbol{G})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}\boldsymbol{u}$, hence

$$\begin{aligned}
&2\tilde{D}^{-1}\,\mathrm{Tr}\left[\boldsymbol{K}\frac{\partial\,\mathbf{vec}(\boldsymbol{G})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}\mathrm{Rem}\right] \\
&= 2\tilde{D}^{-1}\frac{\partial\tilde{D}^{-1}}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}\frac{\partial\,\mathbf{vec}(\boldsymbol{G})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}\,\mathbf{vec}(\boldsymbol{G}) \\
&= -2\tilde{D}^{-4}\,\mathbf{vec}(\boldsymbol{F})^\top\frac{\partial\,\mathbf{vec}(\boldsymbol{F})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}(\boldsymbol{e}_{t_0}\boldsymbol{e}_{t_0'}^\top \otimes \boldsymbol{I}_n)\frac{\partial\,\mathbf{vec}(\boldsymbol{F})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}(\boldsymbol{e}_{t_0}\boldsymbol{e}_{t_0'}^\top \otimes \boldsymbol{I}_n)\,\mathbf{vec}(\boldsymbol{F}) \\
&\le 2\tilde{D}^{-4}\|(\boldsymbol{e}_{t_0'}^\top \otimes \boldsymbol{I}_n)\,\mathbf{vec}(\boldsymbol{F})\,\mathbf{vec}(\boldsymbol{F})^\top\frac{\partial\,\mathbf{vec}(\boldsymbol{F})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}(\boldsymbol{e}_{t_0} \otimes \boldsymbol{I}_n)\|_{\mathrm{F}} \\
&\quad \|(\boldsymbol{e}_{t_0'}^\top \otimes \boldsymbol{I}_n)\frac{\partial\,\mathbf{vec}(\boldsymbol{F})}{\partial\,\mathbf{vec}(\boldsymbol{E})}\boldsymbol{K}(\boldsymbol{e}_{t_0} \otimes \boldsymbol{I}_n)\|_{\mathrm{F}},
\end{aligned}$$

where the last inequality uses Cauchy-Schwarz inequality.

Summing over all $(t_0, t_0') \in [T] \times [T]$ for these three terms in (62), using $\|\frac{\partial \mathbf{vec}(\boldsymbol{F})}{\partial \mathbf{vec}(\boldsymbol{E})}\|_{\mathrm{op}} \le 1$, $\boldsymbol{K} = \boldsymbol{S} \otimes \boldsymbol{I}_n$, and $\|\boldsymbol{S}\|_{\mathrm{F}} \le \|\boldsymbol{S}^{\frac{1}{2}}\|_{\mathrm{F}}^2 = \mathrm{Tr}(\boldsymbol{S})$, we obtain

$$\begin{aligned}
&\sum_{t_0, t_0'} \mathrm{Tr}\left[\left(\boldsymbol{K} \nabla \boldsymbol{f}(z)\right)^2\right] \\
&\le \tilde{D}^{-2} \|\boldsymbol{K}\|_{\mathrm{F}}^2 + \tilde{D}^{-6} \|\boldsymbol{F}\|_{\mathrm{F}}^2 \|\boldsymbol{K}\|_{\mathrm{F}}^2 \|\boldsymbol{F}\|_{\mathrm{F}}^2 + 2\tilde{D}^{-4} \|\boldsymbol{F}\|_{\mathrm{F}}^2 \|\boldsymbol{K}\|_{\mathrm{F}}^2 \\
&= \tilde{D}^{-2} n \|\boldsymbol{S}\|_{\mathrm{F}}^2 + \tilde{D}^{-6} \|\boldsymbol{F}\|_{\mathrm{F}}^4 n \|\boldsymbol{S}\|_{\mathrm{F}}^2 + 2\tilde{D}^{-4} \|\boldsymbol{F}\|_{\mathrm{F}}^2 n \|\boldsymbol{S}\|_{\mathrm{F}}^2 \\
&\le \left[\tilde{D}^{-2} n \,\mathrm{Tr}(\boldsymbol{S}) + \tilde{D}^{-6} \|\boldsymbol{F}\|_{\mathrm{F}}^4 n \,\mathrm{Tr}(\boldsymbol{S}) + 2\tilde{D}^{-4} \|\boldsymbol{F}\|_{\mathrm{F}}^2 n \,\mathrm{Tr}(\boldsymbol{S})\right] \mathrm{Tr}(\boldsymbol{S}).
\end{aligned} \tag{63}$$

Second term in (59). Recall that $\mathrm{Rem} = \boldsymbol{u}\boldsymbol{v}^\top$, hence $[\mathrm{Tr}(\boldsymbol{K}\mathrm{Rem})]^2 = \mathrm{Tr}[(\boldsymbol{K}\mathrm{Rem})^2]$. By calculation of second term in (62), we obtain

$$\sum_{t_0, t_0'} [\mathrm{Tr}(\boldsymbol{K}\mathrm{Rem})]^2 = \sum_{t_0, t_0'} \mathrm{Tr}[(\boldsymbol{K}\mathrm{Rem})^2] \le \tilde{D}^{-6} \|\boldsymbol{F}\|_{\mathrm{F}}^4 n \,\mathrm{Tr}(\boldsymbol{S})^2. \tag{64}$$

Combining the results (59), (60), (61), (63), (64), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{E}^\top \boldsymbol{F}/\tilde{D} - \boldsymbol{S}(n\boldsymbol{I}_T - \widehat{\mathbf{A}})/\tilde{D}\|_{\mathrm{F}}^2\right] \\
&\le 2\left[\tilde{D}^{-2} \|\boldsymbol{F}\|_{\mathrm{F}}^2 + \tilde{D}^{-2} n \,\mathrm{Tr}(\boldsymbol{S}) + 2\tilde{D}^{-6} \|\boldsymbol{F}\|_{\mathrm{F}}^4 n \,\mathrm{Tr}(\boldsymbol{S}) + 2\tilde{D}^{-4} \|\boldsymbol{F}\|_{\mathrm{F}}^2 n \,\mathrm{Tr}(\boldsymbol{S})\right] \mathrm{Tr}(\boldsymbol{S}) \\
&\le 4 \,\mathrm{Tr}(\boldsymbol{S}),
\end{aligned}$$

thanks to $\tilde{D}^2 = \|\boldsymbol{F}\|_{\mathrm{F}}^2 + n \,\mathrm{Tr}(\boldsymbol{S})$. ■

Proof of Lemma E.10. Apply (Bellec, 2020, Proposition 6.3) with $\boldsymbol{\rho} = \boldsymbol{U}\boldsymbol{e}_t$, $\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{e}_{t'}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{U}^\top \boldsymbol{Z} \boldsymbol{V} - \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{\partial}{\partial z_{ij}} \left(\boldsymbol{U}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{V}\right)\right\|_{\mathrm{F}}^2\right] \\
&= \sum_{t,t'=1}^{T} \mathbb{E}\left(\boldsymbol{e}_t^\top \boldsymbol{U}^\top \boldsymbol{Z} \boldsymbol{V} \boldsymbol{e}_{t'} - \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{\partial}{\partial z_{ij}} \boldsymbol{e}_t \boldsymbol{U}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{V} \boldsymbol{e}_{t'}\right)^2 \\
&\le \sum_{t,t'=1}^{T} \left[\mathbb{E}\left[\|\boldsymbol{U}\boldsymbol{e}_t\|^2 \|\boldsymbol{V}\boldsymbol{e}_{t'}\|^2\right] + \mathbb{E} \sum_{ij} \left[2\|\boldsymbol{V}\boldsymbol{e}_{t'}\|^2 \left\|\frac{\partial \boldsymbol{U}\boldsymbol{e}_t}{\partial z_{ij}}\right\|_{\mathrm{F}} + 2\|\boldsymbol{U}\boldsymbol{e}_t\|^2 \left\|\frac{\partial \boldsymbol{V}\boldsymbol{e}_{t'}}{\partial z_{ij}}\right\|_{\mathrm{F}}\right]\right] \\
&= \mathbb{E}\|\boldsymbol{U}\|_{\mathrm{F}}^2 \|\boldsymbol{V}\|_{\mathrm{F}}^2 + \mathbb{E} \sum_{ij} \left[2\|\boldsymbol{V}\|_{\mathrm{F}}^2 \left\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\right\|_{\mathrm{F}}^2 + 2\|\boldsymbol{U}\|_{\mathrm{F}}^2 \left\|\frac{\partial \boldsymbol{V}}{\partial z_{ij}}\right\|_{\mathrm{F}}^2\right].
\end{aligned}$$

■

Proof of Corollary E.11. By Kirszbraun's theorem, there exists an L_1-Lipschitz function $\bar{\boldsymbol{U}} : \mathbb{R}^{n \times p} \to \mathbb{R}^{n \times T}$ such that $\bar{\boldsymbol{U}} = \boldsymbol{U}$ on Ω, and an L_2-Lipschitz function $\bar{\boldsymbol{V}} : \mathbb{R}^{n \times p} \to \mathbb{R}^{n \times T}$ such that $\bar{\boldsymbol{V}} = \boldsymbol{V}$ on Ω. By

projecting $\bar{\boldsymbol{U}}$ and $\bar{\boldsymbol{V}}$ onto the Euclidean ball of radius 1 and K if necessary, we assume without loss of generality that $\|\bar{\boldsymbol{U}}\|_{\mathrm{F}} \le 1$ and $\|\bar{\boldsymbol{V}}\|_{\mathrm{F}} \le K$. Therefore,

$$
\begin{aligned}
&\mathbb{E}\Big[I(\Omega)\|\boldsymbol{U}^\top \boldsymbol{Z}\boldsymbol{V} - \sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\partial}{\partial z_{ij}}\Big(\boldsymbol{U}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{V}\Big)\|_{\mathrm{F}}^2\Big] \\
&= \mathbb{E}\Big[I(\Omega)\|\bar{\boldsymbol{U}}^\top \boldsymbol{Z}\bar{\boldsymbol{V}} - \sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\partial}{\partial z_{ij}}\Big(\bar{\boldsymbol{U}}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \bar{\boldsymbol{V}}\Big)\|_{\mathrm{F}}^2\Big] \\
&\le \mathbb{E}\Big[\|\bar{\boldsymbol{U}}^\top \boldsymbol{Z}\bar{\boldsymbol{V}} - \sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\partial}{\partial z_{ij}}\Big(\bar{\boldsymbol{U}}^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \bar{\boldsymbol{V}}\Big)\|_{\mathrm{F}}^2\Big] \\
&\le \mathbb{E}\Big[\big(\|\bar{\boldsymbol{U}}\|_{\mathrm{F}}^2\|\bar{\boldsymbol{V}}\|_{\mathrm{F}}^2\big) + 2\sum_{ij}\Big(\|\bar{\boldsymbol{V}}\|_{\mathrm{F}}^2\Big\|\frac{\partial \bar{\boldsymbol{U}}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2 + \|\bar{\boldsymbol{U}}\|_{\mathrm{F}}^2\Big\|\frac{\partial \bar{\boldsymbol{V}}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big)\Big] \\
&\le K^2 + 2\mathbb{E}\Big[\sum_{ij}\Big(K^2\Big\|\frac{\partial \bar{\boldsymbol{U}}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2 + \Big\|\frac{\partial \bar{\boldsymbol{V}}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big)\Big] \\
&= K^2 + 2\mathbb{E}\Big[I(\Omega)\sum_{ij}\Big(K^2\Big\|\frac{\partial \bar{\boldsymbol{U}}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2 + \Big\|\frac{\partial \bar{\boldsymbol{V}}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big) + I(\Omega^c)\sum_{ij}\Big(K^2\Big\|\frac{\partial \bar{\boldsymbol{U}}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2 + \Big\|\frac{\partial \bar{\boldsymbol{V}}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big)\Big] \\
&= K^2 + 2\mathbb{E}\Big[I(\Omega)\sum_{ij}\Big(K^2\Big\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2 + \Big\|\frac{\partial \boldsymbol{V}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big) + I(\Omega^c)\sum_{ij}\Big(K^2\Big\|\frac{\partial \bar{\boldsymbol{U}}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2 + \Big\|\frac{\partial \bar{\boldsymbol{V}}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big)\Big] \\
&\le K^2 + 2\mathbb{E}\Big[I(\Omega)\sum_{ij}\Big(K^2\Big\|\frac{\partial \boldsymbol{U}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2 + \Big\|\frac{\partial \boldsymbol{V}}{\partial z_{ij}}\Big\|_{\mathrm{F}}^2\Big)\Big] + 2C(K^2L_1^2 + L_2^2),
\end{aligned}
$$

where the last inequality uses $\sum_{ij}\|\frac{\partial \bar{\boldsymbol{U}}}{\partial z_{ij}}\|_{\mathrm{F}}^2 \le nT(n^{-1/2}L_1)^2 = TL_1^2$, $\sum_{ij}\|\frac{\partial \bar{\boldsymbol{V}}}{\partial z_{ij}}\|_{\mathrm{F}}^2 \le TL_2^2$ by Lipschitz properties of $\bar{\boldsymbol{U}}$, $\bar{\boldsymbol{V}}$, and $P(\Omega^c) \le C/T$. ■

Proof of Lemma E.12. For each $j \in [p]$, let $\mathbb{E}_j(\cdot)$ denote the conditional expectation $\mathbb{E}[\cdot|\{\boldsymbol{Z}e_k, k \ne j\}]$. The left-hand side of the desired inequality can be rewritten as

$$
\mathbb{E}\Big[\Big\|p\boldsymbol{U}^\top \boldsymbol{V} - \sum_{j=1}^{p}(\mathbb{E}_j\boldsymbol{U}^\top \boldsymbol{Z} - \boldsymbol{L}^\top)\boldsymbol{e}_j\boldsymbol{e}_j^\top(\boldsymbol{Z}^\top\mathbb{E}_j\boldsymbol{V} - \hat{\boldsymbol{L}})\Big\|_{\mathrm{F}}\Big]
$$

with $\boldsymbol{L} \in \mathbb{R}^{p\times T}$ defined by $\boldsymbol{L}^\top\boldsymbol{e}_j = \mathbb{E}_j\boldsymbol{U}^\top\boldsymbol{Z}\boldsymbol{e}_j - \boldsymbol{U}^\top\boldsymbol{Z}\boldsymbol{e}_j + \sum_{i=1}^{n}\partial_{ij}\boldsymbol{U}^\top\boldsymbol{e}_i$ and $\hat{\boldsymbol{L}}$ defined similarly with $\boldsymbol{U}$ replaced by $\boldsymbol{V}$. We develop the terms in the sum over j as follows:

$$
\begin{aligned}
& p\boldsymbol{U}^\top\boldsymbol{V} - \sum_{j}(\mathbb{E}_j\boldsymbol{U}^\top\boldsymbol{Z} - \boldsymbol{L}^\top)\boldsymbol{e}_j\boldsymbol{e}_j^\top(\boldsymbol{Z}^\top\mathbb{E}_j\boldsymbol{V} - \hat{\boldsymbol{L}}) \\
= \quad & \sum_{j}\Big(\boldsymbol{U}^\top\boldsymbol{V} - \mathbb{E}_j[\boldsymbol{U}^\top]\mathbb{E}_j\boldsymbol{V}\Big) && (65) \\
& + \sum_{j}\Big(\mathbb{E}_j[\boldsymbol{U}^\top]\mathbb{E}_j\boldsymbol{V} - \mathbb{E}_j\boldsymbol{U}^\top\boldsymbol{Z}\boldsymbol{e}_j\boldsymbol{e}_j^\top\boldsymbol{Z}^\top\mathbb{E}_j\boldsymbol{V}\Big) && (66)
\end{aligned}
$$

$$- \boldsymbol{L}^\top \hat{\boldsymbol{L}} \tag{67}$$

$$+ \sum_j \Big(\mathbb{E}_j \boldsymbol{U}^\top \boldsymbol{Z} \boldsymbol{e}_j \boldsymbol{e}_j^\top \hat{\boldsymbol{L}} \Big) + \Big(\boldsymbol{L}^\top \boldsymbol{e}_j \boldsymbol{e}_j^\top \boldsymbol{Z}^\top \mathbb{E}_j[\boldsymbol{U}] \Big). \tag{68}$$

First, for (67), by the Cauchy-Schwarz inequality $\mathbb{E}\big[\|\boldsymbol{L}^\top \hat{\boldsymbol{L}}\|_{\mathrm{F}}\big] \le \mathbb{E}\big[\|\boldsymbol{L}\|_{\mathrm{F}}^2\big]^{\frac{1}{2}} \mathbb{E}\big[\|\hat{\boldsymbol{L}}\|_{\mathrm{F}}^2\big]^{\frac{1}{2}}$. For a fixed $j \in [p]$ and $t \in [T]$,

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{e}_j^\top \boldsymbol{L} e_t)^2] &\le \sum_{i=1}^n \mathbb{E}[(\boldsymbol{e}_i^\top (\mathbb{E}_j[\boldsymbol{U}] - \boldsymbol{U}) e_t)^2] + \mathbb{E} \sum_{i=1}^n \sum_{l=1}^n \Big(\frac{\boldsymbol{e}_i^\top \partial \boldsymbol{U} e_t}{\partial z_{lj}} \Big)^2 \\
&\le 2\mathbb{E} \sum_{i=1}^n \sum_{l=1}^n \Big(\frac{\boldsymbol{e}_i^\top \partial \boldsymbol{U} e_t}{\partial z_{lj}} \Big)^2,
\end{aligned}$$

where the two inequalities are due to the second-order stein inequality in Lemma E.7, and Gaussian-Poincaré inequality in Lemma E.8, respectively. Summing over $j \in [p]$ and $t \in [T]$ we obtain $\mathbb{E}[\|\boldsymbol{L}\|_{\mathrm{F}}^2] \le 2\mathbb{E} \sum_{lj} \|\partial_{lj} \boldsymbol{U}\|_{\mathrm{F}}^2 = 2\|\boldsymbol{U}\|_\partial^2$. Combined with the same bound for $\hat{\boldsymbol{L}}$, we obtain $\mathbb{E}[\|(67)\|_{\mathrm{F}}^2] \le 2\|\boldsymbol{U}\|_\partial \|\boldsymbol{V}\|_\partial$. We now turn to the two terms in (68). By the triangle inequality for the Frobenius norm,

$$\begin{aligned}
\mathbb{E}\Big[\| \sum_j \mathbb{E}_j \boldsymbol{U}^\top \boldsymbol{Z} \boldsymbol{e}_j \boldsymbol{e}_j^\top \hat{\boldsymbol{L}} \|_{\mathrm{F}} \Big] &\le \sum_j \mathbb{E}\Big[\|\mathbb{E}_j \boldsymbol{U}^\top \boldsymbol{Z} \boldsymbol{e}_j\|_2 \|\boldsymbol{e}_j^\top \hat{\boldsymbol{L}}\|_2 \Big] \\
&\le \mathbb{E}[\sum_j \|\mathbb{E}_j \boldsymbol{U}^\top \boldsymbol{Z} \boldsymbol{e}_j\|_2^2]^{\frac{1}{2}} \mathbb{E}[\sum_j \|\boldsymbol{e}_j^\top \hat{\boldsymbol{L}}\|_2^2]^{\frac{1}{2}} \\
&\le (p\mathbb{E}[\|\boldsymbol{U}\|_{\mathrm{F}}^2])^{\frac{1}{2}} \mathbb{E}[\|\hat{\boldsymbol{L}}\|_{\mathrm{F}}^2]^{\frac{1}{2}},
\end{aligned}$$

where we used that $\|\boldsymbol{a}\boldsymbol{b}^\top\|_{\mathrm{F}} = \|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2$ for two vectors $\boldsymbol{a}, \boldsymbol{b}$, the Cauchy-Schwarz inequality, $\mathbb{E}[\|\boldsymbol{A}\boldsymbol{z}_j\|_2^2 | \boldsymbol{A}] = \|\boldsymbol{A}\|_{\mathrm{F}}^2$ if matrix $\boldsymbol{A}$ is independent of $\boldsymbol{z}_j \sim \mathcal{N}(0, I_n)$ (set $\boldsymbol{z}_j = \boldsymbol{Z}\boldsymbol{e}_j$), and Jensen's inequality.

Next, we decompose (65) as $\sum_j \boldsymbol{U}^\top (\boldsymbol{V} - \mathbb{E}_j \boldsymbol{V}) + \sum_j (\boldsymbol{U} - \mathbb{E}_j \boldsymbol{U})^\top \mathbb{E}_j \boldsymbol{V}$. We have by the submultiplicativity of the Frobenius norm and the Cauchy-Schwarz inequality

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{U}^\top \sum_j (\boldsymbol{V} - \mathbb{E}_j \boldsymbol{V})\|_{\mathrm{F}}] &\le \mathbb{E}[\sum_j \|\boldsymbol{U}\|_{\mathrm{F}} \|\boldsymbol{V} - \mathbb{E}_j \boldsymbol{V}\|_{\mathrm{F}}] \\
&\le \mathbb{E}[p\|\boldsymbol{U}\|_{\mathrm{F}}^2]^{\frac{1}{2}} \mathbb{E}[\sum_j \|\boldsymbol{V} - \mathbb{E}_j \boldsymbol{V}\|_{\mathrm{F}}^2]^{\frac{1}{2}}.
\end{aligned}$$

By the Gaussian Poincaré inequality applied p times, $\mathbb{E}[\sum_j \|\boldsymbol{V} - \mathbb{E}_j \boldsymbol{V}\|_{\mathrm{F}}^2] \le \|\boldsymbol{V}\|_\partial^2$, so that the previous display is bounded from above by $\sqrt{p}\|\boldsymbol{V}\|_\partial$. Similarly, $\mathbb{E}[\|\sum_j (\mathbb{E}_j[\boldsymbol{U}] - \boldsymbol{U})^\top \mathbb{E}_j \boldsymbol{V}\|_{\mathrm{F}}] \le \sqrt{p}\|\boldsymbol{U}\|_\partial$ and $\mathbb{E}[\|(65)\|_{\mathrm{F}}] \le \sqrt{p}(\|\boldsymbol{U}\|_\partial + \|\boldsymbol{V}\|_\partial)$.

For the last remaining term, (66), we first use $\mathbb{E}[\|(66)\|_{\mathrm{F}}] \le \mathbb{E}[\|(66)\|_{\mathrm{F}}^2]^{\frac{1}{2}}$ by Jensen's inequality and now proceed to bound $\|(66)\|_{\mathrm{F}}^2$. We have

$$\|(66)\|_{\mathrm{F}}^2 = \| \sum_j \mathbb{E}_j \boldsymbol{U}^\top \mathbb{E}_j \boldsymbol{V} - \mathbb{E}_j \boldsymbol{U}^\top \boldsymbol{X} \boldsymbol{e}_j \boldsymbol{e}_j^\top \boldsymbol{Z}^\top \mathbb{E}_j \boldsymbol{V} \|_{\mathrm{F}}^2 = \sum_{j,k} \mathrm{Tr}[\boldsymbol{M}_j^\top \boldsymbol{M}_k],$$

where $\boldsymbol{M}_j = \mathbb{E}_j \boldsymbol{U}^\top \mathbb{E}_j \boldsymbol{V} - \mathbb{E}_j \boldsymbol{U}^\top \boldsymbol{Z} \boldsymbol{e}_j \boldsymbol{e}_j^\top \boldsymbol{Z}^\top \mathbb{E}_j \boldsymbol{V}$. We first bound $\sum_j \|\boldsymbol{M}_j\|_{\mathrm{F}}^2$. Since the variance of $\boldsymbol{a}^\top \boldsymbol{b} - \boldsymbol{a}^\top \boldsymbol{g}\boldsymbol{g}^\top \boldsymbol{b}$ for standard normal $\boldsymbol{g} \sim \mathcal{N}(0, I_p)$ is $2\|(\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top)/2\|_{\mathrm{F}}^2 \le 2\|\boldsymbol{a}\|_2^2 \|\boldsymbol{b}\|_2^2$, applying this variance bound on each pair of coordinates $(t, t') \in [T] \times [T]$ gives $\sum_j \|\boldsymbol{M}_j\|_{\mathrm{F}}^2 \le \sum_j 2\|\mathbb{E}_j[\boldsymbol{U}]\|_{\mathrm{F}}^2 \|\mathbb{E}_j \boldsymbol{V}\|_{\mathrm{F}}^2 \le 2p$.

We now bound $\sum_{j \neq k} \mathrm{Tr}[\boldsymbol{M}_j^\top \boldsymbol{M}_k]$. Setting $\boldsymbol{z}_j = \boldsymbol{Z}\boldsymbol{e}_j \sim \mathcal{N}(0, I_n)$ for every $j \in [p]$, we will use many times the identity

$$\mathbb{E}[(\boldsymbol{z}_j^\top f(\boldsymbol{Z}) - \sum_i \partial_{ij} f(\boldsymbol{Z})^\top \boldsymbol{e}_i) g(\boldsymbol{Z}) = \mathbb{E}[\sum_i f(\boldsymbol{Z})^\top \boldsymbol{e}_i \partial_{ij} g(\boldsymbol{Z})] \tag{69}$$

which follows from Stein's formula for $f : \mathbb{R}^{n \times p} \to \mathbb{R}^n$ and $g : \mathbb{R}^{n \times p} \to \mathbb{R}$. With $f^{tt'}(\boldsymbol{Z}) = (\boldsymbol{z}_j^\top \mathbb{E}_j[\boldsymbol{U}] \boldsymbol{e}_{t'}) \mathbb{E}_j \boldsymbol{V} e_t$ and $g^{tt'}(\boldsymbol{Z}) = \boldsymbol{e}_{t'}^\top \boldsymbol{M}_k e_t$, we find

$$\begin{aligned}
&\mathbb{E}\,\mathrm{Tr}[\boldsymbol{M}_j^\top \boldsymbol{M}_k] = \mathbb{E}\,\mathrm{Tr}[\boldsymbol{M}_j^\top \sum_t \boldsymbol{e}_{t'} \boldsymbol{e}_{t'}^\top \boldsymbol{M}_k] = \mathbb{E}[\sum_{tt'} e_t^\top \boldsymbol{M}_j^\top \boldsymbol{e}_{t'} \boldsymbol{e}_{t'}^\top \boldsymbol{M}_k e_t] \\
&= \mathbb{E} \sum_{tt'} \Big(\boldsymbol{z}_j^\top f^{tt'}(\boldsymbol{Z}) - \sum_i \boldsymbol{e}_i^\top \partial_{ij} f^{tt'}(\boldsymbol{Z}) \Big) g^{tt'}(\boldsymbol{Z}) \\
&= \mathbb{E} \sum_{tt'} \sum_i \boldsymbol{e}_i^\top f^{tt'}(\boldsymbol{Z}) \partial_{ij} g^{tt'}(\boldsymbol{Z}).
\end{aligned}$$

where $g_{tt'}(\boldsymbol{Z}) = (e_t^\top \mathbb{E}_k \boldsymbol{V}^\top \mathbb{E}_k \boldsymbol{V} e_t' - e_t^\top \mathbb{E}_k \boldsymbol{U}^\top \boldsymbol{z}_k \boldsymbol{z}_k^\top \mathbb{E}_k \boldsymbol{V} e_t')$ and

$$\partial_{ij} g_{tt'} = \partial_{ij} \boldsymbol{e}_{t'}^\top \boldsymbol{M}_k e_t = \boldsymbol{e}_{t'}^\top \partial_{ij} [\mathbb{E}_k \boldsymbol{U}^\top \mathbb{E}_k \boldsymbol{V}] e_t - \boldsymbol{z}_k^\top \partial_{ij} [\mathbb{E}_k \boldsymbol{U} \boldsymbol{e}_{t'} \boldsymbol{e}_t^\top \mathbb{E}_k \boldsymbol{U}^\top] \boldsymbol{z}_k.$$

Now define $\tilde{f}^{tt'}(\boldsymbol{Z}) = \partial_{ij} [\mathbb{E}_k \boldsymbol{U} \boldsymbol{e}_{t'} \boldsymbol{e}_t^\top \mathbb{E}_k \boldsymbol{U}^\top] \boldsymbol{z}_k$ and $\tilde{g}^{tt'}(\boldsymbol{Z}) = \sum_i \boldsymbol{e}_i^\top f^{tt'}(\boldsymbol{Z})$. By definition of $\tilde{f}^{tt'}(\boldsymbol{Z})$, the previous display is equal to $\boldsymbol{z}_k^\top \tilde{f}^{tt'}(\boldsymbol{Z}) - \sum_l \partial_{lk} \boldsymbol{e}_l^\top \tilde{f}^{tt'}(\boldsymbol{Z})$. We apply (69) again with respect to $\boldsymbol{z}_k$, so that

$$\begin{aligned}
\mathbb{E}\,\mathrm{Tr}[\boldsymbol{M}_j^\top \boldsymbol{M}_k] &= \sum_{il,tt'} \boldsymbol{e}_i^\top \partial_{lk} [f^{tt'}(\boldsymbol{Z})] \boldsymbol{e}_l^\top \tilde{f}^{tt'}(\boldsymbol{Z}) \\
&= \sum_{il,tt'} \Big(\boldsymbol{e}_i^\top \partial_{lk} \Big[\mathbb{E}_j \boldsymbol{V} e_t \boldsymbol{e}_{t'}^\top \mathbb{E}_j \boldsymbol{U}^\top \Big] \boldsymbol{z}_j \Big) \Big(\boldsymbol{e}_l^\top \partial_{ij} \Big[\mathbb{E}_k[\boldsymbol{U}] \boldsymbol{e}_{t'} e_t^\top \mathbb{E}_k[\boldsymbol{V}]^\top \Big] \boldsymbol{z}_k \Big).
\end{aligned}$$

To remove the indices t, t', we rewrite the above using $\sum_t \boldsymbol{e}_t \boldsymbol{e}_t^\top = I_T$ and $\sum_{t'} \boldsymbol{e}_{t'} \boldsymbol{e}_{t'}^\top = I_T$ so that it equals

$$\mathbb{E} \sum_{il} \mathrm{Tr} \Big\{ \partial_{lk} \Big[\mathbb{E}_j \boldsymbol{U}^\top \boldsymbol{z}_j \boldsymbol{e}_i^\top \mathbb{E}_j \boldsymbol{V} \Big] \partial_{ij} \Big[\mathbb{E}_k[\boldsymbol{V}]^\top \boldsymbol{z}_k \boldsymbol{e}_l^\top \mathbb{E}_k[\boldsymbol{U}] \Big] \Big\}.$$

Summing over j, k, using $\mathrm{Tr}[\boldsymbol{A}^\top \boldsymbol{B}] \le \|\boldsymbol{A}\|_{\mathrm{F}} \|\boldsymbol{B}\|_{\mathrm{F}}$) and the Cauchy-Schwarz inequality, the above is bounded from above by

$$\Big\{ \mathbb{E} \sum_{jk,il} \Big\| \partial_{lk} \Big[\mathbb{E}_j \boldsymbol{U}^\top \boldsymbol{z}_j \boldsymbol{e}_i^\top \mathbb{E}_j[\boldsymbol{V}] \Big] \Big\|_{\mathrm{F}}^2 \Big\}^{\frac{1}{2}} \Big\{ \mathbb{E} \sum_{jk,il} \Big\| \partial_{ij} \Big[\mathbb{E}_k[\boldsymbol{V}]^\top \boldsymbol{z}_k \boldsymbol{e}_l^\top \mathbb{E}_k[\boldsymbol{U}] \Big] \Big\|_{\mathrm{F}}^2 \Big\}^{\frac{1}{2}}.$$

At this point the two factors are symmetric, with $(\boldsymbol{V}, \boldsymbol{U})$ in the left factor replaced with $(\boldsymbol{U}, \boldsymbol{V})$ on the right factor. We focus on the left factor; similar bound will apply to the right one by exchanging the

roles of $\boldsymbol{V}$ and $\boldsymbol{U}$. If $\boldsymbol{z}_j$ is independent of matrices $A^{(q)}$ $\mathbb{E}_j[\|\sum_{q=1}^n (\boldsymbol{e}_q^\top \boldsymbol{z}_j) A^{(q)}\|_{\mathrm{F}}^2 = \sum_{q=1}^n \|A^{(q)}\|_{\mathrm{F}}^2$ so that with $A^{(q)} = \partial_{lk}[\mathbb{E}_j \boldsymbol{U}^\top \boldsymbol{e}_q \boldsymbol{e}_i^\top \mathbb{E}_j \boldsymbol{U}]$, the first factor in the above display is equal to

$$
\begin{aligned}
&\Big\{\mathbb{E}\sum_{jk,ilq}\Big\|\partial_{lk}\Big(\mathbb{E}_j\boldsymbol{U}^\top\boldsymbol{e}_q\boldsymbol{e}_i^\top\mathbb{E}_j\boldsymbol{V}\Big)\Big\|_{\mathrm{F}}^2\Big\}^{\frac{1}{2}}\\
&\overset{\text{(i)}}{=}\Big\{\mathbb{E}\sum_{jk,ilq}\Big\|\partial_{lk}\Big(\mathbb{E}_j\boldsymbol{U}^\top\Big)\boldsymbol{e}_q\boldsymbol{e}_i^\top\mathbb{E}_j[\boldsymbol{V}]+\mathbb{E}_j[\boldsymbol{U}]^\top\boldsymbol{e}_q\boldsymbol{e}_i^\top\partial_{lk}\Big(\mathbb{E}_j[\boldsymbol{V}]\Big)\Big\|_{\mathrm{F}}^2\Big\}^{\frac{1}{2}}\\
&\overset{\text{(ii)}}{\le}\Big\{\mathbb{E}\sum_{jk,ilq}\Big\|\partial_{lk}\Big(\mathbb{E}_j[\boldsymbol{U}^\top]\Big)\boldsymbol{e}_q\boldsymbol{e}_i^\top\mathbb{E}_j[\boldsymbol{V}]\Big\|_{\mathrm{F}}^2\Big\}^{\frac{1}{2}}+\Big\{\mathbb{E}\sum_{jk,ilq}\Big\|\mathbb{E}_j\boldsymbol{U}^\top\boldsymbol{e}_q\boldsymbol{e}_i^\top\partial_{lk}\Big(\mathbb{E}_j[\boldsymbol{V}]\Big)\Big\|_{\mathrm{F}}^2\Big\}^{\frac{1}{2}}\\
&\overset{\text{(iii)}}{=}\Big\{\mathbb{E}\sum_{jk,ilq}\Big\|\mathbb{E}_j[\partial_{lk}\boldsymbol{U}]^\top\boldsymbol{e}_q\Big\|_2^2\Big\|\boldsymbol{e}_i^\top\mathbb{E}_j\boldsymbol{V}\Big\|_2^2\Big\}^{\frac{1}{2}}+\Big\{\mathbb{E}\sum_{jk,ilq}\Big\|\mathbb{E}_j\boldsymbol{U}^\top\boldsymbol{e}_q\Big\|_2^2\Big\|\boldsymbol{e}_i^\top\mathbb{E}_j[\partial_{lk}\boldsymbol{V}]\Big\|_2^2\Big\}^{\frac{1}{2}}\\
&\overset{\text{(iv)}}{=}\Big\{\mathbb{E}\sum_{jk,l}\Big\|\mathbb{E}_j[\partial_{lk}\boldsymbol{U}]^\top\Big\|_{\mathrm{F}}^2\Big\|\mathbb{E}_j\boldsymbol{V}\Big\|_{\mathrm{F}}^2\Big\}^{\frac{1}{2}}+\Big\{\mathbb{E}\sum_{jk,l}\Big\|\mathbb{E}_j\boldsymbol{U}^\top\Big\|_{\mathrm{F}}^2\Big\|\mathbb{E}_j[\partial_{lk}\boldsymbol{V}]\Big\|_{\mathrm{F}}^2\Big\}^{\frac{1}{2}},
\end{aligned}
$$

where (i) is the chain rule, (ii) the triangle inequality, (iii) holds provided that the order of the derivation ∂_{lk} and the expectation sign $\mathbb{E}_j$ can be switched and using $\|\boldsymbol{a}\boldsymbol{b}^\top\|_{\mathrm{F}}^2 = \|\boldsymbol{a}\|_2^2\|\boldsymbol{b}\|_2^2$ for vectors $\boldsymbol{a}, \boldsymbol{b}$, and (iv) holds using $\sum_i \|\boldsymbol{A}\boldsymbol{e}_i\|_2^2 = \|\boldsymbol{A}\|_{\mathrm{F}}^2 = \sum_q \|\boldsymbol{A}\boldsymbol{e}_q\|_2^2$ for a matrix $\boldsymbol{A}$ with n columns. Finally, by Jensen's inequality, the above display is bounded by

$$
\Big\{\mathbb{E}\sum_{k,l}\Big\|\partial_{lk}\boldsymbol{U}\Big\|_{\mathrm{F}}^2\sum_j\Big\|\mathbb{E}_j[\boldsymbol{V}]\Big\|_{\mathrm{F}}^2\Big\}^{\frac{1}{2}}+\Big\{\mathbb{E}\sum_{k,l}\Big\|\partial_{lk}\boldsymbol{V}\Big\|_{\mathrm{F}}^2\sum_j\Big\|\mathbb{E}_j[\boldsymbol{U}]\Big\|_{\mathrm{F}}^2\Big\}^{\frac{1}{2}}.
$$

Since $\|\boldsymbol{U}\|_{\mathrm{F}} \vee \|\boldsymbol{V}\|_{\mathrm{F}} \le 1$ almost surely, the previous display is bounded by $\sqrt{p}(\|\boldsymbol{U}\|_\partial + \|\boldsymbol{V}\|_\partial)$. In summary,

$$
\mathbb{E}[\|(66)\|_{\mathrm{F}}^2]^{\frac{1}{2}} \le (2p + [2\sqrt{p}(\|\boldsymbol{U}\|_\partial + \|\boldsymbol{V}\|_\partial)]^2)^{\frac{1}{2}} \le \sqrt{2p} + 2\sqrt{p}(\|\boldsymbol{U}\|_\partial + \|\boldsymbol{V}\|_\partial).
$$

Combining the bounds on the terms (65)-(66)-(67)-(68) with the triangle inequality completes the proof. ■

Proof of Lemma E.13. Since $\boldsymbol{M}^\dagger \preceq \boldsymbol{M}_1^\dagger = \boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + \tau n \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger$,

$$
\begin{aligned}
\|\boldsymbol{N}\|_{\mathrm{op}} &= \|(\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{X}^\top)\|_{\mathrm{op}}\\
&\le \|(\boldsymbol{I}_T \otimes \boldsymbol{X})(\boldsymbol{I}_T \otimes (\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + \tau n \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger)(\boldsymbol{I}_T \otimes \boldsymbol{X}^\top)\|_{\mathrm{op}}\\
&= \|\boldsymbol{X}_{\hat{\mathcal{S}}}(\boldsymbol{X}_{\hat{\mathcal{S}}}^\top \boldsymbol{X}_{\hat{\mathcal{S}}} + \tau n \boldsymbol{P}_{\hat{\mathcal{S}}})^\dagger \boldsymbol{X}_{\hat{\mathcal{S}}}^\top\|_{\mathrm{op}}\\
&\le 1,
\end{aligned}
$$

where the first inequality uses $\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^\top\|_{\mathrm{op}} \le \|\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top\|_{\mathrm{op}}$ for $0 \preceq \boldsymbol{B} \preceq \boldsymbol{C}$. ■

Proof of Lemma E.14. By Lemma E.4, $\frac{\partial F_{lt}}{\partial z_{ij}} = D_{ij}^{lt} + \Delta_{ij}^{lt}$, where

$$
D_{ij}^{lt} = -(\boldsymbol{e}_j^\top \boldsymbol{H} \otimes \boldsymbol{e}_i^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{e}_t \otimes \boldsymbol{e}_l), \tag{70}
$$

$$\Delta_{ij}^{lt} = -(\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\big(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p\big)(\boldsymbol{e}_i \otimes \boldsymbol{e}_j). \tag{71}$$

For the first equality, since $\boldsymbol{e}_i^\top \frac{\partial \boldsymbol{F}}{\partial z_{ij}} = \sum_t (D_{ij}^{it} + \Delta_{ij}^{it})\boldsymbol{e}_t^\top$, we have

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{p} \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j \boldsymbol{e}_i^\top \frac{\partial \boldsymbol{F}}{\partial z_{ij}} \\
=& \sum_{ij} \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j \sum_{t=1}^{T} (D_{ij}^{it} + \Delta_{ij}^{it})\boldsymbol{e}_t^\top \\
=& \sum_{ij} \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j \sum_{t} D_{ij}^{it}\boldsymbol{e}_t^\top + \underbrace{\sum_{ij} \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j \sum_{t} \Delta_{ij}^{it}\boldsymbol{e}_t^\top}_{\boldsymbol{J}_1},
\end{aligned}$$

where the first term can be simplified as below

$$\begin{aligned}
&\sum_{j=1}^{p}\sum_{i=1}^{n} \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j \sum_{t=1}^{T} D_{ij}^{it}\boldsymbol{e}_t^\top \\
=& -\sum_{j=1}^{p}\sum_{i=1}^{n} \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j \sum_{t=1}^{T} (\boldsymbol{e}_j^\top \boldsymbol{H} \otimes \boldsymbol{e}_i^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{e}_t \otimes \boldsymbol{e}_i)\boldsymbol{e}_t^\top \\
=& -\boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{H}\Big[\sum_i (\boldsymbol{I}_T \otimes \boldsymbol{e}_i^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{I}_T \otimes \boldsymbol{e}_i)\Big] \\
=& -\boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{H}(n\boldsymbol{I}_T - \widehat{\mathbf{A}}).
\end{aligned}$$

For the second equality, since $\frac{\partial \boldsymbol{F}}{\partial z_{ij}} = \sum_{lt} \boldsymbol{e}_l (D_{ij}^{lt} + \Delta_{ij}^{lt})\boldsymbol{e}_t^\top$,

$$\sum_{ij}\Big(\frac{\partial \boldsymbol{F}}{\partial z_{ij}}\Big)^\top \boldsymbol{Z}\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{F} = \underbrace{\sum_{ijtl} \boldsymbol{e}_t D_{ij}^{lt} \boldsymbol{e}_l^\top \boldsymbol{Z}\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{F}}_{\boldsymbol{J}_2} + \sum_{ijtl} \boldsymbol{e}_t \Delta_{ij}^{lt} \boldsymbol{e}_l^\top \boldsymbol{Z}\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{F},$$

where the second term can be simplified as below,

$$\begin{aligned}
&\sum_{ijtl} \boldsymbol{e}_t \Delta_{ij}^{lt} \boldsymbol{e}_l^\top \boldsymbol{Z}\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{F} \\
=& \sum_{ijtl} \boldsymbol{e}_t \Delta_{ij}^{lt} (\boldsymbol{e}_i \otimes \boldsymbol{e}_j)^\top (\boldsymbol{F} \otimes \boldsymbol{Z}^\top \boldsymbol{e}_l) \\
=& -\sum_{ijtl} \boldsymbol{e}_t (\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})\big(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p\big)(\boldsymbol{e}_i \otimes \boldsymbol{e}_j)(\boldsymbol{e}_i \otimes \boldsymbol{e}_j)^\top (\boldsymbol{F} \otimes \boldsymbol{Z}^\top \boldsymbol{e}_l) \\
=& -\sum_{l} (\boldsymbol{I}_T \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}})(\boldsymbol{F}^\top \otimes \boldsymbol{I}_p)(\boldsymbol{F} \otimes \boldsymbol{Z}^\top \boldsymbol{e}_l)
\end{aligned}$$

$$
\begin{aligned}
&= -\sum_l (\boldsymbol{I}_T \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger \big(\boldsymbol{F}^\top \boldsymbol{F} \otimes \boldsymbol{X}^\top \boldsymbol{e}_l\big) \\
&= -\sum_l (\boldsymbol{I}_T \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_T \otimes \boldsymbol{X})\boldsymbol{M}^\dagger \big(\boldsymbol{I}_T \otimes \boldsymbol{X}^\top \boldsymbol{e}_l\big)\boldsymbol{F}^\top \boldsymbol{F} \\
&= -\widehat{\mathbf{A}}\boldsymbol{F}^\top \boldsymbol{F},
\end{aligned}
$$

where the last line uses the expression of $\widehat{\mathbf{A}}$ in (10).

It remains to bound the norm of $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$. To bound $\|\boldsymbol{J}_2\|_{\mathrm{F}}$, recall the definition of $\boldsymbol{J}_2$,

$$
\begin{aligned}
\boldsymbol{J}_2 &= \sum_{ijlt} \boldsymbol{e}_t D_{ij}^{lt} \boldsymbol{e}_l^\top \boldsymbol{Z}\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{F} \\
&= -\sum_{ijlt} \boldsymbol{e}_t (\boldsymbol{e}_j^\top \boldsymbol{H} \otimes \boldsymbol{e}_i^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{e}_t \otimes \boldsymbol{e}_l)\boldsymbol{e}_l^\top \boldsymbol{Z}\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{F} \\
&= -\sum_{ijlt} \boldsymbol{e}_t (\boldsymbol{e}_t^\top \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{H}^\top \boldsymbol{e}_j \otimes \boldsymbol{e}_i)\boldsymbol{e}_l^\top \boldsymbol{Z}\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{F} \\
&= -\sum_{ijl} (\boldsymbol{I}_T \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{H}^\top \boldsymbol{e}_j \otimes \boldsymbol{e}_i)\boldsymbol{e}_l^\top \boldsymbol{Z}\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{F} \\
&= -\sum_{l} (\boldsymbol{I}_T \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{H}^\top \boldsymbol{Z}^\top \boldsymbol{e}_l \otimes \boldsymbol{F}).
\end{aligned}
$$

Since $\boldsymbol{N}$ is non-negative definite with $\|\boldsymbol{N}\|_{\mathrm{op}} \le 1$, $\|\boldsymbol{I}_{nT} - \boldsymbol{N}\|_{\mathrm{op}} \le 1$,

$$
\begin{aligned}
\|\boldsymbol{J}_2\|_{\mathrm{F}} &\le \sum_l \big\|(\boldsymbol{I}_T \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})(\boldsymbol{H}^\top \boldsymbol{Z}^\top \boldsymbol{e}_l \otimes \boldsymbol{F})\big\|_{\mathrm{F}} \\
&\le \sum_l \big\|(\boldsymbol{I}_T \otimes \boldsymbol{e}_l^\top)(\boldsymbol{I}_{nT} - \boldsymbol{N})\big\|_{\mathrm{op}} \big\|(\boldsymbol{H}^\top \boldsymbol{Z}^\top \boldsymbol{e}_l \otimes \boldsymbol{F})\big\|_{\mathrm{F}} \\
&\le \sum_l \big\|(\boldsymbol{H}^\top \boldsymbol{Z}^\top \boldsymbol{e}_l \otimes \boldsymbol{F})\big\|_{\mathrm{F}} \\
&= \sum_l \big\|\boldsymbol{H}^\top \boldsymbol{Z}^\top \boldsymbol{e}_l\big\|_{\mathrm{F}} \|\boldsymbol{F}\|_{\mathrm{F}} \\
&\le n^{\frac{1}{2}} \|\boldsymbol{Z}\boldsymbol{H}\|_{\mathrm{F}} \|\boldsymbol{F}\|_{\mathrm{F}} \\
&\le n^{\frac{1}{2}} \|\boldsymbol{Z}\|_{\mathrm{op}} \|\boldsymbol{H}\|_{\mathrm{F}} \|\boldsymbol{F}\|_{\mathrm{F}}.
\end{aligned}
$$

To bound $\|\boldsymbol{J}_1\|_{\mathrm{F}}$, recall the definition of $\boldsymbol{J}_1$,

$$
\boldsymbol{J}_1 = \sum_{ij} \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j \sum_t \Delta_{ij}^{it} \boldsymbol{e}_t^\top.
$$

For each $t, t' \in [T]$,

$$
\boldsymbol{e}_{t'}^\top \boldsymbol{J}_1 \boldsymbol{e}_t = \boldsymbol{e}_{t'}^\top \Big[\sum_{ij} \boldsymbol{F}^\top \boldsymbol{Z}\boldsymbol{e}_j \sum_t \Delta_{ij}^{it} \boldsymbol{e}_t^\top \Big] \boldsymbol{e}_t
$$

$$
\begin{aligned}
&= \sum_{j=1}^{p} \sum_{i=1}^{n} \boldsymbol{e}_{t'}^{\top} \boldsymbol{F}^{\top} \boldsymbol{Z} \boldsymbol{e}_j \Delta_{ij}^{it} \\
&= -\sum_{i} (\boldsymbol{e}_i^{\top} \boldsymbol{F} \otimes \boldsymbol{e}_{t'}^{\top} \boldsymbol{F}^{\top}) \boldsymbol{N} (\boldsymbol{e}_t \otimes \boldsymbol{e}_i) \\
&= -\boldsymbol{e}_{t'}^{\top} \sum_{i} (\boldsymbol{e}_i^{\top} \boldsymbol{F} \otimes \boldsymbol{F}^{\top}) \boldsymbol{N} (\boldsymbol{I}_T \otimes \boldsymbol{e}_i) \boldsymbol{e}_t .
\end{aligned}
$$

Thus, $\boldsymbol{J}_1 = \sum_i (\boldsymbol{e}_i^{\top} \boldsymbol{F} \otimes \boldsymbol{F}^{\top}) \boldsymbol{N} (\boldsymbol{I}_T \otimes \boldsymbol{e}_i)$, and hence

$$
\begin{aligned}
\|\boldsymbol{J}_1\|_{\mathrm{F}} &= \|\sum_{i} (\boldsymbol{e}_i^{\top} \boldsymbol{F} \otimes \boldsymbol{F}^{\top}) \boldsymbol{N} (\boldsymbol{I}_T \otimes \boldsymbol{e}_i)\|_{\mathrm{F}} \\
&\leq \sum_{i} \|(\boldsymbol{e}_i^{\top} \boldsymbol{F} \otimes \boldsymbol{F}^{\top}) \boldsymbol{N} (\boldsymbol{I}_T \otimes \boldsymbol{e}_i)\|_{\mathrm{F}} \\
&\leq \sum_{i} \|(\boldsymbol{e}_i^{\top} \boldsymbol{F} \otimes \boldsymbol{F}^{\top})\|_{\mathrm{F}} \|\boldsymbol{N} (\boldsymbol{I}_T \otimes \boldsymbol{e}_i)\|_{\mathrm{op}} \\
&\leq \sum_{i} \|\boldsymbol{e}_i^{\top} \boldsymbol{F}\| \|\boldsymbol{F}\|_{\mathrm{F}} \\
&\leq n^{\frac{1}{2}} \|\boldsymbol{F}\|_{\mathrm{F}}^2,
\end{aligned}
$$

where the first inequality is by sub-additivity of Frobenius norm, the second inequality uses $\|\boldsymbol{A}_1 \boldsymbol{A}_2\|_{\mathrm{F}} \leq \|\boldsymbol{A}_1\|_{\mathrm{F}} \|\boldsymbol{A}_2\|_{\mathrm{op}}$ for any matrices $\boldsymbol{A}_1, \boldsymbol{A}_2$ with appropriate dimensions, the third inequality is by $\|\boldsymbol{N}\|_{\mathrm{op}} \leq 1$ from Lemma E.13, the last inequality is by Cauchy-Schwarz inequality. ■