Journal of Machine Learning Research 27 (2026) 1-84 Submitted 1/25; Revised 7/26; Published 8/26

Canonical Correlation Analysis as Reduced Rank Regression in High Dimensions

Claire Donnat CDONNAT@UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago, IL, 60637, USA

Elena Tuzhilina ELENA.TUZHILINA@UTORONTO.CA
Department of Statistical Sciences
University of Toronto
Toronto, ON, M5S1A1, Canada

Editor: Pradeep Ravikumar

Abstract

Canonical correlation analysis is a widespread technique for discovering linear relationships between two sets of variables. In high dimensions, however, standard estimates of the canonical directions cease to be consistent without assuming further structure. In this setting, a possible solution consists in leveraging the presumed sparsity of the solution: only a subset of the covariates span the canonical directions. While the last decade has seen a proliferation of sparse canonical correlation analysis methods, practical challenges regarding the scalability and adaptability of these methods still persist. To circumvent these issues, this paper suggests an alternative strategy that uses reduced rank regression to estimate the canonical directions when one of the datasets is high-dimensional while the other remains low-dimensional. By casting the problem of estimating the canonical direction as a regression problem, our estimator is able to leverage the rich statistics literature on high-dimensional regression and is easily adaptable to accommodate a wider range of structural priors. Our proposed solution maintains computational efficiency and accuracy, even in the presence of very high-dimensional data. We demonstrate the advantages of our approach through a series of simulated experiments, achieving a substantial increase in accuracy compared to existing methods. Additionally, we highlight its practical utility by applying it to three real-world datasets, where we show that our method is able to uncover scientifically meaningful associations between variables.

Keywords: canonical correlation analysis; structured sparsity; high-dimensional data; reduced rank regression;

1. Introduction

Consider two random vectors x in $\mathbb{R}^p$ and y in $\mathbb{R}^q$ with respective covariance matrices $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$, and cross-covariance matrix $\mathbf{\Sigma}_{XY}$. The objective of *canonical correlation analysis* (CCA) is to identify a set of $r \leq \min(p, q)$ linear combinations of x and y that are maximally correlated. Mathematically, for $i = 1, \ldots, r$, the i^{th} solution pair (u_i, v_i) of CCA can be

expressed as the maximizer of the following constrained optimization problem:

$$\begin{aligned} &\underset{u\in\mathbb{R}^p,\ v\in\mathbb{R}^q}{\text{maximize}} \quad u^\top \mathbf{\Sigma}_{XY} v \\ &\text{subject to} \quad u^\top \mathbf{\Sigma}_X u = v^\top \mathbf{\Sigma}_Y v = 1, \quad u^\top \mathbf{\Sigma}_X u_j = v^\top \mathbf{\Sigma}_Y v_j = 0 \text{ for } j < i. \end{aligned} \tag{1}$$

The pairs (xu_i, yv_i) are called the *canonical variates*. The individual vectors u_i and v_i are called the *canonical directions*, and the corresponding correlation values $\lambda_i = u_i^\top \mathbf{\Sigma}_{XY} v_i$ are called the *canonical correlations*.

Since its introduction by Hotelling (1936), canonical correlation analysis has become an indispensable tool for analyzing relationships between sets of measurements. Given its versatility and interpretability, it is unsurprising that CCA has found applications in a wide spectrum of domains, ranging from psychology and social sciences (Thorndike, 2000; Fan and Konold, 2018) to genomics (Witten and Tibshirani, 2009; Parkhomenko et al., 2009; Lin et al., 2013; Gossmann et al., 2018) or neuroscience (Zhuang et al., 2020).

Under the canonical pair model (Chen et al., 2013; Gao et al., 2017), it can be shown that the CCA model corresponds to a reparameterization of the cross covariance $\mathbf{\Sigma}_{XY}$ as

$$\mathbf{\Sigma}_{XY} = \mathbf{\Sigma}_X \mathbf{U} \mathbf{\Lambda} \mathbf{V}^\top \mathbf{\Sigma}_Y, \quad \text{where} \quad \mathbf{U}^\top \mathbf{\Sigma}_X \mathbf{U} = \mathbf{V}^\top \mathbf{\Sigma}_Y \mathbf{V} = \mathbf{I}_r. \tag{2}$$

Here $\mathbf{U} = [u_1|u_2|\dots|u_r] \in \mathbb{R}^{p\times r}$ and $\mathbf{V} = [v_1|v_2|\dots|v_r] \in \mathbb{R}^{q\times r}$ represent the r canonical directions, and $\mathbf{\Lambda} \in \mathbb{R}^{r\times r}$ is a diagonal matrix with canonical correlations λ_i on the diagonal ranked by decreasing order of magnitude. It is a classical exercise to show that the canonical directions (u_i, v_i) can be obtained through a transformation of the right and left singular vectors of $\mathbf{\Sigma}_X^{-1/2} \mathbf{\Sigma}_{XY} \mathbf{\Sigma}_Y^{-1/2}$. Specifically, given the singular value decomposition (SVD) $\mathbf{\Sigma}_X^{-1/2} \mathbf{\Sigma}_{XY} \mathbf{\Sigma}_Y^{-1/2} = \mathbf{U}_0 \mathbf{\Lambda} \mathbf{V}_0^\top$, the canonical directions u_i and v_i are simply the i^{th} columns of the matrices $\mathbf{U} = \mathbf{\Sigma}_X^{-1/2} \mathbf{U}_0$ and $\mathbf{V} = \mathbf{\Sigma}_Y^{-1/2} \mathbf{V}_0$ respectively. In practice, the canonical directions are typically obtained by replacing the population covariance matrices by their sample estimators, i.e. $\widehat{\mathbf{\Sigma}}_X, \widehat{\mathbf{\Sigma}}_Y$ and $\widehat{\mathbf{\Sigma}}_{XY}$, respectively.

Recent studies have become increasingly interested in applying CCA to the analysis of high-dimensional datasets where at least one of the dimensions is significantly larger than the number of observations n. It is a well-established fact that the conventional CCA framework breaks down in this regime, both from a theoretical and computational perspective (Gao et al., 2017). From a computational standpoint, if, for instance, $p > n$, the sample covariance matrix $\widehat{\mathbf{\Sigma}}_X$ is no longer invertible, thereby preventing the straightforward calculation of the CCA directions using a singular value decomposition of the transformed cross-covariance matrix $\widehat{\mathbf{\Sigma}}_X^{-1/2} \widehat{\mathbf{\Sigma}}_{XY} \widehat{\mathbf{\Sigma}}_Y^{-1/2}$. From a theoretical point of view, the eigenvector estimates obtained from the sample covariance matrices $\widehat{\mathbf{\Sigma}}_X$ and $\widehat{\mathbf{\Sigma}}_{XY}$ are no longer guaranteed to be consistent (Gao et al., 2015, 2017), and replacing the sample estimator $\widehat{\mathbf{\Sigma}}_X^{-1/2}$ by its pseudoinverse may not lead to an accurate estimate of the canonical directions.

To overcome the challenge posed by the high-dimensionality of data, several adaptations of the original CCA problem have recently been developed. To keep our discussion concise, a comprehensive discussion of these methods is deferred to Section 6, which we summarize here. All of these variants incorporate some form of regularization or constraint on the estimators of $\mathbf{U}$ and $\mathbf{V}$, and can be classified into one of two categories depending on the type of regularization they employ: *ridge-regularized CCA*, which mitigates ill-conditioning

by adding ℓ_2 penalties, and *sparse CCA*, which incorporates ℓ_1 penalties to enforce sparsity, focusing on selecting only the most informative variables. While sparse CCA methods have proliferated (Witten and Tibshirani, 2009; Wilms and Croux, 2016; Parkhomenko et al., 2009; Waaijenborg and Zwinderman, 2009), many rely on heuristics, lack guarantees on the accuracy of the estimated directions, or make simplifying assumptions (for instance, that the covariances $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$ are diagonal to enhance computational speed, e.g. Witten and Tibshirani (2009)). In contrast, theoretical efforts (Gao et al., 2015, 2017; Gao and Ma, 2023) establish algorithms with provable guarantees and high-probability bounds on the error of the associated estimators of $\mathbf{U}$ and $\mathbf{V}$, along with the corresponding minimax rates. However, these methods rely on a prohibitively computationally intensive initialization step (hereafter referred to as the "Fantope initialization", detailed in Appendix B.1) that first estimates the joint product $\mathbf{U}\mathbf{V}^\top$. Because it scales cubically with the dataset dimensions (Gao and Ma, 2023), this initialization step effectively makes this entire class of methods unusable on most real-world datasets.

This highlights an important gap in current CCA methodologies. Existing heuristic methods are fast but tend to yield poor estimates of the canonical directions. Conversely, while theoretical approaches guarantee a more accurate estimation, their computational demands are so substantial that they become impractical even on moderately sized datasets. Moreover, the methods developed by Gao et al. (2015, 2017) and Gao and Ma (2023) all assume symmetry in the roles played by $\mathbf{X}$ and $\mathbf{Y}$, thus imposing sparsity in both. However, many contemporary applications of CCA exhibit substantial asymmetry between the two views. Examples include neuroimaging studies, where thousands of imaging-derived features are related to a relatively small number of cognitive, behavioral, or clinical measurements (Smith et al., 2015; Zhuang et al., 2020); climate applications, where high-dimensional spatial fields are associated with a small number of climate indices or forecasting targets (Barnett and Preisendorfer, 1987; Bretherton et al., 1992); and genomics studies, where high-dimensional molecular profiles are linked to a limited number of phenotypic or clinical outcomes (Chi et al., 2013; Jiang et al., 2023). Such dimensional imbalance is particularly prevalent in neuroscience applications, where modern imaging technologies routinely generate tens of thousands of features from a relatively modest number of subjects and phenotypic observations; see Table 4 in Appendix A.1 for representative examples. In these settings, regularization is primarily needed on the high-dimensional side, whereas imposing sparsity on the low-dimensional view may offer limited benefit while increasing both statistical and computational complexity. This observation suggests that the symmetric regularization strategies adopted by existing sparse CCA methods may not be optimal in many practical settings and motivates the development of methods that explicitly exploit this dimensional imbalance to improve scalability while retaining strong statistical performance.

Our approach adopts this perspective by imposing structural constraints on only one of the two canonical loading matrices. By regularizing only the high-dimensional view, we substantially reduce the computational burden associated with theoretically grounded sparse CCA methods while retaining statistical guarantees. At the same time, this asymmetric formulation achieves improved estimation accuracy relative to existing heuristic approaches. As a result, our method effectively leverages the imbalance between the two views to improve scalability without sacrificing statistical performance.

Contributions. In this paper, we study asymmetric high-dimensional CCA, where one

view is structured and high-dimensional while the other remains low-dimensional. Our main contributions are as follows:

1. **A scalable alternative to Fantope-based sparse CCA.** In Section 3, we develop a two-step estimation procedure inspired by existing theoretically grounded sparse CCA methods (Gao et al., 2015, 2017; Gao and Ma, 2023), while avoiding the computationally intensive Fantope initialization. Specifically, we first estimate the low-rank signal $\mathbf{U}\mathbf{\Lambda}\mathbf{V}_0^\top$ via a convex reduced-rank regression formulation and subsequently recover the canonical directions $\mathbf{U}$ and $\mathbf{V}$. The resulting method achieves substantially improved computational scalability while preserving strong statistical guarantees.

2. **Theoretical guarantees under asymmetric structural constraints.** In Section 4, we establish finite-sample error bounds for the estimated canonical subspaces under structural constraints imposed on only one view (i.e. one of the datasets). Our analysis shows that consistent estimation is possible in this asymmetric setting and provides theoretical guarantees for a broad class of structured regularization schemes.

3. **A flexible framework for structured canonical directions.** Our methodology accommodates a broad class of structural assumptions on the high-dimensional view, including sparsity, group sparsity, and graph-based smoothness penalties (Section 4.2). This allows flexible structure to be incorporated into the estimation procedure. For instance, features may be partitioned into predefined groups (thus motivating group-sparse loadings), or they may form a spatial or network graph (thereby motivating smoothness constraints that encourage neighboring features to have similar loadings). This flexibility can lead to improved and more interpretable canonical directions in applications where the high-dimensional features exhibit known structure. Through simulations and applications to neuroscience and ecological data (Section 5), we show that our proposed estimator recovers interpretable structured loadings, achieve competitive statistical performance compared to existing sparse CCA methods, and are substantially faster to compute.

2. Notations & Parameter Space

For any $t \in \mathbb{Z}_+$, $[t]$ denotes the set $\{1, 2, ..., t\}$. For any set S, $|S|$ denotes its cardinality and S^c its complement. For any $a, b \in \mathbb{R}$, $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For a vector u, $\|u\| = \sqrt{\sum_i u_i^2}$, $\|u\|_0 = \sum_i \mathbf{1}_{\{u_i \neq 0\}}$, and $\|u\|_1 = \sum_i |u_i|$. For any matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times k}$ (denoted by a capital letter and bold font), $\mathbf{A}_{i\cdot}$ denotes its i-th row and $\text{supp}(\mathbf{A}) = \{i \in [p] : \|\mathbf{A}_{i\cdot}\| > 0\}$, the index set of nonzero rows, is called its support. Throughout this paper, $\sigma_i(\mathbf{A})$ stands for the i^{th} largest singular value of the matrix $\mathbf{A}$ and $\sigma_{\max}(\mathbf{A}) = \sigma_1(\mathbf{A})$, $\sigma_{\min}(\mathbf{A}) = \sigma_{p\wedge \text{rank}(\mathbf{A})}(\mathbf{A})$. The Frobenius norm and the operator norm of $\mathbf{A}$ are $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ and $\|\mathbf{A}\|_{op} = \sigma_{\max}(\mathbf{A})$, respectively. The ℓ_{21} norm and the nuclear norm are $\|\mathbf{A}\|_{21} = \sum_{i\in[p]} \|\mathbf{A}_{i\cdot}\|_2$ and $\|\mathbf{A}\|_* = \sum_i \sigma_i(\mathbf{A})$, respectively. For any positive semi-definite matrix $\mathbf{A}$, $\mathbf{A}^{1/2}$ denotes its principal square root, which is positive semi-definite and satisfies $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$. The trace inner product of two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times k}$ is denoted

by $\langle \mathbf{A}, \mathbf{B} \rangle = \mathsf{Tr}(\mathbf{A}^\top \mathbf{B})$. The notation $\mathbf{A}^\dagger$ is used here to denote the pseudo-inverse of the matrix $\mathbf{A}$.

We consider here the canonical pair model of Gao et al. (2017), where the observed n pairs of measurement vectors $(x_i, y_i)_{i=1}^n$ are i.i.d. from a multivariate Gaussian distribution $\mathcal{N}_{p+q}(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_X & \mathbf{\Sigma}_{XY} \\ \mathbf{\Sigma}_{YX} & \mathbf{\Sigma}_Y \end{pmatrix}$ with the cross-covariance matrix $\mathbf{\Sigma}_{XY}$ satisfying (2). We are interested in the situation where the leading canonical coefficient vectors of $\mathbf{U}$ are sparse, without assuming any particular structure on $\mathbf{V}$. Sparsity is here understood in terms of the support of $\mathbf{U}$. More specifically, we define $\mathcal{F}(s_u, p, q, r, \lambda; M)$ to be the collection of all covariance matrices $\mathbf{\Sigma}$ satisfying (2) and

$$\begin{aligned} &\text{(a)} \quad \mathbf{U} \in \mathbb{R}^{p \times r} \text{ and } \mathbf{V} \in \mathbb{R}^{q \times r} \text{ with } |\mathrm{supp}(\mathbf{U})| \le s_u; \\ &\text{(b)} \quad \sigma_{\min}(\mathbf{\Sigma}_X) \wedge \sigma_{\min}(\mathbf{\Sigma}_Y) \ge \tfrac{1}{M} \text{ and } \sigma_{\max}(\mathbf{\Sigma}_X) \vee \sigma_{\max}(\mathbf{\Sigma}_Y) \le M; \\ &\text{(c)} \quad \lambda_r \ge \lambda \text{ and } \lambda_1 \le 1 - \tfrac{1}{M}. \end{aligned} \tag{3}$$

The probability space that we consider throughout this paper is thus

$$\big\{(X_i, Y_i) \overset{i.i.d.}{\sim} \mathcal{N}_{p+q}(0, \mathbf{\Sigma}) \quad \text{with } \mathbf{\Sigma} \in \mathcal{F}(s_u, p, q, r, \lambda; M)\big\}. \tag{4}$$

We shall allow s_u, p, λ to vary with n, while $M > 1$ is restricted to be an absolute constant. We will also assume that q and r are fixed. Note that this latter assumption is in contrast to the approach by Gao et al. (2017) that supposes that q and r can also vary with n.

3. CCA via reduced rank regression

In this section, we begin by highlighting a link between reduced rank regression (RRR) and CCA. As we will see in Section 4, this link will then allow us to formulate a theoretically sound and practical solution for canonical correlation analysis in high dimensions. We note here that, while this connection has already been recognized in the existing literature (Izenman, 1975; De la Torre, 2012), this formulation has not been previously used in the context of high-dimensional CCA.

Here we assume that $p, q \le n$ are fixed, as n is allowed to grow, so that $\widehat{\mathbf{\Sigma}}_X$ and $\widehat{\mathbf{\Sigma}}_Y$ are invertible and consistent estimators of $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$, respectively. We first consider the CCA formulation as a prediction problem, as noted by Gao et al. (2017):

$$\underset{\mathbf{U} \in \mathbb{R}^{p \times r},\ \mathbf{V} \in \mathbb{R}^{q \times r}}{\text{minimize}} \ \|\mathbf{Y}\mathbf{V} - \mathbf{X}\mathbf{U}\|_F^2 \quad \text{subject to} \quad \mathbf{U}^\top \widehat{\mathbf{\Sigma}}_X \mathbf{U} = \mathbf{V}^\top \widehat{\mathbf{\Sigma}}_Y \mathbf{V} = \mathbf{I}_r. \tag{5}$$

Lemma 1 *Letting $\mathbf{Y}_0 = \mathbf{Y}\widehat{\mathbf{\Sigma}}_Y^{-1/2}$, the estimators for the first r canonical directions can be recovered as $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}} = \widehat{\mathbf{\Sigma}}_Y^{-1/2}\widehat{\mathbf{V}}_0$ from the solution of the following problem:*

$$\widehat{\mathbf{U}}, \widehat{\mathbf{V}}_0 = \underset{\substack{\mathbf{U} \in \mathbb{R}^{p \times r},\ \mathbf{V}_0 \in \mathbb{R}^{q \times r} \\ \mathbf{U}^\top \widehat{\mathbf{\Sigma}}_X \mathbf{U} = \mathbf{V}_0^\top \mathbf{V}_0 = \mathbf{I}_r}}{\mathrm{argmin}} \ \|\mathbf{Y}_0 - \mathbf{X}\mathbf{U}\mathbf{V}_0^\top\|_F^2 \tag{6}$$

The proof of this statement can be found in Appendix C.1. Using this objective function, it is evident that the recovery of the matrix $\mathbf{U}\mathbf{V}_0^\top$ in (6) is a constrained instance of the more general reduced-rank regression problem:

$$\underset{\mathbf{B}_0 \in \mathbb{R}^{p\times q}}{\text{minimize}} \quad \|\mathbf{Y}_0 - \mathbf{X}\mathbf{B}_0\|_F^2 \quad \text{subject to} \quad \mathsf{rank}(\mathbf{B}_0) = r, \tag{7}$$

where $\mathbf{B}_0$ must have the additional constraint that ${\mathbf{B}_0}^\top \widehat{\mathbf{\Sigma}}_X \mathbf{B}_0$ and $\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{B}_0{\mathbf{B}_0}^\top\widehat{\mathbf{\Sigma}}_X^{1/2}$ are projection matrices. The following theorem allows us to tie the solution of the reduced rank regression problem (7) to the solution of the CCA problem (5).

Theorem 2 (Low-dimensional CCAR^3 subspace consistency) *Let* $\mathbf{Y}_0 = \mathbf{Y}\widehat{\mathbf{\Sigma}}_Y^{-1/2}$, *where* $\widehat{\mathbf{\Sigma}}_Y$ *is the sample covariance matrix of* $\mathbf{Y}$. *Assume that the cross-covariance matrix follows the canonical pair model*

$$\mathbf{\Sigma}_{XY} = \mathbf{\Sigma}_X \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}_Y, \qquad \mathbf{U}^\top\mathbf{\Sigma}_X\mathbf{U} = \mathbf{V}^\top\mathbf{\Sigma}_Y\mathbf{V} = \mathbf{I}_r,$$

with $\lambda_r > 0$. *Let*

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}\in\mathbb{R}^{p\times q}}{\operatorname{argmin}} \|\mathbf{Y}_0 - \mathbf{X}\mathbf{B}\|_F^2 = \widehat{\mathbf{\Sigma}}_X^{-1}\widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_Y^{-1/2}$$

and let $\operatorname{svd}_r(\widehat{\mathbf{B}}) = \widehat{\mathbf{U}}_{\hat{B}}\widehat{\mathbf{\Lambda}}_{\hat{B}}\widehat{\mathbf{V}}_{\hat{B}}^\top$ *denote the rank-r truncated singular value decomposition of* $\widehat{\mathbf{B}}$, *where* $\widehat{\mathbf{U}}_{\hat{B}}$ *and* $\widehat{\mathbf{V}}_{\hat{B}}$ *contain the leading r left and right singular vectors of* $\widehat{\mathbf{B}}$. *Define*

$$\widehat{\mathbf{U}} = \widehat{\mathbf{U}}_{\hat{B}}\left(\widehat{\mathbf{U}}_{\hat{B}}^\top\widehat{\mathbf{\Sigma}}_X\widehat{\mathbf{U}}_{\hat{B}}\right)^{-1/2}, \qquad \widehat{\mathbf{V}} = \widehat{\mathbf{\Sigma}}_Y^{-1/2}\widehat{\mathbf{V}}_{\hat{B}}.$$

Then $\widehat{\mathbf{U}}$ *and* $\widehat{\mathbf{V}}$ *are consistent estimators of the canonical subspaces spanned by* $\mathbf{U}$ *and* $\mathbf{V}$, *respectively:*

$$\min_{\mathbf{O}\in\mathcal{O}_r} \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{O}\|_F \xrightarrow{a.s.} 0, \qquad \min_{\widetilde{\mathbf{O}}\in\mathcal{O}_r} \|\widehat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}}\|_F \xrightarrow{a.s.} 0.$$

The proof of this theorem is provided in Appendix C.2.

Lemma 1 and Theorem 2 imply that canonical subspaces can be estimated via a simple two-step procedure. First, compute the ordinary least squares (OLS) estimator $\widehat{\mathbf{B}}$. Second, obtain the leading r left and right singular vectors of $\widehat{\mathbf{B}}$ and normalize them appropriately to recover estimates of the canonical directions. Under the conditions of Theorem 2, these estimators consistently recover the population canonical subspaces. The full procedure, which we will refer to as CCAR^3, is outlined in Algorithm 1.

The proposed two-step approach differs from traditional CCA solutions in low dimensions. Initially, given the original data matrix $\mathbf{X}$ and the normalized matrix $\mathbf{Y}_0 = \mathbf{Y}\widehat{\mathbf{\Sigma}}_Y^{-1/2}$, we identify the joint subspace $\widehat{\mathbf{B}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{V}}_0^\top$. This is done by casting the estimation of $\mathbf{B}$ as a convex optimization problem, much in the same spirit as Gao et al. (2017) and Gao and Ma (2023). At the same time, this optimization problem is considerably easier than the initialization procedure of Gao et al. (2017). We then use the solution $\widehat{\mathbf{B}}$ to further estimate the canonical directions. Compared to alternative approaches, this method is therefore less prone to initialization issues.

As dimensions increase, particularly when p increases and exceeds n, this approach becomes infeasible. The inverse of the matrix $\widehat{\mathbf{\Sigma}}_X$ ceases to be well-defined, rendering the

Algorithm 1 CCAR[3]

Input: $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{Y} \in \mathbb{R}^{n\times q}$ with $p, q \leq n$; number of components $r \leq \min(p, q)$.

1: Normalize $\mathbf{Y}_0 = \mathbf{Y}\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}}$.
2: Compute the OLS solution $\widehat{\mathbf{B}} = \operatorname{argmin}_{\mathbf{B}\in\mathbb{R}^{p\times q}} \|\mathbf{Y}_0 - \mathbf{X}\mathbf{B}\|_F^2$.
3: Compute the singular value decomposition $\operatorname{svd}_r(\widehat{\mathbf{B}}) = \widehat{\mathbf{U}}_{\widehat{B}}\widehat{\mathbf{\Lambda}}_{\widehat{B}}\widehat{\mathbf{V}}_{\widehat{B}}^\top$.
4: Normalize directions $\widehat{\mathbf{U}} = \widehat{\mathbf{U}}_{\widehat{B}}(\widehat{\mathbf{U}}_{\widehat{B}}^\top\widehat{\mathbf{\Sigma}}_X\widehat{\mathbf{U}}_{\widehat{B}})^{-1/2}$ and $\widehat{\mathbf{V}} = \widehat{\mathbf{\Sigma}}_Y^{-1/2}\widehat{\mathbf{V}}_{\widehat{B}}$.

Output: CCA directions $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$.

OLS solution to problem (5) is non-unique and inconsistent. It is important to remember that in our discussion, q is considered low-dimensional, so calculating $\widehat{\mathbf{\Sigma}}_Y$ remains manageable and consistent for the estimation of $\mathbf{\Sigma}_Y$. Unlike $\mathbf{X}$, computing $\mathbf{Y}_0 = \mathbf{Y}\widehat{\mathbf{\Sigma}}_Y^{-1/2}$ is still well-defined. In the following sections, we will show, nonetheless, that this formulation of the CCA problem is easily extendable to the high-dimensional case.

4. CCA in high dimensions

As emphasized in the previous section, in high dimensions (more specifically, when p is on the order of or exceeds n), our proposed CCA estimator breaks down as the solution of the OLS problem is not necessarily unique. Imposing some low-dimensional structure on the estimators thus becomes indispensable to reliably estimate the CCA directions. A number of such constraints can be stated in a general form as $\operatorname{Pen}(\mathbf{U}) \leq s$, where Pen is a penalty function on the matrix $\mathbf{U}$, thus transforming the original CCA objective into the following constrained problem:

$$\underset{\mathbf{U}\in\mathbb{R}^{p\times r},\mathbf{V}\in\mathbb{R}^{q\times r}}{\text{maximize}} \ \mathsf{Tr}(\mathbf{U}^\top\widehat{\mathbf{\Sigma}}_{XY}\mathbf{V}) \text{ subject to } \mathbf{U}^\top\widehat{\mathbf{\Sigma}}_X\mathbf{U} = \mathbf{V}^\top\widehat{\mathbf{\Sigma}}_Y\mathbf{V} = \mathbf{I}_r \text{ and } \operatorname{Pen}(\mathbf{U}) \leq s. \quad (8)$$

Note that in the traditional CCA model, the directions are obtained up to rotations of the columns. Consequently, the penalty used here is always a row-penalty. Below we specialize our analysis to the different types of constraints that can be imposed on $\mathbf{U}$: sparsity, group-sparsity, and graph-sparsity. Importantly, we will highlight how our formulation of the CCA problem enables the flexible adoption of a wide variety of constraints.

4.1 CCA with sparsity

As outlined in the introduction, one strategy for estimating the canonical directions in a high-dimensional context consists in imposing sparsity. The usual hypothesis consists of assuming that $\mathbf{U}$ is row sparse, indicating that only a limited number of covariates of $\mathbf{X}$ are involved in determining the canonical directions. Mathematically, this concept is often expressed as $|\operatorname{supp}(\mathbf{U})| \leq s$, which constrains the number of nonzero rows in $\mathbf{U}$ to be smaller than s. This immediately implies that $\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}_0^\top$ also exhibits row sparsity: $|\operatorname{supp}(\mathbf{B})| \leq s$.

To allow consistent estimation of $\mathbf{U}$ and $\mathbf{V}$, we propose replacing the OLS step in Algorithm 1 by

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}\in\mathbb{R}^{p\times q}}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{Y}_0 - \mathbf{X}\mathbf{B}\|_F^2 + \rho\|\mathbf{B}\|_{21}. \tag{9}$$

Note here that (9) is a convex relaxation of the OLS problem with sparsity constraint $|\operatorname{supp}(\mathbf{B})| \leq s$. This ensures the tractability of the objective and returns a solution that is row-sparse, i.e., for which only a subset I of the rows of $\widehat{\mathbf{B}}$ are nonzero. We can then resume the algorithm proposed in the previous section. The complete procedure is described in Algorithm 2. In practice, to efficiently solve the problem (9) we propose a simple implementation relying on the Alternating Direction Method of Multipliers (ADMM; Boyd et al., 2011), which we describe explicitly in Appendix B.2.

Algorithm 2 Sparse CCAR[3] in high dimensions

Input: $\mathbf{X}\in\mathbb{R}^{n\times p}$, $\mathbf{Y}\in\mathbb{R}^{n\times q}$ with $q\leq n$ and $p\gg n$; number of components $r\leq\min(p,q)$.

1: Normalize $\mathbf{Y}_0 = \mathbf{Y}\widehat{\mathbf{\Sigma}}_Y^{-1/2}$.

2: Solve the penalized least squares using ADMM:

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}\in\mathbb{R}^{p\times q}}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{Y}_0 - \mathbf{X}\mathbf{B}\|_F^2 + \rho\|\mathbf{B}\|_{21}.$$

3: Compute the singular value decomposition $\operatorname{svd}_r(\widehat{\mathbf{B}}) = \widehat{\mathbf{U}}_{\widehat{B}}\widehat{\mathbf{\Lambda}}_{\widehat{B}}\widehat{\mathbf{V}}_{\widehat{B}}^\top$.

4: Normalize directions $\widehat{\mathbf{U}} = \widehat{\mathbf{U}}_{\widehat{B}}(\widehat{\mathbf{U}}_{\widehat{B}}^\top\widehat{\mathbf{\Sigma}}_X\widehat{\mathbf{U}}_{\widehat{B}})^{-1/2}$ and $\widehat{\mathbf{V}} = \widehat{\mathbf{\Sigma}}_Y^{-1/2}\widehat{\mathbf{V}}_{\widehat{B}}$.

Output: CCA directions $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$.

It is possible to show that the addition of the ℓ_{21}-penalty ensures that the learned $\widehat{\mathbf{B}}$ is not too far from its population version $\mathbf{B}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}_Y^{1/2}$ (Theorem 3). This allows in turn to ensure that the estimates $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ are good approximations of $\mathbf{U}$ and $\mathbf{V}$, respectively (Theorem 4). To keep this manuscript concise, all proofs are deferred to Appendix D.

Theorem 3 *Consider the family $\mathcal{F}(s_u,p,q,r,\lambda,M)$ of covariance matrices satisfying assumptions (3). Assume $(q+s_u\log(ep/s_u))/n\leq c_0$ for some sufficiently small constant $c_0\in(0,1)$. We write $\mathbf{\Delta} = \widehat{\mathbf{B}}-\mathbf{B}^*$ with $\mathbf{B}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}_Y^{1/2}$, and choose $\rho\geq C_uM^2\sqrt{(q+\log p)/n}$ for some large constant C_u. Then the solution $\widehat{\mathbf{B}}$ of problem (9) is such that, for any $C'>0$, there exist constants C_1 and C_2 that solely depend on C' and c_0 such that:*

$$\left\|\widehat{\mathbf{\Sigma}}_X^{\frac{1}{2}}\mathbf{\Delta}\right\|_F^2 \leq C_1M^5s_u\frac{q+\log p}{n} \quad \text{and} \quad \|\widehat{\mathbf{B}}-\mathbf{B}^*\|_F^2 \lesssim M^6s_u(q+\log p)/n \tag{10}$$

with probability at least $1-\exp(-C'q)-\exp(-C's_u\log(ep/s_u))-\exp(-C'(q+\log p))-\exp(-C'(r+\log p))$. Moreover, with the same probability, the size of the support of $\widehat{\mathbf{B}}$ is of the order of s_u, that is $|\operatorname{supp}(\widehat{\mathbf{B}})|\lesssim M^2s_u$.

Theorem 3 implies that, under certain regularity conditions for the matrix $\mathbf{\Sigma}_X$ (more specifically, that it is well conditioned and $\sigma_{\max}(\mathbf{\Sigma}_X)\leq M$), we know that the support of the solution $\widehat{\mathbf{B}}$ is roughly of the order of the actual support of $\mathbf{B}^*$. This means in particular

that the number of discoveries — and in particular, false discoveries, where a false discovery is understood as a nonzero row in $\widehat{\mathbf{B}}$ that is not in the support of $\mathbf{B}^*$ — is of constant order and does not grow with n.

To put this bound into perspective, we compare it with the results of Gao et al. (2017) (Theorem 4.1), who show that the error in the joint estimation of the subspace $\mathbf{U}\mathbf{V}^\top$ from their Fantope initialization procedure is $\|\widehat{\mathbf{A}} - \mathbf{U}\mathbf{V}^\top\|_F^2 \leq C s_u s_v \log(p+q)/n$, where s_v is the number of non-zero rows of $\mathbf{V}$. The error bound of our procedure for estimating the matrix $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}_Y^{1/2}$ is therefore of the same order, since we do not assume any sparsity structure on $\mathbf{V}$ (i.e. $s_v = q$). Moreover, since we do not have to identify the rows of $\mathbf{V}$ relevant to the canonical directions, we do not have to pay the price of the additional logarithmic factor of q that the setting considered by Gao et al. (2017) incurs. Theorem 4 provides the analysis of the error bounds on both $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ from Algorithm 2.

Theorem 4 *Assume the conditions of Theorem 3. Then, for any $C' > 0$, the procedure outlined in Algorithm 2 yields normalized bases $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ such that there exist orthogonal matrices $\mathbf{O}, \widetilde{\mathbf{O}} \in \mathbb{R}^{r\times r}$ with*

$$\left\|\widehat{\mathbf{U}} - \mathbf{U}\widetilde{\mathbf{O}}\right\|_F \leq C\frac{M^6}{\lambda}\sqrt{r s_u \frac{(q+\log p)}{n}} \quad \text{and} \quad \left\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{O}\right\|_F \leq C\frac{M^{\frac{11}{2}}}{\lambda}\sqrt{r s_u \frac{(q+\log p)}{n}} \tag{11}$$

with probability at least $1 - \exp(-C'q) - \exp(-C' s_u \log(ep/s_u)) - \exp(-C'(q+\log p)) - \exp(-C'(r+\log p))$. Here C is a constant that depends solely on C' and c_0.

We note that if the optional $r \times r$ rotation in Algorithm 2 is applied and the canonical correlations are separated, ordered-direction consistency holds up to diagonal sign changes, with the displayed rate multiplied by a constant depending on the inverse canonical-correlation eigengap.

Observation 1: comparison with existing results. These results have to be compared to the theoretical results obtained by Gao et al. (2017), where the authors derive the following bound for the error of the second step of their procedure:

$$\min_{\mathbf{O}\in\mathcal{P}} \left\|\mathbf{\Sigma}_X^{1/2}(\widehat{\mathbf{U}} - \mathbf{U}\mathbf{O})\right\|_F^2 \leq C\frac{s_u(r+\log p)}{\lambda^2 n}.$$

One of the remarkable properties of this bound is that it decouples the dimensions of $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$: the bound on $\widehat{\mathbf{U}}$ does not depend on q. In our approach, by contrast, the estimation is performed jointly, and the estimation errors of $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ are therefore coupled. However, contrary to the method proposed by Gao et al. (2017), our approach does not require splitting the data into 3 folds, and is able to operate on the entire dataset at once — a situation, that, in practice, can be more efficient from an estimation and computational standpoint (see details in Appendix B.1). Our bound is closer to that for the joint estimation procedure of Gao and Ma (2023), which is shown to be upper bounded as

$$\min_{\mathbf{O}\in\mathcal{P}} \left\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{O}\right\|_F \vee \min_{\widetilde{\mathbf{O}}\in\mathcal{P}} \left\|\widehat{\mathbf{U}} - \mathbf{U}\widetilde{\mathbf{O}}\right\|_F \leq C\left(\frac{s'}{s_u+q}\right)^{\frac{3}{2}} \frac{(3+\frac{2}{M^2})}{\lambda}\sqrt{\frac{r(s_u+q)\log(p+q)}{n}}.$$

Here C is a constant that depends on the spectrum of $\mathbf{\Sigma}_{XY}$, and s' is a parameter controlling the hard row-sparsity of their initial estimator of the matrix $\mathbf{A} = (\mathbf{U}^\top|\mathbf{V}^\top)$. This bound is

more efficient than ours for the estimation of $\mathbf{V}$, as ours scales with the product qs_u, rather than the sum $q + s_u$.

Observation 2: Computational efficiency. The method of Gao and Ma (2023) relies on a Fantope-based initialization, which, as discussed in the introduction, scales as $O\big((p+q)^3\big)$ and can incur substantial computation times. By contrast, the computational cost of our procedure when $n << p$ is $O(n^2q + npq)$, which is linear in the larger dimension (see Appendix B.2 for more details). While an ADMM formulation is used for both CCAR^3 and Fantope-initialization approaches, each ADMM step in our method involves only simple matrix multiplications and shrinkage operations. In comparison, each ADMM step for the Fantope projection requires solving a large Lasso problem and computing the SVD of a $(p+q)\times(p+q)$ matrix (see Appendix B.1). Consequently, our approach provides an efficient alternative that substantially reduces computation while retaining theoretical guarantees.

Simulation experiments. We generate a set of synthetic experiments that allow us to test the behavior of the different methods in the high-dimensional setting. To this end, we first construct a block diagonal covariance matrix as $\mathbf{\Sigma}_X = \begin{pmatrix} \mathbf{U}_X\mathbf{U}_X^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$, where $\mathbf{U}_X \in \mathbb{R}^{p_1 \times r_{pca}}$ is a randomly generated matrix from orthonormal columns, whose diagonal we then set to all ones. This creates a covariance structure for $\mathbf{X}$ that is not trivial, but such that the first r_{pca} principal components are sparse and are only spanned by p_1 out of p rows. In our experiments, we set $p_1 = 20$. We generate a covariance matrix for $\mathbf{Y}$ in a similar fashion as $\mathbf{\Sigma}_Y = \mathbf{U}_Y\mathbf{U}_Y^\top$. Next, we construct the cross-covariance matrix $\mathbf{\Sigma}_{XY}$ as $\mathbf{\Sigma}_{XY} = \mathbf{\Sigma}_X\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}_Y$. Here $\mathbf{U} \in \mathbb{R}^{p\times r},\ \mathbf{V} \in \mathbb{R}^{q\times r}$ are sampled uniformly at random and chosen to have support of size n_{nnz} with $n_{nnz} < p$ (row-sparse), and q respectively. Both $\mathbf{U}$ and $\mathbf{V}$ are then normalized so that $\mathbf{U}^\top\mathbf{\Sigma}_X\mathbf{U} = \mathbf{V}^\top\mathbf{\Sigma}_Y\mathbf{V} = \mathbf{I}_r$. Further, $\mathbf{\Lambda}$ is a diagonal matrix with r entries $(\lambda_i)_{i=1}^r$ equally spaced in the following intervals, depending on the desired signal strength: (0.75, 0.9) for "high", (0.55, 0.7) for "medium", and (0.35, 0.5) for a "low" signal strength. Finally, we generate the observations $(x_i, y_i)_{i=1}^n$ by sampling from the normal joint distribution $\mathcal{N}_{p+q}\left(0, \begin{pmatrix} \mathbf{\Sigma}_X & \mathbf{\Sigma}_{XY} \\ \mathbf{\Sigma}_{YX} & \mathbf{\Sigma}_Y \end{pmatrix}\right)$.

We use the synthetic data to compare the performance of our sparse CCAR^3 method with six alternative approaches of CCA with sparsity. These include a set of "heuristic-based" methods: SAR (Wilms and Croux, 2015), CCA-Lasso (Witten and Tibshirani, 2009), and the SCCA approach of Parkhomenko et al. (2009). We also compare our method with the "theory-grounded" approaches: Fantope CCA (the initialization step of the algorithm from Gao et al. (2017) and Gao and Ma (2023)), SCCA with CoLaR (the full algorithm from Gao et al. (2017)), and SGCA from Gao and Ma (2023). We compare all methods to an "oracle" solution, which is simply the estimated CCA directions (using a standard low-dimensional CCA solver) on the true support of $\mathbf{U}$. In our method, we select an optimal regularization parameter ρ through 5-fold cross validation. To robustify the estimates of the canonical variates, we further shrink $\widehat{\mathbf{\Sigma}}_Y$ using its Ledoit-Wolf estimate (Ledoit and Wolf, 2022). Similarly, we select the regularization parameters in all heuristic-based approaches using either 5-fold cross validation, or the permutation approach proposed by Witten and Tibshirani (2009) for their method ("permuted" in Figure 1). For the Fantope initialization of the methods in Gao et al. (2017) and Gao and Ma (2023), we used the theoretical values for ρ as proposed in the respective papers. However, in the second stage of the CCA with CoLaR, which involves solving a group-lasso problem, we applied cross-validation. To

evaluate the resulting CCA reconstructions, we measure the distances between subspaces spanning canonical variates, which can be done by computing the principal angle (Björck and Golub, 1973) between the subspaces associated with the matrices $\begin{pmatrix} \widehat{\mathbf{U}} \\ \widehat{\mathbf{V}} \end{pmatrix}$ and $\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}$.

We examine the dependence of the subspace distance between the estimated and true canonical subspaces on the number of covariates p, the number of canonical components r, and the sample size n (Figure 1). Results are averaged over 100 simulation replications.

The top panel of Figure 1 investigates the effect of increasing dimensionality p. Across the medium- and high-correlation settings, sparse CCAR^3 consistently achieves the smallest subspace distance among the competing methods. The advantage becomes increasingly pronounced as p grows, demonstrating the ability of sparse CCAR^3 to accurately recover the canonical subspace in high-dimensional regimes. The only method exhibiting comparable performance is SAR (Wilms and Croux, 2015), particularly when p is small and the problem remains effectively low-dimensional. However, the performance gap widens substantially as p increases. In the low-correlation setting, sparse CCAR^3 remains competitive and typically outperforms the remaining benchmarks, with only a few isolated cases in which Fantope CCA attains slightly lower subspace distance.

The middle panel of Figure 1 examines the effect of the number of canonical components r in a representative high-dimensional setting ($p = 1000$, $q = 30$). Sparse CCAR^3 exhibits remarkable stability in high and medium correlation settings as r increases, maintaining nearly constant subspace distance across all values of r. By contrast, most other baselines' performances deteriorate with increasing rank. If SAR achieves similar performance for small r, sparse CCAR^3 generally provides the most accurate recovery of the canonical subspace over the full range of ranks considered. In low-correlation signals, however, the performance is very similar across competitors.

The bottom-left panel studies the effect of sample size n. As expected, the subspace distance decreases for all methods as more observations become available. Sparse CCAR^3 displays one of the fastest convergence rates. SAR exhibits a similar rate of improvement, whereas the remaining methods improve more gradually and remain substantially farther from the CCAR^3 solution even for large sample sizes.

To further assess practical applicability, the bottom-right panel of Figure 1 compares computational costs. To provide an idea of the applicability of our method in real-life settings, we used parallelization across 5 workers to perform the cross-validation methods. We used the efficient code for CCA-Lasso (Witten and Tibshirani, 2009) in the `PMA` package. However, the SAR method by Wilms and Croux (2015) was not parallelized, as the code for this was not optimized by the authors.

We observe that sparse CCAR^3 is slower than the heuristic methods CCA-Lasso and SCCA when n is moderate. However, these methods are consistently less accurate than sparse CCAR^3 across all simulation settings considered above. Furthermore, while the computational cost of competing heuristic methods appears to increase exponentially with the sample size n, the running time of sparse CCAR^3 is largely insensitive to n and, in fact, decreases slightly for larger sample sizes as the ADMM algorithm converges more rapidly. Consequently, for large sample sizes, the computational cost of sparse CCAR^3 becomes comparable to CCA-Lasso and SCCA.

Among all competing methods, SAR is the closest competitor in terms of accuracy. Nevertheless, its computational cost is substantially higher. For example, when $n = 10{,}000$,

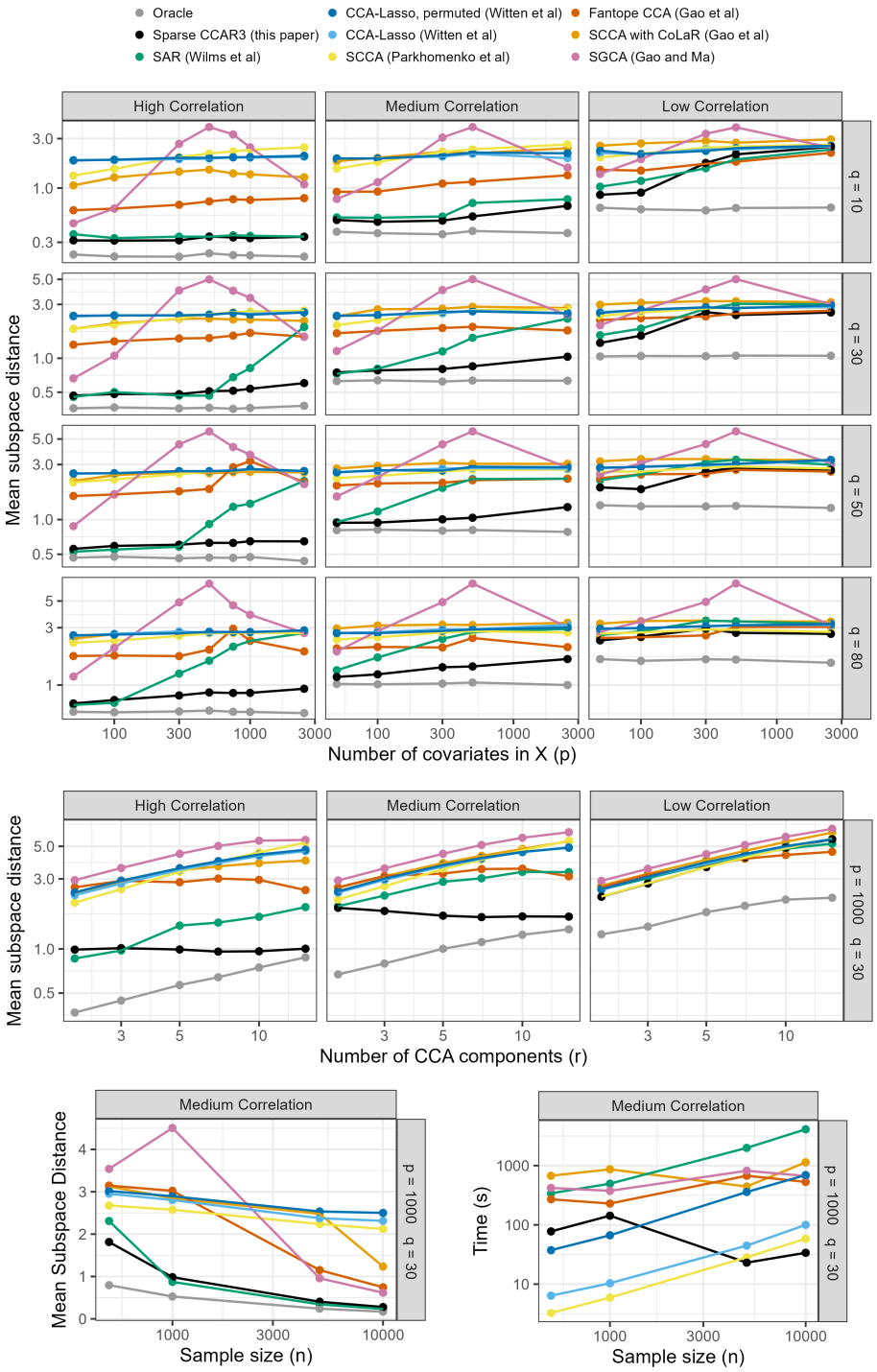


Figure 1: Simulation results for sparse model with $r_{pca} = 5$. Top panel: subspace distance as a function of p, with $n = 500$, $r = 3$ and $n_{nnz} = 6$. Middle panel: subspace distance as a function of the rank r, with $n = 500$, $p = 1,000$, $q = 30$ and $n_{nnz} = 20$. Bottom panel: subspace distance (left) and computation time (right) as a function of n, with $q = 30$, $p = 1000$, $r = 3$ and $n_{nnz} = 20$. Results are averaged over 100 simulations.

SAR requires approximately 4,000 seconds, whereas sparse CCAR^3 requires only 30 seconds. Even if SAR were fully parallelized, our method would remain approximately 25 times faster. We also note that fitting sparse CCAR^3 on a single cross-validation fold is considerably less expensive than solving the optimization problem underlying Fantope CCA: sparse CCAR^3 achieves roughly a 100-fold reduction in computational time relative to Fantope CCA, despite the fact that Fantope CCA does not require cross-validation and is evaluated using the theoretically optimal tuning parameter. These results demonstrate that sparse CCAR^3 effectively balances statistical and computational efficiency, combining state-of-the-art estimation accuracy with computational scalability.

4.2 CCA with structured penalties

The row-sparse formulation from Section 4.1 can be extended when prior structural information on the covariates of $\mathbf{X}$ is available. We consider the estimator

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}\in\mathbb{R}^{p\times q}}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{Y}_0 - \mathbf{X}\mathbf{B}\|_F^2 + \rho\ \mathrm{Pen}(\mathbf{B}). \tag{12}$$

Two choices of penalty are particularly useful.

Group penalty. Suppose that $\mathbf{X}$ covariates are partitioned into known non-overlapping groups

$$\mathcal{G} = \{g_1, \ldots, g_M\}, \qquad \bigcup_{m=1}^{M} g_m = [p], \qquad g_m \cap g_{m'} = \varnothing \text{ for } m \neq m'.$$

Writing $\mathbf{B}_g$ for the block of $\mathbf{B}$ corresponding to the rows in g, we take

$$\mathrm{Pen}(\mathbf{B}) = \sum_{g\in\mathcal{G}} \sqrt{|g|}\,\|\mathbf{B}_g\|_F.$$

This encourages the joint selection of entire groups of covariates. In neuroimaging, for example, the groups may correspond to regions defined by a reference atlas. The row-sparse penalty is recovered as the special case $\mathcal{G} = \{\{1\}, \ldots, \{p\}\}$.

Graph penalty. Suppose instead that the covariates lie on a known graph with vertex set $[p]$ and edge set E. Let $\mathbf{\Gamma} \in \mathbb{R}^{m\times p}$ be the associated edge-incidence matrix, where $m = |E|$. For an edge $(k, k') \in E$, the corresponding row of $\mathbf{\Gamma}$ has entries 1 and -1 in columns k and k', and zeros elsewhere. We then take

$$\mathrm{Pen}(\mathbf{B}) = \|\mathbf{\Gamma}\mathbf{B}\|_{21}. \tag{13}$$

This is an instance of the multivariate total-variation penalty (Hütter and Rigollet, 2016). It encourages neighboring covariates to have similar coefficients while still allowing a small number of jumps across edges, i.e. graph sparsity (Tran et al., 2022). This is again natural in neuroimaging, where nearby voxels are expected to exhibit similar activation patterns.

Both versions of (12) can be solved efficiently using the same ADMM framework as in the sparse case. The updates for the group and graph-penalized version are described in Appendix B.2.

Theoretical results. For the group penalty, let $m_u = \big|\{g \in \mathcal{G} : \|\mathbf{U}_g\|_F \neq 0\}\big|$ denote the number of active groups, and let $s_u = \sum_{g:\|\mathbf{U}_g\|_F \neq 0} |g|$ be their total size. The proof of Theorem 3 extends directly to this setting with the row-sparse complexity $s_u(q + \log p)$ replaced by the group complexity $s_u q + m_u \log |\mathcal{G}|$, up to constants and group-size factors when the groups are balanced. Consequently, the corresponding direction-subspace error scales as

$$\frac{\sqrt{r}}{\lambda}\sqrt{\frac{s_u q + m_u \log |\mathcal{G}|}{n}}$$

up to powers of the conditioning constant M.

For the graph penalty, the relevant structural assumption is that $\mathbf{\Gamma U}$ is row-sparse. Let $s_u = |\mathrm{supp}(\mathbf{\Gamma U})|$, let n_c be the number of connected components, let $\mathbf{L} = \mathbf{\Gamma}^\top \mathbf{\Gamma}$ be the graph Laplacian, and let κ_2 denote the smallest nonzero eigenvalue of $\mathbf{L}$. The graph direction bounds depend on the rate

$$\delta_G = \sqrt{n_c \frac{q + \log n_c}{n}} + \left(1 + \frac{1}{\kappa_2}\right)\sqrt{\sigma_{\max}(\mathbf{L}) s_u \frac{q + \log m}{n}}.$$

The dependence on n_c reflects the fact that $\|\mathbf{\Gamma U}\|_{21}$ does not penalize componentwise constants, so each connected component contributes one unpenalized degree of freedom. The dependence on the smallest nonzero Laplacian eigenvalue κ_2 reflects connectivity: smaller κ_2 corresponds to a more weakly connected graph (Chung, 1996), and therefore to weaker regularization for a fixed total-variation budget. The rigorous theorem statements, along with their proofs, are provided in Appendices E and F.

Simulation experiments. We adopt a data generation mechanism similar to that of the previous subsection, modifying only the generation of $\mathbf{U}$. Specifically, we generate $\mathbf{U}$ either according to a known group-sparse structure or according to a graph-sparse structure in $\mathbf{X}$. Additional details on the simulation setup are provided in Appendix A.2.

The results for the data with group and graph structure are presented in Figures 5 and 7 of the Appendix A.2. We observe that both modifications perform well in terms of the resulting subspace distances between the estimated canonical directions and their population counterpart. SAR is once again the best competitor, achieving better results compared to our methods in settings where p is small ($p \ll n$) and the signal is high. However, as soon as p increases, our methods outperform all competitors. These results are consistent across different values of q.

5. Real-World Experiments

In this section, we underscore the advantages of our approach by conducting a comparative analysis of our method against established methods on three real-world datasets.[1]

5.1 The Nutrimouse data

The Nutrimouse dataset is a common benchmark in multivariate analysis (Gonzalez et al., 2012; Rodosthenous et al., 2020) that contains genomic and nutritional data from a 2007

1. Our method is available on CRAN under the package ccar3. The code for the simulations and real-world examples can be found on https://github.com/donnate/CCAR3code.

study on 40 mice (Martin et al., 2007). This dataset includes the expression profiles of 120 pre-selected genes in the liver, as well as measurements of 21 hepatic fatty acid concentrations (expressed as percentages) observed under various dietary treatments. We denote the genomic and fatty acid measurements by $\mathbf{X}$ and $\mathbf{Y}$, respectively. Each of the 40 mice was assigned to one of five diets, and its genotype was classified as either wildtype or genetically modified (PPAR). We set the number of canonical directions to $r = 5$ upon examining the scree plot of the singular values of $\widehat{\mathbf{\Sigma}}_{XY}$, and compare our method's performance to the benchmarks introduced in the previous section.

To evaluate the performance of the methods, we divide the data into eight folds (5 mice per fold) at random, and use seven folds for training, and keep one for testing. For sparse CCAR^3, the optimal ρ is chosen using cross-validation on the training data, based on the mean squared error (MSE) between the canonical variates $\mathbf{X}\widehat{\mathbf{U}}$ and $\mathbf{Y}\widehat{\mathbf{V}}$ as per equation (5). For the other methods, we use their default implementations to determine the appropriate amount of regularization (e.g. using cross-validation, BIC or a permutation-based analysis, depending on the methods). Performance across methods is measured on the test folds using the MSE between $\mathbf{X}\widehat{\mathbf{U}}$ and $\mathbf{Y}\widehat{\mathbf{V}}$, and the average correlation between directions.

Results. The results, averaged over 10 random 8-fold splits of the data, are presented in Table 1. Sparse CCAR^3 achieves the lowest test MSE and the highest test correlation among all competing methods. In particular, it reduces the average test MSE by approximately 36% relative to SAR (Wilms and Croux, 2015), 38–41% relative to the CCA-Lasso methods of Witten and Tibshirani (2009), and 39% relative to SCCA with CoLaR (Gao and Ma, 2023). The improvement is also reflected in the substantially higher test correlation, with sparse CCAR^3 achieving a correlation of 0.62, compared to at most 0.53 for the competing methods.

Sparse CCAR^3 is also substantially more computationally efficient than existing theoretically grounded sparse CCA methods. For instance, fitting the full sparse CCAR^3 model requires only 35 seconds, including 8-fold cross-validation over 30 candidate values of ρ, compared to 777 seconds for Fantope CCA with no cross-validation. Sparse CCAR^3 thus achieves better predictive performance while requiring only a small fraction of the computational effort of existing theory-based approaches.

We further evaluate the quality of the estimated canonical directions by examining whether the resulting canonical variates separate the different diet groups and genotypes. The canonical variates obtained by SAR (Wilms and Croux, 2015), SCCA with CoLaR (Gao et al., 2017), and sparse CCAR^3 are displayed in Figure 2. All three methods successfully separate the two genotypes using the first two canonical directions (Figure 2, left panels). However, differences emerge when considering the dietary groups. While SAR identifies the COC and SUN diets as distinct clusters (Figure 2b), both SCCA with CoLaR and sparse CCAR^3 achieve substantially better separation among the dietary groups. In particular, sparse CCAR^3 provides the clearest separation of the COC cluster, which appears less distinct under SCCA with CoLaR (Figures 2a and 2c).

To quantify these observations, we classify samples into the five diet groups using the estimated canonical variates $(\mathbf{X}\widehat{u}_1, \ldots, \mathbf{X}\widehat{u}_5)$. Classification is performed using linear discriminant analysis (LDA), with 3/4 of the data used for training and 1/4 for testing, and the results averaged across the four folds. As reported in Table 1, sparse CCAR^3 achieves

	Test MSE	Test Correlation	Diet Label Accuracy (LDA)	Genotype Label / Cluster Accuracy (LDA / cluster)	Running Time (s)
sparse CCAR^3 (this paper)	**0.57 ± 0.10**	**0.62 ± 0.04**	**0.9**	1 / 0.575	34.63
SAR (Wilms et al.)	0.89 ± 0.11	0.53 ± 0.07	0.6	1 / 0.6	6.02
CCA-Lasso with CV (Witten et al.)	0.96 ± 0.12	0.41 ± 0.08	0.7	1 / 0.575	2.86
CCA-Lasso permuted (Witten et al.)	0.92 ± 0.09	0.43 ± 0.06	0.7	1 / 0.575	2.44
SCCA (Parkhomenko et al.)	0.86 ± 0.12	0.48 ± 0.08	0.8	1 / 0.625	0.23
Fantope CCA (Gao et al.)	2.37 ± 2.15	0.41 ± 0.07	0.7	1 / 0.5	777
SCCA with CoLaR (Gao et al.)	0.94 ± 0.14	0.49 ± 0.04	0.7	1 / 0.7	801
SGCA (Gao & Ma)	1.98 ± 1.18	0.38 ± 0.07	0.7	1 / 0.675	1,356

Table 1: Results of the 8-fold cross-validation on the Nutrimouse dataset. Reported values correspond to the average test MSE and correlation over 10 random splits of the data in 8 folds. The ability of CCA methods to yield directions that can separate genotypes is evaluated through both LDA, and, to allow more discrimination between methods, Gaussian clustering.

the highest classification accuracy, improving diet classification accuracy by 50% relative to SAR and by 28% relative to SCCA with CoLaR.

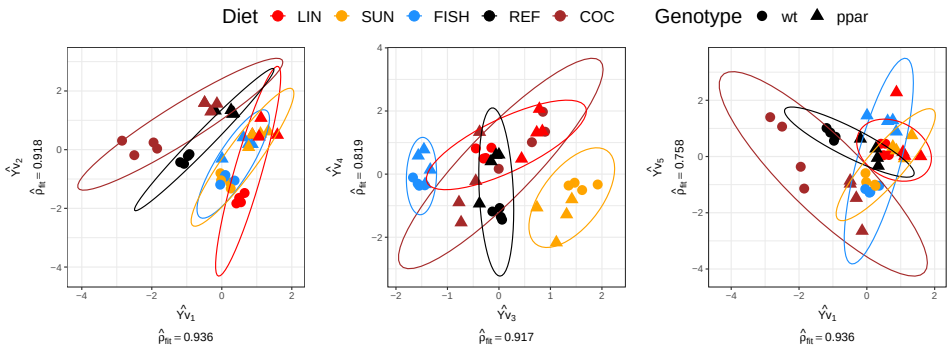
Turning to the loadings, we investigate whether the canonical directions recovered by sparse CCAR^3 are consistent with the biological findings of Martin et al. (2007) (see Figure 8 in Appendix A.3). For loadings obtained by competing methods, we refer the interested reader to the online supplement[2]

The first canonical variate separates mice fed the coconut-oil diet from those fed the fish- and linseed-oil diets (Figure 2a, left panel). The dominant loadings involve fatty acids C20:3n6, C20:3n9, C16:0, C16:1n9, and C18:2n6 together with genes *CAR1*, *CYP3A11*, *PLTP*, *AOX*, and *CYP2c29*. Notably, *AOX*, *CYP3A11*, *CYP2c29*, and *PLTP* were among the variables discussed by Martin et al. (2007) as contributing to differences across dietary interventions. The recovery of these variables suggests that the first canonical direction captures the primary diet-related variation in the data.

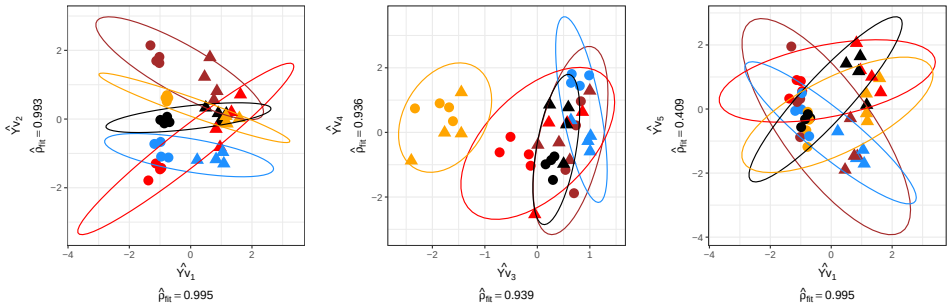
The second canonical variate separates wild-type and PPARα-deficient mice and is characterized by large loadings on long-chain polyunsaturated fatty acids, including C22:5n6, C22:5n3, and C20:5n3 (EPA), together with genes *ACAT2*, *Lpin*, *PLTP*, *CYP3A11*, and *CYP2c29*. Several of these variables, including *Lpin*, *PLTP*, *CYP3A11*, and *CYP2c29*, were highlighted by Martin et al. (2007) as being associated with PPARα activity. Likewise, EPA and related long-chain polyunsaturated fatty acids were among the lipid species exhibiting substantial variation across genotypes in the original study. The recovery of these variables suggests that the second canonical direction captures the primary genotype-related variation in the data.

The third canonical variate separates the reference diet from the fish- and sunflower-oil diets, indicating an additional dietary effect beyond the contrast represented by the first

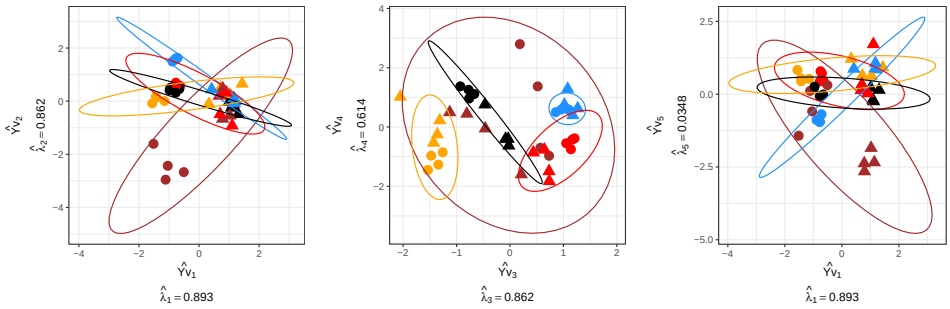
2. `https://github.com/donnate/CCAR3code/blob/main/additional_plots`.



(a) sparse CCAR[3] (our method)



(b) SAR (Wilms and Croux, 2015)



(c) SCCA with CoLaR (Gao et al., 2017)

Figure 2: Scatter plots for canonical variates produced for the Nutrimouse dataset: 1^{st} vs 2^{nd} (left panel), 3^{rd} vs 4^{th} (middle panel), and 1^{st} vs 5^{th} (right panel). Sample points are colored by diet, different shapes correspond to the genotype. The plots include 95% confidence ellipses to emphasize diet clustering.

canonical variate. The corresponding loadings involve DHA (C22:6n3), C22:4n6, C22:5n6, and C20:3n9 together with genes *CYP2c29*, *ACAT2*, *PLTP*, and *CYP3A11*. Notably, both DHA and the C22 polyunsaturated fatty acids featured prominently in the lipid profiles analyzed by Martin et al. (2007), while *PLTP*, *CYP3A11*, and *CYP2c29* were among the genes reported to vary across dietary and genotype conditions. The recovery of these variables suggests that the third canonical direction captures an additional dietary signal present in the data. Overall, the recovered canonical directions closely align with the two principal sources of variation identified by Martin et al. (2007): dietary fatty-acid composition and PPARα activity.

5.2 Neuroscience data: grouping brain regions with group-sparse CCAR^3

In this section, we re-investigate the dataset from Tozzi et al. (2021). The matrix $\mathbf{X}$ consists of the fMRI brain activations from 229 distinct brain regions of interest (ROI) for 153 individuals in response to monetary gains versus losses during a gambling task. These regions include the 210 cortical areas as identified by the Brainetome atlas and the 19 subcortical areas of the Desikan-Kellamy atlas. The matrix $\mathbf{Y}$ contains self-reported measures on nine variables: drive, fun seeking, reward responsiveness, and overall behavioral inhibition (evaluated via the Behavioral Approach System/Behavioral Inhibition Scale), as well as distress, anhedonia, and anxious arousal (as defined through the Mood and Anxiety Symptom Questionnaire), and positive and negative emotional states (derived from the Positive and Negative Affect Schedule). Both datasets $\mathbf{X}$ and $\mathbf{Y}$ were adjusted for sex differences and standardized across each column. We refer the reader to the original papers by Tozzi et al. (2021); Tuzhilina et al. (2021) for more details on the data collection and preprocessing of these datasets.

We propose a joint analysis of the brain activation and questionnaire datasets by calculating the first two canonical directions based on an initial assessment of the singular values from the cross-covariance matrix. We evaluate sparse and group-sparse penalty types for CCAR^3 discussed in Section 4. For the group-sparse method, we assign each of the 229 ROIs to 29 groups of brain regions (which we further split into two, depending on whether they are in the right or left hemisphere) corresponding to the main gyri or subcortical regions, as well as the brain stem. We employ the same cross-validation approach as described in the previous section (using 20 folds here). Table 2 summarizes the performance of our two CCAR^3 variants on the validation folds, along with that of existing benchmarks. Overall, we observe that group CCAR^3 outperforms all other benchmarks, attaining the lowest MSE and highest correlation. We also note that our group modification reduces by 8.4% the MSE of the sparse CCAR^3, the second best performing method.

In Figure 3, we present the estimated canonical variates and directions obtained using group CCAR^3. Corresponding results for sparse CCAR^3, SCCA (Parkhomenko et al., 2009), and SGCA (Gao and Ma, 2023) are provided in the online supplement.[3] As expected, the group-sparse formulation yields a more parsimonious solution for $(\widehat{u}_i)_{i=1}^2$, selecting only 11 of the 30 brain regions, compared with a larger number selected by the sparse method.

The bottom bar plot of Figure 3 shows that the first canonical direction identified by group CCAR^3 is significantly associated with anhedonia, whereas drive contributes strongly

3. `https://github.com/donnate/CCAR3code/tree/main/additional_plots`.

	Test MSE	Test Correlation
sparse CCAR^3 (this paper)	1.67 ± 0.10	0.05 ± 0.03
group CCAR^3 (59 L/R groups; this paper)	**1.53 ± 0.03**	**0.11 ± 0.01**
SAR (Wilms et al.)	2.03 ± 0.08	0.05 ± 0.05
CCA-Lasso (Witten et al.)	2.02 ± 0.13	-0.01 ± 0.06
SCCA (Parkhomenko et al.)	1.75 ± 0.07	0.05 ± 0.02
Fantope CCA (Gao et al.)	2.15 ± 0.10	0.02 ± 0.04
SCCA with CoLaR (Gao et al.)	1.90 ± 0.13	0.02 ± 0.03
SGCA (Gao & Ma)	1.81 ± 0.03	0.05 ± 0.01

Table 2: Results of the 20-fold cross-validation on the Neuroscience dataset with rank $r = 2$. Values correspond to the test MSE over the test folds, averaged over five random splits of the data into 20 folds, ± standard deviation.

	Drive	Funseeking	Reward Resp.	Total	Distress	Anhedonia	Anx. Arousal	Pos. Affect	Neg. Affect
sparse CCAR^3	0.10	**0.41**	0.05	0.14	0.11	0.15	0.06	0.15	0.12
group CCAR^3	**0.38**	0.08	**0.28**	0.04	0.22	**0.46**	0.01	0.29	0.22
SAR	0.10	0.28	0.00	0.04	0.00	0.02	0.04	-0.01	0.03
CCA-Lasso	0.22	0.07	0.10	0.10	0.12	0.13	0.10	0.12	0.16
SCCA	0.11	0.07	0.11	0.12	0.23	0.38	0.08	0.28	0.14
Fantope CCA	-0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
SGCA	0.07	0.07	0.06	**0.36**	**0.52**	0.45	**0.35**	**0.32**	**0.51**

Table 3: Adjusted R^2 for the regression of $\mathbf{X}\widehat{\mathbf{U}}$ onto each of the columns of $\mathbf{Y}$.

to both the first and second canonical directions. These behavioral dimensions induce a clear structure in the canonical variate space $\mathbf{X}\widehat{\mathbf{U}}$, where subjects with different levels of anhedonia and drive occupy distinct regions (see the scatter plot of Figure 3).

To assess the extent to which the estimated canonical variates capture variation in individual questionnaire measures, we regress each of the nine behavioral outcomes in $\mathbf{Y}$ on $\mathbf{X}\widehat{\mathbf{U}}$. The adjusted R^2 values are reported in Table 3. Group CCAR^3 achieves the strongest associations with anhedonia and drive, indicating that these variables are the dominant behavioral correlates of the brain connectivity patterns extracted by the method.

From a neuroscientific perspective, group CCAR^3 identifies the thalamus as an important contributor to the canonical directions. The thalamus plays a central role in arousal, attention, and consciousness, making its involvement plausible in the context of a reward-based gambling task. In addition, group CCAR^3 assigns large weight to the nucleus accumbens, a key component of the brain's reward circuitry that has been strongly linked to reward anticipation and processing (Day and Carelli, 2007). Interestingly, the Parahippocampal Gyrus is also highlighted by our method. While the Parahippocampal Gyrus is primarily known for its role in memory encoding and contextual associative processing, it has also been implicated in the integration of contextual information during value-based learning and decision making. This suggests that contextual and memory-related processes may contribute to the neural mechanisms underlying reward-related behavior in this task.

5.3 Climate data: exploring spatial structure using the graph-sparse model

In this section, we illustrate the proposed graph CCAR^3 method using a climate dataset. The predictor variables $\mathbf{X}$ consist of sea surface temperature (SST) measurements from the Extended Reconstructed Sea Surface Temperature dataset, Version 5 (ERSSTv5), produced

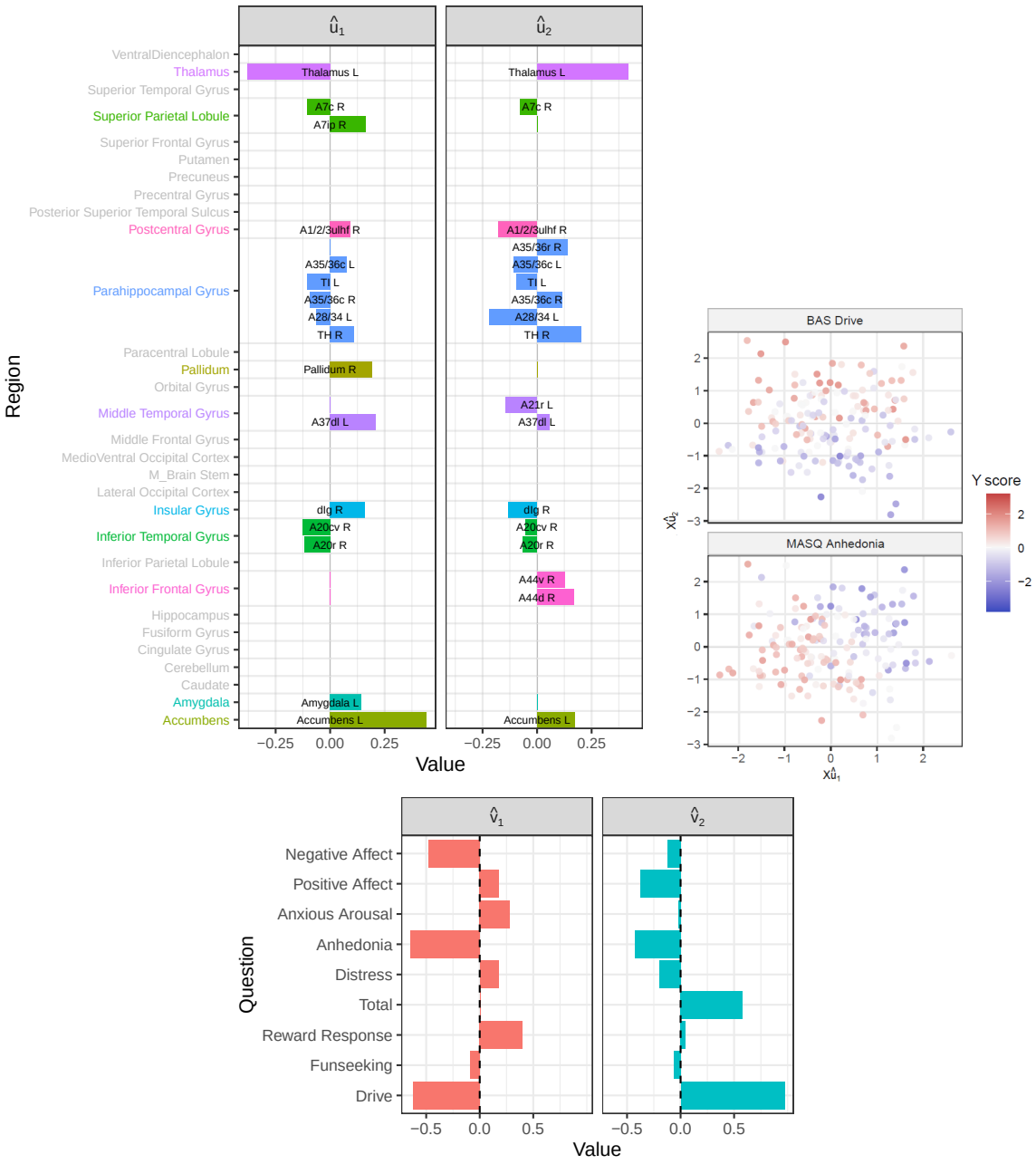


Figure 3: Results on the Neuroscience dataset using group sparse CCAR[3]. Bar plots: canonical directions $(\widehat{v}_i)_{i=1}^2$ representing the questionnaire and top 20 brain activation directions $(\widehat{u}_i)_{i=1}^2$ grouped along the y-axis and colored according to the brain gyri. Scatter plots: canonical variates $(\mathbf{X}\widehat{u}_i)_{i=1}^2$, colored by drive and anhedonia scores.

by the National Oceanic and Atmospheric Administration (NOAA). The dataset provides monthly SST observations on a $2^\circ \times 2^\circ$ latitude–longitude grid covering the global ocean, with records available since January 1854. We focus on the equatorial Pacific region, selecting grid points with longitudes between 120°E and 280°E and latitudes between 20°S and 20°N. To reduce spatial resolution, the grid is coarsened to 4°, resulting in a time-by-space matrix $\mathbf{X}$ where each row corresponds to a monthly observation and each column to a spatial grid location. The seasonal cycle is removed prior to analysis.

The response variables $\mathbf{Y}$ comprise several large-scale climate indices obtained from NOAA datasets, including the Arctic Oscillation (AO), North Atlantic Oscillation (NAO), North Pacific index (NP), outgoing longwave radiation (OLR), Pacific–North American pattern (PNA), and the Southern Oscillation Index (SOI). These indices summarize major modes of atmospheric variability and are widely used to characterize large-scale climate dynamics.

After removing missing values, the dataset contains $n = 589$ observations, $p = 441$ predictors, and $q = 6$ response variables. We construct train–test–validation splits using consecutive time points: 250 for training, 100 for validation, and 100 for testing. To reduce temporal dependence between the sets, a gap of 25 time points is introduced between consecutive segments. By shifting the starting time point forward by 25 at each iteration, we obtain four distinct train–test–validation splits.

We fit sparse CCA methods with $r = 3$ components. The graph CCAR^3 method uses the spatial locations of features in $\mathbf{X}$ to construct a graph by connecting neighboring grid points within a 4° distance. After selecting the optimal penalty parameter for each method using the test set, we evaluate their performance on the test set.

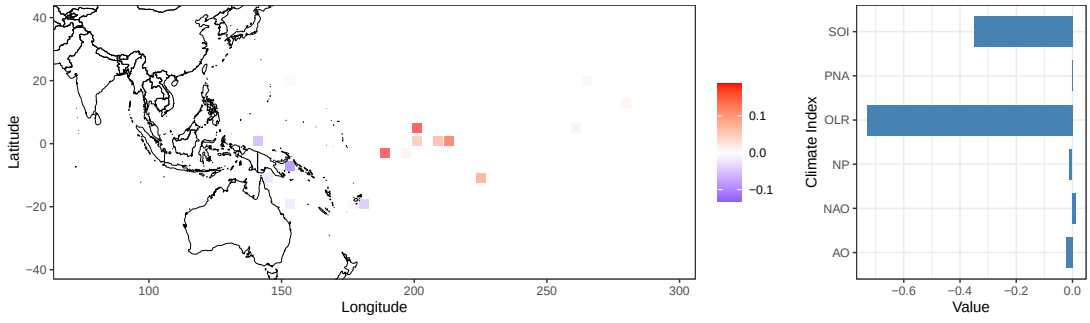
Sparse CCAR^3 achieves a validation correlation of 0.383 with an MSE of 1.09. In contrast, graph CCAR^3 performs better, attaining a substantially higher test correlation of 0.474 and a lower MSE of 0.946. By comparison, SAR achieves a correlation of 0.370 with an MSE of 1.27, and CCA-Lasso attains a correlation of 0.420 with an MSE of 1.08, both underperforming relative to the graph-based approach. Notably, CCA-Lasso fails to recover more than two components for most penalty choices.

In terms of computational efficiency, both CCAR^3 variants are substantially faster than SAR: training graph CCAR^3 and sparse CCAR^3 requires, on average, 2.30 and 1.70 seconds, respectively, compared to 25.7 seconds for SAR, while CCA-Lasso is the fastest at 0.0583 seconds.

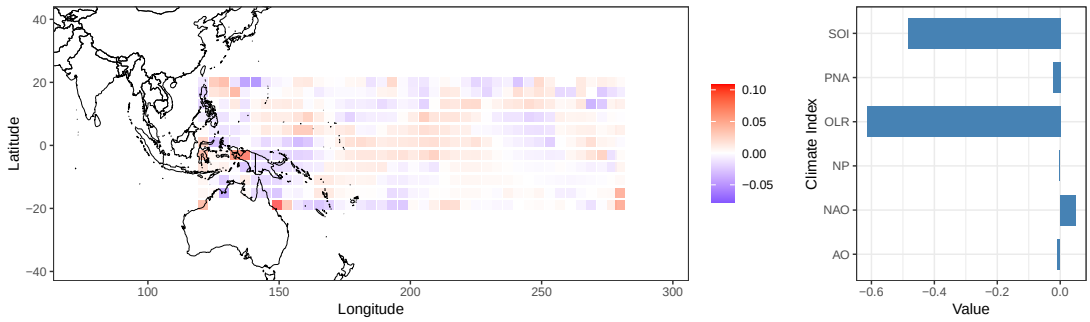
Figure 4 compares the first canonical directions obtained by sparse and graph CCAR^3. The $\widehat{v}_1$ vectors exhibit similar patterns, highlighting the influence of SOI and OLR. In contrast, the $\widehat{u}_1$ vectors differ substantially: the sparse method identifies a few isolated locations with non-zero values (positive near longitude 200°E and latitude 0°, and negative near longitude 150°E and latitude -10°), whereas the graph-based method reveals clusters of significant regions, with similar overall positions of positive and negative important features.

These positions correspond to physically meaningful SST regions: the central/eastern equatorial Pacific (around 200°E) and the western Pacific warm pool (around 150°E) are key areas influencing El Niñ-Southern Oscillation variability, consistent with Southern Oscillation Index (SOI) and outgoing longwave radiation (OLR) patterns. The graph-regularization captures spatially coherent SST structures, reflecting connected regions that

jointly influence large-scale climate indices, whereas the sparse method highlights only individual grid points, potentially missing broader regional effects. Overall, the graph-based approach provides a more interpretable and physically meaningful representation of ocean-atmosphere interactions.



(a) sparse CCAR^3



(b) graph CCAR^3

Figure 4: Comparison of the first canonical directions obtained by sparse and graph CCAR^3. Left panels represent the canonical directions $\widehat{u}_1$ corresponding to SST features. Right panels represent the canonical directions $\widehat{v}_1$ corresponding to climate indices.

6. Related Works

To overcome the challenge posed by the high dimensionality of the data, several adaptations of the original CCA problem have recently been developed. All of these modifications incorporate some form of regularization or constraint on the estimators of $\mathbf{U}$ and $\mathbf{V}$, and can be classified into one of two categories depending on the type of regularization they employ.

Ridge-regularized CCA methods attempt to resolve the issue of the ill-conditioning of $\widehat{\mathbf{\Sigma}}_X$ and $\widehat{\mathbf{\Sigma}}_Y$ by adjusting their diagonals (Leurgans et al., 1993). In this setting, the CCA problem becomes:

$$\begin{aligned} \underset{\mathbf{U}\in\mathbb{R}^{p\times r},\ \mathbf{V}\in\mathbb{R}^{q\times r}}{\text{maximize}} \quad & \mathsf{Tr}\left(\mathbf{U}^\top(\widehat{\mathbf{\Sigma}}_X+\rho_1\mathbf{I})^{-1}\widehat{\mathbf{\Sigma}}_{XY}(\widehat{\mathbf{\Sigma}}_Y+\rho_2\mathbf{I})^{-1}\mathbf{V}\right) \\ \text{subject to} \quad & \mathbf{U}^\top\widehat{\mathbf{\Sigma}}_X\mathbf{U}=\mathbf{V}^\top\widehat{\mathbf{\Sigma}}_Y\mathbf{V}=\mathbf{I}_r. \end{aligned} \tag{14}$$

While ℓ_2-regularized CCA methods have demonstrated some empirical success, to the best of our knowledge, there are no theoretical guarantees as to the accuracy of the proposed estimators of $\mathbf{U}$ and $\mathbf{V}$. Moreover, while these methods shrink coefficients, they often yield dense solutions that do not necessarily lend themselves well to subsequent interpretation and analysis.

Sparse CCA methods, on the other hand, suggest imposing sparsity constraints on the CCA directions. This is typically done by introducing a variant of the ℓ_1 penalty, which effectively reduces the complexity of the model by identifying and using only the most informative variables in $\mathbf{X}$ and $\mathbf{Y}$ to estimate the directions $\mathbf{U}$ and $\mathbf{V}$. Building on the success of the lasso in the regression setting (Tibshirani, 1996), initial approaches to sparse canonical correlation analysis focused primarily on imposing a sparsity penalty on the individual canonical directions. For instance, Witten and Tibshirani (2009) and Parkhomenko et al. (2009) propose algorithms imposing a lasso penalty on each of the canonical directions for the case where $r = 1$. However, their proposed methods are not straightforwardly applicable to the estimation of several directions, particularly when the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$ are not diagonal and it is no longer possible to neglect the normalization constraints $\mathbf{U}^\top \widehat{\mathbf{\Sigma}}_X \mathbf{U} = \mathbf{I}_r$ and $\mathbf{V}^\top \widehat{\mathbf{\Sigma}}_Y \mathbf{V} = \mathbf{I}_r$. As demonstrated by our experiments in Section 4.1, in this case, these methods struggle to estimate the canonical directions even when n is large. To allow the estimation of multiple canonical directions, Wilms and Croux (2015) propose an iterative algorithm based on an alternating minimization scheme to solve for each pair (u_i, v_i). This method allows the computation of more than one pair of canonical directions by progressively deflating the matrices $\mathbf{X}$ and $\mathbf{Y}$. Since the sparsity of canonical directions obtained on the deflated matrices does not necessarily translate into sparse canonical directions expressed in terms of the original data matrices $\mathbf{X}$ and $\mathbf{Y}$, the authors propose adding a postprocessing step that sparsifies their estimated canonical directions in the original covariate basis. Although the proposed technique is efficient, the alternating nature of this method makes it sensitive to initialization. Moreover, while each of the resulting CCA directions $(u_i)_{i=1}^r$ and $(v_i)_{i=1}^r$ are typically sparse, the matrices $\mathbf{U}$ and $\mathbf{V}$ are not necessarily row-sparse, which may lead to non-sparse canonical subspaces (defined by the span of the columns of $\mathbf{U}$ or $\mathbf{V}$).

While sparse CCA variants have proliferated over the years, one of the few theoretical treatments is provided by Gao and coauthors (Chen et al., 2013; Gao et al., 2015, 2017). More specifically, Gao et al. (2015) consider the setting where the support sizes, i.e., the number of nonzero rows of $\mathbf{U}$ and $\mathbf{V}$, are small: $|\mathrm{supp}(\mathbf{U})| \leq s_u$ and $|\mathrm{supp}(\mathbf{V})| \leq s_v$ for some $s_u \leq p$, $s_v \leq q$. Under this assumption, they show that the minimax rate for estimators of $\mathbf{U}$ and $\mathbf{V}$ under the joint loss $\|\widehat{\mathbf{U}}\widehat{\mathbf{V}}^\top - \mathbf{U}\mathbf{V}^\top\|_F^2$ is

$$\frac{1}{n\lambda_r^2}\Big(r(s_u + s_v) + s_u \log(ep/s_u) + s_v \log(eq/s_v)\Big),$$

where λ_r denotes the r^{th} canonical correlation.

Since the estimator proposed by Gao et al. (2015) is not tractable in practice, Gao et al. (2017) introduce a three-step algorithm that achieves the minimax rate. This procedure requires splitting the data into three batches. In the first step, the cross-product $\mathbf{F} = \mathbf{U}\mathbf{V}^\top$ is estimated from the first batch, transforming the original non-convex CCA problem into

a convex optimization problem:

$$\underset{\mathbf{F}\in\mathbb{R}^{p\times q}}{\text{maximize}}\ \langle\widehat{\mathbf{\Sigma}}_{XY},\mathbf{F}\rangle-\rho\|\mathbf{F}\|_{11}\ \text{ subject to }\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{F}\widehat{\mathbf{\Sigma}}_Y^{1/2}\|_*\le r,\ \|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{F}\widehat{\mathbf{\Sigma}}_Y^{1/2}\|_{\text{op}}\le 1, \quad (15)$$

where ρ is a sparsity-controlling penalty parameter, and $\|\cdot\|_{11}$, $\|\cdot\|_*$, and $\|\cdot\|_{\text{op}}$ denote the usual matrix ℓ_1, nuclear, and operator norms, respectively. This step provides initial estimates of the canonical directions $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$. In the second step, using the second batch of data, the estimators are refined by solving separate group-lasso regression problems for $\mathbf{U}$ and $\mathbf{V}$, enforcing sparsity via the ℓ_{21} penalty. The third step uses the third batch to normalize the resulting estimates.

While theoretically sound, this procedure has important practical limitations. The first step relies on an ADMM algorithm that requires repeated SVDs of $p\times q$ matrices and solving lasso problems with pq parameters at each iteration, resulting in substantial computational costs. Moreover, splitting the data into three batches can be problematic in small-sample settings, further limiting the method's practical applicability.

Building on this work, Gao and Ma (2023) propose an alternative two-step procedure. The first step serves as an initialization for $\mathbf{F}=\left(\begin{smallmatrix}\mathbf{U}\\ \mathbf{V}\end{smallmatrix}\right)\left(\begin{smallmatrix}\mathbf{U}^\top & \mathbf{V}^\top\end{smallmatrix}\right)\in\mathbb{R}^{(p+q)\times(p+q)}$ and is based on a Fantope projection problem:

$$\underset{\mathbf{F}\in\mathcal{F}}{\text{minimize}}\ -\langle\widehat{\mathbf{\Sigma}},\mathbf{F}\rangle+\rho\|\mathbf{F}\|_1\quad\text{subject to}\quad\widehat{\mathbf{\Sigma}}_0^{1/2}\mathbf{F}\widehat{\mathbf{\Sigma}}_0^{1/2}\in\mathcal{F},$$

where

$$\mathcal{F}=\left\{\mathbf{X}\in\mathbb{R}^{(p+q)\times(p+q)}:0\preceq\mathbf{X}\preceq\mathbf{I}_{p+q},\ \mathsf{Tr}(\mathbf{X})=r\right\}.$$

The second step refines the estimates using a thresholded gradient descent algorithm. While this approach improves estimation, the initialization still relies on an ADMM algorithm that involves repeated SVDs of a $(p+q)\times(p+q)$ matrix, which scales poorly with dimension and becomes computationally expensive for large datasets.

While these methods achieve minimax-optimal statistical guarantees, the quality of the estimators critically depends on the initializer obtained in the first step, which is computationally expensive (see Appendix B.1 for details). Moreover, selecting the regularization parameter ρ via cross-validation requires repeatedly solving this initialization problem, further amplifying the computational burden.

7. Conclusion

In this paper we present CCAR^3, a novel technique for estimating canonical directions based on the link between CCA and reduced rank regression when one of the datasets is high-dimensional while the dimension of the other remains small. Our method can easily incorporate different types of penalties, allowing a good compromise between computational efficiency, statistical accuracy, and practical flexibility in a wide range of practical settings. A potential direction for future research consists in extending CCAR^3 to settings where both p and q are large. This will require further investigation of computationally efficient alternatives to the Fantope optimization problem, which is both the crux of the success and the computational bottleneck of Gao et al. (2017) and Gao and Ma (2023).

Acknowledgments

C.D. was supported by the National Science Foundation under Award Number 2238616, as well as the resources provided by the University of Chicago's Research Computing Center. E.T. was supported by the Natural Sciences and Engineering Research Council of Canada under grant RGPIN-2023-04727; the University of Toronto Data Science Institute Catalyst grant; and the University of Toronto McLaughlin Center under grant MC-2023-05. This research was supported in part by grants from the NSF (DMS-2235451) and Simons Foundation (MPS-NITMB-00005320) to the NSF-Simons National Institute for Theory and Mathematics in Biology (NITMB).

References

Tim P. Barnett and Rudolf Preisendorfer. Origins and Levels of Monthly and Seasonal Forecast Skill for United States Surface Air Temperatures Determined by Canonical Correlation Analysis. *Monthly Weather Review*, 115(9):1825–1850, 1987.

Ake Björck and Gene H. Golub. Numerical Methods for Computing Angles Between Linear Subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2011.

Christopher S. Bretherton, Catherine Smith, and John M. Wallace. An Intercomparison of Methods for Finding Coupled Patterns in Climate Data. *Journal of Climate*, 5(6): 541–560, 1992.

Mengjie Chen, Chao Gao, Zhao Ren, and Harrison H Zhou. Sparse CCA via precision adjusted iterative thresholding. *arXiv preprint arXiv:1311.6186*, 2013.

Eric C. Chi, Genevera I. Allen, Hua Zhou, Omid Kohannim, Kenneth Lange, and Paul M. Thompson. Imaging genetics via sparse canonical correlation analysis. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 740–743, 2013.

Fan RK Chung. Lectures on spectral graph theory. *CBMS Lectures, Fresno*, 6(92):17–21, 1996.

Jeremy J. Day and Regina M. Carelli. The Nucleus Accumbens and Pavlovian Reward Learning. *The Neuroscientist*, 13(2):148–159, 2007.

Fernando De la Torre. A Least-Squares Framework for Component Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1041–1055, 2012.

Xitao Fan and Timothy R Konold. Canonical correlation analysis. In *The reviewer's guide to quantitative methods in the social sciences*, pages 29–41. Routledge, 2018.

Chao Gao, Zongming Ma, Zhao Ren, and Harrison H Zhou. Minimax estimation in sparse canonical correlation analysis. *The Annals of Statistics*, 43(5):2168–2197, 2015.

Chao Gao, Zongming Ma, and Harrison H Zhou. Sparse CCA: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017.

Sheng Gao and Zongming Ma. Sparse GCA and thresholded gradient descent. *Journal of Machine Learning Research*, 24(123):1–61, 2023.

Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

Ignacio Gonzalez, Kim-Anh Le Cao, Melissa J. Davis, and Sebastien Dejean. Visualising associations between paired 'omics' data sets. *BioData Mining*, 5(19), 2012.

Alexej Gossmann, Pascal Zille, Vince Calhoun, and Yu-Ping Wang. FDR-corrected sparse canonical correlation analysis with applications to imaging genomics. *IEEE transactions on medical imaging*, 37(8):1761–1774, 2018.

Harold Hotelling. Relations between two sets of variates. In *Biometrika, 28*, pages 321–377. Springer, 1936.

Jan-Christian Hütter and Philippe Rigollet. Optimal rates for total variation denoising. In *Conference on Learning Theory*, pages 1115–1146. PMLR, 2016.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975. ISSN 0047-259X.

Min-Zhi Jiang, François Aguet, Kristin Ardlie, Jiawen Chen, Elaine Cornell, Dan Cruz, Peter Durda, Stacey B. Gabriel, Robert E. Gerszten, Xiuqing Guo, Craig W. Johnson, Silva Kasela, Leslie A. Lange, Tuuli Lappalainen, Yongmei Liu, Alex P. Reiner, Josh Smith, Tamar Sofer, Kent D. Taylor, Russell P. Tracy, David J. VanDenBerg, James G. Wilson, Stephen S. Rich, Jerome I. Rotter, Michael I. Love, Laura M. Raffield, Yun Li, and TOPMed Analysis Working Group NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. Canonical correlation analysis for multi-omics: Application to cross-cohort analysis. *PLOS Genetics*, 19(5):1–22, 05 2023.

Olivier Ledoit and Michael Wolf. The power of (non-) linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics*, 20(1):187–218, 2022.

S.E. Leurgans, R.A. Moyeed, and B.W. Silverman. Canonical Correlation Analysis when the Data are Curves. *Journal of the Royal Statistical Society*, 55(3):725–740, 1993.

Dongdong Lin, Jigang Zhang, Jingyao Li, Vince D Calhoun, Hong-Wen Deng, and Yu-Ping Wang. Group sparse canonical correlation analysis for genomic data integration. *BMC bioinformatics*, 14(1):1–16, 2013.

Tristan Looden, Dorothea L Floris, Alberto Llera, Roselyne J Chauvin, Tony Charman, Tobias Banaschewski, Declan Murphy, Andre F Marquand, Jan K Buitelaar, Christian F Beckmann, et al. Patterns of connectome variability in autism across five functional activation tasks: findings from the LEAP project. *Molecular autism*, 13(1):53, 2022.

Lekai Luo, Lizhou Chen, Yuxia Wang, Qian Li, Ning He, Yuanyuan Li, Wanfang You, Yaxuan Wang, Fenghua Long, Lanting Guo, and et al. Patterns of brain dynamic functional connectivity are linked with attention-deficit/hyperactivity disorder-related behavioral and cognitive dimensions. *Psychological Medicine*, 53(14):6666–6677, 2023.

Pascal GP Martin, Hervé Guillou, Frédéric Lasserre, Sébastien Déjean, Annaig Lan, Jean-Marc Pascussi, Magali SanCristobal, Philippe Legrand, Philippe Besse, and Thierry Pineau. Novel aspects of ppar α-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. *Hepatology*, 45(3):767–777, 2007.

Elena Parkhomenko, David Tritchler, and Joseph Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical applications in genetics and molecular biology*, 8(1), 2009.

Grace Pigeau, Manuela Costantino, Gabriel Devenyi, Aurelie Bussy, Olivier Parent, and Mallar Chakravarty. The impact of study design choices on significance and generalizability of canonical correlation analysis in neuroimaging studies. *bioRxiv*, pages 2022–12, 2022.

Theodoulos Rodosthenous, Vahid Shahrezaei, and Marina Evangelou. Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study. *Bioinformatics*, 36(17):4616–4625, 2020.

Stephen M Smith, Thomas E Nichols, Diego Vidaurre, Anderson M Winkler, Timothy EJ Behrens, Matthew F Glasser, Kamil Ugurbil, Deanna M Barch, David C Van Essen, and Karla L Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature neuroscience*, 18(11):1565–1567, 2015.

Robert M Thorndike. Canonical correlation analysis. In *Handbook of applied multivariate statistics and mathematical modeling*, pages 237–263. Elsevier, 2000.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Leonardo Tozzi, Elena Tuzhilina, Matthew F Glasser, Trevor J Hastie, and Leanne M Williams. Relating whole-brain functional connectivity to self-reported negative emotion in a large sample of young adults using group regularized canonical correlation analysis. *Neuroimage*, 237:118137, 2021.

Huy Tran, Sansen Wei, and Claire Donnat. The Generalized Elastic Net for least squares regression with network-aligned signal and correlated design. *arXiv preprint arXiv:2211.00292*, 2022.

Elena Tuzhilina, Leonardo Tozzi, and Trevor Hastie. Canonical correlation analysis in high dimensions with structured regularization. *Statistical Modelling*, 23(3):203–227, 2021.

Sandra Waaijenborg and Aeilko H. Zwinderman. Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. *BMC Bioinformatics*, 10, 2009.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Zishan Wang, Ruqiang Huang, Ye Yan, Zhiguo Luo, Shaokai Zhao, Bei Wang, Jing Jin, Liang Xie, and Erwei Yin. An Improved Canonical Correlation Analysis for EEG Inter-Band Correlation Extraction. *Bioengineering*, 10(10):1200, 2023.

Ines Wilms and Christophe Croux. Sparse canonical correlation analysis from a predictive point of view. *Biometrical Journal*, 57(5):834–851, 2015.

Ines Wilms and Christophe Croux. Robust sparse canonical correlation analysis. *BMC Systems Biology*, 10(72), 2016.

Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1), 2009.

Xiaowei Zhuang, Zhengshi Yang, and Dietmar Cordes. A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping*, 41(13):3807–3833, 2020.

Appendix A. Additional plots and tables

A.1 A review of CCA in recent applications

Study	Objective	Number of Observations	Size of dataset 1	Size of dataset 2
Wang et al. (2023)	**X**: features extracted from EEG data **Y**: Emotion Data	$n = 15$	$p = 310$	$q = 3$
Looden et al. (2022)	**X**: Connectivity features extracted from fMRI **Y**: Behavioral Measurements	$n = 299$	$p = 30, 3081$	$q = 3$
Luo et al. (2023)	**X**: Connectivity features extracted from fMRI measurements **Y**: Clinical and Behavioural Measurements	$n = 122$	$p = 25, 651$	$q \in [30, 50]$
Tozzi et al. (2021)	**X**: fMRI activation of each voxel in the brain **Y**: Questionnaire Data	$n = 269$	$p = 90, 000$	$q = 89$
Pigeau et al. (2022)	**X**: vertex-level cortical thickness measurements **Y**: Behavioural Data	$n = 25, 043$	$p = 38, 561$	$q = 52$

Table 4: Short survey of recent applications of high-dimensional CCA.

A.2 Simulations

Group Sparsity. We simulate here a setting where we know the groups. We adopt a similar data generation mechanism as in the previous subsection, adapting only the procedure for generating **U**. In particular, here, we split the p covariates into groups of size 10, and select 5 of these groups to be nonzero. We then select at random each of the entries of these groups and normalize as for the simulations in the previous section.

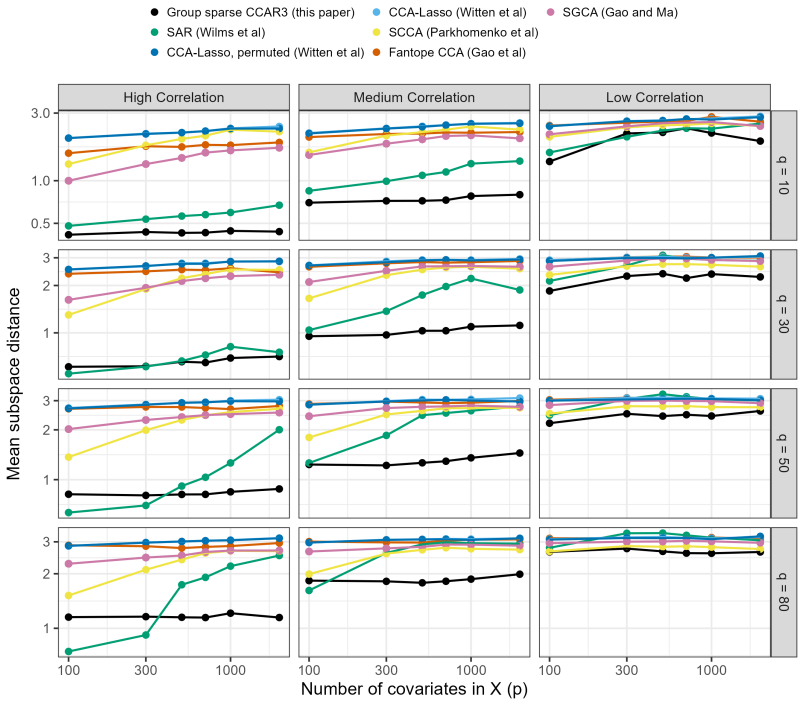


Figure 5: Simulation results for group model with $r_{pca} = 5$, 5 nonzero groups and $n = 500$. Subspace distance as a function of p. Results are averaged over 100 simulations.

In Figure 6 the true discovery is understood as an overlap between the supports of the estimator and the true support of U^*, and a false discovery is understood as the complement of this intersection in the support of $\widehat{U}$.

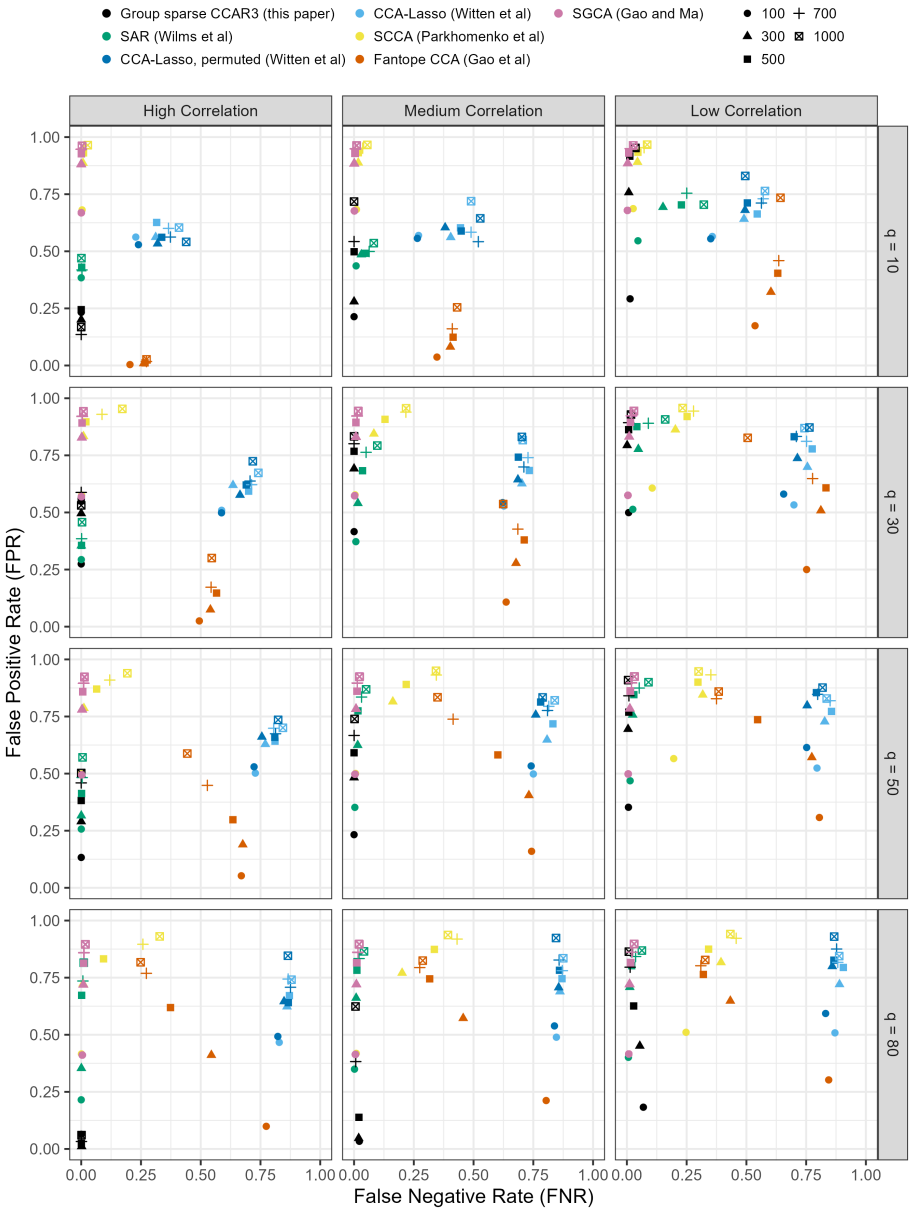


Figure 6: Simulation results for sparse model with $r_{pca} = 5$. False Negative Rate vs False Positive Rate with $n = 500$, $r = 3$ and $n_{nnz} = 6$. The shape of each point represents X data dimensionality (p). Results are averaged over 100 simulations.

Graph Sparsity. We reiterate the previous set of synthetic experiments, changing only the data generation mechanism for $\mathbf{U}$ to accommodate a graph-structure. More specifically, we first generate a graph (here a 2D grid) with p nodes and m edges. We sample $\widetilde{\mathbf{U}} \in \mathbb{R}^{m \times r}$ in the same way as in our sparse CCA experiments. We then apply a transformation to $\widetilde{\mathbf{U}}$ as $\mathbf{U} = \mathbf{\Gamma}^{\dagger}\widetilde{\mathbf{U}}$, where $\mathbf{\Gamma}^{\dagger}$ denotes the pseudoinverse of $\mathbf{\Gamma}$. This results in a $\mathbf{U}$ such that $\mathbf{\Gamma U}$ is sparse and $\mathbf{\Pi} \boldsymbol{U} = 0$, where $\mathbf{\Pi} = \boldsymbol{I}_p - \mathbf{\Gamma}^{\dagger}\mathbf{\Gamma}$. Similar to the previous experiments, we vary the signal strength and sizes of p and q.

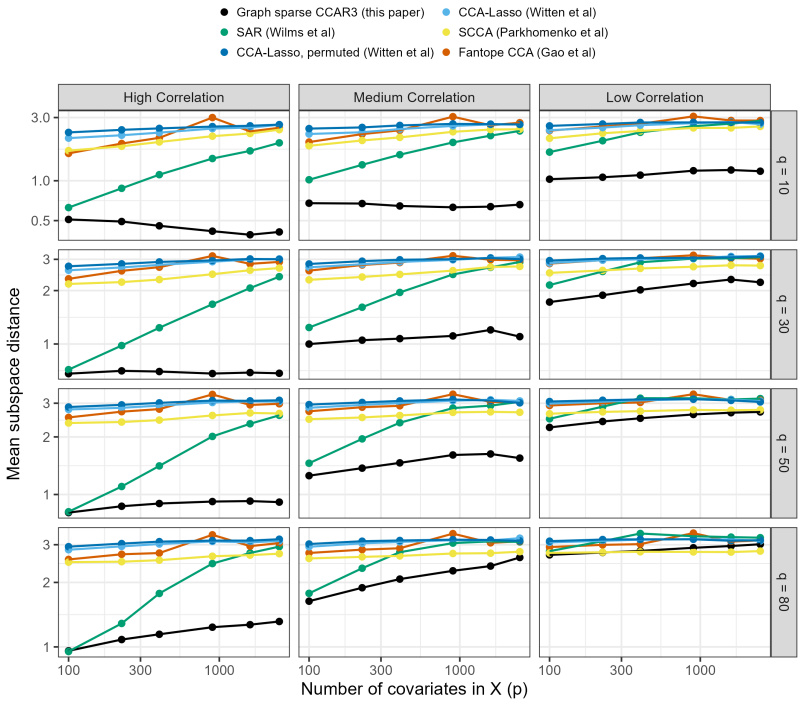


Figure 7: Simulation results for graph model with $r_{pca} = 5$, $\text{supp}(\mathbf{\Gamma B}) = 5$ and $n = 500$. Subspace distance as a function of p. Results are averaged over 100 simulations.

A.3 Nutrimouse dataset

Below, we present the loading vectors recovered by sparse CCAR^3.

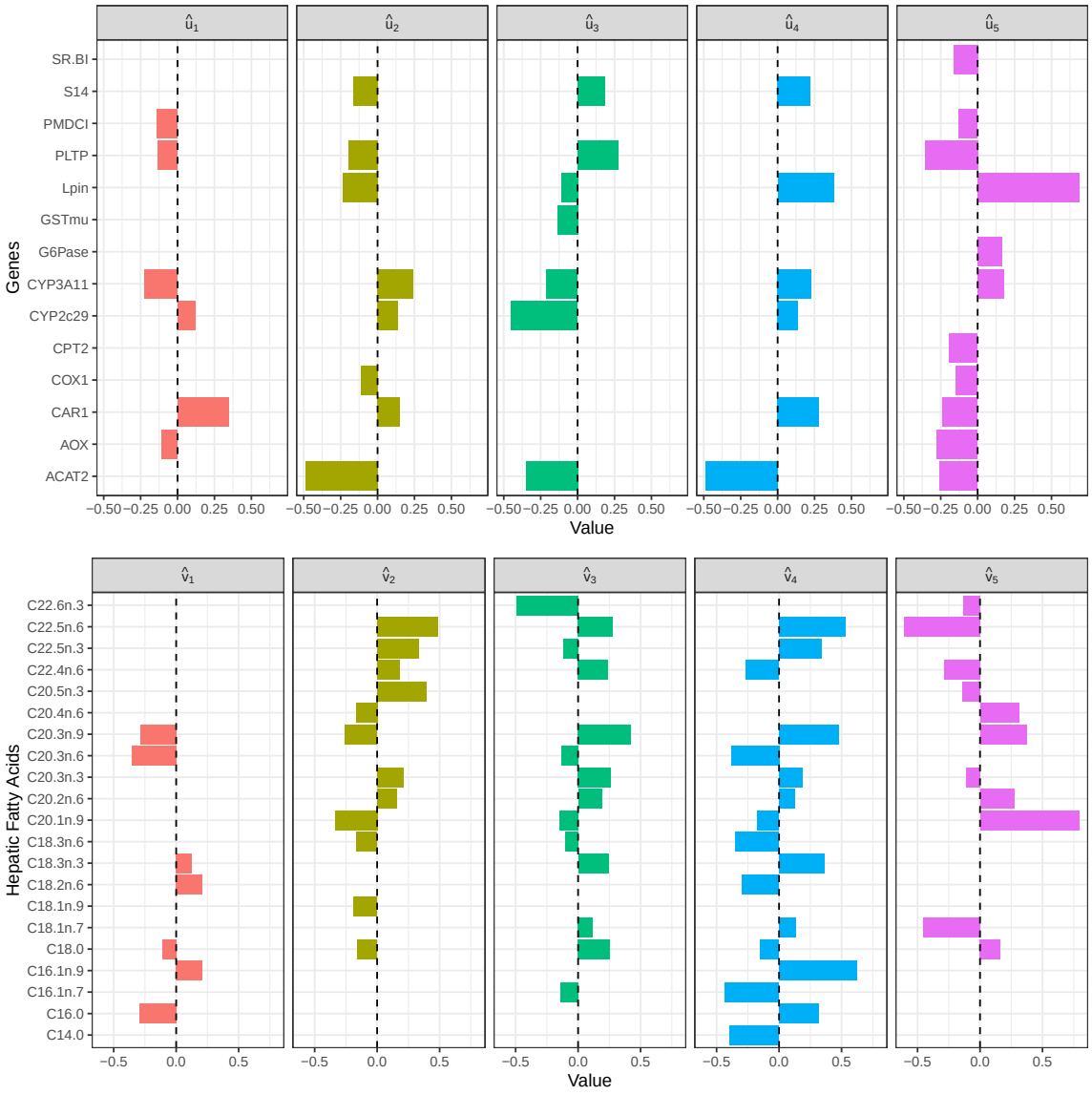


Figure 8: Sparse CCAR^3 presented in this paper. Five canonical directions corresponding to genes, i.e. $(\widehat{\boldsymbol{u}}_i)_{i=1}^5$, and fatty acids, i.e. $(\widehat{\boldsymbol{v}}_i)_{i=1}^5$, produced for the Nutrimouse dataset. For the convenience of visualization, the plot represents only values greater than the 0.1 threshold.

Appendix B. Algorithms

B.1 Computational complexity of theoretical methods

We review the computational complexity of the three-stage procedure in *Sparse CCA: Adaptive Estimation and Computational Barriers* by Gao et al. (2017) and the two-stage procedure in the *Thresholded Gradient Descent method for Sparse GCA* by Gao and Ma (2023). In both methods, the main computational burden arises from the convex ADMM-based initialization. The subsequent refinement stages are significantly cheaper when $r \ll \min(p, q)$.

Sparse CCA by Gao et al. (2017) The data are split into three independent batches indexed by $0, 1, 2$.

Stage 1: Initialization. The initialization solves

$$\max_{\mathbf{F}\in\mathbb{R}^{p\times q}} \langle \widehat{\mathbf{\Sigma}}_{XY}^{(0)}, \mathbf{F}\rangle - \rho\|\mathbf{F}\|_1 \quad \text{s.t.} \quad \|(\widehat{\mathbf{\Sigma}}_X^{(0)})^{1/2}\mathbf{F}(\widehat{\mathbf{\Sigma}}_Y^{(0)})^{1/2}\|_* \le r, \quad \|(\widehat{\mathbf{\Sigma}}_X^{(0)})^{1/2}\mathbf{F}(\widehat{\mathbf{\Sigma}}_Y^{(0)})^{1/2}\|_{op} \le 1.$$

The optimization variable $\mathbf{F}$ is a $p \times q$ matrix, so the problem has pq parameters. The algorithm is based on ADMM and each iteration consists of three main steps.

1. *Lasso.* Solve an ℓ_1-penalized quadratic problem with pq parameters.
2. *Singular value capped soft-thresholding.* This requires computing the SVD of a $p \times q$ matrix.
3. *Matrix multiplications.* Multiply $p \times p$, $p \times q$, and $q \times q$ matrices.

Stage 2: Group Lasso Refinement. Let $\mathbf{V}^{(0)} \in \mathbb{R}^{q\times r}$ store the top r right singular vectors of the solution $\widehat{\mathbf{F}}$ from Stage 1. The second stage solves a group lasso problem with pr parameters to obtain $\mathbf{U}^{(1)}$. The analogous update for $\mathbf{V}$ uses the estimated left singular vectors and has qr parameters.

Stage 3: Normalization. The final estimators are normalized on the third data split, e.g. $\widehat{\mathbf{U}} = \widehat{\mathbf{U}}^{(1)} \left(\widehat{\mathbf{U}}^{(1)\top}\widehat{\mathbf{\Sigma}}_X^{(2)}\widehat{\mathbf{U}}^{(1)}\right)^{-1/2}$ and analogously for $\widehat{\mathbf{V}}$.

Overall, the computational bottleneck of Sparse CCA is the convex ADMM initialization, dominated by repeated lasso solves and SVDs of $p \times q$ matrices. Additionally, this approach requires splitting data into three batches, which is not ideal for many applications with limited data.

Sparse GCA by Gao and Ma (2023) This approach requires two stages.

Stage 1: Initialization via Fantope Relaxation. Let $\mathbf{F} = \mathbf{A}\mathbf{A}^\top \in \mathbb{R}^{(p+q)\times(p+q)}$, where $\mathbf{A} = \begin{pmatrix}\mathbf{U}\\ \mathbf{V}\end{pmatrix} \in \mathbb{R}^{(p+q)\times r}$. The initialization is based on a Fantope projection problem, which solves

$$\min_{\mathbf{F}\in\mathcal{F}} \quad -\langle\widehat{\mathbf{\Sigma}}, \mathbf{F}\rangle + \rho\|\mathbf{F}\|_1 \quad \text{subject to} \qquad \widehat{\mathbf{\Sigma}}_0^{1/2}\mathbf{F}\widehat{\mathbf{\Sigma}}_0^{1/2} \in \mathcal{F},$$

where $\mathcal{F} = \left\{\mathbf{X} \in \mathbb{R}^{(p+q)\times(p+q)} : 0 \preceq \mathbf{X} \preceq I_{p+q},\ \mathrm{Tr}(\mathbf{X}) = r\right\}$. This problem is also solved using ADMM. Like in Gao et al. (2017), it requires computing an SVD at each iteration,

but the matrix size is now $(p+q)\times(p+q)$. The per-iteration SVD cost is therefore $O((p+q)^3)$.

Stage 2: Thresholded Gradient Descent. After initialization, refinement is performed via thresholded gradient descent. The per-iteration complexity is $O(n(p+q)r + (p+q)\log(p+q))$. The first term corresponds to gradient computation, and the second to thresholding/sorting.

Overall, the bottleneck of SGCA is the repeated SVDs of $(p+q)\times(p+q)$ matrices.

B.2 Algorithm for CCAR^3

The sparse and group-sparse versions of Step 2, and the profiled graph-sparse version described below, can be written as

$$\widehat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}\in\mathcal{C}}{\operatorname{argmin}} \frac{1}{n}\|\widetilde{\mathbf{Y}}_0 - \widetilde{\mathbf{X}}\mathbf{\Theta}\|_F^2 + \rho\sum_{m=1}^{M} w_m \|\mathbf{\Theta}_{\mathcal{G}_m\cdot}\|_F, \tag{16}$$

where $\mathcal{G}_1,\dots,\mathcal{G}_M$ form a partition of $[\tilde{p}]$, $w_m>0$ are prespecified weights, and $\mathcal{C}\subseteq\mathbb{R}^{\tilde{p}\times q}$ is a closed linear subspace. For the sparse and group-sparse cases, $\mathcal{C}=\mathbb{R}^{\tilde{p}\times q}$. For the graph case, $\mathcal{C}$ enforces that the edge variable belongs to the column space of the incidence matrix. The three special cases are as follows.

Sparse CCAR^3. Take $\tilde{p}=p$, $\widetilde{\mathbf{X}}=\mathbf{X}$, $\widetilde{\mathbf{Y}}_0=\mathbf{Y}_0$, $\mathcal{C}=\mathbb{R}^{p\times q}$, $\mathcal{G}_m=\{m\}$, $w_m=1$. Then $\widehat{\mathbf{B}}=\widehat{\mathbf{\Theta}}$.

Group-sparse CCAR^3. Take $\tilde{p}=p$, $\widetilde{\mathbf{X}}=\mathbf{X}$, $\widetilde{\mathbf{Y}}_0=\mathbf{Y}_0$, $\mathcal{C}=\mathbb{R}^{p\times q}$, $\mathcal{G}_m=g_m$, $w_m=\sqrt{|g_m|}$. Then $\widehat{\mathbf{B}}=\widehat{\mathbf{\Theta}}$.

Graph-sparse CCAR^3. Let $\mathbf{\Gamma}\in\mathbb{R}^{m\times p}$ be the graph incidence matrix, where $m=|E|$, and let $\mathbf{\Pi}=\mathbf{I}_p-\mathbf{\Gamma}^\dagger\mathbf{\Gamma}$ be the orthogonal projector onto $\ker(\mathbf{\Gamma})$. Any coefficient matrix can be written as $\mathbf{B}=\mathbf{B}_{\text{null}}+\mathbf{\Gamma}^\dagger\mathbf{\Theta}$, where $\mathbf{B}_{\text{null}}\in\ker(\mathbf{\Gamma})$ and $\mathbf{\Theta}=\mathbf{\Gamma}\mathbf{B}\in\operatorname{col}(\mathbf{\Gamma})^{\oplus q}$. Define the projection onto the fitted null-space design by

$$\mathbf{P}_\Pi = \mathbf{X}\mathbf{\Pi}\big\{(\mathbf{X}\mathbf{\Pi})^\top(\mathbf{X}\mathbf{\Pi})\big\}^\dagger(\mathbf{X}\mathbf{\Pi})^\top.$$

Profiling out the null-space component gives the exact profiled problem (16) with

$$\tilde{p}=m,\qquad \widetilde{\mathbf{X}}=(\mathbf{I}_n-\mathbf{P}_\Pi)\mathbf{X}\mathbf{\Gamma}^\dagger,\qquad \widetilde{\mathbf{Y}}_0=(\mathbf{I}_n-\mathbf{P}_\Pi)\mathbf{Y}_0,\qquad \mathcal{C}=\{\mathbf{\Theta}:(\mathbf{\Gamma}\mathbf{\Gamma}^\dagger)\mathbf{\Theta}=\mathbf{\Theta}\},$$

and singleton edge groups $\mathcal{G}_m=\{m\}$ with $w_m=1$. After obtaining $\widehat{\mathbf{\Theta}}$, the profiled null-space component is

$$\widehat{\mathbf{B}}_{\text{null}} = \mathbf{\Pi}\big\{(\mathbf{X}\mathbf{\Pi})^\top(\mathbf{X}\mathbf{\Pi})\big\}^\dagger(\mathbf{X}\mathbf{\Pi})^\top\left(\mathbf{Y}_0-\mathbf{X}\mathbf{\Gamma}^\dagger\widehat{\mathbf{\Theta}}\right),$$

and the final estimator is

$$\widehat{\mathbf{B}} = \widehat{\mathbf{B}}_{\text{null}} + \mathbf{\Gamma}^\dagger\widehat{\mathbf{\Theta}}.$$

The constraint $\widehat{\mathbf{\Theta}}\in\operatorname{col}(\mathbf{\Gamma})^{\oplus q}$ is part of the exact equivalence with the original penalty $\|\mathbf{\Gamma}\mathbf{B}\|_{21}$.

To solve (16), introduce a splitting variable $\mathbf{Z} \in \mathbb{R}^{\tilde{p}\times q}$ and consider the constrained problem

$$\underset{\mathbf{\Theta}\in\mathcal{C},\,\mathbf{Z}}{\text{minimize}}\, \frac{1}{n}\|\widetilde{\mathbf{Y}}_0 - \widetilde{\mathbf{X}}\mathbf{\Theta}\|_F^2 + \rho\sum_{m=1}^{M} w_m \|\mathbf{Z}_{\mathcal{G}_m\cdot}\|_F \quad \text{subject to} \quad \mathbf{\Theta} = \mathbf{Z}.$$

Let $\mathbf{P}_{\mathcal{C}}$ denote the orthogonal projector onto $\mathcal{C}$. Thus $\mathbf{P}_{\mathcal{C}} = \mathbf{I}_{\tilde{p}}$ for the sparse and group-sparse cases, while $\mathbf{P}_{\mathcal{C}}\mathbf{\Theta} = (\mathbf{\Gamma}\mathbf{\Gamma}^{\dagger})\mathbf{\Theta}$ for the graph-sparse case. Using the scaled augmented Lagrangian with scaled dual variable $\mathbf{D}$,

$$\mathcal{L}_\delta(\mathbf{\Theta},\mathbf{Z},\mathbf{D}) = \frac{1}{n}\|\widetilde{\mathbf{Y}}_0 - \widetilde{\mathbf{X}}\mathbf{\Theta}\|_F^2 + \rho\sum_{m=1}^{M} w_m\|\mathbf{Z}_{\mathcal{G}_m\cdot}\|_F + \frac{\delta}{2}\|\mathbf{\Theta}-\mathbf{Z}+\mathbf{D}\|_F^2 - \frac{\delta}{2}\|\mathbf{D}\|_F^2,$$

we obtain the ADMM updates given in Algorithm 3.

Algorithm 3 ADMM for (16)

Input: $\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times\tilde{p}}$, $\widetilde{\mathbf{Y}}_0 \in \mathbb{R}^{n\times q}$, groups $\mathcal{G}_1,\dots,\mathcal{G}_M$, weights $w_1,\dots,w_M$, regularization parameter $\rho > 0$, ADMM parameter $\delta > 0$, projector $\mathbf{P}_{\mathcal{C}}$, stopping tolerances $\epsilon_{\text{pri}}, \epsilon_{\text{dual}} > 0$.
Initialize: $\mathbf{\Theta}^{(0)}, \mathbf{Z}^{(0)}, \mathbf{D}^{(0)} \in \mathbb{R}^{\tilde{p}\times q}$.
Precompute: $\mathbf{K}_\delta = \frac{2}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \delta\mathbf{I}_{\tilde{p}}$ and $\mathbf{C} = \frac{2}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{Y}}_0$. When $\tilde{p}$ is moderate, factorize $\mathbf{K}_\delta$ once and reuse the factorization throughout; when $\tilde{p} \gg n$, the Woodbury identity described below is more efficient.
Repeat until convergence:

1: Update $\mathbf{\Theta}$ by solving $\mathbf{K}_\delta\mathbf{\Theta}^{(t+1)} = \mathbf{P}_{\mathcal{C}}\{\mathbf{C} + \delta(\mathbf{Z}^{(t)} - \mathbf{D}^{(t)})\}$.
2: **for** $m = 1,\dots,M$ **do**
3: Update the block $\mathbf{Z}_{\mathcal{G}_m\cdot}$ via group soft-thresholding $\mathbf{Z}_{\mathcal{G}_m\cdot}^{(t+1)} = \mathcal{S}_{\rho w_m/\delta}\Big(\mathbf{\Theta}_{\mathcal{G}_m\cdot}^{(t+1)} + \mathbf{D}_{\mathcal{G}_m\cdot}^{(t)}\Big)$, where $\mathcal{S}_\tau(\mathbf{V}) = \Big(1 - \frac{\tau}{\|\mathbf{V}\|_F}\Big)_+ \mathbf{V}$ with the convention $\mathcal{S}_\tau(0) = 0$.
4: **end for**
5: Update the scaled dual variable: $\mathbf{D}^{(t+1)} = \mathbf{D}^{(t)} + \mathbf{\Theta}^{(t+1)} - \mathbf{Z}^{(t+1)}$.

Output: $\widehat{\mathbf{\Theta}}$, and hence $\widehat{\mathbf{B}}$ through the case-specific reconstruction above.

Define the primal and dual residuals by $\mathbf{r}^{(t+1)} = \mathbf{\Theta}^{(t+1)} - \mathbf{Z}^{(t+1)}$ and $\mathbf{s}^{(t+1)} = \delta\big(\mathbf{Z}^{(t+1)} - \mathbf{Z}^{(t)}\big)$. We terminate the iterations when $\|\mathbf{r}^{(t+1)}\|_F \le \epsilon_{\text{pri}}$ and $\|\mathbf{s}^{(t+1)}\|_F \le \epsilon_{\text{dual}}$.

Choice of δ. The ADMM parameter $\delta > 0$ does not affect the optimizer of (16), but it influences convergence speed and numerical stability. In the $\mathbf{\Theta}$-update, a larger δ improves conditioning of $\mathbf{K}_\delta$, which is beneficial when $\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$ is ill-conditioned. In the $\mathbf{Z}$-update, a larger δ gives a smaller proximal threshold $\rho w_m/\delta$ for that iteration, but this changes only the ADMM trajectory, not the limiting optimizer. Overall, δ controls the balance between the primal and dual residuals. In all experiments, we use $\delta = 1$, though adaptive residual-balancing schemes can lead to faster convergence.

Computational cost. For fixed δ, the matrices $\mathbf{K}_\delta$ and $\mathbf{C}$ are constant across iterations and may be precomputed once. With a dense factorization of $\mathbf{K}_\delta$, the one-time cost is

$O(n\tilde{p}^2 + \tilde{p}^3 + n\tilde{p}q)$, and each subsequent $\mathbf{\Theta}$-update costs $O(\tilde{p}^2 q)$. The $\mathbf{Z}$- and $\mathbf{D}$-updates cost $O(\tilde{p}q)$, so the overall per-iteration cost is $O(\tilde{p}^2 q)$ after factorization.

When $\tilde{p} \gg n$, the Woodbury identity gives

$$\mathbf{K}_\delta^{-1} = \frac{1}{\delta}\mathbf{I}_{\tilde{p}} - \frac{2}{n\delta^2}\widetilde{\mathbf{X}}^\top \left(\mathbf{I}_n + \frac{2}{n\delta}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\right)^{-1} \widetilde{\mathbf{X}}.$$

Thus it suffices to factorize an $n \times n$ matrix. In that case the precomputation cost is $O(n^2\tilde{p} + n^3 + n\tilde{p}q)$, and each $\mathbf{\Theta}$-update costs $O(n^2 q + n\tilde{p}q)$. This is preferable, for example, in graph-sparse CCAR^3 when $\tilde{p} = |E| \gg n$.

Appendix C. Proofs: ordinary CCAR^3

C.1 Proof of Lemma 1

Proof Using the change of variables $\mathbf{V}_0 = \widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}}\mathbf{V}$ and $\mathbf{Y}_0 = \mathbf{Y}\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}}$, and the orthonormality of the columns of $\mathbf{V}_0$ (i.e. $\mathbf{V}_0^\top \mathbf{V}_0 = \mathbf{I}_r$), we transform the objective in Equation (5) into

$$\begin{aligned}
\|\mathbf{YV} - \mathbf{XU}\|_F^2 &= \|\mathbf{Y}_0\mathbf{V}_0 - \mathbf{XU}\|_F^2 \\
&= \underbrace{\mathsf{Tr}(\mathbf{V}_0^\top \mathbf{Y}_0^\top \mathbf{Y}_0 \mathbf{V}_0)}_{=rn} - 2\,\mathsf{Tr}(\mathbf{U}^\top \mathbf{X}^\top \mathbf{Y}_0 \mathbf{V}_0) + \underbrace{\mathsf{Tr}(\mathbf{U}^\top \mathbf{X}^\top \mathbf{XU})}_{=rn} \\
&= \underbrace{\mathsf{Tr}(\mathbf{Y}_0^\top \mathbf{Y}_0)}_{=qn} - 2\,\mathsf{Tr}(\mathbf{V}_0 \mathbf{U}^\top \mathbf{X}^\top \mathbf{Y}_0) + \underbrace{\mathsf{Tr}(\mathbf{V}_0 \mathbf{U}^\top \mathbf{X}^\top \mathbf{XU}\mathbf{V}_0^\top)}_{=rn} + (r-q)n \\
&= \|\mathbf{Y}_0 - \mathbf{XUV}_0^\top\|_F^2 + (r-q)n.
\end{aligned}$$

The additive constant $(r-q)n$ does not affect the minimizer, which proves the claim. ■

C.2 Proof of Theorem 2

Proof By the Law of Large Numbers, since p and q are fixed,

$$\widehat{\mathbf{\Sigma}}_X \xrightarrow{a.s.} \mathbf{\Sigma}_X, \qquad \widehat{\mathbf{\Sigma}}_Y \xrightarrow{a.s.} \mathbf{\Sigma}_Y, \qquad \widehat{\mathbf{\Sigma}}_{XY} \xrightarrow{a.s.} \mathbf{\Sigma}_{XY}.$$

Because $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$ are positive definite, the sample covariance matrices are positive definite for all sufficiently large n, almost surely. Moreover, by continuity of the inverse and inverse-square-root maps on the positive definite cone,

$$\widehat{\mathbf{\Sigma}}_X^{-1} \xrightarrow{a.s.} \mathbf{\Sigma}_X^{-1}, \qquad \widehat{\mathbf{\Sigma}}_Y^{-1/2} \xrightarrow{a.s.} \mathbf{\Sigma}_Y^{-1/2}.$$

Therefore,

$$\widehat{\mathbf{B}} = \widehat{\mathbf{\Sigma}}_X^{-1}\widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_Y^{-1/2} \xrightarrow{a.s.} \mathbf{B}^\star := \mathbf{\Sigma}_X^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-1/2}.$$

Under the canonical pair model,

$$\mathbf{B}^\star = \mathbf{\Sigma}_X^{-1}\mathbf{\Sigma}_X \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top \mathbf{\Sigma}_Y \mathbf{\Sigma}_Y^{-1/2} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}_0^\top, \qquad \mathbf{V}_0 := \mathbf{\Sigma}_Y^{1/2}\mathbf{V}.$$

Since

$$\mathbf{V}_0^\top \mathbf{V}_0 = \mathbf{V}^\top \mathbf{\Sigma}_Y \mathbf{V} = \mathbf{I}_r,$$

the columns of $\mathbf{V}_0$ are orthonormal. We also note that $\mathbf{U}$ has full column rank because $\mathbf{U}^\top \mathbf{\Sigma}_X \mathbf{U} = \mathbf{I}_r$. Hence

$$\operatorname{rank}(\mathbf{B}^\star) = r.$$

Moreover,

$$\sigma_r(\mathbf{B}^\star) = \sigma_r(\mathbf{U}\mathbf{\Lambda}\mathbf{V}_0^\top) \geq \sigma_r(\mathbf{U})\lambda_r > 0.$$

Thus the r-dimensional left and right singular subspaces of $\mathbf{B}^\star$ are separated from the null singular subspace by a positive gap.

Let $\mathbf{B}^\star = \mathbf{U}_B \mathbf{\Lambda}_B \mathbf{V}_B^\top$ be the rank-r singular value decomposition of $\mathbf{B}^\star$, where $\mathbf{U}_B \in \mathbb{R}^{p\times r}$ and $\mathbf{V}_B \in \mathbb{R}^{q\times r}$ have orthonormal columns. Since

$$\mathbf{B}^\star = \mathbf{U}\mathbf{\Lambda}\mathbf{V}_0^\top,$$

we have

$$\operatorname{span}(\mathbf{U}_B) = \operatorname{span}(\mathbf{U}), \qquad \operatorname{span}(\mathbf{V}_B) = \operatorname{span}(\mathbf{V}_0).$$

Because both $\mathbf{V}_B$ and $\mathbf{V}_0$ have orthonormal columns and span the same subspace, there exists an orthogonal matrix $\mathbf{Q}_Y \in \mathcal{O}_r$ such that

$$\mathbf{V}_B = \mathbf{V}_0 \mathbf{Q}_Y.$$

Similarly, define

$$\mathbf{U}_\diamond := \mathbf{U}_B \left(\mathbf{U}_B^\top \mathbf{\Sigma}_X \mathbf{U}_B\right)^{-1/2}.$$

Then

$$\mathbf{U}_\diamond^\top \mathbf{\Sigma}_X \mathbf{U}_\diamond = \mathbf{I}_r,$$

and $\operatorname{span}(\mathbf{U}_\diamond) = \operatorname{span}(\mathbf{U})$. Since $\mathbf{U}$ is also $\mathbf{\Sigma}_X$-orthonormal, there exists $\mathbf{Q}_X \in \mathcal{O}_r$ such that

$$\mathbf{U}_\diamond = \mathbf{U}\mathbf{Q}_X.$$

To prove consistency, we apply a standard Davis–Kahan/Wedin singular-subspace perturbation bound to $\widehat{\mathbf{B}}$ and $\mathbf{B}^\star$. Since $\widehat{\mathbf{B}} \to \mathbf{B}^\star$ in Frobenius norm almost surely and $\sigma_r(\mathbf{B}^\star) > 0$, there exist sequences $\mathbf{O}_n, \widetilde{\mathbf{O}}_n \in \mathcal{O}_r$ such that

$$\|\widehat{\mathbf{U}}_{\hat{B}} - \mathbf{U}_B \mathbf{O}_n\|_F \xrightarrow{a.s.} 0, \qquad \|\widehat{\mathbf{V}}_{\hat{B}} - \mathbf{V}_B \widetilde{\mathbf{O}}_n\|_F \xrightarrow{a.s.} 0.$$

We show that this implies the right singular-vector convergence to the estimator $\widehat{\mathbf{V}}$. Since $\mathbf{V}_B = \mathbf{V}_0\mathbf{Q}_Y$ and $\mathbf{V}_0 = \mathbf{\Sigma}_Y^{1/2}\mathbf{V}$, we have

$$\mathbf{\Sigma}_Y^{-1/2}\mathbf{V}_B = \mathbf{V}\mathbf{Q}_Y.$$

Therefore,

$$\begin{aligned}
\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{Q}_Y\widetilde{\mathbf{O}}_n\|_F &= \|\widehat{\mathbf{\Sigma}}_Y^{-1/2}\widehat{\mathbf{V}}_{\hat{B}} - \mathbf{\Sigma}_Y^{-1/2}\mathbf{V}_B\widetilde{\mathbf{O}}_n\|_F \\
&\leq \|\widehat{\mathbf{\Sigma}}_Y^{-1/2}\|_{op}\|\widehat{\mathbf{V}}_{\hat{B}} - \mathbf{V}_B\widetilde{\mathbf{O}}_n\|_F \\
&\quad + \|\widehat{\mathbf{\Sigma}}_Y^{-1/2} - \mathbf{\Sigma}_Y^{-1/2}\|_{op}\|\mathbf{V}_B\|_F \xrightarrow{a.s.} 0.
\end{aligned}$$

Since $\mathbf{Q}_Y\widetilde{\mathbf{O}}_n \in \mathcal{O}_r$, this implies

$$\min_{\widetilde{\mathbf{O}}\in\mathcal{O}_r} \|\widehat{\mathbf{V}} - \mathbf{V}\widetilde{\mathbf{O}}\|_F \xrightarrow{a.s.} 0.$$

We now turn to the proof of consistency for $\mathbf{U}$. Let

$$\widehat{\mathbf{A}}_n := \widehat{\mathbf{U}}_{\hat{B}}^\top \widehat{\mathbf{\Sigma}}_X \widehat{\mathbf{U}}_{\hat{B}}, \qquad \mathbf{A} := \mathbf{U}_B^\top \mathbf{\Sigma}_X \mathbf{U}_B.$$

Since

$$\widehat{\mathbf{U}}_{\hat{B}} - \mathbf{U}_B\mathbf{O}_n \xrightarrow{a.s.} 0 \quad \text{and} \quad \widehat{\mathbf{\Sigma}}_X \xrightarrow{a.s.} \mathbf{\Sigma}_X,$$

we have

$$\widehat{\mathbf{A}}_n - \mathbf{O}_n^\top \mathbf{A}\mathbf{O}_n \xrightarrow{a.s.} 0.$$

The matrix $\mathbf{A}$ is positive definite. Hence, for all sufficiently large n, $\widehat{\mathbf{A}}_n$ is also positive definite almost surely, and by continuity of the inverse-square-root map,

$$\widehat{\mathbf{A}}_n^{-1/2} - \mathbf{O}_n^\top \mathbf{A}^{-1/2}\mathbf{O}_n \xrightarrow{a.s.} 0.$$

Equivalently,

$$\mathbf{O}_n\widehat{\mathbf{A}}_n^{-1/2} - \mathbf{A}^{-1/2}\mathbf{O}_n \xrightarrow{a.s.} 0.$$

Now

$$\widehat{\mathbf{U}} = \widehat{\mathbf{U}}_{\hat{B}}\widehat{\mathbf{A}}_n^{-1/2}, \qquad \mathbf{U}_\diamond = \mathbf{U}_B\mathbf{A}^{-1/2}.$$

Thus,

$$\begin{aligned}\|\widehat{\mathbf{U}} - \mathbf{U}_\diamond\mathbf{O}_n\|_F &= \|\widehat{\mathbf{U}}_{\hat{B}}\widehat{\mathbf{A}}_n^{-1/2} - \mathbf{U}_B\mathbf{A}^{-1/2}\mathbf{O}_n\|_F \\ &\le \|\widehat{\mathbf{U}}_{\hat{B}} - \mathbf{U}_B\mathbf{O}_n\|_F\|\widehat{\mathbf{A}}_n^{-1/2}\|_{op} \\ &\quad + \|\mathbf{U}_B\|_{op}\|\mathbf{O}_n\widehat{\mathbf{A}}_n^{-1/2} - \mathbf{A}^{-1/2}\mathbf{O}_n\|_F \xrightarrow{a.s.} 0.\end{aligned}$$

Since $\mathbf{U}_\diamond = \mathbf{U}\mathbf{Q}_X$, we obtain

$$\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{Q}_X\mathbf{O}_n\|_F \xrightarrow{a.s.} 0.$$

Because $\mathbf{Q}_X\mathbf{O}_n \in \mathcal{O}_r$, it follows that

$$\min_{\mathbf{O}\in\mathcal{O}_r} \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{O}\|_F \xrightarrow{a.s.} 0.$$

This completes the proof. ■

Appendix D. Proofs: sparse CCAR^3

D.1 Useful lemmas

This subsection provides a list of the technical lemmas adapted from Gao et al. (2017) and Gao et al. (2015) that we used in this manuscript to derive our theoretical results.

Lemma 5 (Adaptation of Lemma 3 in Gao et al. (2015)) *Assume* $\frac{s_u+q}{n} < c_0$ *for* $c_0 \in (0,1)$. *Consider a deterministic set* $T_u \in [p]$ *with* $|T_u| = s_u$. *For any constant* $C' > 0$ *and assuming that* c_0 *is small enough, there exists a constant* $C > 0$ *that depends solely on* C' *and* c_0, *such that:*

$$\begin{aligned}
\left\| [\widehat{\mathbf{\Sigma}}_X]_{T_u T_u} - [\mathbf{\Sigma}_X]_{T_u T_u} \right\|_{op}^2 &\leq \frac{CM^2}{n} s_u \\
\left\| [\widehat{\mathbf{\Sigma}}_X]^{\frac{1}{2}}_{T_u T_u} - [\mathbf{\Sigma}_X]^{\frac{1}{2}}_{T_u T_u} \right\|_{op}^2 &\leq \frac{CM^2}{n} s_u \\
\left\| \widehat{\mathbf{\Sigma}}_Y - \mathbf{\Sigma}_Y \right\|_{op}^2 &\leq \frac{CM^2}{n} q \\
\left\| \widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}} - \mathbf{\Sigma}_Y^{\frac{1}{2}} \right\|_{op}^2 &\leq \frac{CM^2}{n} q
\end{aligned} \tag{17}$$

with probability at least $1 - \exp\{-C' s_u\} - \exp\{-C' q\}$.

Proof We adapt Lemma 3 of Gao et al. (2015) to explicit the dependency on M. Note that the set T_u considered here is deterministic. Therefore, by Lemma 6.5 of Wainwright (2019), there are positive universal constants $\{c_j\}_{j=0}^3$ such that, for any $\delta \in (0,1)$:

$$\mathbb{P}\left[\frac{\left\| [\widehat{\mathbf{\Sigma}}_X]_{T_u T_u} - [\mathbf{\Sigma}_X]_{T_u T_u} \right\|_{op}}{\|\mathbf{\Sigma}_X\|_{op}} > c_1 \Big(\sqrt{\frac{s_u}{n}} + \frac{s_u}{n} \Big) + \delta \right] \leq c_2 \exp\big\{ - c_3 n (\delta \wedge \delta^2) \big\}.$$

Denote $c_2 = e^{s_u \gamma_2}$ and choose $\delta^2 = \frac{s_u}{n} \frac{C' + \gamma_2}{c_3}$. Assuming that c_0 is small enough (i.e. $c_0 \leq \frac{c_3}{C' + \gamma_2}$) so that $\delta \leq 1$ (and thus, $\delta^2 \leq \delta$) we have:

$$\mathbb{P}\left[\left\| [\widehat{\mathbf{\Sigma}}_X]_{T_u T_u} - [\mathbf{\Sigma}_X]_{T_u T_u} \right\|_{op} > c_1' M \sqrt{\frac{s_u}{n}} \right] \leq \exp\{-C s_u\}$$

for a choice of $c_1' = 2c_1 + \sqrt{\frac{C' + \gamma_2}{c_3}} \leq 2c_1 + \frac{1}{\sqrt{c_0}}$. The same result holds for $\mathbf{\Sigma}_Y$ replacing s_u by q. Finally, for the operator norm of the square roots, by Lemma 2 of the supplementary material of Gao et al. (2015), we have:

$$\left\| [\widehat{\mathbf{\Sigma}}_X]^{\frac{1}{2}}_{T_u T_u} - [\mathbf{\Sigma}_X]^{\frac{1}{2}}_{T_u T_u} \right\|_{op} \leq \frac{\left\| [\widehat{\mathbf{\Sigma}}_X]_{T_u T_u} - [\mathbf{\Sigma}_X]_{T_u T_u} \right\|_{op}}{\sigma_{\min}\Big([\widehat{\mathbf{\Sigma}}_X]^{\frac{1}{2}}_{T_u T_u} \Big) + \sigma_{\min}\Big([\mathbf{\Sigma}_X]^{\frac{1}{2}}_{T_u T_u} \Big)}$$

Let us assume for now that $\|\mathbf{\Sigma}_X\|_{op} = 1$, and take $M = 1$. Let $\mathcal{A}_X$ denote the event $\mathcal{A}_X = \Big\{ \left\| [\widehat{\mathbf{\Sigma}}_X]_{T_u T_u} - [\mathbf{\Sigma}_X]_{T_u T_u} \right\|_{op} \leq c_1' \sqrt{\frac{s_u}{n}} \Big\}$, with c_1' as previously defined. Then, on $\mathcal{A}_X$, by Weyl's inequality (Giraud, 2021),

$$\sigma_{\min}\Big([\widehat{\mathbf{\Sigma}}_X]^{\frac{1}{2}}_{T_u T_u} \Big) = \sigma_{\min}\Big([\widehat{\mathbf{\Sigma}}_X]_{T_u T_u} \Big)^{\frac{1}{2}} \geq \sqrt{\sigma_{\min}\Big([\mathbf{\Sigma}_X]_{T_u T_u} \Big) - c_1' \sqrt{\frac{s_u}{n}}}.$$

Therefore, as long as $c_1'\sqrt{\frac{s_u}{n}} \leq \frac{3}{4}\sigma_{\min}(\mathbf{\Sigma}_X)$, which happens as soon as $c_0 \leq \left(\frac{\frac{3}{4}\sigma_{\min}(\mathbf{\Sigma}_X)}{c_1'}\right)^2$, we have:

$$\left\|[\widehat{\mathbf{\Sigma}}_X]^{\frac{1}{2}}_{T_uT_u} - [\mathbf{\Sigma}_X]^{\frac{1}{2}}_{T_uT_u}\right\|_{op} \leq \frac{c_1'\sqrt{\frac{s_u}{n}}}{\frac{3}{2}\sigma_{\min}\left([\mathbf{\Sigma}_X]^{\frac{1}{2}}_{T_uT_u}\right)}.$$

Scaling the previous inequality by $\left\|\mathbf{\Sigma}_X^{\frac{1}{2}}\right\|_{op}$, and using the fact that $\sigma_{\min}\left(\mathbf{\Sigma}_X^{\frac{1}{2}}\right) \geq \frac{1}{\sqrt{M}}$, we have:

$$\left\|[\widehat{\mathbf{\Sigma}}_X]^{\frac{1}{2}}_{T_uT_u} - [\mathbf{\Sigma}_X]^{\frac{1}{2}}_{T_uT_u}\right\|_{op} \leq Mc_1''\sqrt{\frac{s_u}{n}}.$$

The proof for $\mathbf{\Sigma}_Y^{\frac{1}{2}}$ is similar and considers instead the event

$$\mathcal{A}_Y = \Big\{ \left\|\widehat{\mathbf{\Sigma}}_Y - \mathbf{\Sigma}_Y\right\|_{op} > c_1'\sqrt{\frac{q}{n}}\Big\}.$$

■

The following lemma is a direct consequence of the previous lemma by applying a union bound.

Lemma 6 (Adapted from Lemma 12 in Gao et al. (2015)) *Let S_u be a set of indices of size $|S_u| = s_u$. Assume $\frac{q+s_u\log\left(\frac{ep}{s_u}\right)}{n} < c_0$ for some constant $c_0 \in (0,1)$. For any constant $C' > 0$, there exists some constant $C > 0$ only depending on C' and c_0 such that:*

$$\left\|[\widehat{\mathbf{\Sigma}}_X]_{S_uS_u} - [\mathbf{\Sigma}_X]_{S_uS_u}\right\|^2_{op} \leq \frac{CM^2}{n}s_u\log\left(\frac{ep}{s_u}\right)$$

$$\left\|[\widehat{\mathbf{\Sigma}}_X]^{\frac{1}{2}}_{S_uS_u} - [\mathbf{\Sigma}_X]^{\frac{1}{2}}_{S_uS_u}\right\|^2_{op} \leq \frac{CM^2}{n}s_u\log\left(\frac{ep}{s_u}\right)$$

$$\left\|\widehat{\mathbf{\Sigma}}_Y - \mathbf{\Sigma}_Y\right\|^2_{op} \leq \frac{CM^2}{n}q$$

$$\left\|\widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}} - \mathbf{\Sigma}_Y^{\frac{1}{2}}\right\|^2_{op} \leq \frac{CM^2}{n}q$$

with probability at least $1 - \exp\{-C'q\} - \exp\big\{-C's_u\log\left(\frac{ep}{s_u}\right)\big\}$.

Proof Bounding over all possible sets $T_u \subset [p]$ with cardinality at most s_u, we obtain:

$$\begin{aligned}
\mathbb{P}\left[\left\|[\widehat{\mathbf{\Sigma}}_X]_{S_uS_u} - [\mathbf{\Sigma}_X]_{S_uS_u}\right\|_{op} \geq t\right] &\leq \mathbb{P}\left[\max_{T_u\in[p]:\ |T_u|=s_u}\left\|[\widehat{\mathbf{\Sigma}}_X]_{T_uT_u} - [\mathbf{\Sigma}_X]_{T_uT_u}\right\|_{op} \geq t\right] \\
&\leq \sum_{T_u\in[p]}\mathbb{P}\left[\left\|[\widehat{\mathbf{\Sigma}}_X]_{T_uT_u} - [\mathbf{\Sigma}_X]_{T_uT_u}\right\|_{op} \geq t\right] \qquad (18) \\
&\leq \binom{p}{s_u}\mathbb{P}\left[\left\|[\widehat{\mathbf{\Sigma}}_X]_{T_uT_u} - [\mathbf{\Sigma}_X]_{T_uT_u}\right\|_{op} \geq t\right]
\end{aligned}$$

Therefore, noting that $\binom{p}{s_u} \leq \left(\frac{ep}{s_u}\right)^{s_u}$, for any $\delta > 0$, by the proof of Lemma 5:

$$\begin{aligned}\mathbb{P}\left[\left\|[\widehat{\mathbf{\Sigma}}_X]_{S_uS_u} - [\mathbf{\Sigma}_X]_{S_uS_u}\right\|_{op} \geq \|\mathbf{\Sigma}_X\|_{op}\left(c_1\Big(\sqrt{\frac{s_u}{n}} + \frac{s_u}{n}\Big) + \delta\right)\right] \\ \leq c_2 \exp\left\{s_u \log\left(\frac{ep}{s_u}\right) - c_3 n(\delta \wedge \delta^2)\right\}\end{aligned} \tag{19}$$

Writing $c_2 = e^{\gamma_2 s_u \log\left(\frac{ep}{s_u}\right)}$ for some constant γ_2, we choose $\delta^2 = \frac{C'+1+(\gamma_2)_+}{c_3} \cdot \frac{s_u \log\left(\frac{ep}{s_u}\right)}{n}$ for a constant $C' > 0$. Assuming that c_0 is small enough so that $\delta \leq 1$, which happens as soon as $c_0 \leq \frac{c_3}{C'+1+(\gamma_2)_+}$, we have:

$$\mathbb{P}\left[\left\|[\widehat{\mathbf{\Sigma}}_X]_{S_uS_u} - [\mathbf{\Sigma}_X]_{S_uS_u}\right\|_{op} \geq \|\mathbf{\Sigma}_X\|_{op}\left(c_1\Big(\sqrt{\frac{s_u}{n}} + \frac{s_u}{n}\Big) + \sqrt{\frac{c_3 + 1 + (\gamma_2)_+}{c_3} \cdot \frac{s_u \log\left(\frac{ep}{s_u}\right)}{n}}\right)\right]$$
$$\leq \exp\Big\{ - \big(C' + 1 + (\gamma_2)_+ - 1 - \gamma_2\big)\ s_u \log\left(\frac{ep}{s_u}\right)\Big\}$$

implying that

$$\mathbb{P}\left[\left\|[\widehat{\mathbf{\Sigma}}_X]_{S_uS_u} - [\mathbf{\Sigma}_X]_{S_uS_u}\right\|_{op} \geq c_1' \|\mathbf{\Sigma}_X\|_{op} \sqrt{\frac{s_u \log\left(\frac{ep}{s_u}\right)}{n}}\right] \leq \exp\Big\{ - C' s_u \log\left(\frac{ep}{s_u}\right)\Big\}$$

for $c_1' = 2c_1 + \sqrt{\frac{C'+1+(\gamma_2)_+}{c_3}} \leq 2c_1 + \frac{1}{\sqrt{c_0}}$. The proof for the square root matrices is identical to the proof of the previous lemma (Lemma 5), and also requires that c_0 is small enough in the sense that $c_0 \leq \left(\frac{\frac{3}{4}\sigma_{\min}(\mathbf{\Sigma}_X)}{c_1'}\right)^2$. ■

Lemma 7 (Adapted from Lemma 6.7 of Gao et al. (2017)) *Assume $\frac{q+\log p}{n} \leq c_0$ for some sufficiently small constant $c_0 \in (0,1)$. For $C' > 0$, there exists a constant $C > 0$ only depending on C' and c_0 such that*

$$\max_{j\in[p]} \left\| \big[\widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \widehat{\mathbf{\Sigma}}_X\mathbf{B}^*\big]_{j\cdot}\right\|_2 \leq CM^2\sqrt{\frac{q+\log p}{n}},$$

with probability at least

$$1 - \exp\{-C'q\} - \exp\{-C's_u\log(ep/s_u)\} - \exp\{-C'(q+\log p)\} - \exp\{-C'(r+\log p)\}.$$

Proof Recall that we defined $\mathbf{B}^*$ as $\mathbf{B}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}_Y^{\frac{1}{2}}$. By definition of $\mathbf{B}^*$, we know that $\mathbf{\Sigma}_X\mathbf{B}^* = \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}$. We thus have

$$\begin{aligned}\max_{j\in[p]} \left\| \big[\widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \widehat{\mathbf{\Sigma}}_X\mathbf{B}^*\big]_{j\cdot}\right\| \leq \underbrace{\max_{j\in[p]} \left\| \big[(\widehat{\mathbf{\Sigma}}_{XY} - \mathbf{\Sigma}_{XY})\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}}\big]_{j\cdot}\right\|}_{(A)} \\ + \underbrace{\max_{j\in[p]} \left\| \big[(\mathbf{\Sigma}_{XY}(\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_Y^{-\frac{1}{2}})\big]_{j\cdot}\right\|}_{(B)} + \underbrace{\max_{j\in[p]} \left\| \big[\widehat{\mathbf{\Sigma}}_X\mathbf{B}^* - \mathbf{\Sigma}_X\mathbf{B}^*\big]_{j\cdot}\right\|}_{(C)}\end{aligned} \tag{20}$$

For a given constant $C' > 0$, let $\mathcal{A}$ be the event:

$$\mathcal{A} = \Big\{ \Big\| [\widehat{\mathbf{\Sigma}}_X]^{\frac{1}{2}}_{S_u S_u} - [\mathbf{\Sigma}_X]^{\frac{1}{2}}_{S_u S_u} \Big\|^2_{op} \leq \frac{CM^2}{n} s_u \\ \Big\| \widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}} - \mathbf{\Sigma}_Y^{\frac{1}{2}} \Big\|^2_{op} \leq \frac{CM^2}{n} q \Big\} \tag{21}$$

where $S_u = \text{supp}(\mathbf{U})$, and where the constant C was chosen appropriately such that, by Lemma 5, we know that $\mathbb{P}[\mathcal{A}] \geq 1 - \exp\{-C'q\} - \exp\{-C's_u\}$.

Bound for (A). We note that

$$\begin{aligned} \max_{j\in[p]} \Big\| \big[(\widehat{\mathbf{\Sigma}}_{XY} - \mathbf{\Sigma}_{XY})\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}}\big]_{j\cdot} \Big\| &= \max_{j\in[p]} \Big\| \big[(\widehat{\mathbf{\Sigma}}_{XY} - \mathbf{\Sigma}_{XY})\mathbf{\Sigma}_Y^{-\frac{1}{2}}\mathbf{\Sigma}_Y^{\frac{1}{2}}\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}}\big]_{j\cdot} \Big\| \\ &\leq \max_{j\in[p]} \Big\| \big[(\widehat{\mathbf{\Sigma}}_{XY} - \mathbf{\Sigma}_{XY})\mathbf{\Sigma}_Y^{-\frac{1}{2}}\big]_{j\cdot} \Big\| \ \Big\| \mathbf{\Sigma}_Y^{\frac{1}{2}}\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} \Big\|_{op} \end{aligned}$$

By Weyl's theorem (Giraud, 2021), we know that on the event $\mathcal{A}$:

$$\forall i \in [q], \quad \big|\sigma_i\big(\widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}}\big) - \sigma_i\big(\mathbf{\Sigma}_Y^{\frac{1}{2}}\big)\big| \leq \Big\| \widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}} - \mathbf{\Sigma}_Y^{\frac{1}{2}} \Big\|_{op} \leq \sqrt{CM^2\frac{q}{n}}.$$

Therefore, writing $\delta_{q,n} = \sqrt{\frac{q}{n}}$ and assuming $\sqrt{C}M\delta_{q,n} \leq \frac{1}{2}\sigma_{\min}\big(\mathbf{\Sigma}_Y^{\frac{1}{2}}\big)$, we have:

$$\Big\| \widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} \Big\|_{op} = \frac{1}{\sigma_{\min}\big(\widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}}\big)} \leq \frac{1}{\sigma_{\min}\big(\mathbf{\Sigma}_Y^{\frac{1}{2}}\big) - \sqrt{C}M\delta_{q,n}} \leq \frac{2}{\sigma_{\min}\big(\mathbf{\Sigma}_Y^{\frac{1}{2}}\big)}. \tag{22}$$

Let $\mathbf{Z}_i \sim \mathcal{N}_p(0, \mathbf{\Sigma}_X)$, $\widetilde{\mathbf{Y}}_i \sim \mathcal{N}_q(0, \mathbf{I}_q)$, and $\mathbf{1}_j^\top \in \mathbb{R}^{1\times p}$ the vector indicator of row j (i.e. $[\mathbf{1}_j]_i = 0$ for all $i \neq j$, and $[\mathbf{1}_j]_j = 1$). Thus, we may write

$$\mathbf{1}_j^\top\big(\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\big) = \frac{1}{n}\sum_{i=1}^n \big(\mathbf{1}_j^\top\mathbf{Z}_i\widetilde{\mathbf{Y}}_i^\top - \mathbb{E}(\mathbf{1}_j^\top\mathbf{Z}_i\widetilde{\mathbf{Y}}_i^\top)\big).$$

Define for each $j \in [p]$ the matrix $\mathbf{H}_i^{(j)} = \begin{pmatrix} \mathbf{1}_j^\top\mathbf{Z}_i \\ \widetilde{\mathbf{Y}}_i \end{pmatrix}$. Since $\mathbf{1}_j^\top\mathbf{Z}_i\widetilde{\mathbf{Y}}_i^\top$ is a submatrix of $\mathbf{H}_i^{(j)}(\mathbf{H}_i^{(j)})^\top$, we have:

$$\max_{j\in[p]} \Big\| \mathbf{1}_j^\top(\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}) \Big\| \leq \max_{j\in[p]} \Big\| \frac{1}{n}\sum_{i=1}^n \big(\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top - \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top\big) \Big\|_{op} \tag{23}$$

Denoting $\mathcal{H}^{(j)} = \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top = \begin{pmatrix} \mathbf{1}_j^\top \mathbf{\Sigma}_X \mathbf{1}_j & \mathbf{1}_j^\top \mathbf{\Sigma}_{XY} \mathbf{\Sigma}_Y^{-\frac{1}{2}} \\ \mathbf{\Sigma}_Y^{-\frac{1}{2}} \mathbf{\Sigma}_{YX} \mathbf{1}_j & \mathbf{I}_q \end{pmatrix}$, this implies:

$$\begin{aligned} \left\|\mathcal{H}^{(j)}\right\|_{op} &= \sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q \\ u_1^2+\|u_2\|^2=1}} \mathbf{1}_j^\top \mathbf{\Sigma}_X \mathbf{1}_j u_1^2 + 2u_1 \mathbf{1}_j^\top \mathbf{\Sigma}_{XY} \mathbf{\Sigma}_Y^{-\frac{1}{2}} u_2 + u_2^\top \mathbf{I}_q u_2 \\ &\le \sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q \\ u_1^2+\|u_2\|^2=1}} M u_1^2 + 2\lambda_1 \sqrt{M} u_1 \|u_2\| + \|u_2\|^2 \\ &\le \sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q \\ u_1^2+\|u_2\|^2=1}} M u_1^2 + \lambda_1 \sqrt{M}(u_1^2 + \|u_2\|^2) + \|u_2\|^2 \\ &= M(1+\lambda_1) \le 2M, \end{aligned} \tag{24}$$

since by assumption $\|\mathbf{\Sigma}_X\|_{op} \le M$ and $\lambda_1 \le 1$. By Theorem 6.5 of Wainwright (2019), there exists universal positive constants $\{c_j\}_{j=1}^3$ such that, for any $t > 0$:

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n \big(\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top - \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top\big)\right\|_{op} \ge 4Mc_1\sqrt{\frac{1+q}{n}} + 2Mt\right] \le c_2 \exp\big\{-c_3 n(t \wedge t^2)\big\}.$$

Thus, by a simple union bound, we conclude that:

$$\mathbb{P}\left[\max_{j\in[p]}\left\|\frac{1}{n}\sum_{i=1}^n \big(\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top - \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top\big)\right\|_{op} \ge 4Mc_1\sqrt{\frac{1+q}{n}} + 2Mt\right] \le pc_2 \exp\big\{-c_3 n(t \wedge t^2)\big\}.$$

We write $c_2 = e^{\gamma_2(q+\log p)}$, for γ_2 a constant. Let $C' > 0$ be a constant and taking $t^2 = \frac{C'+1+(\gamma_2)_+}{c_3} \cdot \frac{q+\log p}{n}$, and assuming that c_0 is small enough (i.e. $c_0 \le \frac{c_3}{C'+1+(\gamma_2)_+}$) so that $t \le 1$, the previous equation implies

$$\begin{aligned} &\mathbb{P}\left[\max_{j\in[p]}\left\|\frac{1}{n}\sum_{i=1}^n \big(\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top - \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top\big)\right\|_{op}\right. \\ &\qquad \left. \ge 4Mc_1\sqrt{\frac{1+q}{n}} + 2M\sqrt{\frac{C'+1+(\gamma_2)_+}{c_3}\cdot\frac{q+\log p}{n}}\right] \\ &\qquad \le \exp\big\{\gamma_2(q+\log p) + \big(1 - C' - 1 - (\gamma_2)_+\big)\log p - \big(C'+1+(\gamma_2)_+\big)q\big\} \\ &\qquad = \exp\big\{-C'\log p - (C'+1)q\big\} \\ &\qquad \le \exp\big\{-C'(\log p + q))\big\}. \end{aligned} \tag{25}$$

Therefore, we can choose $c_1' = 4c_1 + 2\sqrt{\frac{C'+1+(\gamma_2)_+}{c_3}}$ such that, as soon as $p \ge 3$:

$$\begin{aligned} &\mathbb{P}\left[\max_{j\in[p]}\left\|\frac{1}{n}\sum_{i=1}^n \big(\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top - \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top\big)\right\|_{op} \ge c_1' M\sqrt{\frac{q+\log p}{n}}\right] \\ &\le \exp\big\{-C'(q+\log p)\big\} \end{aligned} \tag{26}$$

Let us denote as $\mathcal{A}_1$ the event:

$$\mathcal{A}_1 = \left\{ \max_{j\in[p]} \left\| \frac{1}{n}\sum_{i=1}^{n} (\mathbf{H}_i^{(j)}(\mathbf{H}_i^{(j)})^\top - \mathbb{E}\mathbf{H}_i^{(j)}(\mathbf{H}_i^{(j)})^\top) \right\|_{op} \geq c_1' M \sqrt{\frac{q+\log p}{n}} \right\}.$$

Therefore, with probability at least $1-\exp\{-C'(q+\log p)\}-\exp\{-C'q\}-\exp\{-C's_u\}$, the event $\mathcal{A}\cap\mathcal{A}_1^c$ occurs and

$$\max_{j\in[p]} \left\| \left[(\widehat{\mathbf{\Sigma}}_{XY}-\mathbf{\Sigma}_{XY})\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}}\right]_{j\cdot} \right\| \leq M c_1' \sqrt{\frac{q+\log p}{n}} \cdot \frac{\sigma_{\max}(\mathbf{\Sigma}_Y^{\frac{1}{2}})}{\sigma_{\min}(\mathbf{\Sigma}_Y^{\frac{1}{2}})} \leq M^2 c_1' \sqrt{\frac{q+\log p}{n}} \tag{27}$$

Bound (B). We have, on the event $\mathcal{A}$ (and provided that c_0 is small enough, in the sense of the proof of Lemma 7):

$$\begin{aligned}
\max_{j\in[p]} \left\| \left[(\mathbf{\Sigma}_{XY}(\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_Y^{-\frac{1}{2}})\right]_{j\cdot} \right\| &\leq \max_{j\in[p]} \left\| \left[\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right]_{j\cdot} \right\| \ \left\| \mathbf{\Sigma}_Y^{\frac{1}{2}}\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{I}_q \right\|_{op} \\
&\leq \left\| \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}} \right\|_{op} \ \left\| \mathbf{\Sigma}_Y^{\frac{1}{2}} - \widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}} \right\|_{op} \ \left\| \widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} \right\|_{op}. \\
&\leq \sqrt{M}\lambda_1 \cdot \sqrt{C} M \sqrt{\frac{q}{n}} \cdot \frac{2}{\sigma_{\min}(\mathbf{\Sigma}_Y^{\frac{1}{2}})} \\
&= 2\sqrt{C}\sqrt{\frac{q}{n}} M^{\frac{3}{2}} \cdot \frac{\lambda_1}{\sigma_{\min}(\mathbf{\Sigma}_Y^{\frac{1}{2}})}
\end{aligned} \tag{28}$$

Here we have used Inequality (22) to bound $\left\| \widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} \right\|_{op}$ on $\mathcal{A}$. We thus obtain:

$$\max_{j\in[p]} \left\| \left[(\mathbf{\Sigma}_{XY}(\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_Y^{-\frac{1}{2}})\right]_{j\cdot} \right\| \leq 2M^2\sqrt{C}\lambda_1 \cdot \sqrt{\frac{q}{n}} \tag{29}$$

Bound (C). By the proof of Lemma 6.7 in Gao et al. (2017):

$$\begin{aligned}
\max_{j\in[p]} \left\| \left[(\widehat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X)\mathbf{B}^*\right]_{j\cdot} \right\| &= \max_{j\in[p]} \left\| \left[(\widehat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X)\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}_Y^{\frac{1}{2}}\right]_{j\cdot} \right\| \\
&\leq \max_{j\in[p]} \frac{\left\| \left[(\widehat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X)\mathbf{U}\right]_{j\cdot} \right\|}{\|\mathbf{\Sigma}_X\|_{op}} \cdot \|\mathbf{\Sigma}_X\|_{op} \ \left\| \mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}_Y^{\frac{1}{2}} \right\|_{op} \\
&\leq C_2\sqrt{\frac{r+\log p}{n}} M\lambda_1
\end{aligned}$$

with probability at least $1-\exp\{-C'(r+\log p)\}$, for C' a constant that can be chosen to be arbitrarily large, and C_2 a constant that depends solely on C'. Let

$$\mathcal{A}_2 = \left\{ \max_{j\in[p]} \left\| \left[(\widehat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X)\mathbf{B}^*\right]_{j\cdot} \right\| \geq C_2\lambda_1 M \sqrt{\frac{r+\log p}{n}} \right\}.$$

Combined bound (A) + (B) + (C). Therefore, since $r < q$, and $\lambda < 1$, on the event $\mathcal{A} \cap \mathcal{A}_1^c \cap \mathcal{A}_2^c$, Inequality (20) becomes:

$$\begin{aligned}\max_{j\in[p]} \left\| \left[\widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \widehat{\mathbf{\Sigma}}_X\mathbf{B}^*\right]_{j\cdot}\right\| &\leq c_1' M^2\sqrt{\frac{q+\log p}{n}} + 2\sqrt{C}M^2\sqrt{\frac{q}{n}} + C_5 M\sqrt{\frac{r+\log p}{n}} \\ &\leq \widetilde{C}M^2\sqrt{\frac{q+\log p}{n}}\end{aligned} \tag{30}$$

for an appropriate choice of $\widetilde{C} = 3(c_1' \vee 2\sqrt{C} \vee C_2)$. This occurs with probability at least $1 - \exp\{-C'q\} - \exp\{-C's_u \log(ep/s_u)\} - \exp\{-C'(q+\log p)\} - \exp\{-C'(r+\log p)\}$. ■

D.2 Proof of Theorem 3

Proof The proof of Theorem 3 is based on the same regression argument as before; the only population object that must be used throughout is

$$\mathbf{B}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}_Y^{1/2} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}_0^\top,$$

because $\mathbf{Y}_0 = \mathbf{Y}\widehat{\mathbf{\Sigma}}_Y^{-1/2}$. In particular, under our model assumptions:

$$\mathbf{\Sigma}_X\mathbf{B}^* = \mathbf{\Sigma}_X\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}_Y^{1/2} = \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-1/2}.$$

Let $S_u = \text{supp}(\mathbf{B}^*) = \text{supp}(\mathbf{U})$ and define the high-probability score event

$$\mathcal{G} = \left\{ \max_{j\in[p]} \left\| \left[\widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_Y^{-1/2} - \widehat{\mathbf{\Sigma}}_X\mathbf{B}^*\right]_{j\cdot}\right\|_2 \leq CM^2\sqrt{\frac{q+\log p}{n}}\right\}.$$

let $\mathcal{A}$ be the event:

$$\begin{aligned}\mathcal{A} = \Big\{ \left\|[\widehat{\mathbf{\Sigma}}_X]_{S_uS_u} - [\mathbf{\Sigma}_X]_{S_uS_u}\right\|_{op}^2 &\leq \frac{CM^2}{n}s_u \\ \left\|\widehat{\mathbf{\Sigma}}_Y - \mathbf{\Sigma}_Y\right\|_{op}^2 &\leq \frac{CM^2}{n}q \\ \left\|\widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}} - \mathbf{\Sigma}_Y^{\frac{1}{2}}\right\|_{op}^2 &\leq \frac{CM^2}{n}q\Big\}\end{aligned} \tag{31}$$

By Lemma 5 and the uniform sparse-spectrum event $\mathcal{E}_{\text{sp}}$ of Lemma 6 used below, since the set S_u is deterministic, we may work on an event $\mathcal{A} \cap \mathcal{E}_{\text{sp}}$ satisfying

$$\mathbb{P}[\mathcal{A} \cap \mathcal{E}_{\text{sp}}] \geq 1 - \exp\{-C'q\} - \exp\{-C's_u\log(ep/s_u)\}.$$

Bound on $\left\|\widehat{\mathbf{\Sigma}}_X^{\frac{1}{2}}\mathbf{\Delta}\right\|_F$ and $\|\mathbf{\Delta}\|_F$. By the Basic Inequality:

$$\frac{1}{n}\left\|\mathbf{Y}_0 - \mathbf{X}\widehat{\mathbf{B}}\right\|_F^2 + \rho\|\widehat{\mathbf{B}}\|_{21} \leq \frac{1}{n}\left\|\mathbf{Y}_0 - \mathbf{X}\mathbf{B}^*\|_F^2 + \rho\|\mathbf{B}^*\|_{21}\right.$$

Therefore, writing $\mathbf{\Delta} = \widehat{\mathbf{B}} - \mathbf{B}^*$:

$$\begin{aligned}
\frac{1}{n}\|\mathbf{X}\mathbf{\Delta}\|_F^2 &\leq \frac{2}{n}\langle \mathbf{X}^\top \mathbf{Y}_0 - \mathbf{X}^\top \mathbf{X}\mathbf{B}^*, \mathbf{\Delta}\rangle + \rho(\|\mathbf{B}^*\|_{21} - \|\mathbf{\Delta} + \mathbf{B}^*\|_{21}) \\
\left\|\widehat{\mathbf{\Sigma}}_X^{\frac{1}{2}}\mathbf{\Delta}\right\|_F^2 &\leq 2\langle \widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \widehat{\mathbf{\Sigma}}_X\mathbf{B}^*, \mathbf{\Delta}\rangle + \rho(\|\mathbf{B}^*_{S_u\cdot}\|_{21} - \|\mathbf{B}^*_{S_u\cdot} + \mathbf{\Delta}_{S_u\cdot}\|_{21} - \|\mathbf{\Delta}_{S_u^c\cdot}\|_{21}) \\
&\leq 2\max_{j\in[p]}\left\|\left[\widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \widehat{\mathbf{\Sigma}}_X\mathbf{B}^*\right]_{j\cdot}\right\| \cdot \sum_{j=1}^{p}\|\mathbf{\Delta}_{j\cdot}\| + \rho(\|\mathbf{\Delta}_{S_u\cdot}\|_{21} - \|\mathbf{\Delta}_{S_u^c\cdot}\|_{21})
\end{aligned}$$

where the last line follows by an application of Hölder's inequality to tackle the inner product term, and the reverse triangle inequality for the ℓ_{21} terms. If we select ρ such that

$$\rho > 4\max_{j\in[p]}\left\|\left[\widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \widehat{\mathbf{\Sigma}}_X\mathbf{B}^*\right]_{j\cdot}\right\| \tag{32}$$

(and we will provide a value of ρ for which this holds in a later paragraph), then

$$\left\|\widehat{\mathbf{\Sigma}}_X^{\frac{1}{2}}\mathbf{\Delta}\right\|_F^2 \leq \frac{3\rho}{2}\|\mathbf{\Delta}_{S_u\cdot}\|_{21} - \frac{\rho}{2}\|\mathbf{\Delta}_{S_u^c\cdot}\|_{21} \tag{33}$$

Since $\sum_{j\in S_u}\|\mathbf{\Delta}_{j\cdot}\| \leq \left(s_u\sum_{j\in S_u}\|\mathbf{\Delta}_{j\cdot}\|^2\right)^{\frac{1}{2}}$, the upper bound becomes

$$\left\|\widehat{\mathbf{\Sigma}}_X^{\frac{1}{2}}\mathbf{\Delta}\right\|_F^2 \leq \frac{3\rho\sqrt{s_u}}{2}\left(\sum_{j\in S_u}\|\mathbf{\Delta}_{j\cdot}\|^2\right)^{\frac{1}{2}} \tag{34}$$

Note that Equation (34) also provides a cone inequality:

$$\sum_{j\in S_u^c}\|\mathbf{\Delta}_{j\cdot}\| \leq 3\sum_{j\in S_u}\|\mathbf{\Delta}_{j\cdot}\| \tag{35}$$

For an integer t, let the set $J_1 = \{j_1, \cdots, j_t\}$ in S_u^c correspond to the t rows with the largest ℓ_2 norm in $\mathbf{\Delta}$, and define the extended support $\widetilde{S}_u = S_u \cup J_1$. To proceed, we adopt the peeling strategy of Gao et al. (2017). We partition $\widetilde{S}_u^c$ into disjoint subsets $J_2, ..., J_K$ of size t (with $|J_K| \leq t$), such that J_k is the set of (row) indices corresponding to the t largest row ℓ_2 norms in $\mathbf{\Delta}$ outside $\widetilde{S}_u \bigcup_{j=2}^{k-1} J_j$. By the triangle inequality,

$$\begin{aligned}
\left\|\widehat{\mathbf{\Sigma}}_X^{\frac{1}{2}}\mathbf{\Delta}\right\|_F &\geq \left\|\widehat{\mathbf{\Sigma}}_X^{\frac{1}{2}}\mathbf{\Delta}_{\widetilde{S}_u\cdot}\right\|_F - \sum_{k=2}^{K}\left\|\widehat{\mathbf{\Sigma}}_X^{\frac{1}{2}}\mathbf{\Delta}_{J_k\cdot}\right\|_F \\
&\geq \sqrt{\sigma_{\min}^{\widehat{\mathbf{\Sigma}}_X}(s_u + t)} \cdot \left\|\mathbf{\Delta}_{\widetilde{S}_u\cdot}\right\|_F - \sqrt{\sigma_{\max}^{\widehat{\mathbf{\Sigma}}_X}(t)} \cdot \sum_{k=2}^{K}\|\mathbf{\Delta}_{J_k\cdot}\|_F\,.
\end{aligned}$$

where $\sigma_{\min}^{\widehat{\mathbf{\Sigma}}_X}(t + s_u)$ indicates the minimum eigenvalue of $\widehat{\mathbf{\Sigma}}_X$ on any set of size at most $s_u + t$, while $\sigma_{\max}^{\widehat{\mathbf{\Sigma}}_X}(t)$ denotes the maximum eigenvalue of $\widehat{\mathbf{\Sigma}}_X$ on any set of size at most t.

By construction, we also have:

$$\begin{aligned}\sum_{k=2}^{K}\|\mathbf{\Delta}_{J_k\cdot}\|_F &\leq \sqrt{t}\sum_{k=2}^{K}\max_{j\in J_k}\|\mathbf{\Delta}_{j\cdot}\| \leq \frac{1}{\sqrt{t}}\sum_{k=2}^{K}\sum_{j\in J_{k-1}}\|\mathbf{\Delta}_{j\cdot}\| \qquad (36)\\ &\leq \frac{1}{\sqrt{t}}\sum_{j\in S_u^c}\|\mathbf{\Delta}_{j\cdot}\| \leq \frac{3}{\sqrt{t}}\sum_{j\in S_u}\|\mathbf{\Delta}_{j\cdot}\|\\ &\leq 3\sqrt{\frac{s_u}{t}}\left(\sum_{j\in S_u}\|\mathbf{\Delta}_{j\cdot}\|^2\right)^{\frac{1}{2}} \leq 3\sqrt{\frac{s_u}{t}}\left\|\mathbf{\Delta}_{\widetilde{S}_u\cdot}\right\|_F. \qquad (37)\end{aligned}$$

In the above derivation, we have used the construction of J_k and the cone condition (35). Hence, $\left\|\widehat{\mathbf{\Sigma}}_X^{\frac{1}{2}}\mathbf{\Delta}\right\|_F \geq \kappa\left\|\mathbf{\Delta}_{\widetilde{S}_u\cdot}\right\|_F$ with $\kappa = \sqrt{\sigma_{\min}^{\widehat{\mathbf{\Sigma}}_X}(s_u+t)} - 3\sqrt{\frac{s_u}{t}\sigma_{\max}^{\widehat{\mathbf{\Sigma}}_X}(t)}$.

In view of Lemma 6 (on the event $\mathcal{E}_{\text{sp}}$) and by direct application of Weyl's inequality (Giraud, 2021), taking $t = c_1 s_u$ for some sufficiently large constant c_1, on the event $\mathcal{E}_{\text{sp}}$, κ can be lower-bounded by a positive constant κ_0 only depending on $\sigma_{\min}(\mathbf{\Sigma}_X)$, c_0 and c_1:

$$\begin{aligned}\kappa &\geq \sqrt{\frac{1}{M}} - M\sqrt{C\frac{(c_1+1)s_u\log\left(\frac{ep}{(c_1+1)s_u}\right)}{n}} - 3\sqrt{\frac{1}{c_1}}\cdot\left(\sqrt{M} + M\sqrt{C\frac{c_1 s_u\log\left(\frac{ep}{c_1 s_u}\right)}{n}}\right)\\ &\geq \sqrt{\frac{1}{M}} - 3\sqrt{\frac{M}{c_1}} - M\left(\sqrt{C\frac{(c_1+1)s_u\log\left(\frac{ep}{s_u}\right)}{n}} + 3\sqrt{C\frac{s_u\log\left(\frac{ep}{s_u}\right)}{n}}\right)\\ &\geq \frac{1}{2\sqrt{M}} - M\left(\sqrt{C\frac{(c_1+1)s_u\log\left(\frac{ep}{s_u}\right)}{n}} + 3\sqrt{C\frac{s_u\log\left(\frac{ep}{s_u}\right)}{n}}\right)\\ &\quad \text{choosing } c_1 \text{ sufficiently large depending only on } M\\ &\geq \frac{1}{2\sqrt{M}} - M\sqrt{Cc_0}\left(\sqrt{c_1+1}+3\right)\\ &\geq \frac{1}{4\sqrt{M}} \qquad \text{for } c_0 \text{ sufficiently small, depending only on } M \text{ and } c_1.\end{aligned}$$

Now, assuming that c_1 is chosen as above and c_0 is small enough,

$$\kappa \geq \gamma_0\sqrt{\sigma_{\min}(\mathbf{\Sigma}_X)} \geq \gamma_0\frac{1}{\sqrt{M}}$$

where $\gamma_0 \geq \frac{1}{4}$ is a constant. Combined with (34), this lower bound means that we have:

$$\left\|\mathbf{\Delta}_{\widetilde{S}_u\cdot}\right\|_F \leq C\frac{\rho\sqrt{s_u}}{2\kappa_0^2} \leq C\frac{\rho\sqrt{s_u}}{2\gamma_0^2\sigma_{\min}(\mathbf{\Sigma}_X)}. \qquad (38)$$

By (36)-(37), we have

$$\left\|\mathbf{\Delta}_{\widetilde{S}_u^c\cdot}\right\|_F \leq \sum_{k=2}^{K}\|\mathbf{\Delta}_{J_k\cdot}\|_F \leq 3\sqrt{\frac{s_u}{t}}\left\|\mathbf{\Delta}_{\widetilde{S}_u\cdot}\right\|_F \leq \frac{3}{\sqrt{c_1}}\left\|\mathbf{\Delta}_{\widetilde{S}_u\cdot}\right\|_F. \qquad (39)$$

Summing (38) and (39), we have $\|\mathbf{\Delta}\|_F \leq C_0 \frac{\sqrt{s_u}\rho}{\sigma_{\min}(\mathbf{\Sigma}_X)}$, therefore, Equation (34) becomes

$$\left\|\widehat{\mathbf{\Sigma}}_X^{\frac{1}{2}}\mathbf{\Delta}\right\|_F^2 \leq C_0 \frac{3\rho^2 s_u}{2\sigma_{\min}(\mathbf{\Sigma}_X)} \leq C_0 \frac{3M\rho^2 s_u}{2} \tag{40}$$

where C_0 is a generic constant that depends on γ_0.

By Lemma 7, we may choose $\rho \geq M^2 C_u \sqrt{\frac{q+\log p}{n}}$ for some large C_u so that (32) holds on the score event $\mathcal{G}$ defined above. We work on $\mathcal{A} \cap \mathcal{E}_{\text{sp}} \cap \mathcal{G}$; by a union bound, this event has at least the probability stated in the theorem. In this case:

$$\begin{aligned} \left\|\widehat{\mathbf{\Sigma}}_X^{\frac{1}{2}}\mathbf{\Delta}\right\|_F^2 &\leq C_0 M^5 \frac{3C_u^2 s_u}{2} \cdot \frac{q+\log p}{n} \leq C' M^5 s_u \cdot \frac{q+\log p}{n} \\ \|\mathbf{\Delta}\|_F &\leq C'' M^3 \sqrt{\frac{s_u(q+\log p)}{n}} \end{aligned} \tag{41}$$

with high probability, with $C' = \frac{3}{2}C_0 C_u^2$, and $C'' = C_0 C_u$. The second line implies $\|\mathbf{\Delta}\|_F^2 \leq (C'')^2 M^6 s_u(q+\log p)/n$. This concludes the proof of Equation 10.

Bound on $|\text{supp}\widehat{\mathbf{B}}|$. We note that, by the KKT conditions for all j, it holds

$$\frac{2}{n}\left[\mathbf{X}^\top(\mathbf{X}\widehat{\mathbf{B}} - \mathbf{Y}_0)\right]_{j\cdot} \in -\rho\, \partial\|\widehat{\mathbf{B}}_{j\cdot}\|_2$$

This implies that:

$$\begin{cases} \frac{1}{n}\left\|\left[\mathbf{X}^\top(\mathbf{X}\widehat{\mathbf{B}} - \mathbf{Y}_0)\right]_{j\cdot}\right\| = \frac{\rho}{2} & \text{if} \quad \left\|\widehat{\mathbf{B}}_{j\cdot}\right\| \neq 0 \\ \frac{1}{n}\left\|\left[\mathbf{X}^\top(\mathbf{X}\widehat{\mathbf{B}} - \mathbf{Y}_0)\right]_{j\cdot}\right\| \leq \frac{\rho}{2} & \text{if} \quad \left\|\widehat{\mathbf{B}}_{j\cdot}\right\| = 0 \end{cases} \tag{42}$$

Let $T = \text{supp}\widehat{\mathbf{B}}$. The following inequality therefore holds for any $j \in T$ on $\mathcal{G}$:

$$\frac{1}{n}\left\|\left[\mathbf{X}^\top(\mathbf{X}\mathbf{\Delta})\right]_{j\cdot}\right\| \geq \frac{1}{n}\left\|\left[\mathbf{X}^\top(\mathbf{X}\widehat{\mathbf{B}} - \mathbf{Y}_0)\right]_{j\cdot}\right\| - \frac{1}{n}\left\|\left[\mathbf{X}^\top(\mathbf{Y}_0 - \mathbf{X}\mathbf{B}^*)\right]_{j\cdot}\right\| \geq \frac{\rho}{4} \tag{43}$$

Thus

$$\sum_{j\in\text{supp}\widehat{\mathbf{B}}} \frac{1}{n^2}\left\|\left[\mathbf{X}^\top(\mathbf{X}\mathbf{\Delta})\right]_{j\cdot}\right\|^2 \geq \frac{\rho^2}{16}|T|.$$

Let

$$\widehat{M} = \max_{J\subset[p]:\ |J|=|T|} \left\|[\widehat{\mathbf{\Sigma}}_X]_{JJ}\right\|_{op}.$$

Let $\mathbf{X}_T$ denote the submatrix of $\mathbf{X}$ with columns indexed by T. The previous inequality implies

$$\begin{aligned} |T| &\leq \frac{16}{\rho^2}\frac{1}{n^2}\left\|\mathbf{X}_T^\top\mathbf{X}\mathbf{\Delta}\right\|_F^2 \leq \frac{16}{\rho^2}\frac{\left\|\mathbf{X}_T^\top\mathbf{X}_T\right\|_{op}}{n^2}\|\mathbf{X}\mathbf{\Delta}\|_F^2 \\ &\leq \frac{16\widehat{M}}{\rho^2}\left\|\widehat{\mathbf{\Sigma}}_X^{\frac{1}{2}}\mathbf{\Delta}\right\|_F^2 \leq \frac{16\widehat{M}}{\rho^2}C_0\frac{3M\rho^2 s_u}{2} \leq C_1\widehat{M}Ms_u \end{aligned} \tag{44}$$

for a choice of constant C_1 large enough. By Weyl's inequality, combined with the uniform sparse-spectrum event in Lemma 6, we have, simultaneously for all relevant sets J,

$$\widehat{M} \leq M + CM\sqrt{|T|\frac{\log\left(\frac{ep}{|T|}\right)}{n}}.$$

This yields

$$|T| \leq C_1 M^2 s_u + C_1 M^2 s_u \sqrt{|T|\frac{\log\left(\frac{ep}{|T|}\right)}{n}}.$$

Solving this one-dimensional inequality on the same sparse-spectrum event gives

$$|T| \leq CM^2 s_u,$$

provided c_0 in $s_u \log(ep/s_u)/n \leq c_0$ is sufficiently small. Using that $x/\log(ep/x)$ is increasing on the relevant range, if $|T| > 2C_1 M^2 s_u$, then the preceding display implies

$$|T| \leq 4C_1^2 M^4 s_u^2 \frac{\log(ep/|T|)}{n},$$

which contradicts $|T| > CM^2 s_u$ for a sufficiently large constant C under the same smallness assumption. Therefore, on the same event as above, we obtain:

$$|\mathrm{supp}\widehat{\mathbf{B}}| \lesssim M^2 s_u. \tag{45}$$

■

D.3 Proof of Theorem 4

Proof The first part of this argument follows the exact same path as the proof of Theorem 2. Let

$$\mathbf{V}_0 := \mathbf{\Sigma}_Y^{1/2}\mathbf{V}, \qquad \mathbf{B}^* := \mathbf{U}\mathbf{\Lambda}\mathbf{V}_0^\top, \qquad S := \mathrm{supp}(\mathbf{B}^*) = \mathrm{supp}(\mathbf{U}).$$

Then $\mathbf{V}_0^\top \mathbf{V}_0 = \mathbf{I}_r$ and

$$\mathrm{col}(\mathbf{B}^*) = \mathrm{col}(\mathbf{U}), \qquad \mathrm{row}(\mathbf{B}^*) = \mathrm{col}(\mathbf{V}_0).$$

Let

$$\mathbf{B}^* = \mathbf{U}_B \mathbf{D}_B \mathbf{V}_B^\top$$

be the compact singular value decomposition of $\mathbf{B}^*$, where $\mathbf{U}_B$ and $\mathbf{V}_B$ have orthonormal columns. Since $\mathbf{V}_0$ is orthonormal and spans the same subspace as $\mathbf{V}_B$, there exists $\mathbf{Q}_Y \in \mathcal{O}_r$ such that

$$\mathbf{V}_B = \mathbf{V}_0 \mathbf{Q}_Y.$$

Similarly, define

$$\mathbf{U}_\diamond := \mathbf{U}_B (\mathbf{U}_B^\top \mathbf{\Sigma}_X \mathbf{U}_B)^{-1/2}.$$

Then $\mathbf{U}_\diamond^\top \mathbf{\Sigma}_X \mathbf{U}_\diamond = \mathbf{I}_r$ and $\mathrm{col}(\mathbf{U}_\diamond) = \mathrm{col}(\mathbf{U})$. Since $\mathbf{U}^\top \mathbf{\Sigma}_X \mathbf{U} = \mathbf{I}_r$, there exists $\mathbf{Q}_X \in \mathcal{O}_r$ such that

$$\mathbf{U}_\diamond = \mathbf{U}\mathbf{Q}_X. \tag{46}$$

Let $\mathbf{\Delta} := \widehat{\mathbf{B}} - \mathbf{B}^*$. We first make the probability bookkeeping explicit. Fix the target constant $C' > 0$ in the theorem statement and apply the concentration bounds below with an auxiliary constant $C_\star > C'$ chosen sufficiently large; after the final union bound we relabel the resulting exponent by C'.

Let $\mathcal{A}_Y$ denote the event:

$$\mathcal{A}_Y = \Big\{ \Big\| \widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}} - \mathbf{\Sigma}_Y^{\frac{1}{2}} \Big\|_{op}^2 \le \frac{CM^2 q}{n}, \quad \Big\| \widehat{\mathbf{\Sigma}}_Y^{-1/2} \Big\|_{op} \le 2\sqrt{M}, \quad \Big\| \widehat{\mathbf{\Sigma}}_Y^{-1/2} - \mathbf{\Sigma}_Y^{-1/2} \Big\|_{op} \le CM^2 \sqrt{\frac{q}{n}} \Big\}. \tag{47}$$

By Lemma 5 and the identity

$$\widehat{\mathbf{\Sigma}}_Y^{-1/2} - \mathbf{\Sigma}_Y^{-1/2} = \widehat{\mathbf{\Sigma}}_Y^{-1/2} (\mathbf{\Sigma}_Y^{1/2} - \widehat{\mathbf{\Sigma}}_Y^{1/2}) \mathbf{\Sigma}_Y^{-1/2},$$

we have $\mathbb{P}(\mathcal{A}_Y^c) \le \exp\{-C_\star q\}$, provided c_0 is sufficiently small.

Let $\mathcal{G}$ denote the event:

$$\mathcal{G} = \Big\{ \Big\| \widehat{\mathbf{\Sigma}}_X^{\frac{1}{2}} \mathbf{\Delta} \Big\|_F^2 \le C_1 M^5 s_u \frac{q + \log p}{n}, \quad \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_F^2 \le C_2 M^6 s_u \frac{q + \log p}{n}, \quad |\mathrm{supp}(\widehat{\mathbf{B}})| \le C_J M^2 s_u \Big\}.$$

By Theorem 3, applied with exponent $C_\star$,

$$\mathbb{P}(\mathcal{G}^c) \le \exp\{-C_\star q\} + \exp\left\{ -C_\star s_u \log\left(\frac{ep}{s_u}\right) \right\} + \exp\{-C_\star(q + \log p)\} + \exp\{-C_\star(r + \log p)\}.$$

Let $s_J := \lceil (C_J M^2 + 1) s_u \rceil$ and define the uniform sparse-covariance event

$$\mathcal{H}_X = \left\{ \max_{J \subset [p] : |J| \le s_J} \Big\| [\widehat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X]_{JJ} \Big\|_{op} \le C_4 M^2 \sqrt{\frac{s_u \log(ep/s_u)}{n}} \right\}.$$

By Lemma 5 and a union bound over all subsets of size at most s_J,

$$\mathbb{P}(\mathcal{H}_X^c) \le \exp\{-C_\star s_J \log(ep/s_J)\} \le \exp\left\{ -C'_\star s_u \log\left(\frac{ep}{s_u}\right) \right\},$$

where the last inequality uses that M is treated as an absolute constant. Hence, for

$$\mathcal{E} := \mathcal{G} \cap \mathcal{A}_Y \cap \mathcal{H}_X,$$

choosing $C_\star$ large enough and decreasing constants if necessary gives

$$\mathbb{P}(\mathcal{E}) \ge 1 - \exp\{-C' q\} - \exp\left\{ -C' s_u \log\left(\frac{ep}{s_u}\right) \right\} - \exp\{-C'(q + \log p)\} - \exp\{-C'(r + \log p)\}.$$

All bounds below are proved on the event $\mathcal{E}$.

Step 1: perturbation of the singular subspaces of $\widehat{\mathbf{B}}$. By Theorem 3, on the event $\mathcal{E}$,

$$\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_F \le C_1 M^3 \sqrt{s_u \frac{q + \log p}{n}}, \qquad |\mathrm{supp}(\widehat{\mathbf{B}})| \lesssim M^2 s_u. \tag{48}$$

Moreover,

$$\sigma_r(\mathbf{B}^*) = \sigma_r(\mathbf{U}\mathbf{\Lambda}\mathbf{V}_0^\top) \ge \sigma_r(\mathbf{U})\lambda_r \ge \frac{\lambda}{\sqrt{M}},$$

because $\mathbf{V}_0$ has orthonormal columns and $\mathbf{U}^\top \mathbf{\Sigma}_X \mathbf{U} = \mathbf{I}_r$ with $\|\mathbf{\Sigma}_X\|_{op} \le M$ implies $\mathbf{I}_r = \mathbf{U}^\top \mathbf{\Sigma}_X \mathbf{U} \preceq M \mathbf{U}^\top \mathbf{U}$, hence $\sigma_r(\mathbf{U}) \ge M^{-1/2}$. For c_0 sufficiently small, the right-hand side of (48) is smaller than a fixed fraction of $\sigma_r(\mathbf{B}^*)$. Therefore Wedin's theorem (Lemma 8) applied directly to $\widehat{\mathbf{B}}$ and $\mathbf{B}^*$ gives orthogonal matrices $\mathbf{O}_X, \mathbf{O}_Y \in \mathcal{O}_r$ such that

$$\|\widehat{\mathbf{U}}_{\widehat{B}} - \mathbf{U}_B \mathbf{O}_X\|_F \vee \|\widehat{\mathbf{V}}_{\widehat{B}} - \mathbf{V}_B \mathbf{O}_Y\|_F \le C_2 \frac{M^{7/2}}{\lambda} \sqrt{r s_u \frac{q + \log p}{n}}. \tag{49}$$

Step 2: Error bound on $\widehat{\mathbf{V}}$. Since $\mathbf{V}_B = \mathbf{V}_0 \mathbf{Q}_Y$ and $\mathbf{V}_0 = \mathbf{\Sigma}_Y^{1/2} \mathbf{V}$,

$$\mathbf{\Sigma}_Y^{-1/2} \mathbf{V}_B = \mathbf{V} \mathbf{Q}_Y.$$

Furthermore, on $\mathcal{A}_Y$,

$$\left\|\widehat{\mathbf{\Sigma}}_Y^{-1/2}\right\|_{op} \le 2\sqrt{M}, \qquad \left\|\widehat{\mathbf{\Sigma}}_Y^{-1/2} - \mathbf{\Sigma}_Y^{-1/2}\right\|_{op} \le C M^2 \sqrt{\frac{q}{n}}. \tag{50}$$

Thus, using (49),

$$\begin{aligned}
\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{Q}_Y\mathbf{O}_Y\|_F &= \|\widehat{\mathbf{\Sigma}}_Y^{-1/2}\widehat{\mathbf{V}}_{\widehat{B}} - \mathbf{\Sigma}_Y^{-1/2}\mathbf{V}_B\mathbf{O}_Y\|_F \\
&\le \|\widehat{\mathbf{\Sigma}}_Y^{-1/2}\|_{op}\|\widehat{\mathbf{V}}_{\widehat{B}} - \mathbf{V}_B\mathbf{O}_Y\|_F + \|\widehat{\mathbf{\Sigma}}_Y^{-1/2} - \mathbf{\Sigma}_Y^{-1/2}\|_{op}\|\mathbf{V}_B\|_F \\
&\le 2\sqrt{M}\|\widehat{\mathbf{V}}_{\widehat{B}} - \mathbf{V}_B\mathbf{O}_Y\|_F + CM^2\sqrt{\frac{q}{n}}\,\|\mathbf{V}_B\|_F \\
&\le C_3 \frac{M^{11/2}}{\lambda}\sqrt{r s_u \frac{q + \log p}{n}}.
\end{aligned}$$

Taking $\mathbf{O} = \mathbf{Q}_Y \mathbf{O}_Y \in \mathcal{O}_r$ proves the desired bound for $\widehat{\mathbf{V}}$ on the event $\mathcal{E}$.

Step 3: Error bound on $\widehat{\mathbf{U}}$. Let

$$\widehat{\mathbf{A}} := \widehat{\mathbf{U}}_{\widehat{B}}^\top \widehat{\mathbf{\Sigma}}_X \widehat{\mathbf{U}}_{\widehat{B}}, \qquad \mathbf{A} := \mathbf{U}_B^\top \mathbf{\Sigma}_X \mathbf{U}_B.$$

The matrix $\mathbf{A}$ is positive definite, with eigenvalues in $[1/M, M]$. Let $J := \mathrm{supp}(\widehat{\mathbf{B}}) \cup S$. On the event $\mathcal{E}$, $|J| \le s_J$. Since the columns of $\widehat{\mathbf{U}}_{\widehat{B}}$ are supported on $\mathrm{supp}(\widehat{\mathbf{B}})$ and the columns of $\mathbf{U}_B$ are supported on S, the event $\mathcal{H}_X$ gives

$$\|[\widehat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X]_{JJ}\|_{op} \le C_4 M^2 \sqrt{\frac{s_u \log(ep/s_u)}{n}}.$$

Together with (49) and $\left\|[\widehat{\mathbf{\Sigma}}_X]_{JJ}\right\|_{op} \le CM$ on $\mathcal{E}$, this implies

$$\|\widehat{\mathbf{A}} - \mathbf{O}_X^\top \mathbf{A}\mathbf{O}_X\|_F \le C_5 \frac{M^{9/2}}{\lambda}\sqrt{r s_u \frac{q+\log p}{n}}. \tag{51}$$

Let

$$\widetilde{\mathbf{A}} := \mathbf{O}_X^\top \mathbf{A}\mathbf{O}_X.$$

Since the eigenvalues of $\mathbf{A}$ lie in $[1/M, M]$, the same is true for $\widetilde{\mathbf{A}}$. Moreover, by (51) and the smallness assumption on c_0, we have

$$\lambda_{\min}(\widehat{\mathbf{A}}) \ge \frac{1}{2M}, \qquad \lambda_{\min}(\widetilde{\mathbf{A}}) \ge \frac{1}{M}.$$

In particular, for c_0 sufficiently small, $\widehat{\mathbf{A}}$ is therefore positive definite. We now use the standard inverse-square-root perturbation bound for positive definite matrices: if $\mathbf{H}_1, \mathbf{H}_2 \succ 0$ and

$$\lambda_{\min}(\mathbf{H}_1) \wedge \lambda_{\min}(\mathbf{H}_2) \ge a,$$

then

$$\|\mathbf{H}_1^{-1/2} - \mathbf{H}_2^{-1/2}\|_F \le \frac{1}{2a^{3/2}}\|\mathbf{H}_1 - \mathbf{H}_2\|_F.$$

For completeness, this follows from the resolvent representation

$$\mathbf{H}^{-1/2} = \frac{1}{\pi}\int_0^\infty t^{-1/2}(\mathbf{H} + t\mathbf{I})^{-1}\,dt,$$

which, using the identity $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$ applied to $\mathbf{A} = \mathbf{H}_1 + t\mathbf{I}$ and $\mathbf{B} = \mathbf{H}_2 + t\mathbf{I}$, gives

$$\mathbf{H}_1^{-1/2} - \mathbf{H}_2^{-1/2} = \frac{1}{\pi}\int_0^\infty t^{-1/2}(\mathbf{H}_1 + t\mathbf{I})^{-1}(\mathbf{H}_2 - \mathbf{H}_1)(\mathbf{H}_2 + t\mathbf{I})^{-1}\,dt.$$

Therefore,

$$\begin{aligned}\|\mathbf{H}_1^{-1/2} - \mathbf{H}_2^{-1/2}\|_F &\le \frac{1}{\pi}\|\mathbf{H}_1 - \mathbf{H}_2\|_F \int_0^\infty \frac{t^{-1/2}}{(a+t)^2}\,dt \\ &= \frac{1}{2a^{3/2}}\|\mathbf{H}_1 - \mathbf{H}_2\|_F.\end{aligned}$$

Applying this bound with $\mathbf{H}_1 = \widehat{\mathbf{A}}$, $\mathbf{H}_2 = \widetilde{\mathbf{A}}$, and $a = 1/(2M)$ yields

$$\|\widehat{\mathbf{A}}^{-1/2} - \widetilde{\mathbf{A}}^{-1/2}\|_F \le CM^{3/2}\|\widehat{\mathbf{A}} - \widetilde{\mathbf{A}}\|_F.$$

Finally, by orthogonal invariance of the functional calculus,

$$\widetilde{\mathbf{A}}^{-1/2} = (\mathbf{O}_X^\top \mathbf{A}\mathbf{O}_X)^{-1/2} = \mathbf{O}_X^\top \mathbf{A}^{-1/2}\mathbf{O}_X.$$

Hence

$$\|\widehat{\mathbf{A}}^{-1/2} - \mathbf{O}_X^\top \mathbf{A}^{-1/2}\mathbf{O}_X\|_F \le CM^{3/2}\|\widehat{\mathbf{A}} - \mathbf{O}_X^\top \mathbf{A}\mathbf{O}_X\|_F.$$

Combining this with (51) gives

$$\|\widehat{\mathbf{A}}^{-1/2} - \mathbf{O}_X^\top \mathbf{A}^{-1/2}\mathbf{O}_X\|_F \le C_6 \frac{M^6}{\lambda}\sqrt{r s_u \frac{q+\log p}{n}},$$

provided that (51) is stated with the corresponding $M^{9/2}/\lambda$ rate.

By the Lipschitz continuity of the inverse-square-root map on matrices whose eigenvalues are bounded below by $1/(2M)$,

$$\|\widehat{\mathbf{A}}^{-1/2} - \mathbf{O}_X^\top \mathbf{A}^{-1/2}\mathbf{O}_X\|_F \le C_6 \frac{M^6}{\lambda}\sqrt{r s_u \frac{q+\log p}{n}}. \tag{52}$$

Now

$$\widehat{\mathbf{U}} = \widehat{\mathbf{U}}_{\widehat{B}}\widehat{\mathbf{A}}^{-1/2}, \qquad \mathbf{U}_\diamond = \mathbf{U}_B\mathbf{A}^{-1/2}.$$

Using (49) and (52),

$$\begin{aligned}\|\widehat{\mathbf{U}} - \mathbf{U}_\diamond\mathbf{O}_X\|_F &= \|\widehat{\mathbf{U}}_{\widehat{B}}\widehat{\mathbf{A}}^{-1/2} - \mathbf{U}_B\mathbf{A}^{-1/2}\mathbf{O}_X\|_F \\ &\le \|\widehat{\mathbf{U}}_{\widehat{B}} - \mathbf{U}_B\mathbf{O}_X\|_F\|\widehat{\mathbf{A}}^{-1/2}\|_{op} + \|\mathbf{U}_B\|_{op}\|\mathbf{O}_X\widehat{\mathbf{A}}^{-1/2} - \mathbf{A}^{-1/2}\mathbf{O}_X\|_F \\ &\le C_7\frac{M^6}{\lambda}\sqrt{r s_u \frac{q+\log p}{n}}.\end{aligned}$$

Finally, by (46),

$$\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{Q}_X\mathbf{O}_X\|_F \le C_7\frac{M^6}{\lambda}\sqrt{r s_u \frac{q+\log p}{n}}.$$

The last line follows from $\|U_B\|_{op} = 1$. Taking $\widetilde{\mathbf{O}} = \mathbf{Q}_X\mathbf{O}_X \in \mathcal{O}_r$ proves the bound for $\widehat{\mathbf{U}}$. Since all estimates above hold on $\mathcal{E}$ and $\mathcal{E}$ has the probability lower bound displayed above, the theorem follows. This completes the proof. ∎

Lemma 8 (Wedin's sin–Θ theorem, rank-r signal form) *Let $\mathbf{A}, \widehat{\mathbf{A}} \in \mathbb{R}^{p\times q}$, and suppose $\mathbf{A}$ has rank r with compact singular value decomposition*

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \qquad \sigma_r(\mathbf{A}) > 0,$$

where $\mathbf{U} \in \mathbb{R}^{p\times r}$ and $\mathbf{V} \in \mathbb{R}^{q\times r}$ have orthonormal columns. Let

$$\widehat{\mathbf{A}} = \widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^\top + \widehat{\mathbf{A}}_\perp$$

be a singular value decomposition of $\widehat{\mathbf{A}}$, where $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ contain the leading r left and right singular vectors. Set

$$\mathbf{E} = \widehat{\mathbf{A}} - \mathbf{A}.$$

If

$$\|\mathbf{E}\|_{op} \le \frac{1}{2}\sigma_r(\mathbf{A}),$$

then

$$\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\|_F \vee \|\sin\Theta(\widehat{\mathbf{V}}, \mathbf{V})\|_F \le C_W\frac{\|\mathbf{E}\|_F}{\sigma_r(\mathbf{A})},$$

for a universal numerical constant C_W. Consequently, there exist orthogonal matrices $\mathbf{O}_U, \mathbf{O}_V \in \mathcal{O}_r$ such that

$$\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{O}_U\|_F \vee \|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{O}_V\|_F \leq C'_W \frac{\|\mathbf{E}\|_F}{\sigma_r(\mathbf{A})},$$

where C'_W is another universal numerical constant.

Proof Let

$$\sigma := \sigma_r(\mathbf{A}) > 0, \qquad \mathbf{E} = \widehat{\mathbf{A}} - \mathbf{A}.$$

Let

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

be the compact singular value decomposition of $\mathbf{A}$, and extend $\mathbf{U}$ and $\mathbf{V}$ to orthogonal matrices $[\mathbf{U}, \mathbf{U}_\perp]$ and $[\mathbf{V}, \mathbf{V}_\perp]$. Then

$$\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\|_F = \|\mathbf{U}_\perp^\top \widehat{\mathbf{U}}\|_F, \qquad \|\sin\Theta(\widehat{\mathbf{V}}, \mathbf{V})\|_F = \|\mathbf{V}_\perp^\top \widehat{\mathbf{V}}\|_F.$$

By Weyl's inequality,

$$\sigma_r(\widehat{\mathbf{A}}) \geq \sigma_r(\mathbf{A}) - \|\widehat{\mathbf{A}} - \mathbf{A}\|_{op} \geq \frac{\sigma}{2}.$$

Thus the leading r singular values collected in $\widehat{\mathbf{D}}$ are bounded below by $\sigma/2$, and $\widehat{\mathbf{D}}$ is invertible with

$$\|\widehat{\mathbf{D}}^{-1}\|_{op} \leq \frac{2}{\sigma}.$$

We first bound the left singular subspace. Since

$$\widehat{\mathbf{A}}\widehat{\mathbf{V}} = \widehat{\mathbf{U}}\widehat{\mathbf{D}},$$

we have

$$\mathbf{A}\widehat{\mathbf{V}} = \widehat{\mathbf{U}}\widehat{\mathbf{D}} - \mathbf{E}\widehat{\mathbf{V}}.$$

Because the columns of $\mathbf{A}\widehat{\mathbf{V}}$ lie in $\mathrm{col}(\mathbf{U})$, multiplying by $\mathbf{U}_\perp^\top$ gives

$$\mathbf{0} = \mathbf{U}_\perp^\top \widehat{\mathbf{U}}\widehat{\mathbf{D}} - \mathbf{U}_\perp^\top \mathbf{E}\widehat{\mathbf{V}}.$$

Therefore,

$$\mathbf{U}_\perp^\top \widehat{\mathbf{U}} = \mathbf{U}_\perp^\top \mathbf{E}\widehat{\mathbf{V}}\widehat{\mathbf{D}}^{-1}.$$

Hence

$$\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\|_F = \|\mathbf{U}_\perp^\top \widehat{\mathbf{U}}\|_F \leq \|\mathbf{U}_\perp^\top \mathbf{E}\widehat{\mathbf{V}}\|_F \|\widehat{\mathbf{D}}^{-1}\|_{op} \leq \frac{2\|\mathbf{E}\|_F}{\sigma_r(\mathbf{A})}.$$

The right singular subspace is handled in the same way. Since

$$\widehat{\mathbf{A}}^\top \widehat{\mathbf{U}} = \widehat{\mathbf{V}}\widehat{\mathbf{D}},$$

we have

$$\mathbf{A}^\top \widehat{\mathbf{U}} = \widehat{\mathbf{V}}\widehat{\mathbf{D}} - \mathbf{E}^\top \widehat{\mathbf{U}}.$$

The columns of $\mathbf{A}^\top \widehat{\mathbf{U}}$ lie in $\mathrm{col}(\mathbf{V})$, so multiplying by $\mathbf{V}_\perp^\top$ yields

$$\mathbf{0} = \mathbf{V}_\perp^\top \widehat{\mathbf{V}} \widehat{\mathbf{D}} - \mathbf{V}_\perp^\top \mathbf{E}^\top \widehat{\mathbf{U}}.$$

Thus

$$\mathbf{V}_\perp^\top \widehat{\mathbf{V}} = \mathbf{V}_\perp^\top \mathbf{E}^\top \widehat{\mathbf{U}} \widehat{\mathbf{D}}^{-1},$$

and consequently

$$\|\sin\Theta(\widehat{\mathbf{V}}, \mathbf{V})\|_F = \|\mathbf{V}_\perp^\top \widehat{\mathbf{V}}\|_F \le \frac{2\|\mathbf{E}\|_F}{\sigma_r(\mathbf{A})}.$$

Combining the two displays gives

$$\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\|_F \vee \|\sin\Theta(\widehat{\mathbf{V}}, \mathbf{V})\|_F \le 2\frac{\|\mathbf{E}\|_F}{\sigma_r(\mathbf{A})}.$$

Thus the first conclusion holds with, for example, $C_W = 2$.

It remains to convert the sin–Θ bound into a bound after orthogonal alignment. For matrices with orthonormal columns, the Procrustes inequality gives

$$\min_{\mathbf{O}\in\mathcal{O}_r} \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{O}\|_F \le \sqrt{2}\,\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\|_F,$$

and similarly

$$\min_{\mathbf{O}\in\mathcal{O}_r} \|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{O}\|_F \le \sqrt{2}\,\|\sin\Theta(\widehat{\mathbf{V}}, \mathbf{V})\|_F.$$

Indeed, if the singular values of $\mathbf{U}^\top \widehat{\mathbf{U}}$ are $\cos\theta_1, \dots, \cos\theta_r$, then

$$\min_{\mathbf{O}\in\mathcal{O}_r} \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{O}\|_F^2 = 2\sum_{j=1}^r (1 - \cos\theta_j) \le 2\sum_{j=1}^r \sin^2\theta_j.$$

The same argument applies to $\mathbf{V}$ and $\widehat{\mathbf{V}}$.

Therefore, there exist $\mathbf{O}_U, \mathbf{O}_V \in \mathcal{O}_r$ such that

$$\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{O}_U\|_F \vee \|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{O}_V\|_F \le 2\sqrt{2}\,\frac{\|\mathbf{E}\|_F}{\sigma_r(\mathbf{A})}.$$

Thus the second conclusion holds with, for example, $C'_W = 2\sqrt{2}$. ■

Appendix E. Proofs: group-sparse CCAR^3

We first introduce notation for group supports. Let

$$\mathcal{G} = \{g_1, \dots, g_K\}$$

be a non-overlapping partition of $[p]$, and write

$$d_g = |g|, \qquad d_{\min} = \min_{g\in\mathcal{G}} d_g, \qquad d_{\max} = \max_{g\in\mathcal{G}} d_g, \qquad \eta = \frac{d_{\max}}{d_{\min}}.$$

For a collection of groups $\mathcal{H} \subseteq \mathcal{G}$, define

$$T(\mathcal{H}) = \bigcup_{g \in \mathcal{H}} g, \qquad d(\mathcal{H}) = \sum_{g \in \mathcal{H}} d_g.$$

For a matrix $\mathbf{B} \in \mathbb{R}^{p \times q}$, let

$$\mathrm{supp}_{\mathcal{G}}(\mathbf{B}) = \{g \in \mathcal{G} : \|\mathbf{B}_g\|_F > 0\},$$

where $\mathbf{B}_g$ denotes the submatrix of $\mathbf{B}$ formed by the rows indexed by g. We use the weighted group norm

$$\|\mathbf{B}\|_{\mathcal{G},21} = \sum_{g \in \mathcal{G}} \sqrt{d_g}\, \|\mathbf{B}_g\|_F.$$

Let

$$\mathbf{B}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{V}_0^\top, \qquad \mathbf{V}_0 = \mathbf{\Sigma}_Y^{1/2}\mathbf{V},$$

and define the true active group set

$$\mathcal{A} = \mathrm{supp}_{\mathcal{G}}(\mathbf{B}^*) = \{g \in \mathcal{G} : \|\mathbf{B}^*_g\|_F > 0\}.$$

Let

$$m_u = |\mathcal{A}|, \qquad s_u = d(\mathcal{A}) = \sum_{g \in \mathcal{A}} d_g.$$

Note that $m_u \le s_u / d_{\min}$. Define the group rate

$$R_{\mathcal{G}} = \frac{s_u q + (s_u/d_{\min}) \log K}{n}.$$

Theorem 9 (Group-regularized regression error with support control) *Consider the group-penalized estimator*

$$\widehat{\mathbf{B}} = \underset{\mathbf{B} \in \mathbb{R}^{p \times q}}{\mathrm{argmin}} \frac{1}{n} \|\mathbf{Y}_0 - \mathbf{X}\mathbf{B}\|_F^2 + \rho \sum_{g \in \mathcal{G}} \sqrt{d_g}\, \|\mathbf{B}_g\|_F.$$

Assume that $R_{\mathcal{G}} \le c_0$ for a sufficiently small constant $c_0 \in (0,1)$. Assume also that $\eta = d_{\max}/d_{\min}$ is bounded by an absolute constant; otherwise the constants below depend on η. Let

$$\mathbf{\Delta} = \widehat{\mathbf{B}} - \mathbf{B}^*.$$

Choose

$$\rho \ge C_u M^2 \left(\sqrt{\frac{q}{n}} + \sqrt{\frac{\log K}{d_{\min} n}} \right)$$

for a sufficiently large constant C_u. Then, for any $C' > 0$, with a probability of at least

$$1 - \exp(-C'q) - \exp\{-C'(s_u/d_{\min}) \log(eK d_{\min}/s_u)\} - \exp\{-C'(q + \log K)\} - \exp\{-C'(r + \log K)\},$$

there exist constants $C_1, C_2, C_3 > 0$, depending only on C', c_0, and the bounded group-size ratio, such that

$$\left\| \widehat{\mathbf{\Sigma}}_X^{1/2} \mathbf{\Delta} \right\|_F^2 \le C_1 M^5 R_{\mathcal{G}},$$

and

$$\|\mathbf{\Delta}\|_F^2 \le C_2 M^6 R_{\mathcal{G}}.$$

Moreover, with the same probability, the selected group support satisfies

$$d(\widehat{\mathcal{A}}) = \sum_{g\in\widehat{\mathcal{A}}} d_g \le C_3 M^2 s_u, \qquad |\widehat{\mathcal{A}}| \le C_3 M^2 \frac{s_u}{d_{\min}},$$

where

$$\widehat{\mathcal{A}} = \mathrm{supp}_{\mathcal{G}}(\widehat{\mathbf{B}}).$$

Proof The proof follows the same regression argument as in the sparse case, but the cone and peeling steps are performed at the group level. Let

$$\mathbf{Z} = \widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_Y^{-1/2} - \widehat{\mathbf{\Sigma}}_X \mathbf{B}^*.$$

We work on the group score event

$$\mathcal{E}_{\text{score}} = \left\{ \max_{g\in\mathcal{G}} \frac{\|\mathbf{Z}_g\|_F}{\sqrt{d_g}} \le \frac{\rho}{4} \right\}.$$

We first verify the group score event. Recall that

$$\mathbf{Z} = \widehat{\mathbf{\Sigma}}_{XY}\widehat{\mathbf{\Sigma}}_Y^{-1/2} - \widehat{\mathbf{\Sigma}}_X \mathbf{B}^*, \qquad \mathbf{B}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{V}_0^\top, \qquad \mathbf{V}_0 = \mathbf{\Sigma}_Y^{1/2}\mathbf{V}.$$

Since

$$\mathbf{\Sigma}_X \mathbf{B}^* = \mathbf{\Sigma}_X \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top \mathbf{\Sigma}_Y^{1/2} = \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-1/2},$$

we may decompose, for every group $g \in \mathcal{G}$,

$$\mathbf{Z}_g = \underbrace{\left[(\widehat{\mathbf{\Sigma}}_{XY} - \mathbf{\Sigma}_{XY})\widehat{\mathbf{\Sigma}}_Y^{-1/2}\right]_g}_{\mathbf{A}_g} + \underbrace{\left[\mathbf{\Sigma}_{XY}(\widehat{\mathbf{\Sigma}}_Y^{-1/2} - \mathbf{\Sigma}_Y^{-1/2})\right]_g}_{\mathbf{B}_g} - \underbrace{\left[(\widehat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X)\mathbf{B}^*\right]_g}_{\mathbf{C}_g}. \tag{E.0}$$

We work on the covariance event

$$\mathcal{E}_Y = \left\{ \|\widehat{\mathbf{\Sigma}}_Y^{1/2} - \mathbf{\Sigma}_Y^{1/2}\|_{\text{op}} \le CM\sqrt{\frac{q}{n}}, \quad \|\widehat{\mathbf{\Sigma}}_Y^{-1/2}\|_{\text{op}} \le 2\sqrt{M} \right\}.$$

By Lemma 5,

$$\mathbb{P}(\mathcal{E}_Y^c) \le \exp(-C'q),$$

provided $q/n \le c_0$ and c_0 is sufficiently small. On $\mathcal{E}_Y$,

$$\|\mathbf{\Sigma}_Y^{1/2}\widehat{\mathbf{\Sigma}}_Y^{-1/2}\|_{\text{op}} \le CM,$$

and

$$\|\mathbf{\Sigma}_Y^{1/2}\widehat{\mathbf{\Sigma}}_Y^{-1/2} - \mathbf{I}_q\|_{\text{op}} = \|(\mathbf{\Sigma}_Y^{1/2} - \widehat{\mathbf{\Sigma}}_Y^{1/2})\widehat{\mathbf{\Sigma}}_Y^{-1/2}\|_{\text{op}} \le CM^{3/2}\sqrt{\frac{q}{n}}. \tag{E.0a}$$

We now bound the three terms in $(E.0)$.

First, set

$$\widetilde{\mathbf{Y}}_i = \mathbf{\Sigma}_Y^{-1/2}\mathbf{Y}_i.$$

For a fixed group g, and for any matrix $\mathbf{H} \in \mathbb{R}^{d_g \times q}$ with $\|\mathbf{H}\|_F = 1$, the random variables

$$\xi_i(\mathbf{H}) = \left\langle \mathbf{X}_{i,g}\widetilde{\mathbf{Y}}_i^\top - \mathbb{E}(\mathbf{X}_{i,g}\widetilde{\mathbf{Y}}_i^\top), \mathbf{H} \right\rangle$$

are centered sub-exponential with

$$\|\xi_i(\mathbf{H})\|_{\psi_1} \leq CM.$$

Indeed, $(\mathbf{X}_{i,g}, \widetilde{\mathbf{Y}}_i)$ is jointly Gaussian and its covariance operator norm is bounded by CM. Bernstein's inequality, followed by a 1/4-net argument over the Frobenius unit sphere in $\mathbb{R}^{d_g \times q}$, gives, for every $t > 0$,

$$\mathbb{P}\left[\left\| \left[(\widehat{\mathbf{\Sigma}}_{XY} - \mathbf{\Sigma}_{XY})\mathbf{\Sigma}_Y^{-1/2} \right]_g \right\|_F > CM\sqrt{\frac{d_g q + t}{n}} \right] \leq \exp\{-c(d_g q + t)\}, \tag{E.0b}$$

where the linear Bernstein term is absorbed under the small-rate condition. Taking $t = \log K$ and union-bounding over $g \in \mathcal{G}$ gives

$$\max_{g \in \mathcal{G}} \frac{\left\| \left[(\widehat{\mathbf{\Sigma}}_{XY} - \mathbf{\Sigma}_{XY})\mathbf{\Sigma}_Y^{-1/2} \right]_g \right\|_F}{\sqrt{d_g}} \leq CM\left(\sqrt{\frac{q}{n}} + \sqrt{\frac{\log K}{d_{\min} n}} \right) \tag{E.0c}$$

with probability at least

$$1 - \exp\{-C'(q + \log K)\}.$$

Combining $(E.0c)$ with $\|\mathbf{\Sigma}_Y^{1/2}\widehat{\mathbf{\Sigma}}_Y^{-1/2}\|_{\text{op}} \leq CM$, we obtain

$$\max_{g \in \mathcal{G}} \frac{\|\mathbf{A}_g\|_F}{\sqrt{d_g}} \leq CM^2\left(\sqrt{\frac{q}{n}} + \sqrt{\frac{\log K}{d_{\min} n}} \right). \tag{E.0d}$$

Second, for the term $\mathbf{B}_g$, note that

$$\mathbf{B}_g = \left[\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-1/2} \right]_g \left(\mathbf{\Sigma}_Y^{1/2}\widehat{\mathbf{\Sigma}}_Y^{-1/2} - \mathbf{I}_q \right).$$

Moreover,

$$\frac{\|[\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-1/2}]_g\|_F}{\sqrt{d_g}} \leq \|\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-1/2}\|_{\text{op}} = \|\mathbf{\Sigma}_X \mathbf{U}\mathbf{\Lambda}\mathbf{V}_0^\top\|_{\text{op}} \leq \sqrt{M}.$$

Therefore, using $(E.0a)$,

$$\max_{g \in \mathcal{G}} \frac{\|\mathbf{B}_g\|_F}{\sqrt{d_g}} \leq CM^2\sqrt{\frac{q}{n}}. \tag{E.0e}$$

Third, for the term $\mathbf{C}_g$, since

$$\mathbf{B}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{V}_0^\top \qquad \text{and} \qquad \left\| \mathbf{\Lambda}\mathbf{V}_0^\top \right\|_{op} \leq 1,$$

we have

$$\|\mathbf{C}_g\|_F \le \left\| \left[(\widehat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X)\mathbf{U} \right]_g \right\|_F .$$

For a fixed group g, and for any $\mathbf{H} \in \mathbb{R}^{d_g \times r}$ with $\|\mathbf{H}\|_F = 1$, the variables

$$\zeta_i(\mathbf{H}) = \left\langle \mathbf{X}_{i,g}(\mathbf{U}^\top \mathbf{X}_i)^\top - \mathbb{E}\{\mathbf{X}_{i,g}(\mathbf{U}^\top \mathbf{X}_i)^\top\}, \mathbf{H} \right\rangle$$

are centered sub-exponential with

$$\|\zeta_i(\mathbf{H})\|_{\psi_1} \le CM,$$

because $(\mathbf{X}_{i,g}, \mathbf{U}^\top \mathbf{X}_i)$ is jointly Gaussian, the covariance of $\mathbf{X}_{i,g}$ has operator norm at most M, and

$$\mathrm{Cov}(\mathbf{U}^\top \mathbf{X}_i) = \mathbf{U}^\top \mathbf{\Sigma}_X \mathbf{U} = I_r.$$

Another Bernstein plus net argument gives, for every $t > 0$,

$$\mathbb{P}\left[\left\| \left[(\widehat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X)\mathbf{U} \right]_g \right\|_F > CM\sqrt{\frac{d_g r + t}{n}} \right] \le \exp\{-c(d_g r + t)\}. \tag{E.0f}$$

Taking $t = \log K$, union-bounding over the K groups, and using $r \le q$, we get

$$\max_{g \in \mathcal{G}} \frac{\|\mathbf{C}_g\|_F}{\sqrt{d_g}} \le CM\left(\sqrt{\frac{r}{n}} + \sqrt{\frac{\log K}{d_{\min} n}} \right) \le CM^2 \left(\sqrt{\frac{q}{n}} + \sqrt{\frac{\log K}{d_{\min} n}} \right) \tag{E.0g}$$

with probability at least

$$1 - \exp\{-C'(r + \log K)\}.$$

Combining $(E.0d)$, $(E.0e)$, and $(E.0g)$, we conclude that

$$\max_{g \in \mathcal{G}} \frac{\|\mathbf{Z}_g\|_F}{\sqrt{d_g}} \le CM^2 \left(\sqrt{\frac{q}{n}} + \sqrt{\frac{\log K}{d_{\min} n}} \right) \tag{E.0h}$$

with probability at least

$$1 - \exp(-C'q) - \exp\{-C'(q + \log K)\} - \exp\{-C'(r + \log K)\}.$$

Thus, choosing

$$\rho \ge C_\rho M^2 \left(\sqrt{\frac{q}{n}} + \sqrt{\frac{\log K}{d_{\min} n}} \right)$$

with C_ρ sufficiently large ensures the group score event

$$\mathcal{E}_{\text{score}} = \left\{ \max_{g \in \mathcal{G}} \frac{\|\mathbf{Z}_g\|_F}{\sqrt{d_g}} \le \frac{\rho}{4} \right\}.$$

Thus, the stated choice of ρ ensures $\mathcal{E}_{\text{score}}$ with high probability.

We also work on a uniform group-sparse covariance event. Specifically, for all collections $H \subseteq \mathcal{G}$,

$$\left\| [\widehat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X]_{T(\mathcal{H}),T(\mathcal{H})} \right\|_{\text{op}} \leq CM\sqrt{\frac{d(\mathcal{H}) + |\mathcal{H}| \log(eK/|\mathcal{H}|)}{n}}.$$

This follows by applying the usual sub-Gaussian sample-covariance bound to each fixed group union and union-bounding over group subsets. On the group-sparse sizes used below, the right-hand side is small because $R_{\mathcal{G}} \leq c_0$. Hence, for all such group unions,

$$\lambda_{\min}\left([\widehat{\mathbf{\Sigma}}_X]_{T(\mathcal{H}),T(\mathcal{H})} \right) \geq \frac{1}{2M}, \qquad \lambda_{\max}\left([\widehat{\mathbf{\Sigma}}_X]_{T(\mathcal{H}),T(\mathcal{H})} \right) \leq 2M,$$

provided c_0 is small enough.

We now prove the regression error bound. By optimality of $\widehat{\mathbf{B}}$,

$$\frac{1}{n}\|\mathbf{Y}_0 - \mathbf{X}\widehat{\mathbf{B}}\|_F^2 + \rho\|\widehat{\mathbf{B}}\|_{\mathcal{G},21} \leq \frac{1}{n}\|\mathbf{Y}_0 - \mathbf{X}\mathbf{B}^*\|_F^2 + \rho\|\mathbf{B}^*\|_{\mathcal{G},21}.$$

Writing $\mathbf{\Delta} = \widehat{\mathbf{B}} - \mathbf{B}^*$, this gives

$$\left\| \widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta} \right\|_F^2 \leq 2\langle \mathbf{Z}, \mathbf{\Delta} \rangle + \rho\left(\|\mathbf{B}^*\|_{\mathcal{G},21} - \|\mathbf{B}^* + \mathbf{\Delta}\|_{\mathcal{G},21} \right).$$

On $\mathcal{E}_{\text{score}}$,

$$2\langle \mathbf{Z}, \mathbf{\Delta} \rangle \leq 2\max_{g\in\mathcal{G}} \frac{\|\mathbf{Z}_g\|_F}{\sqrt{d_g}} \sum_{g\in\mathcal{G}} \sqrt{d_g}\,\|\mathbf{\Delta}_g\|_F \leq \frac{\rho}{2}\sum_{g\in\mathcal{G}} \sqrt{d_g}\,\|\mathbf{\Delta}_g\|_F.$$

By decomposability of the group norm,

$$\|\mathbf{B}^*\|_{\mathcal{G},21} - \|\mathbf{B}^* + \mathbf{\Delta}\|_{\mathcal{G},21} \leq \sum_{g\in\mathcal{A}} \sqrt{d_g}\,\|\mathbf{\Delta}_g\|_F - \sum_{g\notin\mathcal{A}} \sqrt{d_g}\,\|\mathbf{\Delta}_g\|_F.$$

Therefore,

$$\left\| \widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta} \right\|_F^2 \leq \frac{3\rho}{2}\sum_{g\in\mathcal{A}} \sqrt{d_g}\,\|\mathbf{\Delta}_g\|_F - \frac{\rho}{2}\sum_{g\notin\mathcal{A}} \sqrt{d_g}\,\|\mathbf{\Delta}_g\|_F. \tag{E.1}$$

In particular, since the left-hand side is nonnegative, we obtain the group cone condition

$$\sum_{g\notin\mathcal{A}} \sqrt{d_g}\,\|\mathbf{\Delta}_g\|_F \leq 3\sum_{g\in\mathcal{A}} \sqrt{d_g}\,\|\mathbf{\Delta}_g\|_F. \tag{E.2}$$

Moreover,

$$\sum_{g\in\mathcal{A}} \sqrt{d_g}\,\|\mathbf{\Delta}_g\|_F \leq \left(\sum_{g\in\mathcal{A}} d_g \right)^{1/2} \left(\sum_{g\in\mathcal{A}} \|\mathbf{\Delta}_g\|_F^2 \right)^{1/2} = \sqrt{s_u}\,\|\mathbf{\Delta}_{\mathcal{A}}\|_F.$$

Thus (E.1) implies

$$\left\| \widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta} \right\|_F^2 \leq \frac{3\rho\sqrt{s_u}}{2}\|\mathbf{\Delta}_{\mathcal{A}}\|_F. \tag{E.3}$$

We now use a group peeling argument to lower-bound $\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}\|_F$ by $\|\mathbf{\Delta}\|_F$.

For each inactive group $g \notin \mathcal{A}$, define the normalized group size

$$a_g = \frac{\|\mathbf{\Delta}_g\|_F}{\sqrt{d_g}}.$$

Order the inactive groups in decreasing order of a_g. Let m_* be an integer of the form

$$m_* = \left\lceil c_* \, M^2 \eta \, \frac{s_u}{d_{\min}} \right\rceil,$$

where $c_* > 0$ is a sufficiently large universal constant. Let

$$\mathcal{J}_1, \mathcal{J}_2, \ldots$$

be consecutive blocks of m_* inactive groups in this ordering, with the final block possibly smaller. Define the enlarged active group set

$$\mathcal{S}_e = \mathcal{A} \cup \mathcal{J}_1.$$

For $k \geq 2$, by construction of the ordering,

$$\max_{g \in \mathcal{J}_k} a_g \leq \frac{1}{m_* d_{\min}} \sum_{h \in \mathcal{J}_{k-1}} d_h a_h = \frac{1}{m_* d_{\min}} \sum_{h \in \mathcal{J}_{k-1}} \sqrt{d_h} \, \|\mathbf{\Delta}_h\|_F.$$

Therefore,

$$\|\mathbf{\Delta}_{\mathcal{J}_k}\|_F = \left(\sum_{g \in \mathcal{J}_k} d_g a_g^2 \right)^{1/2} \leq \sqrt{m_* d_{\max}} \max_{g \in \mathcal{J}_k} a_g \leq \sqrt{\frac{\eta}{m_* d_{\min}}} \sum_{h \in \mathcal{J}_{k-1}} \sqrt{d_h} \, \|\mathbf{\Delta}_h\|_F.$$

Summing over $k \geq 2$ and using the cone condition (E.2),

$$\sum_{k \geq 2} \|\mathbf{\Delta}_{\mathcal{J}_k}\|_F \leq \sqrt{\frac{\eta}{m_* d_{\min}}} \sum_{g \notin \mathcal{A}} \sqrt{d_g} \, \|\mathbf{\Delta}_g\|_F \leq 3 \sqrt{\frac{\eta}{m_* d_{\min}}} \sum_{g \in \mathcal{A}} \sqrt{d_g} \, \|\mathbf{\Delta}_g\|_F.$$

Using Cauchy–Schwarz again,

$$\sum_{k \geq 2} \|\mathbf{\Delta}_{\mathcal{J}_k}\|_F \leq 3 \sqrt{\frac{\eta s_u}{m_* d_{\min}}} \, \|\mathbf{\Delta}_{\mathcal{A}}\|_F. \tag{E.4}$$

By the definition of m_*, choosing c_* large enough gives

$$3 \sqrt{\frac{\eta s_u}{m_* d_{\min}}} \leq \frac{1}{4M}. \tag{E.5}$$

Now write

$$\mathbf{\Delta} = \mathbf{\Delta}_{\mathcal{S}_e} + \sum_{k \geq 2} \mathbf{\Delta}_{\mathcal{J}_k}.$$

By the triangle inequality,

$$\left\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}\right\|_F \geq \left\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}_{\mathcal{S}_e}\right\|_F - \sum_{k\geq 2}\left\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}_{\mathcal{J}_k}\right\|_F.$$

On the group-sparse covariance event,

$$\left\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}_{\mathcal{S}_e}\right\|_F \geq \frac{1}{\sqrt{2M}}\|\mathbf{\Delta}_{\mathcal{S}_e}\|_F,$$

and

$$\left\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}_{\mathcal{J}_k}\right\|_F \leq \sqrt{2M}\,\|\mathbf{\Delta}_{\mathcal{J}_k}\|_F.$$

Hence, using (E.4)–(E.5) and $\|\mathbf{\Delta}_{\mathcal{A}}\|_F \leq \|\mathbf{\Delta}_{\mathcal{S}_e}\|_F$,

$$\left\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}\right\|_F \geq \left(\frac{1}{\sqrt{2M}} - \frac{\sqrt{2M}}{4M}\right)\|\mathbf{\Delta}_{\mathcal{S}_e}\|_F \geq \frac{c}{\sqrt{M}}\|\mathbf{\Delta}_{\mathcal{S}_e}\|_F \tag{E.6}$$

for a universal constant $c > 0$. In addition, (E.4)–(E.5) also imply

$$\|\mathbf{\Delta}_{\mathcal{S}_e^c}\|_F \leq \sum_{k\geq 2}\|\mathbf{\Delta}_{\mathcal{J}_k}\|_F \leq \frac{1}{4M}\|\mathbf{\Delta}_{\mathcal{S}_e}\|_F.$$

Therefore,

$$\|\mathbf{\Delta}\|_F \leq \|\mathbf{\Delta}_{\mathcal{S}_e}\|_F + \|\mathbf{\Delta}_{\mathcal{S}_e^c}\|_F \leq C\|\mathbf{\Delta}_{\mathcal{S}_e}\|_F \leq C\sqrt{M}\left\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}\right\|_F. \tag{E.7}$$

Combining (E.3) with (E.6), we get

$$\left\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}\right\|_F^2 \leq \frac{3\rho\sqrt{s_u}}{2}\|\mathbf{\Delta}_{\mathcal{A}}\|_F \leq \frac{3\rho\sqrt{s_u}}{2}\|\mathbf{\Delta}_{\mathcal{S}_e}\|_F \leq C\sqrt{M}\,\rho\sqrt{s_u}\left\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}\right\|_F.$$

Thus

$$\left\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}\right\|_F^2 \leq CM\rho^2 s_u. \tag{E.8}$$

By (E.7),

$$\|\mathbf{\Delta}\|_F^2 \leq CM^2\rho^2 s_u. \tag{E.9}$$

Since

$$\rho^2 \leq CM^4\left(\frac{q}{n} + \frac{\log K}{d_{\min}n}\right),$$

the bounds (E.8)–(E.9) give

$$\left\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}\right\|_F^2 \leq CM^5\left(\frac{s_u q}{n} + \frac{(s_u/d_{\min})\log K}{n}\right) = CM^5 R_{\mathcal{G}},$$

and

$$\|\mathbf{\Delta}\|_F^2 \leq CM^6\left(\frac{s_u q}{n} + \frac{(s_u/d_{\min})\log K}{n}\right) = CM^6 R_{\mathcal{G}}.$$

It remains to prove the selected-support bound. Let

$$\widehat{\mathcal{A}} = \mathrm{supp}_{\mathcal{G}}(\widehat{\mathbf{B}}), \qquad \widehat{T} = T(\widehat{\mathcal{A}}), \qquad \widehat{s} = d(\widehat{\mathcal{A}}).$$

The KKT conditions for the group-penalized estimator give, for every group g,

$$\frac{2}{n}\mathbf{X}_g^\top(\mathbf{X}\widehat{\mathbf{B}} - \mathbf{Y}_0) \in -\rho\sqrt{d_g}\,\partial\|\widehat{\mathbf{B}}_g\|_F.$$

Therefore, if $g \in \widehat{\mathcal{A}}$, then

$$\frac{1}{n}\left\|\mathbf{X}_g^\top(\mathbf{X}\widehat{\mathbf{B}} - \mathbf{Y}_0)\right\|_F = \frac{\rho\sqrt{d_g}}{2}.$$

On the score event,

$$\frac{1}{n}\left\|\mathbf{X}_g^\top(\mathbf{Y}_0 - \mathbf{X}\mathbf{B}^*)\right\|_F = \|\mathbf{Z}_g\|_F \le \frac{\rho\sqrt{d_g}}{4}.$$

Hence, for every $g \in \widehat{\mathcal{A}}$,

$$\frac{1}{n}\left\|\mathbf{X}_g^\top\mathbf{X}\boldsymbol{\Delta}\right\|_F \ge \frac{\rho\sqrt{d_g}}{4}.$$

Squaring and summing over $g \in \widehat{\mathcal{A}}$ yields

$$\frac{\rho^2}{16}\sum_{g\in\widehat{\mathcal{A}}} d_g \le \frac{1}{n^2}\sum_{g\in\widehat{\mathcal{A}}}\|\mathbf{X}_g^\top\mathbf{X}\boldsymbol{\Delta}\|_F^2.$$

Because the groups are disjoint,

$$\sum_{g\in\widehat{\mathcal{A}}}\|\mathbf{X}_g^\top\mathbf{X}\boldsymbol{\Delta}\|_F^2 = \|\mathbf{X}_{\widehat{T}}^\top\mathbf{X}\boldsymbol{\Delta}\|_F^2.$$

Therefore,

$$\frac{\rho^2}{16}\widehat{s} \le \frac{1}{n^2}\|\mathbf{X}_{\widehat{T}}^\top\mathbf{X}\boldsymbol{\Delta}\|_F^2 \le \left\|[\widehat{\boldsymbol{\Sigma}}_X]_{\widehat{T},\widehat{T}}\right\|_{\mathrm{op}}\left\|\widehat{\boldsymbol{\Sigma}}_X^{1/2}\boldsymbol{\Delta}\right\|_F^2.$$

Using (E.8),

$$\widehat{s} \le CMs_u\left\|[\widehat{\boldsymbol{\Sigma}}_X]_{\widehat{T},\widehat{T}}\right\|_{\mathrm{op}}. \tag{E.10}$$

On the uniform group-sparse covariance event,

$$\left\|[\widehat{\boldsymbol{\Sigma}}_X]_{\widehat{T},\widehat{T}}\right\|_{\mathrm{op}} \le M + CM\sqrt{\frac{\widehat{s} + (\widehat{s}/d_{\min})\log(eK)}{n}}.$$

Substituting this into (E.10),

$$\widehat{s} \le CM^2s_u + CM^2s_u\sqrt{\frac{\widehat{s} + (\widehat{s}/d_{\min})\log(eK)}{n}}. \tag{E.11}$$

We now solve this one-dimensional inequality. If

$$\widehat{s} \le 2CM^2s_u,$$

we are done. Otherwise, the second term in (E.11) must dominate, and hence

$$\widehat{s} \le CM^4 s_u^2 \frac{1 + \log(eK)/d_{\min}}{n}.$$

Since

$$\frac{s_u}{n}\left(1 + \frac{\log(eK)}{d_{\min}}\right) \le CR_{\mathcal{G}} \le Cc_0,$$

choosing c_0 sufficiently small gives a contradiction to $\widehat{s} > 2CM^2 s_u$. Therefore,

$$\widehat{s} = d(\widehat{\mathcal{A}}) \le CM^2 s_u.$$

Finally, since each selected group has size at least $d_{\min}$,

$$|\widehat{\mathcal{A}}| \le \frac{d(\widehat{\mathcal{A}})}{d_{\min}} \le CM^2 \frac{s_u}{d_{\min}}.$$

This completes the proof. ∎

Theorem 10 (Group-sparse CCAR3 direction error) *Let*

$$\mathcal{G} = \{g_1, \dots, g_K\}$$

be a non-overlapping partition of $[p]$*. Write*

$$d_g = |g|, \qquad d_{\min} = \min_{g \in \mathcal{G}} d_g, \qquad d_{\max} = \max_{g \in \mathcal{G}} d_g,$$

and assume that $d_{\max}/d_{\min}$ *is bounded by an absolute constant.*

Assume the canonical pair model

$$\mathbf{\Sigma}_{XY} = \mathbf{\Sigma}_X \mathbf{U} \mathbf{\Lambda} \mathbf{V}^\top \mathbf{\Sigma}_Y, \qquad \mathbf{U}^\top \mathbf{\Sigma}_X \mathbf{U} = \mathbf{V}^\top \mathbf{\Sigma}_Y \mathbf{V} = I_r,$$

with

$$\lambda_r(\mathbf{\Lambda}) \ge \lambda, \qquad \lambda_1(\mathbf{\Lambda}) \le 1 - \frac{1}{M},$$

and

$$\frac{1}{M} \le \lambda_{\min}(\mathbf{\Sigma}_X) \wedge \lambda_{\min}(\mathbf{\Sigma}_Y), \qquad \lambda_{\max}(\mathbf{\Sigma}_X) \vee \lambda_{\max}(\mathbf{\Sigma}_Y) \le M.$$

Let

$$\mathbf{V}_0 = \mathbf{\Sigma}_Y^{1/2} \mathbf{V}, \qquad \mathbf{B}^* = \mathbf{U} \mathbf{\Lambda} \mathbf{V}_0^\top.$$

Define the active group set

$$\mathcal{A} = \{g \in \mathcal{G} : \|\mathbf{U}_g\|_F \neq 0\}, \qquad m_u = |\mathcal{A}|, \qquad s_u = \sum_{g \in \mathcal{A}} |g|.$$

Let

$$R_{\mathcal{G}} = \frac{s_u q + (s_u/d_{\min}) \log K}{n}, \qquad \delta_{\mathcal{G}} = \sqrt{R_{\mathcal{G}}}.$$

Assume $R_{\mathcal{G}} \le c_0$ for a sufficiently small constant $c_0 \in (0,1)$, and choose

$$\rho \ge C_\rho M^2 \left(\sqrt{\frac{q}{n}} + \sqrt{\frac{\log K}{d_{\min} n}} \right)$$

for a sufficiently large constant C_ρ. Let

$$\widehat{\mathbf{B}} = argmin_{\mathbf{B} \in \mathbb{R}^{p \times q}} \frac{1}{n} \|\mathbf{Y}_0 - \mathbf{X}\mathbf{B}\|_F^2 + \rho \sum_{g \in \mathcal{G}} \sqrt{|g|}\, \|\mathbf{B}_g\|_F, \qquad \mathbf{Y}_0 = \mathbf{Y}\widehat{\mathbf{\Sigma}}_Y^{-1/2}.$$

Let

$$\widehat{\mathbf{B}}_r = \widehat{\mathbf{U}}_{\hat{B}} \widehat{\mathbf{\Lambda}}_{\hat{B}} \widehat{\mathbf{V}}_{\hat{B}}^\top$$

be the rank-r truncated singular value decomposition of $\widehat{\mathbf{B}}$, and define

$$\widehat{\mathbf{U}} = \widehat{\mathbf{U}}_{\hat{B}} \left(\widehat{\mathbf{U}}_{\hat{B}}^\top \widehat{\mathbf{\Sigma}}_X \widehat{\mathbf{U}}_{\hat{B}} \right)^{-1/2}, \qquad \widehat{\mathbf{V}} = \widehat{\mathbf{\Sigma}}_Y^{-1/2} \widehat{\mathbf{V}}_{\hat{B}}.$$

If

$$C M^{7/2} \delta_{\mathcal{G}} \le \lambda/2,$$

then, for any $C' > 0$, with probability at least

$$1 - \exp(-C'q) - \exp\left\{ -C' \frac{s_u}{d_{\min}} \log\left(\frac{eK d_{\min}}{s_u} \right) \right\} - \exp\{-C'(q + \log K)\} - \exp\{-C'(r + \log K)\},$$

there exist orthogonal matrices $\mathbf{O}, \widetilde{\mathbf{O}} \in \mathbb{R}^{r \times r}$ such that

$$\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{O}\|_F \le C \frac{M^{11/2}\sqrt{r}}{\lambda} \sqrt{\frac{s_u q + (s_u/d_{\min}) \log K}{n}},$$

and

$$\|\widehat{\mathbf{U}} - \mathbf{U}\widetilde{\mathbf{O}}\|_F \le C \frac{M^{6}\sqrt{r}}{\lambda} \sqrt{\frac{s_u q + (s_u/d_{\min}) \log K}{n}}.$$

In particular, when the active groups are balanced so that $s_u/d_{\min} \asymp m_u$,

$$\delta_{\mathcal{G}} \asymp \sqrt{\frac{s_u q + m_u \log K}{n}}.$$

Proof The proof is identical to the proof of Theorem 4: replace the sparse regression theorem by Theorem 9, replace the sparse rate $\sqrt{s_u(q + \log p)/n}$ by $\delta_{\mathcal{G}}$, and use the selected-support bound $d(\widehat{\mathcal{A}}) \lesssim M^2 s_u$ to obtain the same uniform covariance control in the normalization step. ■

Remark 11 *We note that group sparsity and regular sparsity give very similar bounds. The main improvement comes from the replacement of the term $\sqrt{\frac{s_u(q+\log(p))}{n}}$ in Theorem 4 by $\sqrt{\frac{s_u\left(q + \frac{\log(K)}{d_{min}}\right)}{n}}$. Consequently, when p is large and in the presence of a small number of balanced groups, the group lasso is expected to improve upon the regular lasso term.*

Appendix F. Proofs: graph-sparse CCAR[3]

We consider now the graph-sparse CCA problem. For a graph $\mathcal{G}$ on p nodes and m edges indicating the relationships between the p covariates of $\mathbf{X}$, let $\mathcal{F}_{\mathcal{G}}(s_u, p, m, q, r, \lambda, M)$ be the collection of all covariance matrices $\mathbf{\Sigma}$ satisfying (2) and

$$\begin{aligned} &\text{(a)} \quad \mathbf{U} \in \mathbb{R}^{p\times r} \text{ and } \mathbf{V} \in \mathbb{R}^{q\times r} \text{ with } |\text{supp}(\mathbf{\Gamma U})| \leq s_u \ ; \\ &\text{(b)} \quad \sigma_{\min}(\mathbf{\Sigma}_X) \wedge \sigma_{\min}(\mathbf{\Sigma}_Y) \geq \tfrac{1}{M} \text{ and } \sigma_{\max}(\mathbf{\Sigma}_X) \vee \sigma_{\max}(\mathbf{\Sigma}_Y) \leq M; \\ &\text{(c)} \quad \lambda_r \geq \lambda \text{ and } \lambda_1 \leq 1 - \tfrac{1}{M}. \end{aligned} \tag{53}$$

We denote $\mathbf{\Gamma}$ the incidence matrix of the graph $\mathcal{G}$, and by $\mathbf{\Gamma}^\dagger$ its pseudo inverse. Denoting $\mathbf{\Pi} = \mathbf{I}_p - \mathbf{\Gamma}^\dagger\mathbf{\Gamma}$, we decompose $\mathbf{U} \in \mathbb{R}^{p\times r}$ as a sum of terms carried by the orthogonal subspaces: $\mathbf{U} = \mathbf{\Pi U} + \mathbf{\Gamma}^\dagger\mathbf{\Gamma U}$.

We begin with a few observations on the properties of the matrices $\mathbf{\Gamma}$ and $\mathbf{\Pi}$.

1. By the property of the incidence matrix,

$$\mathbf{\Gamma}^\top\mathbf{\Gamma} = \mathbf{D} - \mathbf{A} = \mathbf{L} \qquad (*)$$

   is the (unnormalized) Laplacian of the graph (Hütter and Rigollet, 2016). Here $\mathbf{A}$ denotes the adjacency of the graph and $\mathbf{D}$ is a diagonal matrix such that $\mathbf{D}_{ii}$ is the degree of node i. We denote by $\kappa_2 = \lambda_{n_c+1}(\mathbf{L})$ the smallest nonzero eigenvalue of $\mathbf{L}$.

2. $\mathbf{\Pi}$ is the projection operator of $\mathbb{R}^p$ onto the nullspace of $\mathbf{\Gamma}^\dagger\mathbf{\Gamma}$, thus, $\mathbf{\Pi}^2 = \mathbf{\Pi}$. Moreover, by $(*)$, $\mathbf{\Pi}$ corresponds to the eigenvectors associated to the eigenvalue 0 of the Laplacian.

3. Consequently, we have an explicit formulation for $\mathbf{\Pi}$ (Chung, 1996). Let n_c be the number of connected components of G, and let C_i denote the i^{th} connected component of G. Let $\mathbb{1}_{C_i} \in \mathbb{R}^{p\times 1}$ be the indicator (column) vector of the i^{th} connected component, i.e $[\mathbb{1}_{C_i}]_j = 1$ if $j \in C_i$ and 0 otherwise. Then

$$\mathbf{\Pi} = \sum_{i\in[n_c]} \frac{\mathbb{1}_{C_i}\mathbb{1}_{C_i}^\top}{|C_i|}.$$

4. Since $\mathbf{\Pi}$ is a projection matrix then for any $\mathbf{B} \in \mathbb{R}^{p\times q}$ it holds $\|\mathbf{\Pi B}\|_F \leq \|\mathbf{B}\|_F$.

5. We also have:

$$\|\mathbf{\Gamma B}\|_F^2 \leq \sigma_{\max}(\mathbf{L})\|\mathbf{B}\|_F^2,$$

   where $\sigma_{\max}(\mathbf{L})$ denotes the maximum eigenvalue of the Laplacian of the graph G. By property of the eigenvalues of the Laplacian, we thus have:

$$\|\mathbf{\Gamma B}\|_F^2 \leq 2d_{\max}\|\mathbf{B}\|_F^2,$$

   where $d_{\max}$ is the maximum degree of the graph (Chung, 1996).

Similarly to the analysis of Hütter and Rigollet (2016), we introduce the scaling constant:

$$\rho(\mathbf{\Gamma}) = \max_{e\in[m]} \left\|\mathbf{\Gamma}^\dagger_{\cdot e}\right\| \tag{54}$$

F.1 Useful Lemmas for the Graph Setting

Lemma 12 *For $i \in [n]$, let $\mathbf{W}_i \sim \mathcal{N}(0, \bar{\mathbf{\Sigma}})$ denote a set of multivariate random variables with covariance matrix $\bar{\mathbf{\Sigma}} = \mathbf{I}_q - \mathbf{V}_0 \mathbf{\Lambda}^2 \mathbf{V}_0^\top$, where $\mathbf{V}_0 = \mathbf{\Sigma}_Y^{\frac{1}{2}} \mathbf{V}$, with the notations introduced in section 4.1. Let $\mathbf{X}_i \sim \mathcal{N}(0, \mathbf{\Sigma}_X)$ denote a set of multivariate normal observations independent of $\mathbf{W}$. Assume*

$$M \frac{q + s_u \log\left(\frac{ep}{s_u}\right)}{n} \leq c_0 \tag{$\mathcal{H}_{1,\mathcal{G}}$}$$

for $c_0 \in (0,1)$. Then, for a constant $C' > 0$, there exists $C > 0$ that solely depends on C' such that:

$$\max_{j \in [m]} \left\| \left[(\mathbf{\Gamma}^\dagger)^\top \frac{\mathbf{X}^\top \mathbf{W}}{n} \right]_{j\cdot} \right\| \leq C \cdot \left(1 + \frac{M}{\kappa_2}\right) \cdot \sqrt{\frac{q + \log m}{n}} \tag{55}$$

with probability at least $1 - \exp\{-C'(q + \log m)\}$. Here we assume that c_0 is small enough so that $C' c_0 \leq c_3 - \log c_2 - 1$ with c_3, c_2 the universal constants of Theorem 6.5 of Wainwright (2019).

Proof We note first that:

$$\max_{j \in [m]} \left\| \left[(\mathbf{\Gamma}^\dagger)^\top \frac{\mathbf{X}^\top \mathbf{W}}{n} \right]_{j\cdot} \right\| \leq \max_{j \in [m]} \left\| \left[(\mathbf{\Gamma}^\dagger)^\top \frac{\mathbf{X}^\top \mathbf{W}}{n} - \mathbf{\Sigma}_{X\Gamma^\dagger, W} \right]_{j\cdot} \right\| + \max_{j \in [m]} \left\| \left[\mathbf{\Sigma}_{X\Gamma^\dagger, W} \right]_{j\cdot} \right\| \tag{56}$$

where $\mathbf{\Sigma}_{X\Gamma^\dagger, W}$ is the cross-covariance matrix between the transformed variables $\mathbf{X}\mathbf{\Gamma}^\dagger$ and $\mathbf{W}$. We also note that, by independence of $\mathbf{W}$ and $\mathbf{X}$:

$$\max_{j \in [m]} \left\| \left[\mathbf{\Sigma}_{X\Gamma^\dagger, W} \right]_{j\cdot} \right\| = 0$$

Let $\mathbf{Z}_i \sim \mathcal{N}_m\big(0, (\mathbf{\Gamma}^\dagger)^\top \mathbf{\Sigma}_X \mathbf{\Gamma}^\dagger\big)$, $\mathbf{W}_i \sim \mathcal{N}_q(0, \bar{\mathbf{\Sigma}})$, and $\mathbb{1}_j^\top \in \mathbb{R}^{1 \times m}$ the vector indicator of row j (i.e. $[\mathbb{1}_j]_i = 0$ for all $i \neq j$, and $[\mathbb{1}_j]_j = 1$). Thus, denoting $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{\Gamma}^\dagger$, we may write

$$\mathbb{1}_j^\top (\widehat{\mathbf{\Sigma}}_{\widetilde{X}W} - \mathbf{\Sigma}_{\widetilde{X}W}) = \frac{1}{n} \sum_{i=1}^n \big(\mathbb{1}_j^\top \mathbf{Z}_i \mathbf{W}_i^\top - \mathbb{E}(\mathbb{1}_j^\top \mathbf{Z}_i \mathbf{W}_i^\top) \big).$$

We also note that

$$\left\| \mathbf{\Sigma}_{\widetilde{X}} \right\|_{op} = \left\| (\mathbf{\Gamma}^\dagger)^\top \mathbf{\Sigma}_X \mathbf{\Gamma}^\dagger \right\|_{op} = \sup_{u \in \mathbb{R}^m} \frac{u^\top (\mathbf{\Gamma}^\dagger)^\top \mathbf{\Sigma}_X \mathbf{\Gamma}^\dagger u}{\|u\|^2} \leq M \frac{\|\mathbf{\Gamma}^\dagger u\|^2}{\|u\|^2} \leq \frac{M}{\kappa_2}.$$

Here, as we defined above, constant κ_2 that appearing above is the smallest non-zero eigenvalue of the Laplacian $\mathbf{L} = \mathbf{\Gamma}^\top \mathbf{\Gamma}$. Moreover we note that

$$\left\| \bar{\mathbf{\Sigma}} \right\|_{op} = \left\| \mathbf{I} - \mathbf{V}_0 \mathbf{\Lambda}^2 \mathbf{V}_0^\top \right\|_{op} \leq 1.$$

Define for each $j \in [m]$ the matrix $\mathbf{H}_i^{(j)} = \begin{pmatrix} \mathbb{1}_j^\top \mathbf{Z}_i \\ \mathbf{W}_i \end{pmatrix}$. Since $\mathbb{1}_j^\top \mathbf{Z}_i \mathbf{W}_i^\top$ is a submatrix of $\mathbf{H}_i^{(j)} \big(\mathbf{H}_i^{(j)}\big)^\top$

$$\max_{j \in [m]} \left\| \mathbb{1}_j^\top (\widehat{\mathbf{\Sigma}}_{\widetilde{X}W} - \mathbf{\Sigma}_{\widetilde{X}W}) \right\| \leq \max_{j \in [m]} \left\| \frac{1}{n} \sum_{i=1}^n \big(\mathbf{H}_i^{(j)} \big(\mathbf{H}_i^{(j)}\big)^\top - \mathbb{E}\mathbf{H}_i^{(j)} \big(\mathbf{H}_i^{(j)}\big)^\top \big) \right\|_{op} \tag{57}$$

Denoting $\mathcal{H}^{(j)} = \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top = \begin{pmatrix} \mathbb{1}_j^\top \mathbf{\Sigma}_{\widetilde{X}} \mathbb{1}_j & \mathbb{1}_j^\top \mathbf{\Sigma}_{\widetilde{X}W} \\ \mathbf{\Sigma}_{W\widetilde{X}} \mathbb{1}_j & \bar{\mathbf{\Sigma}} \end{pmatrix} = \begin{pmatrix} \mathbb{1}_j^\top \mathbf{\Sigma}_{\widetilde{X}} \mathbb{1}_j & 0 \\ 0 & \bar{\mathbf{\Sigma}} \end{pmatrix}$, where the last equality follows from the fact that $\mathbf{W}$ and $\mathbf{X}$ are independent, by assumption. This implies:

$$\begin{aligned} \Big\|\mathcal{H}^{(j)}\Big\|_{op} &= \sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q \\ u_1^2+\|u_2\|^2=1}} \mathbb{1}_j^\top \mathbf{\Sigma}_{\widetilde{X}} \mathbb{1}_j u_1^2 + u_2^\top \bar{\mathbf{\Sigma}} u_2 \\ &\leq \sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q \\ u_1^2+\|u_2\|^2=1}} \frac{M}{\kappa_2} u_1^2 + \|u_2\|^2 \leq \frac{M}{\kappa_2} + 1 \end{aligned} \tag{58}$$

By Theorem 6.5 of Wainwright (2019), there exist universal constants $c_1, c_2, c_3 > 0$ such that:

$$\begin{aligned} &\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n \big(\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top - \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top\big)\right\|_{op} \geq \Big\|\mathcal{H}^{(j)}\Big\|_{op}\left(2c_1\sqrt{\frac{q+1}{n}} + t\right)\right] \\ &\leq c_2 \exp\big\{-c_3 n \min\{t, t^2\}\big\}. \end{aligned}$$

Thus, by a simple union bound, we conclude that

$$\begin{aligned} &\mathbb{P}\left[\max_{j\in[m]}\left\|\frac{1}{n}\sum_{i=1}^n \big(\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top - \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top\big)\right\|_{op} \geq \Big\|\mathcal{H}^{(j)}\Big\|_{op}\left(2c_1\sqrt{\frac{q+1}{n}} + t\right)\right] \\ &\leq c_2 m \exp\big\{-c_3 n \min\{t, t^2\}\big\} \end{aligned}$$

Now, we write $c_2 = e^{\gamma_2(q+\log m)}$ for an appropriate constant γ_2 and take $t^2 = \frac{1+C'+(\gamma_2)_+}{c_3} \cdot \frac{q+\log m}{n}$ for any $C' > 0$ while assuming c_0 small enough so that $t \leq 1$. This implies that

$$\begin{aligned} &\mathbb{P}\left[\max_{j\in[m]}\left\|\frac{1}{n}\sum_{i=1}^n \big(\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top - \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top\big)\right\|_{op} \geq C_3' \times \left(1 + \frac{M}{\kappa_2}\right)\sqrt{\frac{q+\log m}{n}}\right] \\ &\leq \exp\big\{\gamma_2(q+\log m) + \log m - \big(1 + C' + (\gamma_2)_+\big)(q+\log m)\big\} \\ &\leq \exp\{-C'(q+\log m)\} \end{aligned} \tag{59}$$

where $C = 2c_1 + \sqrt{\frac{1+C'+(\gamma_2)_+}{c_3}}$. ■

Lemma 13 *Let $(\mathbf{X}, \mathbf{Y}) \sim \mathcal{N}(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a covariance matrix in $\mathcal{F}_{\mathcal{G}}(s_u, p, m, q, r, \lambda, M)$ as defined in (53). Let $\mathbf{\Gamma}$ be the incidence matrix corresponding to the graph G on p nodes and m edges. Denote the support of $\mathbf{\Gamma U}$ as $\mathrm{supp}(\mathbf{\Gamma U})$. Assume that $\mathcal{H}_{1,\mathcal{G}}$ holds and that $|\mathrm{supp}(\mathbf{\Gamma U})| \leq s_u$. Then, for a constant $C' > 0$, there exists $C > 0$ that solely depends on C' and c_0 such that:*

$$\max_{j\in[m]}\left\|\left[(\mathbf{\Gamma}^\dagger)^\top \frac{\mathbf{X}^\top\mathbf{Y}}{n}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right]_{j\cdot}\right\| \leq C\left(1 + \frac{M}{\kappa_2}\right)\cdot\sqrt{\frac{q+\log m}{n}} + \sqrt{\frac{M}{\kappa_2}} \tag{60}$$

with probability at least $1-\exp\{-C'(q+\log m)\}$. *Here we assume that* c_0 *is small enough so that* $C'c_0 \leq c_3 - \log c_2 - 1$ *with* c_3, c_2 *the universal constants of Theorem 6.5 of Wainwright (2019).*

Proof Consider the term:

$$\begin{aligned}\max_{j\in[m]}\left\|\left[(\mathbf{\Gamma}^\dagger)^\top \frac{\mathbf{X}^\top\mathbf{Y}}{n}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right]_{j.}\right\| \leq &\max_{j\in[m]}\left\|\left[(\mathbf{\Gamma}^\dagger)^\top \frac{\mathbf{X}^\top\mathbf{Y}}{n}\mathbf{\Sigma}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_{X\mathbf{\Gamma}^\dagger, Y\mathbf{\Sigma}_Y^{-\frac{1}{2}}}\right]_{j.}\right\| \\ &+ \max_{j\in[m]}\left\|\left[\mathbf{\Sigma}_{X\mathbf{\Gamma}^\dagger, Y\mathbf{\Sigma}_Y^{-\frac{1}{2}}}\right]_{j.}\right\|\end{aligned} \tag{61}$$

Introducing $\mathbb{1}_j^\top \in \mathbb{R}^{1\times m}$ the vector indicator of row j (i.e. $[\mathbb{1}_j]_i = 0$ for all $i \neq j$, and $[\mathbb{1}_j]_j = 1$), we observe that:

$$\max_{j\in[m]}\left\|\left[\mathbf{\Sigma}_{X\mathbf{\Gamma}^\dagger, Y\mathbf{\Sigma}_Y^{-\frac{1}{2}}}\right]_{j.}\right\| = \max_{j\in[m]}\left\|\mathbb{1}_j^\top\mathbf{\Sigma}_{X\mathbf{\Gamma}^\dagger, Y\mathbf{\Sigma}_Y^{-\frac{1}{2}}}\right\|_F \leq \left\|\mathbf{\Sigma}_{X\mathbf{\Gamma}^\dagger, Y\mathbf{\Sigma}_Y^{-\frac{1}{2}}}\right\|_{op} \leq \sqrt{\frac{M}{\kappa_2}}\lambda_1. \tag{62}$$

We now turn to the first term on the RHS of equation (61). The proof is identical to that of Lemma 12, by introducing the transformed variables $\mathbf{Z}_i = \mathbf{X}_i\mathbf{\Gamma}^\dagger \sim \mathcal{N}_m\big(0, (\mathbf{\Gamma}^\dagger)^\top\mathbf{\Sigma}_X\mathbf{\Gamma}^\dagger\big)$ and $\widetilde{\mathbf{Y}}_i = \mathbf{Y}_i\mathbf{\Sigma}_Y^{-\frac{1}{2}} \sim \mathcal{N}_q(0, \mathbf{I}_q)$, and writing

$$\mathbb{1}_j^\top\left((\mathbf{\Gamma}^\dagger)^\top\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}} - (\mathbf{\Gamma}^\dagger)^\top\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right) = \frac{1}{n}\sum_{i=1}^n\big(\mathbb{1}_j^\top\mathbf{Z}_i\widetilde{\mathbf{Y}}_i^\top - \mathbb{E}(\mathbb{1}_j^\top\mathbf{Z}_i\widetilde{\mathbf{Y}}_i^\top)\big).$$

Define for each $j \in [m]$ the matrix $\mathbf{H}_i^{(j)} = \begin{pmatrix}\mathbb{1}_j^\top\mathbf{Z}_i \\ \widetilde{\mathbf{Y}}_i\end{pmatrix}$. We have then:

$$\max_{j\in[m]}\left\|\mathbb{1}_j^\top\left((\mathbf{\Gamma}^\dagger)^\top\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}} - (\mathbf{\Gamma}^\dagger)^\top\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right)\right\| \leq \max_{j\in[m]}\left\|\frac{1}{n}\sum_{i=1}^n\big(\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top - \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top\big)\right\|_{op} \tag{63}$$

Denoting $\mathcal{H}^{(j)} = \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top = \begin{pmatrix}\mathbb{1}_j^\top(\mathbf{\Gamma}^\dagger)^\top\mathbf{\Sigma}_X(\mathbf{\Gamma}^\dagger)\mathbb{1}_j & \mathbb{1}_j^\top(\mathbf{\Gamma}^\dagger)^\top\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}} \\ \mathbf{\Sigma}_Y^{-\frac{1}{2}}\mathbf{\Sigma}_{YX}(\mathbf{\Gamma}^\dagger)\mathbb{1}_j & \mathbf{I}_q\end{pmatrix}$, this implies:

$$\begin{aligned}\left\|\mathcal{H}^{(j)}\right\|_{op} &= \sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q \\ u_1^2+\|u_2\|^2=1}} \mathbb{1}_j^\top(\mathbf{\Gamma}^\dagger)^\top\mathbf{\Sigma}_X(\mathbf{\Gamma}^\dagger)\mathbb{1}_j u_1^2 + 2u_1\mathbb{1}_j^\top(\mathbf{\Gamma}^\dagger)^\top\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}u_2 + u_2^\top u_2 \\ &\leq \sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q \\ u_1^2+\|u_2\|^2=1}} \frac{M}{\kappa_2}u_1^2 + 2\lambda_1\sqrt{\frac{M}{\kappa_2}}u_1\|u_2\| + \|u_2\|^2 \\ &\leq \sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q \\ u_1^2+\|u_2\|^2=1}} \frac{M}{\kappa_2}u_1^2 + \sqrt{\frac{M}{\kappa_2}}(u_1^2+\|u_2\|^2) + \|u_2\|^2 \\ &= 2\left(1+\frac{M}{\kappa_2}\right)\end{aligned} \tag{64}$$

By the same arguments as for Lemma 12, we can show that for a constant $C' > 0$, there exists $C > 0$ that solely depends on C' such that, for c_0 small enough:

$$\mathbb{P}\left[\max_{j\in[m]}\left\|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{H}_i^{(j)}(\mathbf{H}_i^{(j)})^\top - \mathbb{E}\mathbf{H}_i^{(j)}(\mathbf{H}_i^{(j)})^\top)\right\|_{op} \geq C\cdot 2\left(1+\frac{M}{\kappa_2}\right)\cdot\sqrt{\frac{q+\log m}{n}}\right] \quad (65)$$
$$\leq \exp\big\{-C'(q+\log m)\big\}.$$

This concludes the proof. ■

Lemma 14 *Let $\mathbf{X},\mathbf{Y}\sim\mathcal{F}_{\mathcal{G}}(s_u,p,m,q,r,\lambda,M)$ as defined in (53). Assume that*

$$M\frac{\log n_c + q}{n} \leq c_0, \qquad (\mathcal{H}_{1,\mathcal{G},\Pi})$$

with $c_0\in(0,1)$. Then, for a constant $C'>0$, there exists $C>0$ that solely depends on C' such that

$$\left\|\mathbf{\Pi}^\top\frac{\mathbf{X}^\top\mathbf{Y}}{n}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right\|_F \leq \sqrt{Mn_c}\left(C\sqrt{M}\cdot\sqrt{\frac{q+\log n_c}{n}}+1\right) \quad (66)$$

with probability at least $1-\exp\big\{-C'(q+\log n_c)\big\}$. Here we assume that c_0 is small enough so that $C'c_0\leq c_3-\log c_2-1$ with c_3,c_2 the universal constants of Theorem 6.5 of Wainwright (2019).

Proof In this case:

$$\begin{aligned}\left\|\mathbf{\Pi}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right\|_F^2 &= \sum_{j\in[n_c]}\left\|\mathbb{1}_{C_j}\frac{\mathbb{1}_{C_j}^\top}{|C_j|}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right\|_F^2\\ &= \sum_{j\in[n_c]}\left\|\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right\|^2 \qquad (\star)\\ &\leq n_c\max_{j\in[n_c]}\left\|\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right\|^2\end{aligned}$$

where in $(\star)$, we have used the fact that

$$\left\|\mathbb{1}_{C_j}\frac{\mathbb{1}_{C_j}^\top}{|C_j|}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right\|_F^2 = \sum_{k\in C_j}\frac{1}{|C_j|}\left\|\left[\mathbb{1}_{C_j}\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right]_{k.}\right\|^2 = \left\|\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right\|^2.$$

We now proceed to bound the term:

$$\begin{aligned}\max_{j\in[n_c]}\left\|\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right\|^2 &\leq \max_{j\in[n_c]}\left\|\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\Big(\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}-\mathbf{\Sigma}_{X,Y\mathbf{\Sigma}_Y^{-\frac{1}{2}}}\Big)\right\|^2+\max_{j\in[n_c]}\left\|\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\mathbf{\Sigma}_{X,Y\mathbf{\Sigma}_Y^{-\frac{1}{2}}}\right\|^2\\ &\leq \max_{j\in[n_c]}\left\|\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\Big(\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}-\mathbf{\Sigma}_{X,Y\mathbf{\Sigma}_Y^{-\frac{1}{2}}}\Big)\right\|^2+\left\|\mathbf{\Sigma}_{X,Y\mathbf{\Sigma}_Y^{-\frac{1}{2}}}\right\|_{op}^2\\ &\leq \max_{j\in[n_c]}\left\|\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\Big(\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}-\mathbf{\Sigma}_{X,Y\mathbf{\Sigma}_Y^{-\frac{1}{2}}}\Big)\right\|^2+M\lambda_1^2\end{aligned} \tag{67}$$

We follow once again the outline of the proof of Lemma 7 in Gao et al. (2017). Defining $\mathbf{Z}_i\sim\mathcal{N}(0,\mathbf{\Sigma}_X)$ and $\widetilde{\mathbf{Y}}_i\sim\mathcal{N}(0,\mathbf{I}_q)$, we may write

$$\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\Big(\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}-\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\Big)=\frac{1}{n}\sum_{i=1}^n\left(\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\mathbf{Z}_i\widetilde{\mathbf{Y}}_i^\top-\mathbb{E}\left(\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\mathbf{Z}_i\widetilde{\mathbf{Y}}_i^\top\right)\right).$$

Then, define $\mathbf{H}_i^{(j)}=\begin{pmatrix}\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\mathbf{Z}_i\\ \widetilde{\mathbf{Y}}_i\end{pmatrix}$. Since $\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\Big(\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}-\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\Big)$ is a submatrix of $\mathbf{H}_i^{(j)}(\mathbf{H}_i^{(j)})^\top$, we have

$$\max_{j\in[n_c]}\left\|\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}(\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}-\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}})\right\|\leq\max_{j\in[n_c]}\left\|\frac{1}{n}\sum_{i=1}^n\left(\mathbf{H}_i^{(j)}(\mathbf{H}_i^{(j)})^\top-\mathbb{E}\mathbf{H}_i^{(j)}(\mathbf{H}_i^{(j)})^\top\right)\right\|_{op} \tag{68}$$

where $\mathcal{H}^{(j)}=\mathbb{E}\mathbf{H}_i^{(j)}(\mathbf{H}_i^{(j)})^\top=\begin{pmatrix}\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\mathbf{\Sigma}_X\frac{\mathbb{1}_{C_j}}{\sqrt{|C_j|}} & \frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\\ \mathbf{\Sigma}_Y^{-\frac{1}{2}}\mathbf{\Sigma}_{YX}\frac{\mathbb{1}_{C_j}}{\sqrt{|C_j|}} & \mathbf{I}_q\end{pmatrix}$. We also note that

$$\begin{aligned}\left\|\mathcal{H}^{(j)}\right\|_{op}&=\sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q\\ u_1^2+\|u_2\|^2=1}}\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\mathbf{\Sigma}_X\frac{\mathbb{1}_{C_j}}{\sqrt{|C_j|}}u_1^2+2u_1\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}u_2+u_2^\top u_2\\ &\leq\sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q\\ u_1^2+\|u_2\|^2=1}}Mu_1^2+2\lambda_1\sqrt{M}u_1\|u_2\|+\|u_2\|^2\\ &\leq\sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q\\ u_1^2+\|u_2\|^2=1}}Mu_1^2+\lambda_1\sqrt{M}\left(u_1^2+\|u_2\|^2\right)+M\|u_2\|^2\\ &\leq 2M\end{aligned} \tag{69}$$

since $\lambda_1\leq 1$ and $\sqrt{M}\leq M$. By similar arguments to Lemma 12 (application of a union bound combined with Theorem 6.5 of Wainwright (2019), we conclude that for a constant

$C' > 0$, there exists $C > 0$ such that:

$$\mathbb{P}\left[\max_{j\in[n_c]}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{H}_i^{(j)}(\mathbf{H}_i^{(j)})^\top - \mathbb{E}\mathbf{H}_i^{(j)}(\mathbf{H}_i^{(j)})^\top\right)\right\|_{op} \geq C\cdot 2M\cdot\sqrt{\frac{q+\log n_c}{n}}\right] \leq \exp\left\{-C'(q+\log n_c)\right\} \tag{70}$$

We conclude the proof by noting that this implies

$$\max_{j\in[n_c]}\left\|\frac{\mathbb{1}_{C_j}^\top}{\sqrt{|C_j|}}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right\|^2 \leq C^2\cdot 4M^2\frac{q+\log n_c}{n} + M\lambda_1^2 \tag{71}$$

with the same probability. ■

Lemma 15 *Let $(\mathbf{X},\mathbf{Y}) \sim \mathcal{N}(0,\mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a covariance matrix in $\mathcal{F}_{\mathcal{G}}(s_u,p,m,q,r,\lambda,M)$ as defined in (53). Let $\mathbf{W}\sim\mathcal{N}(0,\mathbf{I}_q)$ and independent of $(\mathbf{X},\mathbf{Y})$. Assume ($\mathcal{H}_{1,\mathcal{G},\Pi}$) for $c_0\in(0,1)$. Then, for a constant $C'>0$, there exists a constant $C>0$ that depends solely on C' such that*

$$\left\|\frac{\mathbf{\Pi}^\top\mathbf{X}^\top\mathbf{W}}{n}\right\|_F \leq CM\sqrt{n_c}\sqrt{\frac{q+\log n_c}{n}} \tag{72}$$

with probability at least $1-\exp\left\{-C'(q+\log n_c)\right\}$. Here we assume that c_0 is small enough so that $C'c_0 \leq c_3 - \log c_2 - 1$, with c_3, c_2 the universal constants of Theorem 6.5 of Wainwright (2019).

Proof The proof is identical to that of Lemma 14, by replacing the matrix $\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}$ with $\mathbf{\Sigma}_{XW}$, which is 0 since $\mathbf{X}$ and $\mathbf{W}$ are assumed to be independent, and adjusting the corresponding $\left\|\mathcal{H}^{(j)}\right\|_{op}$.

■

Lemma 16 *Let $\mathbf{X}_i\sim\mathcal{N}(0,\mathbf{\Sigma}_X)$ where $\mathbf{\Sigma}_X$ is a covariance matrix as defined in (53). Let $\mathbf{W}_i\sim\mathcal{N}(0,\bar{\mathbf{\Sigma}})$ independently of $\mathbf{X}$. Assume*

$$M\frac{\log m + q}{n} \leq c_0 \tag{$\mathcal{H}_{1,\mathcal{G}}$}$$

Then, for a constant $C'>0$, there exists a constant $C>0$ that depends solely on C' such that

$$\max_{j\in[m]}\left\|\mathbb{1}_j^\top\left((\mathbf{\Gamma}^\dagger)^\top\widehat{\mathbf{\Sigma}}_{XW} - (\mathbf{\Gamma}^\dagger)^\top\mathbf{\Sigma}_{XW}\right)\right\| \leq C\left(1+\frac{M}{\kappa_2}\right)\sqrt{\frac{\log m+q}{n}} \tag{73}$$

with probability at least $1-\exp\left\{-C'(q+\log m)\right\}$. Here we assume that c_0 is small enough so that $C'c_0 \leq c_3 - \log c_2 - 1$ with c_3, c_2 the universal constants of Theorem 6.5 of Wainwright (2019)

Proof Let $\mathbf{Z}_i \sim \mathcal{N}_m\big(0, (\mathbf{\Gamma}^\dagger)^\top \mathbf{\Sigma}_X \mathbf{\Gamma}^\dagger\big)$ and $\widetilde{\mathbf{X}}_i \sim \mathcal{N}_q(0, \bar{\mathbf{\Sigma}})$, and $\mathbb{1}_j^\top \in \mathbb{R}^{1\times m}$ the vector indicator of row j (i.e. $[\mathbb{1}_j]_i = 0$ for all $i \neq j$, and $[\mathbb{1}_j]_j = 1$). Thus, denoting $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{\Gamma}^\dagger$, we may write

$$\mathbb{1}_j^\top (\widehat{\mathbf{\Sigma}}_{\widetilde{X}W} - \mathbf{\Sigma}_{\widetilde{X}W}) = \frac{1}{n}\sum_{i=1}^n \big(\mathbb{1}_j^\top \mathbf{Z}_i \mathbf{W}_i^\top - \mathbb{E}(\mathbb{1}_j^\top \mathbf{Z}_i \mathbf{W}_i^\top)\big).$$

We also note that

$$\big\|\mathbf{\Sigma}_{\widetilde{X}}\big\|_{op} = \Big\|(\mathbf{\Gamma}^\dagger)^\top \mathbf{\Sigma}_X \mathbf{\Gamma}^\dagger\Big\|_{op} = \sup_{u\in\mathbb{R}^m} \frac{u^\top (\mathbf{\Gamma}^\dagger)^\top \mathbf{\Sigma}_X \mathbf{\Gamma}^\dagger u}{\|u\|^2} \leq M \frac{\|\mathbf{\Gamma}^\dagger u\|^2}{\|u\|^2} \leq \frac{M}{\kappa_2}.$$

Moreover, one can show that

$$\big\|\bar{\mathbf{\Sigma}}\big\|_{op} = \Big\|\mathbf{I} - \mathbf{V}_0 \mathbf{\Lambda}^2 \mathbf{V}_0^\top\Big\|_{op} \leq 1.$$

Define for each $j \in [m]$ the matrix $\mathbf{H}_i^{(j)} = \begin{pmatrix} \mathbb{1}_j^\top \mathbf{Z}_i \\ \mathbf{W}_i \end{pmatrix}$. We have then

$$\max_{j\in[m]} \Big\|\mathbb{1}_j^\top (\widehat{\mathbf{\Sigma}}_{\widetilde{X}W} - \mathbf{\Sigma}_{\widetilde{X}W})\Big\| \leq \max_{j\in[m]} \left\|\frac{1}{n}\sum_{i=1}^n \big(\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top - \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top\big)\right\|_{op} \tag{74}$$

Denote $\mathcal{H}^{(j)} = \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top = \begin{pmatrix} \mathbb{1}_j^\top \mathbf{\Sigma}_{\widetilde{X}} \mathbb{1}_j & 0 \\ 0 & \bar{\mathbf{\Sigma}} \end{pmatrix}$. This implies:

$$\begin{aligned}
\Big\|\mathcal{H}^{(j)}\Big\|_{op} &= \sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q \\ u_1^2+\|u_2\|^2=1}} \mathbb{1}_j^\top \mathbf{\Sigma}_{\widetilde{X}} \mathbb{1}_j u_1^2 + u_2^\top \bar{\mathbf{\Sigma}} u_2 \\
&\leq \sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q \\ u_1^2+\|u_2\|^2=1}} \frac{M}{\kappa_2} u_1^2 + \|u_2\|^2 \\
&\leq \sup_{\substack{u_1\in\mathbb{R},\ u_2\in\mathbb{R}^q \\ u_1^2+\|u_2\|^2=1}} \frac{M}{\kappa_2} u_1^2 + \sqrt{\frac{M}{\kappa_2}}(u_1^2 + \|u_2\|^2) + \|u_2\|^2 \\
&\leq 1 + \frac{M}{\kappa_2},
\end{aligned} \tag{75}$$

Therefore, by arguments similar to Lemma 12 (union bound and Theorem 6.5 of Wainwright (2019)), we conclude that for any $C' > 0$ such that $C'c_0 \leq c_3 - \log c_2 - 1$, with c_3, c_2 the universal constants of Theorem 6.5 of Wainwright (2019), there exists $C > 0$ such that:

$$\begin{aligned}
&\mathbb{P}\left[\max_{j\in[m]} \left\|\frac{1}{n}\sum_{i=1}^n \big(\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top - \mathbb{E}\mathbf{H}_i^{(j)}\big(\mathbf{H}_i^{(j)}\big)^\top\big)\right\|_{op} \geq C \cdot \left(1 + \frac{M}{\kappa_2}\right)\sqrt{\frac{q + \log m}{n}}\right] \\
&\leq \exp\{-C'(\log m + q)\}.
\end{aligned} \tag{76}$$

This concludes the proof. ■

F.2 Proof of graph CCAR^3 consistency

Theorem 17 *Consider the family $\mathcal{F}_{\mathcal{G}}(s_u, p, m, q, r, \lambda, M)$ of covariance matrices satisfying assumptions. Assume*

$$M\frac{q+\log m}{n} \leq c_0, \qquad M\frac{q+\log n_c}{n} \leq c_0. \tag{$\mathcal{H}_{1,\mathcal{G}}$}$$

Define

$$K_{\Gamma,n} = C_\star\sqrt{n_c\frac{q+\log n_c}{q+\log m}} + 4\sqrt{s_u\sigma_{\max}(\mathbf{L})},$$

where $C_\star$ is a numerical constant depending only on the constants in the score bounds. Assume in addition that

$$n \geq 10 \vee 128M^2\left(72n_c + 576\rho(\mathbf{\Gamma})^2\log p\, K_{\Gamma,n}^2\right).$$

We write $\mathbf{\Delta} = \widehat{\mathbf{B}} - \mathbf{B}^*$ with $\mathbf{B}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{\Sigma}_Y^{\frac{1}{2}}$, and choose $\rho \geq CM^2\left(1+\frac{1}{\kappa_2}\right)\sqrt{\frac{q+\log m}{n}}$ for some large constant C. Then the solution $\widehat{\mathbf{B}}$ of the penalized regression problem is such that, there exist constants C_1 and C_2 that do not depend on the parameters of the problem such that

$$\left\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}\right\|_F^2 \leq C_1M^5\left(n_c\frac{q+\log(n_c)}{n} + \left(1+\frac{1}{\kappa_2}\right)^2\sigma_{\max}(\mathbf{L})\cdot s_u\frac{q+\log m}{n}\right)$$
$$\|\mathbf{\Delta}\|_F \leq C_2M^3\left(\sqrt{n_c\frac{q+\log(n_c)}{n}} + \sqrt{\sigma_{\max}(\mathbf{L})}\left(1+\frac{1}{\kappa_2}\right)\sqrt{s_u\frac{q+\log m}{n}}\right)$$

with probability at least

$$1 - \exp\{-C'q\} - 2\exp\{-C'(q+\log n_c)\} - 2\exp\{-C'(q+\log m)\} - q\exp\{-c_2n\}.$$

Proof We begin by the following observation. Since $(\mathbf{X}, \mathbf{Y})$ are jointly multivariate normal and writing $\mathbf{V}_0 = \mathbf{\Sigma}_Y^{\frac{1}{2}}\mathbf{V}$, conditioned on $\mathbf{X}$, the following holds:

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}\left(\mathbf{X}\mathbf{B}^*\mathbf{\Sigma}_Y^{\frac{1}{2}},\ \mathbf{\Sigma}_Y^{\frac{1}{2}}(\mathbf{I} - \mathbf{V}_0\mathbf{\Lambda}^2\mathbf{V}_0^\top)\mathbf{\Sigma}_Y^{\frac{1}{2}}\right)$$

Therefore

$$\mathbf{Y}\mathbf{\Sigma}_Y^{-\frac{1}{2}}|\mathbf{X} \sim \mathcal{N}\left(\mathbf{X}\mathbf{B}^*,\ (\mathbf{I} - \mathbf{V}_0\mathbf{\Lambda}^2\mathbf{V}_0^\top)\right) \tag{$*$}$$

We work under the sample-size assumptions stated in the theorem.

We now turn to the analysis of the regression problem

$$\widehat{\mathbf{B}} \in \underset{\mathbf{B}\in\mathbb{R}^{p\times q}}{\operatorname{argmin}}\left\{\frac{1}{n}\|\mathbf{Y}_0 - \mathbf{X}\mathbf{B}\|_F^2 + \rho\|\mathbf{\Gamma}\mathbf{B}\|_{21}\right\}, \qquad \mathbf{Y}_0 = \mathbf{Y}\widehat{\mathbf{\Sigma}}_Y^{-1/2}.$$

The convex first-order optimality condition gives, for every $\mathbf{B} \in \mathbb{R}^{p\times q}$,

$$\frac{2}{n}\langle \mathbf{X}(\mathbf{B}-\widehat{\mathbf{B}}), \mathbf{Y}_0 - \mathbf{X}\widehat{\mathbf{B}}\rangle \le \rho\{\|\mathbf{\Gamma B}\|_{21} - \|\mathbf{\Gamma}\widehat{\mathbf{B}}\|_{21}\}.$$

Since

$$\mathbf{Y}\mathbf{\Sigma}_Y^{-1/2} = \mathbf{X}\mathbf{B}^* + \mathbf{W}, \qquad \mathbf{Y}_0 = \mathbf{X}\mathbf{B}^* + \mathbf{W} + \mathbf{Y}(\widehat{\mathbf{\Sigma}}_Y^{-1/2} - \mathbf{\Sigma}_Y^{-1/2}),$$

where $\mathbf{W}$ has independent rows with covariance $\bar{\mathbf{\Sigma}} = \mathbf{I}_q - \mathbf{V}_0\mathbf{\Lambda}^2\mathbf{V}_0^\top$, we obtain

$$\begin{aligned}
&\frac{2}{n}\langle \mathbf{X}(\widehat{\mathbf{B}}-\mathbf{B}),\, \mathbf{X}(\widehat{\mathbf{B}}-\mathbf{B}^*)\rangle \\
&\qquad \le \frac{2}{n}\langle \mathbf{X}(\widehat{\mathbf{B}}-\mathbf{B}), \mathbf{Y}\big(\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_Y^{-\frac{1}{2}}\big) + \mathbf{W}\rangle + \rho\big(\|\mathbf{\Gamma B}\|_{21} - \|\mathbf{\Gamma}\widehat{\mathbf{B}}\|_{21}\big).
\end{aligned} \tag{77}$$

This means that, to proceed further, we need to bound the term:

$$\begin{aligned}
\frac{1}{n}\langle \mathbf{X}(\widehat{\mathbf{B}}-\mathbf{B}), \mathbf{Y}\big(\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_Y^{-\frac{1}{2}}\big) + \mathbf{W}\rangle &= \frac{1}{n}\langle \mathbf{\Pi}(\widehat{\mathbf{B}}-\mathbf{B}), \mathbf{\Pi}^\top\mathbf{X}^\top\big(\mathbf{Y}\big(\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_Y^{-\frac{1}{2}}\big) + \mathbf{W}\big)\rangle \\
&\quad + \frac{1}{n}\langle \mathbf{\Gamma}(\widehat{\mathbf{B}}-\mathbf{B}), (\mathbf{\Gamma}^\dagger)^\top\mathbf{X}^\top\big(\mathbf{Y}\big(\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_Y^{-\frac{1}{2}}\big) + \mathbf{W}\big)\rangle \\
&\le \underbrace{\frac{1}{n}\Big\|\mathbf{\Pi}(\widehat{\mathbf{B}}-\mathbf{B})\Big\|_F \Big\|\mathbf{\Pi}^\top\mathbf{X}^\top\big(\mathbf{Y}\big(\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_Y^{-\frac{1}{2}}\big) + \mathbf{W}\big)\Big\|_F}_{(A)} \\
&\quad + \underbrace{\big\|\mathbf{\Gamma}(\widehat{\mathbf{B}}-\mathbf{B})\big\|_{21} \max_{j\in[m]}\Big\|\Big[(\mathbf{\Gamma}^\dagger)^\top\frac{\mathbf{X}^\top\mathbf{Y}}{n}\big(\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_Y^{-\frac{1}{2}}\big)\Big]_{j\cdot}\Big\|}_{(B)} \\
&\quad + \underbrace{\big\|\mathbf{\Gamma}(\widehat{\mathbf{B}}-\mathbf{B})\big\|_{21} \max_{j\in[m]}\Big\|\Big[(\mathbf{\Gamma}^\dagger)^\top\frac{\mathbf{X}^\top\mathbf{W}}{n}\Big]_{j\cdot}\Big\|}_{(C)}
\end{aligned} \tag{78}$$

Bound (A) in Equation (78). We first consider the term:

$$\begin{aligned}
&\Big\|\mathbf{\Pi}(\widehat{\mathbf{B}}-\mathbf{B})\Big\|_F \Big\|\mathbf{\Pi}^\top\frac{\mathbf{X}^\top\mathbf{Y}}{n}\big(\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_Y^{-\frac{1}{2}}\big) + \frac{\mathbf{\Pi}^\top\mathbf{X}^\top\mathbf{W}}{n}\Big\|_F \\
&\le \|\widehat{\mathbf{B}}-\mathbf{B}\|_F \cdot \Big(\Big\|\frac{1}{n}\mathbf{\Pi}^\top\mathbf{X}^\top\mathbf{Y}\big(\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_Y^{-\frac{1}{2}}\big)\Big\|_F + \Big\|\frac{1}{n}\mathbf{\Pi}^\top\mathbf{X}^\top\mathbf{W}\Big\|_F\Big) \\
&\le \|\widehat{\mathbf{B}}-\mathbf{B}\|_F \cdot \Big(\Big\|\mathbf{\Pi}\widehat{\mathbf{\Sigma}}_{XY}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\Big\|_F \Big\|\mathbf{\Sigma}_Y^{\frac{1}{2}}\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{I}_q\Big\|_{op} + \Big\|\widehat{\mathbf{\Sigma}}_{X\Pi,W}\Big\|_F\Big) \\
&\le \|\widehat{\mathbf{B}}-\mathbf{B}\|_F \cdot \Big(\Big\|\widehat{\mathbf{\Sigma}}_{X\Pi,Y}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\Big\|_F \Big\|\widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}} - \mathbf{\Sigma}_Y^{\frac{1}{2}}\Big\|_{op}\Big\|\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}}\Big\|_{op} + \Big\|\widehat{\mathbf{\Sigma}}_{X\Pi,W}\Big\|_F\Big)
\end{aligned} \tag{79}$$

Let $\mathcal{A}_Y$ be the event: $\mathcal{A}_Y = \Big\{\Big\|\widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}} - \mathbf{\Sigma}_Y^{\frac{1}{2}}\Big\|_{op}^2 < CM^2\frac{q}{n}\Big\}$, with C a constant chosen such that $\mathcal{A}_Y$ holds with probability at least $1-\exp\{-C'q\}$ with $C'>0$ (see Lemma 5).

(i) Bounding the terms in operator norm. By Weyl's inequality, letting $\delta_n = \sqrt{\frac{q}{n}}$, we know that on $\mathcal{A}_Y$

$$\left\|\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}}\right\|_{op} \leq \frac{1}{\sigma_{\min}\left(\mathbf{\Sigma}_Y^{\frac{1}{2}}\right) - \sqrt{CM^2}\cdot\delta_n} \leq 2\sqrt{M} \tag{80}$$

as long as $\sqrt{CM^2}\cdot\sqrt{\frac{q}{n}} \leq \frac{1}{2\sqrt{M}}$. Therefore, on $\mathcal{A}_Y$

$$\left\|\widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}} - \mathbf{\Sigma}_Y^{\frac{1}{2}}\right\|_{op}\left\|\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}}\right\|_{op} \leq 2M^{\frac{3}{2}}\sqrt{C}\sqrt{\frac{q}{n}}.$$

(ii) Bounding the terms in Frobenius norm.

We begin with the term $\left\|\widehat{\mathbf{\Sigma}}_{\Pi X,Y}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right\|_F$. For $C'>0$, we denote as $\mathcal{A}_{X\Pi,Y}$ the event

$$\mathcal{A}_{X\Pi,Y} = \left\{\left\|\widehat{\mathbf{\Sigma}}_{\Pi X,Y}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right\|_F \leq \sqrt{Mn_c}\Big(1 + C\sqrt{M}\sqrt{\frac{q+\log n_c}{n}}\Big)\right\} \tag{81}$$

for some constant C that depends only on C'. As shown in Lemma 14, this event has probability at least $1-\exp\{-C'(\log n_c + q)\}$.

On $\mathcal{A}_Y \cap \mathcal{A}_{X\Pi,Y}$, after readjusting our constant C to be the maximum of the constants in the bounds of $\mathcal{A}_Y$ and $\mathcal{A}_{X\Pi,Y}$ we have:

$$\begin{aligned}
\left\|\widehat{\mathbf{\Sigma}}_{X\Pi,Y}\mathbf{\Sigma}_Y^{-\frac{1}{2}}\right\|_F &\left\|\widehat{\mathbf{\Sigma}}_Y^{\frac{1}{2}} - \mathbf{\Sigma}_Y^{\frac{1}{2}}\right\|_{op}\left\|\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}}\right\|_{op} \\
&\leq \sqrt{Mn_c}\left(1 + C\sqrt{M}\sqrt{\frac{q+\log n_c}{n}}\right)\cdot 2M^{\frac{3}{2}}\sqrt{C}\sqrt{\frac{q}{n}} \\
&\leq \tilde{C}M^2\sqrt{\frac{qn_c}{n}}\Big(1 + \sqrt{M}\cdot\sqrt{\frac{q+\log n_c}{n}}\Big) \\
&\leq 2\tilde{C}M^2\sqrt{\frac{qn_c}{n}} \qquad \text{under assumption } \mathcal{H}_{1,\mathcal{G}}.
\end{aligned} \tag{82}$$

We now turn to the term $\left\|\mathbf{\Pi}\widehat{\mathbf{\Sigma}}_{XW}\right\|_F = \left\|\widehat{\mathbf{\Sigma}}_{X\Pi,W}\right\|_F$. By Lemma 15, for $C'>0$ we know that there exists $C>0$ a constant that depends solely on C' such that:

$$\begin{aligned}
\left\|\mathbf{\Pi}\widehat{\mathbf{\Sigma}}_{XW}\right\|_F &\leq \left\|\mathbf{\Pi}\widehat{\mathbf{\Sigma}}_{X\widetilde{W}}\right\|_F\left\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}\right\|_{op} \quad \text{with} \quad \widetilde{\mathbf{W}} \sim \mathcal{N}(0,\mathbf{I}_q) \\
&\leq CM\sqrt{n_c\cdot\frac{q+\log n_c}{n}} \quad \text{since} \quad \left\|\bar{\mathbf{\Sigma}}^{\frac{1}{2}}\right\|_{op} \leq 1
\end{aligned} \tag{83}$$

with probability at least $1-\exp\{-C'(q+\log n_c)\}$. In the rest of this proof, we denote as $\mathcal{A}_{\Pi XW}$ the event

$$\mathcal{A}_{\Pi XW} = \Big\{\left\|\mathbf{\Pi}^\top\mathbf{X}^\top\mathbf{W}/n\right\|_F \leq CM\sqrt{n_c\cdot\frac{q+\log n_c}{n}}\Big\}. \tag{84}$$

Combining the results together and taking C the maximum of all constants in the corresponding bounds we deduce that on the event $\mathcal{A}_{\Pi XW} \cap \mathcal{A}_Y \cap \mathcal{A}_{X\Pi,Y}$,

$$\begin{aligned}\left\|\boldsymbol{\Pi}^\top \frac{\mathbf{X}^\top \mathbf{Y}}{n}\left(\widehat{\boldsymbol{\Sigma}}_Y^{-\frac{1}{2}} - \boldsymbol{\Sigma}_Y^{-\frac{1}{2}}\right) + \boldsymbol{\Pi}^\top \frac{\mathbf{X}^\top \mathbf{W}}{n}\right\|_F &\leq CM^2\sqrt{\frac{qn_c}{n}} + CM\sqrt{n_c \cdot \frac{q+\log n_c}{n}} \\ &\leq CM^2\sqrt{n_c\frac{q+\log n_c}{n}}\end{aligned} \tag{85}$$

To characterize the corresponding probability, we note that:

$$\begin{aligned}\mathbb{P}\left[\mathcal{A}^c_{\Pi XW} \cup \mathcal{A}^c_Y \cup \mathcal{A}^c_{X\Pi,Y}\right] &\leq \mathbb{P}\left[\mathcal{A}^c_{\Pi XW}\right] + \mathbb{P}\left[\mathcal{A}^c_Y\right] + \mathbb{P}\left[\mathcal{A}^c_{X\Pi,Y}\right] \\ &\leq \exp\{-C'(\log n_c + q)\} + \exp\{-C'q\} + \exp\{-C'(q+\log n_c)\} \\ &\leq 3\exp\{-C'(q+\log n_c)\}\end{aligned} \tag{86}$$

Bound (B) in Equation (78). We now proceed to bound the term:

$$\begin{aligned}\max_{j\in[m]}\left\|\left[(\boldsymbol{\Gamma}^\dagger)^\top \frac{\mathbf{X}^\top \mathbf{Y}}{n}\left(\widehat{\boldsymbol{\Sigma}}_Y^{-\frac{1}{2}} - \boldsymbol{\Sigma}_Y^{-\frac{1}{2}}\right)\right]_{j\cdot}\right\| &\leq \max_{j\in[m]}\left\|\left[(\boldsymbol{\Gamma}^\dagger)^\top \frac{\mathbf{X}^\top \mathbf{Y}}{n}\boldsymbol{\Sigma}_Y^{-\frac{1}{2}}\right]_{j\cdot}\right\| \left\|\boldsymbol{\Sigma}_Y^{\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{I}_q\right\|_{op} \\ &\leq \max_{j\in[m]}\left\|\left[(\boldsymbol{\Gamma}^\dagger)^\top \frac{\mathbf{X}^\top \mathbf{Y}}{n}\boldsymbol{\Sigma}_Y^{-\frac{1}{2}}\right]_{j\cdot}\right\| \left\|\boldsymbol{\Sigma}_Y^{\frac{1}{2}} - \widehat{\boldsymbol{\Sigma}}_Y^{\frac{1}{2}}\right\|_{op} \left\|\widehat{\boldsymbol{\Sigma}}_Y^{-\frac{1}{2}}\right\|_{op}\end{aligned} \tag{87}$$

For $C' > 0$ a constant, let us denote as $\mathcal{A}_{(\Gamma^\dagger)^\top\Sigma_{XY}}$ the event:

$$\mathcal{A}_{(\Gamma^\dagger)^\top\Sigma_{XY}} = \left\{\max_{j\in[m]}\left\|\left[(\boldsymbol{\Gamma}^\dagger)^\top \widehat{\boldsymbol{\Sigma}}_{XY}\boldsymbol{\Sigma}_Y^{-\frac{1}{2}}\right]_{j\cdot}\right\| \geq \sqrt{\frac{M}{\kappa_2}} + C\left(1+\frac{M}{\kappa_2}\right)\cdot\sqrt{\frac{q+\log m}{n}}\right\}.$$

In the previous equation, we've chosen C to be a constant that depends solely on C' such that, with probability at least $1 - \exp\{-C'(q+\log m)\} - \exp\{-C'q\}$, the event $\mathcal{A}^c_{(\Gamma^\dagger)^\top\Sigma_{XY}} \cap \mathcal{A}_Y$ occurs (see Lemma 13). Thus

$$\begin{aligned}&\max_{j\in[m]}\left\|\left[(\boldsymbol{\Gamma}^\dagger)^\top \widehat{\boldsymbol{\Sigma}}_{XY}\left(\widehat{\boldsymbol{\Sigma}}_Y^{-\frac{1}{2}} - \boldsymbol{\Sigma}_Y^{-\frac{1}{2}}\right)\right]_{j\cdot}\right\| \\ &\quad\leq \max_{j\in[m]}\left\|\left[(\boldsymbol{\Gamma}^\dagger)^\top \widehat{\boldsymbol{\Sigma}}_{XY}\boldsymbol{\Sigma}_Y^{-\frac{1}{2}}\right]_{j\cdot}\right\| \cdot \sqrt{CM^2\frac{q}{n}}\cdot 2\sqrt{M} \quad \text{by Eq.(80)} \\ &\quad\leq M^2\left(\sqrt{\frac{1}{\kappa_2}} + C\left(1+\frac{1}{\kappa_2}\right)\sqrt{M}\cdot\sqrt{\frac{q+\log m}{n}}\right)\cdot\sqrt{\frac{q}{n}} \quad \text{on } \mathcal{A}^c_{(\Gamma^\dagger)^\top\Sigma_{XY}} \\ &\quad\leq C'M^2\left(1+\frac{1}{\kappa_2}\right)\sqrt{\frac{q}{n}} \qquad \text{under } H_{1.\mathcal{G}}.\end{aligned} \tag{88}$$

where $C' = 1 \vee C$.

Bound (C) in Equation (78). For $\mathbf{W} \sim \mathcal{N}(0, \bar{\mathbf{\Sigma}})$ with covariance $\bar{\mathbf{\Sigma}} = \mathbf{I}_q - \mathbf{V}_0 \mathbf{\Lambda}^2 \mathbf{V}_0^\top$, we now proceed to bound the term:

$$\max_{j \in [m]} \left\| \left[(\mathbf{\Gamma}^\dagger)^\top \frac{\mathbf{X}^\top \mathbf{W}}{n} \right]_{j\cdot} \right\| \tag{89}$$

For $C' > 0$, let us denote as $\mathcal{A}_{X\Gamma^\dagger, W}$ the event:

$$\mathcal{A}_{X\Gamma^\dagger, W} = \left\{ \max_{j \in [m]} \left\| \left[(\mathbf{\Gamma}^\dagger)^\top \frac{\mathbf{X}^\top \mathbf{W}}{n} \right]_{j\cdot} \right\| \geq C \cdot (1 + \frac{M}{\kappa_2}) \sqrt{\frac{q + \log m}{n}} \right\}.$$

In the previous equation, we've chosen C such that the event $\mathcal{A}^c_{X\Gamma^\dagger, W}$ occurs with probability at least $1 - \exp\{-C'(q + \log m)\}$ (see Lemma 12).

Bounding (A) +(B) + (C) in Equation (78). Define

$$R_\Pi = CM^2 \sqrt{n_c \frac{q + \log n_c}{n}}, \qquad R_\Gamma = CM^2 \left(1 + \frac{1}{\kappa_2}\right) \sqrt{\frac{q + \log m}{n}}.$$

Combining the three previous bounds, we obtain on the event

$$\mathcal{E}_{\text{score}} = \mathcal{A}_Y \cap \mathcal{A}_{X\Pi, Y} \cap \mathcal{A}_{\Pi X W} \cap \mathcal{A}^c_{X\Gamma^\dagger, W} \cap \mathcal{A}^c_{(\Gamma^\dagger)^\top \Sigma_{XY}}$$

that

$$\frac{1}{n} \left| \left\langle \mathbf{X}(\widehat{\mathbf{B}} - \mathbf{B}),\ \mathbf{Y}\left(\widehat{\mathbf{\Sigma}}_Y^{-\frac{1}{2}} - \mathbf{\Sigma}_Y^{-\frac{1}{2}}\right) + \mathbf{W} \right\rangle \right| \leq R_\Pi \|\widehat{\mathbf{B}} - \mathbf{B}\|_F + R_\Gamma \|\mathbf{\Gamma}(\widehat{\mathbf{B}} - \mathbf{B})\|_{21}. \tag{90}$$

Bounding the regression error For any $S \subseteq [m]$:

$$\begin{aligned} \|\mathbf{\Gamma B}\|_{21} - \|\mathbf{\Gamma}\widehat{\mathbf{B}}\|_{21} &\leq \big\|[\mathbf{\Gamma}(\mathbf{B} - \widehat{\mathbf{B}})]_{S\cdot}\big\|_{21} + \big\|[\mathbf{\Gamma B}]_{S^c}\big\|_{21} - \big\|[\mathbf{\Gamma}\widehat{\mathbf{B}}]_{S^c}\big\|_{21} \\ &\leq \big\|[\mathbf{\Gamma}(\mathbf{B} - \widehat{\mathbf{B}})]_{S\cdot}\big\|_{21} + 2\big\|[\mathbf{\Gamma B}]_{S^c}\big\|_{21} - \big\|[\mathbf{\Gamma}(\widehat{\mathbf{B}} - \mathbf{B})]_{S^c}\big\|_{21}. \end{aligned} \tag{91}$$

Since $\rho \geq 4R_\Gamma$, Equations (77) and (90) imply, on $\mathcal{E}_{\text{score}}$,

$$\begin{aligned} \frac{2}{n} \left\langle \mathbf{X}(\widehat{\mathbf{B}} - \mathbf{B}), \mathbf{X}(\widehat{\mathbf{B}} - \mathbf{B}^*) \right\rangle &\leq 2R_\Pi \|\widehat{\mathbf{B}} - \mathbf{B}\|_F + \frac{\rho}{2} \|\mathbf{\Gamma}(\widehat{\mathbf{B}} - \mathbf{B})\|_{21} \\ &+ \rho\Big(\big\|[\mathbf{\Gamma}(\mathbf{B} - \widehat{\mathbf{B}})]_{S\cdot}\big\|_{21} + 2\big\|[\mathbf{\Gamma B}]_{S^c}\big\|_{21} - \big\|[\mathbf{\Gamma}(\widehat{\mathbf{B}} - \mathbf{B})]_{S^c}\big\|_{21} \Big). \end{aligned} \tag{92}$$

Therefore, for $\mathbf{B} = \mathbf{B}^*$ and choosing $S = \text{supp}(\mathbf{\Gamma B}^*) \subseteq \text{supp}(\mathbf{\Gamma U})$, so that $|S| \leq s_u$ and $[\mathbf{\Gamma B}^*]_{S^c\cdot} = 0$, we have:

$$\frac{1}{n} \|\mathbf{X\Delta}\|_F^2 \leq CR_\Pi \|\mathbf{\Delta}\|_F + \rho \left(\frac{3}{2} \big\|[\mathbf{\Gamma\Delta}]_{S\cdot}\big\|_{21} - \frac{1}{2} \big\|[\mathbf{\Gamma\Delta}]_{S^c\cdot}\big\|_{21} \right), \tag{93}$$

where we denote $\mathbf{\Delta} = \widehat{\mathbf{B}} - \mathbf{B}^*$. Note that, by Cauchy-Schwarz,

$$\begin{aligned} \big\|[\mathbf{\Gamma\Delta}]_{S\cdot}\big\|_{21} = \sum_{j \in S} \big\|[\mathbf{\Gamma\Delta}]_{j\cdot}\big\| &\leq \sqrt{s_u} \sqrt{\sum_{j \in S} \big\|[\mathbf{\Gamma\Delta}]_{j\cdot}\big\|^2} \\ &\leq \sqrt{s_u} \cdot \|\mathbf{\Gamma\Delta}\|_F \leq \sqrt{s_u \sigma_{\max}(\mathbf{L})} \cdot \|\mathbf{\Delta}\|_F. \end{aligned} \tag{94}$$

The expression in Equation (93) can be thus simplified to

$$\frac{1}{n}\|\mathbf{X}\mathbf{\Delta}\|_F^2 \leq \|\mathbf{\Delta}\|_F \left(CR_\Pi + \frac{3\rho}{2}\sqrt{s_u \sigma_{\max}(\mathbf{L})} \right) - \frac{\rho}{2}\big\|[\mathbf{\Gamma}\mathbf{\Delta}]_{S^c\cdot}\big\|_{21}. \tag{95}$$

Note that Equation (95) implies the following generalized cone constraint:

$$\big\|[\mathbf{\Gamma}\mathbf{\Delta}]_{S^c\cdot}\big\|_{21} \leq \|\mathbf{\Delta}\|_F \left(\frac{2CR_\Pi}{\rho} + 3\sqrt{s_u \sigma_{\max}(\mathbf{L})} \right). \tag{96}$$

Equation (95) also implies:

$$\frac{1}{n}\|\mathbf{X}\mathbf{\Delta}\|_F^2 \leq \|\mathbf{\Delta}\|_F \left(CR_\Pi + \frac{3\rho}{2}\sqrt{s_u \sigma_{\max}(\mathbf{L})} \right). \tag{97}$$

To proceed any further, we must show a Restricted Eigenvalue property on our generalised cone (Eq. 96). To this end, we leverage Lemma 2.2 in Tran et al. (2022).

Lemma 18 (Lemma 2.2 in Tran et al. (2022)) *If $\mathbf{X} \in \mathbb{R}^{n\times p}$ has i.i.d. $\mathcal{N}(0, \Sigma_X)$ rows and $p \geq 2$, $n \geq 10$, then the event*

$$\left\| \frac{\mathbf{X}v}{\sqrt{n}} \right\|^2 \geq v^\top \left(\frac{1}{64}\mathbf{\Sigma}_X \right) v - 72\sigma_{max}(\mathbf{\Sigma}_X) n_c \frac{\|v\|^2}{n} - 576\rho(\mathbf{\Gamma})^2 \frac{\sigma_{max}(\mathbf{\Sigma}_X)\log p}{n} \|\mathbf{\Gamma}v\|_1^2$$

holds for all $v \in \mathbb{R}^p$ with probability at least $1 - c_1 \exp\{-nc_2\}$, for some universal constants $c_1, c_2 > 0$.

Let $\mathcal{A}_\kappa^j$ and $\mathcal{A}_\kappa$ denote the events:

$$\mathcal{A}_\kappa^j = \left\{ \left\| \frac{\mathbf{X}\mathbf{\Delta}_{\cdot j}}{\sqrt{n}} \right\|^2 \geq \mathbf{\Delta}_{\cdot j}^\top \left(\frac{1}{64}\mathbf{\Sigma}_X \right) \mathbf{\Delta}_{\cdot j} - 72\sigma_{\max}(\mathbf{\Sigma}_X) n_c \frac{\|\mathbf{\Delta}_{\cdot j}\|^2}{n} - 576\rho(\mathbf{\Gamma})^2 \frac{\sigma_{\max}(\mathbf{\Sigma}_X)\log p}{n} \|\mathbf{\Gamma}\mathbf{\Delta}_{\cdot j}\|_1^2 \right\}$$

$$\mathcal{A}_\kappa = \left\{ \frac{\|\mathbf{X}\mathbf{\Delta}\|_F^2}{n} \geq \frac{1}{64}\left\| \mathbf{\Sigma}_X^{\frac{1}{2}}\mathbf{\Delta} \right\|_F^2 - \frac{72Mn_c}{n}\|\mathbf{\Delta}\|_F^2 - 576\frac{\rho(\mathbf{\Gamma})^2 M\log p}{n} \sum_{j\in[q]} \big\|[\mathbf{\Gamma}\mathbf{\Delta}]_{\cdot j}\big\|_1^2 \right\}$$

By Lemma 18, applying a simple union bound for q vectors $\mathcal{A}_\kappa^c \subset \bigcup_{j\in[q]} (\mathcal{A}_\kappa^j)^c$, therefore:

$$\mathbb{P}[\mathcal{A}_\kappa^c] \leq \sum_{j\in[q]} \mathbb{P}\big[(\mathcal{A}_\kappa^j)^c\big] \leq qe^{-c_2 n} \tag{98}$$

This means in particular:

$$\frac{1}{64}\left\| \mathbf{\Sigma}_X^{\frac{1}{2}}\mathbf{\Delta} \right\|_F^2 \leq \frac{\|\mathbf{X}\mathbf{\Delta}\|_F^2}{n} + \frac{72Mn_c}{n}\|\mathbf{\Delta}\|_F^2 + 576\frac{\rho(\mathbf{\Gamma})^2 M\log p}{n} \sum_{j\in[q]} \big\|[\mathbf{\Gamma}\mathbf{\Delta}]_{\cdot j}\big\|_1^2 \tag{99}$$

We note that, for any $v \in \mathbb{R}^{p\times q}$,

$$\begin{aligned}\sum_{j\in[q]} \left\|[\mathbf{\Gamma} v]_{\cdot j}\right\|_1^2 &= \sum_{j\in[q]}\sum_{a\in[m]} \left|[\mathbf{\Gamma} v]_{aj}\right| \sum_{b\in[m]} \left|[\mathbf{\Gamma} v]_{bj}\right| \\ &= \sum_{a\in[m]}\sum_{b\in[m]}\sum_{j\in[q]} \left|[\mathbf{\Gamma} v]_{aj}\right|\left|[\mathbf{\Gamma} v]_{bj}\right| \\ &\leq \sum_{a\in[m]}\sum_{b\in[m]} \left\|[\mathbf{\Gamma} v]_{a\cdot}\right\| \left\|[\mathbf{\Gamma} v]_{b\cdot}\right\| = \|\mathbf{\Gamma} v\|_{21}^2.\end{aligned} \tag{100}$$

This is the point where no support bound on $\widehat{\mathbf{B}}$ is needed: the graph restricted-eigenvalue lemma is uniform in v, and the cone bound controls $\|\mathbf{\Gamma}\Delta\|_{21}$ directly. Combining Equations (94) and (96),

$$\begin{aligned}\|\mathbf{\Gamma}\mathbf{\Delta}\|_{21} &\leq \left\|[\mathbf{\Gamma}\mathbf{\Delta}]_{S^c\cdot}\right\|_{21} + \left\|[\mathbf{\Gamma}\mathbf{\Delta}]_{S\cdot}\right\|_{21} \\ &\leq \|\mathbf{\Delta}\|_F \left(\frac{2CR_\Pi}{\rho} + 4\sqrt{s_u \sigma_{\max}(\mathbf{L})}\right).\end{aligned} \tag{101}$$

Let

$$K_\Gamma = \frac{2CR_\Pi}{\rho} + 4\sqrt{s_u \sigma_{\max}(\mathbf{L})}.$$

By the lower bound on ρ, $K_\Gamma \leq K_{\Gamma,n}$ after increasing the numerical constant $C_\star$ in the theorem statement. Combined with Equations (101) and (97), Equation (99) becomes

$$\begin{aligned}\frac{1}{64}\left\|\mathbf{\Sigma}_X^{\frac{1}{2}}\mathbf{\Delta}\right\|_F^2 &\leq \|\mathbf{\Delta}\|_F \left(CR_\Pi + \frac{3\rho}{2}\sqrt{s_u\sigma_{\max}(\mathbf{L})}\right) \\ &\quad + \frac{M}{n}\|\mathbf{\Delta}\|_F^2 \left(72 n_c + 576\rho(\mathbf{\Gamma})^2 \log p\, K_\Gamma^2\right).\end{aligned} \tag{102}$$

Using $\sigma_{\min}(\mathbf{\Sigma}_X) \geq 1/M$, if

$$n \geq 128 M^2 \left(72 n_c + 576\rho(\mathbf{\Gamma})^2 \log p\, K_\Gamma^2\right),$$

the quadratic term on the right-hand side is absorbed into the left-hand side, and hence

$$\begin{aligned}\|\mathbf{\Delta}\|_F &\leq CM\left(R_\Pi + \rho\sqrt{s_u\sigma_{\max}(\mathbf{L})}\right) \\ &\leq CM^3\left(\sqrt{n_c\frac{q+\log n_c}{n}} + \left(1+\frac{1}{\kappa_2}\right)\sqrt{\sigma_{\max}(\mathbf{L})}\sqrt{s_u\frac{q+\log m}{n}}\right).\end{aligned} \tag{103}$$

Consequently, Equation (97) gives

$$\begin{aligned}\frac{1}{n}\|\mathbf{X}\mathbf{\Delta}\|_F^2 &= \|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{\Delta}\|_F^2 \\ &\leq CM^5\left(n_c\frac{q+\log(n_c)}{n} + \left(1+\frac{1}{\kappa_2}\right)^2 \sigma_{\max}(\mathbf{L})\cdot s_u\frac{q+\log m}{n}\right).\end{aligned} \tag{104}$$

We conclude the proof by characterizing the probability of the event

$$\mathcal{E}_{\text{graph}} = \mathcal{E}_{\text{score}} \cap \mathcal{A}_\kappa.$$

By the union bound and Equation (98),

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{\text{graph}}) &\geq 1 - \mathbb{P}(\mathcal{A}_Y^c) - \mathbb{P}(\mathcal{A}_{X\Pi,Y}^c) - \mathbb{P}(\mathcal{A}_{\Pi XW}^c) \\ &\quad - \mathbb{P}(\mathcal{A}_{X\Gamma^\dagger,W}) - \mathbb{P}(\mathcal{A}_{(\Gamma^\dagger)^\top \Sigma_{XY}}) - \mathbb{P}(\mathcal{A}_\kappa^c) \\ &\geq 1 - \exp\{-C'q\} - 2\exp\{-C'(q+\log n_c)\} - 2\exp\{-C'(q+\log m)\} - q\exp\{-c_2 n\}. \end{aligned} \tag{105}$$

■

F.3 Consistency of the graph-regularized CCA estimator

Lemma 19 (Graph-restricted covariance for the normalization step) *For $R \geq 1$, define*

$$\mathbb{T}_{\mathcal{G}}(R) = \left\{\mathbf{H} \in \mathbb{R}^{p\times r} : \|\mathbf{H}\|_F \leq 1, \|\mathbf{\Gamma H}\|_{21} \leq R\right\}$$

and

$$w_{\mathcal{G}}(R) = \sqrt{rn_c} + R\rho(\mathbf{\Gamma})\sqrt{r+\log m}, \qquad \varepsilon_{\mathcal{G}}(R) = CM\left\{\frac{w_{\mathcal{G}}(R)}{\sqrt{n}} + \frac{w_{\mathcal{G}}^2(R)}{n}\right\}.$$

Then, for a numerical constant $c > 0$, with probability at least $1 - 2\exp\{-cw_{\mathcal{G}}^2(R)\}$,

$$\sup_{\mathbf{H}_1,\mathbf{H}_2\in\mathbb{T}_{\mathcal{G}}(R)} \left\|\mathbf{H}_1^\top(\widehat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X)\mathbf{H}_2\right\|_F \leq \varepsilon_{\mathcal{G}}(R). \tag{106}$$

Proof Let $\mathbf{G} \in \mathbb{R}^{p\times r}$ have i.i.d. $N(0,1)$ entries. For any $\mathbf{H} \in \mathbb{T}_{\mathcal{G}}(R)$, use $\mathbf{H} = \mathbf{\Pi H} + \mathbf{\Gamma}^\dagger\mathbf{\Gamma H}$ to write

$$\langle \mathbf{G}, \mathbf{H}\rangle = \langle \mathbf{\Pi G}, \mathbf{\Pi H}\rangle + \langle (\mathbf{\Gamma}^\dagger)^\top \mathbf{G}, \mathbf{\Gamma H}\rangle.$$

Therefore

$$\sup_{\mathbf{H}\in\mathbb{T}_{\mathcal{G}}(R)} \langle \mathbf{G}, \mathbf{H}\rangle \leq \|\mathbf{\Pi G}\|_F + R\|(\mathbf{\Gamma}^\dagger)^\top \mathbf{G}\|_{\infty,2}.$$

Since $\operatorname{rank}(\mathbf{\Pi}) = n_c$, $\mathbb{E}\|\mathbf{\Pi G}\|_F \leq \sqrt{rn_c}$. Also, by the definition of $\rho(\mathbf{\Gamma})$ and a union bound over the m rows,

$$\mathbb{E}\|(\mathbf{\Gamma}^\dagger)^\top \mathbf{G}\|_{\infty,2} \leq C\rho(\mathbf{\Gamma})\sqrt{r+\log m}.$$

Thus the Gaussian width of $\mathbb{T}_{\mathcal{G}}(R)$ is at most $Cw_{\mathcal{G}}(R)$. The standard Gaussian quadratic-process bound for subgaussian quadratic forms over a star-shaped class then gives

$$\sup_{\mathbf{H}\in\mathbb{T}_{\mathcal{G}}(R)} \left|\|\widehat{\mathbf{\Sigma}}_X^{1/2}\mathbf{H}\|_F^2 - \|\mathbf{\Sigma}_X^{1/2}\mathbf{H}\|_F^2\right| \leq CM\left\{\frac{w_{\mathcal{G}}(R)}{\sqrt{n}} + \frac{w_{\mathcal{G}}^2(R)}{n}\right\}$$

with probability at least $1 - 2\exp\{-cw_{\mathcal{G}}^2(R)\}$. Finally, the bilinear form bound (106) follows by polarization. Indeed, for any matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{r\times r}$ with $\|\mathbf{A}\|_F \leq 1$ and

$\|\mathbf{B}\|_F \leq 1$, the matrices $\mathbf{H}_1\mathbf{A}$ and $\mathbf{H}_2\mathbf{B}$ remain in $\mathbb{T}_{\mathcal{G}}(R)$, and applying the quadratic bound to $(\mathbf{H}_1\mathbf{A} \pm \mathbf{H}_2\mathbf{B})/2$ yields the displayed Frobenius bound. ■

Theorem 20 *Consider the family $\mathcal{F}_{\mathcal{G}}(s_u, p, m, q, r, \lambda, M)$ of covariance matrices satisfying assumptions. We assume as well that n is large enough so that:*

$$n \geq 10 \vee 128M^2\Big(72n_c + 576\rho(\mathbf{\Gamma})^2 \log p\big(96C'\sqrt{n_c} + 4\sqrt{s_u\sigma_{\max}(\mathbf{L})}\big)^2\Big),$$

and such that:

$$M\frac{q + s_u \log\left(\frac{ep}{s_u}\right)}{n} \leq c_0 \qquad \text{and} \qquad M\frac{\log n_c + q}{n} \leq c_0 \tag{$\mathcal{H}_{1,\mathcal{G}}$}$$

Define

$$\delta_{\mathcal{G}} = \sqrt{n_c\frac{q + \log n_c}{n}} + \left(1 + \frac{1}{\kappa_2}\right)\sqrt{\sigma_{\max}(\mathbf{L})s_u\frac{q + \log m}{n}},$$

$$R_{\mathcal{G}} = C_R\frac{M^{7/2}}{\lambda}\left(\sqrt{n_c} + \sqrt{s_u\sigma_{\max}(\mathbf{L})}\right), \qquad \varepsilon_{\mathcal{G}} = \varepsilon_{\mathcal{G}}(R_{\mathcal{G}}),$$

where $\varepsilon_{\mathcal{G}}(\cdot)$ is defined in Lemma 19. We also assume the normalization radius is small enough that

$$C\left\{\frac{M^{9/2}\sqrt{r}}{\lambda}\delta_{\mathcal{G}} + r\varepsilon_{\mathcal{G}}\right\} \leq \frac{1}{2M}.$$

Then, there exist orthogonal matrices $\mathbf{O} \in \mathbb{R}^{r\times r}$ and $\tilde{\mathbf{O}} \in \mathbb{R}^{r\times r}$ such that the canonical directions $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ estimated using the graph algorithm have errors bounded by:

$$\left\|\widehat{\mathbf{V}} - \mathbf{V}\mathbf{O}\right\|_F \leq CM^{\frac{11}{2}}\frac{\sqrt{r}}{\lambda}\left(\sqrt{n_c\frac{q + \log(n_c)}{n}} + \left(1 + \frac{1}{\kappa_2}\right)\sqrt{\sigma_{\max}(\mathbf{L})}\cdot\sqrt{s_u\frac{q + \log m}{n}}\right)$$

$$\|\widehat{\mathbf{U}} - \mathbf{U}\widetilde{\mathbf{O}}\|_F \leq CM^6\frac{\sqrt{r}}{\lambda}\left(\sqrt{n_c\frac{q + \log(n_c)}{n}} + \left(1 + \frac{1}{\kappa_2}\right)\sqrt{\sigma_{\max}(\mathbf{L})}\cdot\sqrt{s_u\frac{q + \log m}{n}}\right) + CM^{3/2}r\varepsilon_{\mathcal{G}}.$$

with probability at least

$$1 - \exp\{-C'q\} - 3\exp\{-C'(q + \log n_c)\} - 2\exp\{-C'(q + \log m)\} - q\exp\{-c_2 n\} - 2\exp\{-cw_{\mathcal{G}}^2(R_{\mathcal{G}})\}.$$

Here C is a constant that depends solely on C' and c_0, and c_2 is a universal constant that does not depend on the dimensions.

Proof Let

$$\delta_{\mathcal{G}} = \sqrt{n_c\frac{q + \log n_c}{n}} + \left(1 + \frac{1}{\kappa_2}\right)\sqrt{\sigma_{\max}(\mathbf{L})s_u\frac{q + \log m}{n}}.$$

Theorem 17 gives

$$\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_F \leq CM^3\delta_{\mathcal{G}}. \tag{107}$$

Let

$$\widehat{\mathbf{B}} = \widehat{\mathbf{U}}_{\widehat{B}} \widehat{\mathbf{D}}_{\widehat{B}} \widehat{\mathbf{V}}_{\widehat{B}}^\top$$

be the rank-r truncated SVD used in the graph Algorithm, and let $\mathbf{B}^* = \mathbf{U}_B \mathbf{D}_B \mathbf{V}_B^\top$ be the compact SVD of $\mathbf{B}^*$. As in the sparse proof, $\mathbf{V}_B = \mathbf{V}_0 \mathbf{Q}_Y$ for some $\mathbf{Q}_Y \in \mathcal{O}_r$, where $\mathbf{V}_0 = \mathbf{\Sigma}_Y^{1/2} \mathbf{V}$, and

$$\sigma_r(\mathbf{B}^*) \geq \lambda / \sqrt{M}.$$

For sufficiently small c_0, (107) is smaller than a fixed fraction of $\sigma_r(\mathbf{B}^*)$. Therefore Wedin's theorem gives orthogonal matrices $\mathbf{O}_X, \mathbf{O}_Y \in \mathcal{O}_r$ such that

$$\|\widehat{\mathbf{U}}_{\widehat{B}} - \mathbf{U}_B \mathbf{O}_X\|_F \vee \|\widehat{\mathbf{V}}_{\widehat{B}} - \mathbf{V}_B \mathbf{O}_Y\|_F \leq C \frac{M^{7/2}}{\lambda} \delta_{\mathcal{G}}. \tag{108}$$

The Y-side normalization follows exactly as in the sparse proof. On the covariance event for $\widehat{\mathbf{\Sigma}}_Y$,

$$\|\widehat{\mathbf{\Sigma}}_Y^{-1/2}\|_{op} \leq C\sqrt{M}, \qquad \|\widehat{\mathbf{\Sigma}}_Y^{-1/2} - \mathbf{\Sigma}_Y^{-1/2}\|_{op} \leq C M^2 \sqrt{q/n}.$$

Since $\mathbf{\Sigma}_Y^{-1/2} \mathbf{V}_B = \mathbf{V} \mathbf{Q}_Y$, we obtain

$$\|\widehat{\mathbf{V}} - \mathbf{V} \mathbf{Q}_Y \mathbf{O}_Y\|_F \leq C \frac{M^{11/2} \sqrt{r}}{\lambda} \delta_{\mathcal{G}},$$

after absorbing the $\sqrt{qr/n}$ covariance-normalization term into $\sqrt{r}\delta_{\mathcal{G}}$.

It remains to justify the X-side normalization. This is where the sparse proof cannot be copied verbatim: the graph penalty controls $\mathbf{\Gamma}\widehat{\mathbf{B}}$, but it does not imply that $\widehat{\mathbf{B}}$ has a small row support. We therefore use Lemma 19 instead of a selected-support covariance bound.

First, we check that the singular-vector matrices to which the covariance event is applied belong to the stated graph class. Let $S = \mathrm{supp}(\mathbf{\Gamma}\mathbf{B}^*)$. Since $\mathrm{col}(\mathbf{U}_B) = \mathrm{col}(\mathbf{U})$, we have $\mathbf{U}_B = \mathbf{U}\mathbf{R}$ for some invertible $r \times r$ matrix $\mathbf{R}$, and therefore $\mathrm{supp}(\mathbf{\Gamma}\mathbf{U}_B) \subseteq S$. Hence

$$\|\mathbf{\Gamma}\mathbf{U}_B\|_{21} \leq \sqrt{s_u} \|\mathbf{\Gamma}\mathbf{U}_B\|_F \leq \sqrt{r s_u \sigma_{\max}(\mathbf{L})}.$$

Thus $\mathbf{U}_B / \sqrt{r} \in \mathbb{T}_{\mathcal{G}}(R_{\mathcal{G}})$. For the estimated left singular vectors, use $\widehat{\mathbf{U}}_{\widehat{B}} = \widehat{\mathbf{B}} \widehat{\mathbf{V}}_{\widehat{B}} \widehat{\mathbf{D}}_{\widehat{B}}^{-1}$ and Weyl's inequality, which gives $\sigma_r(\widehat{\mathbf{B}}) \geq \lambda / (2\sqrt{M})$ on the present event. Therefore

$$\|\mathbf{\Gamma}\widehat{\mathbf{U}}_{\widehat{B}}\|_{21} \leq \frac{2\sqrt{M}}{\lambda} \left(\|\mathbf{\Gamma}\mathbf{B}^*\|_{21} + \|\mathbf{\Gamma}\mathbf{\Delta}\|_{21} \right).$$

Moreover,

$$\|\mathbf{\Gamma}\mathbf{B}^*\|_{21} \leq \sqrt{s_u} \|\mathbf{\Gamma}\mathbf{B}^*\|_F \leq \sqrt{M r s_u \sigma_{\max}(\mathbf{L})},$$

and the cone bound in the regression proof gives

$$\|\mathbf{\Gamma}\mathbf{\Delta}\|_{21} \leq C \|\mathbf{\Delta}\|_F \left(\sqrt{n_c} + \sqrt{s_u \sigma_{\max}(\mathbf{L})} \right) \leq C M^3 \delta_{\mathcal{G}} \left(\sqrt{n_c} + \sqrt{s_u \sigma_{\max}(\mathbf{L})} \right).$$

Since c_0 is chosen small enough, the last display is dominated by the radius defining $R_{\mathcal{G}}$, and hence $\widehat{\mathbf{U}}_{\widehat{B}} / \sqrt{r} \in \mathbb{T}_{\mathcal{G}}(R_{\mathcal{G}})$. Consequently, on the event of Lemma 19,

$$\left\| \widehat{\mathbf{U}}_{\widehat{B}}^\top (\widehat{\mathbf{\Sigma}}_X - \mathbf{\Sigma}_X) \widehat{\mathbf{U}}_{\widehat{B}} \right\|_F \leq r \varepsilon_{\mathcal{G}}. \tag{109}$$

Define

$$\widehat{\mathbf{A}} = \widehat{\mathbf{U}}_{\widehat{B}}^\top \widehat{\mathbf{\Sigma}}_X \widehat{\mathbf{U}}_{\widehat{B}}, \qquad \mathbf{A} = \mathbf{U}_B^\top \mathbf{\Sigma}_X \mathbf{U}_B.$$

The eigenvalues of $\mathbf{A}$ lie in $[1/M, M]$. Combining (108) with (109) gives

$$\|\widehat{\mathbf{A}} - \mathbf{O}_X^\top \mathbf{A} \mathbf{O}_X\|_F \leq C \frac{M^{9/2}\sqrt{r}}{\lambda} \delta_{\mathcal{G}} + Cr\varepsilon_{\mathcal{G}}. \tag{110}$$

Indeed, writing $\mathbf{E}_U = \widehat{\mathbf{U}}_{\widehat{B}} - \mathbf{U}_B \mathbf{O}_X$, the population part is bounded by $CM\|\mathbf{E}_U\|_F$, and the empirical covariance part is bounded by (109). By the small-radius assumption in the statement, $\lambda_{\min}(\widehat{\mathbf{A}}) \geq 1/(2M)$. The inverse-square-root perturbation bound on matrices with eigenvalues bounded below by $1/(2M)$ gives

$$\|\widehat{\mathbf{A}}^{-1/2} - \mathbf{O}_X^\top \mathbf{A}^{-1/2} \mathbf{O}_X\|_F \leq C \frac{M^6\sqrt{r}}{\lambda} \delta_{\mathcal{G}} + CM^{3/2} r\varepsilon_{\mathcal{G}}. \tag{111}$$

Finally, let $\mathbf{U}_\diamond = \mathbf{U}_B(\mathbf{U}_B^\top \mathbf{\Sigma}_X \mathbf{U}_B)^{-1/2}$. As before, $\mathbf{U}_\diamond = \mathbf{U}\mathbf{Q}_X$ for some $\mathbf{Q}_X \in \mathcal{O}_r$. Since

$$\widehat{\mathbf{U}} = \widehat{\mathbf{U}}_{\widehat{B}} \widehat{\mathbf{A}}^{-1/2},$$

we have, using (108) and (111),

$$\|\widehat{\mathbf{U}} - \mathbf{U}_\diamond \mathbf{O}_X\|_F \leq C \frac{M^6\sqrt{r}}{\lambda} \delta_{\mathcal{G}} + CM^{3/2} r\varepsilon_{\mathcal{G}}.$$

Taking $\mathbf{O} = \mathbf{Q}_Y \mathbf{O}_Y$ and $\widetilde{\mathbf{O}} = \mathbf{Q}_X \mathbf{O}_X$ completes the proof. ■