Meta Sparse Principal Component Analysis

Imon Banerjee
Department of Industrial Engineering
and Management Science
Northwestern University

Jean Honorio
Computing and Information Systems
The University of Melbourne, and
ARC OPTIMA

Abstract

We study the meta-learning for support recovery (i.e., non-zero coordinates of the eigenvectors) in high-dimensional Principal Component Analysis. We reduce the sufficient sample complexity in a novel task, with the information that is learned from auxiliary tasks, where a task is defined as a random Principal Component (PC) matrix with its own support. We pool data from all the tasks to execute an improper estimation of a single PC matrix, by maximising the ℓ_1-regularised predictive covariance. With m tasks for p-variate sub-Gaussian random vectors, we establish the sufficient sample complexity for each task to be of the order $O(\sqrt{m^{-1}\log p})$, with high probability. This is very relevant for meta-learning where there are many tasks $m = O(\log p)$, each with very few samples, i.e., $n = O(1)$, in an scenario where multi-task learning fails. For a novel task, we prove that the sufficient sample complexity of successful support recovery can be reduced to $O(\log |J|)$, under an additional constraint that the support of the novel task is a subset of the estimated support union (J) from the auxiliary tasks. This reduces the original sample complexity of $O(\log p)$ for learning a single task. Theoretical claims are validated with numerical simulations and the problem of true covariance estimation in brain-imaging and cancer genetics data sets are considered to validate the proposed methodology.

1 INTRODUCTION

Principal component analysis (PCA) is an important problem in high-dimensional statistical learning with several applications in information theory (Deshpande and Montanari, 2014), image classification (Sun et al., 2019), brain computer interface (Lin et al., 2008), speech recognition (Zheng et al., 2015) among others. Traditionally, PCA focuses on finding the most significant eigenvectors of a covariance matrix calculated from independent and identically distributed (i.i.d.) samples. However, heterogeneous datasets are ubiquitous across various practical fields, and it is reasonable to infer that variables partitioned by categorical attributes such as country, race, gender, and season do not exhibit identical distributions.

Apart from the issue of heterogeneity, high-dimensional PCA brings unique theoretical challenges of its own kind. For instance, it is known that the estimates PCA produces in high-dimensional situations are often inconsistent (Johnstone and Lu, 2009; Nadler, 2008; Paul, 2007). In particular, it has been shown that the eigenvalues of the sample covariance matrix in high-dimensional data can be more spread than those of the population covariance matrix (Kato, 1949; Bickel and Levina, 2008). Moreover, the eigenvectors of a sample covariance matrix are known to be an inconsistent estimator for the true eigenvectors, i.e., the eigenvectors of the population covariance matrix (Bai et al., 2007). Hence, these challenges present a clear case for carefully understanding conditions under which heterogeneity can be efficiently ameliorated.

In this paper, we refer to the matrix of eigenvectors as PC matrix. Literature about the problem of learning multiple PC matrices is sparse; where there is a compelling argument for efficiently *transferring knowledge already gained from a previous learning task to a new one*, given some structural similarity between previous task and the new PC matrix (referred to henceforth as a "novel" task). This transfer of knowledge is also known as "transfer learning"; Pan and Yang (2010) provide a comprehensive discussion of this topic.

This presents to us three questions. Let m be the number of different PC matrices, referred to henceforth as "auxiliary" tasks, each having n samples of p-variate sub-Gaussian random vectors. Then, this paper is mainly concerned with answering the following questions:

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1. What is an optimal number of auxiliary tasks m to estimate the underlying PC matrix?
2. Given m auxiliary tasks, what is the optimal number of samples n?
3. Given a PC matrix estimated from m tasks, what is the optimal number of samples to estimate the PC matrix of a novel task?

While multi-task learning (Yamane et al., 2016; Yu et al., 2006) allows to partially answer the first two questions (albeit not optimally since it requires to learn every single PC matrix), we focus on meta-learning (Hospedales et al., 2022) which allows to simultaneously and optimally address the above three questions. Furthermore, a common situation in the real world is that there are usually many auxiliary tasks (m) but only a small number of samples (n) in each task. For instance, for $m = 5000$ taks but only $n = 2$ samples per task, it is infeasible to learn the support of each of the m PC matrices, rendering the multi-task learning method meaningless. In contrast, our meta-learning approach performs "improper" estimation since it estimates a *single* PC matrix with data from different tasks. This enables us to efficiently recover the true support with an optimal number of samples per task.

There are no prior studies on meta-learning for sparse PCA (Zou et al., 2006; Jolliffe et al., 2003). For other machine learning problems, prior work includes experimental studies, without theoretical guarantees (Lake et al., 2015; Lemke et al., 2015; Vinyals et al., 2016; Ravi and Larochelle, 2017; Finn et al., 2017; Snell et al., 2017; Grant et al., 2018; Yoon et al., 2018; Hospedales et al., 2022). Thus far, theoretical work on meta-learning pertains only to generalisation bounds in learning theory (Maurer, 2005; Pentina and Lampert, 2014; Amit and Meir, 2018; Denevi et al., 2018; Khodak et al., 2019a; Huang et al., 2020; Tripuraneni et al., 2020; Farid and Majumdar, 2021; Chen and Chen, 2022; Guan et al., 2022) and convergence rates in optimisation (Fallah et al., 2020; Finn et al., 2019; Khodak et al., 2019a; Gao and Sener, 2020). In the aforementioned works, performance is only viewed in terms of risks (e.g., misclassification rate, mean squared error) and not in terms of support recovery.

While there has been some work on multi-task PCA (Tiomoko et al., 2023; Yamane et al., 2016), non-asymptotic statistical analyses are not available for multi-task PCA. The closest result for PCA is on structured sparsity (Deng, 2019, Corollary II.5), which sets the sample complexity of multi-task PCA to $n \in O(m \log(mp))$ for each task.[1] The main difference between *multi-task* learning and *meta* learning, is that multi-task learning estimates one PC matrix for each of the different m tasks simultaneously while meta-learning estimates a single PC matrix from different tasks. This allows us to obtain a sample complexity of $n \in O(\sqrt{m^{-1} \log p})$ for each task. For instance, our PCA meta-learning results imply that a regime of many auxiliary tasks $m \in O(\log p)$, only requires few samples $n \in O(1)$ per task. That is, in meta-learning, the number of samples per task (n) is constant with respect to the dimension (p). Under the same regime, PCA multi-task learning would require $n \in O((\log p)^2)$ samples per task. That is, in multi-task learning, the number of samples per task (n) grows poly-logarithmically in the dimension (p). Thus, when there are only few samples $n \in O(1)$ per task, multi-task learning fails.

The superiority of meta-learning can be seen across different statistical tasks like regression (Wang and Honorio, 2021), logistic regression (Xie and Honorio, 2022), precision matrix estimation (Zhang et al., 2021), classification (Wang et al., 2021; Wu et al., 2019) and neural networks (Hospedales et al., 2022). Techniques for incorporating aspects of multi-task learning under the meta-learning framework have been studied by Khodak et al. (2019b). Besides our non-asymptotic theoretical results for estimation of PC matrices under sparsity conditions, we also show experimentally that meta-learning produces better estimates than multi-task learning.

1.1 Summary of our contributions and organisation of the article

Let $\Sigma^{(1)}, \ldots, \Sigma^{(m)}$ be m many covariance matrices with their corresponding PC matrices $\Pi^{(1)}, \ldots, \Pi^{(m)}$, both being random perturbations of a true underlying covariance matrix Σ with PC matrix Π. Let J be the support of Π, i.e., the set of non-zero coordinates the eigenvectors. Our key contributions in this paper can be summarised as follows:

(a) The minimax sample complexity for estimating J is given by $n \in O(|J|\sqrt{m^{-1} \log p})$ for each task (Theorems 1 and 2). Thus, our meta-learning method works on a challenging scenario where there are many tasks $m = O(|J|^2 \log p)$, each with very few samples, i.e., $n = O(1)$. In this same scenario, multi-task learning fails.

(b) Given a correctly estimated support union J, the optimal sample complexity for estimating a novel task $\Pi^{(m+1)}$ is $O(\log |J|)$, where $|J|$ is the cardinality of the set J (Theorem 3). This considerably reduces the original sample complexity of $O(\log p)$ for learning a single task. To the best of our knowledge, this improvement is unique to meta-learning and cannot be replicated by other multi-task learning techniques.

[1] Multi task learning is one example of structured learning, where we consider a $pm \times pm$ matrix with a structural constraint that the matrix is block-diagonal, composed of m blocks of $p \times p$ matrices. In addition, the sparsity level is $m|J|$. Then, we invoke (Deng, 2019, Corollary II.5).

Outline. The rest of this paper is outlined as follows: In Section 2, we formally introduce the meta-learning setup and formulate the estimation as a primal-dual problem. In Sections 3 and 4, we formalise the theorems for recovering the sample complexities of meta-learning and learning the novel task. In Section 5, we provide simulation studies to experimentally validate our claims. We end with a summary followed by discussions in Section 6.

2 PRELIMINARIES

In this section, we introduce the definitions of the concepts important to our meta-learning problem. A summary of notations used in the paper is illustrated in Table 1.

Multivariate sub-Gaussian Distribution. This paper is concerned with meta-learning of multiple principal component matrices, each of which are generated randomly. To make inferences about those matrices in sufficient generality, we must first formally define the class of sub-Gaussian vectors with random covariance matrices.

The class of sub-Gaussian variates includes for instance Gaussian variables, any bounded random variable (e.g. Bernoulli, multinomial, uniform), any random variable with strictly log-concave density, and any finite mixture of sub-Gaussian variables. We have deferred the definition of real-valued sub-Gaussian and sub-exponential distributions to Appendix A.

Definition 1. *Let $n^{(i)}$ be the number of samples for auxiliary task i. For $j \in [n^{(i)}]$ and $i \in [m]$, we say $X_j^{(i)} \in \mathbb{R}^p$ follows a family of random p-dimensional multivariate sub-Gaussian vectors with randomised covariance matrices $\{\Sigma^{(i)}\}$ distribution of size m and parameter σ if $X_j^{(i)}$ are independent random variables such that for each i and all j, $X_j^{(i)}$ is distributed according to a distribution P_i such that $\mathbb{E}[X_j^{(i)}] = 0$, $\mathrm{Cov}\left(X_j^{(i)}\right) = \Sigma^{(i)}$. Furthermore, $X_{l,j}^{(i)}/\Sigma_{l,l}^{(i)}$ is sub-Gaussian with parameter σ, $\forall i \in [m]$ and $j \in [n^{(i)}]$ and $l \in [p]$.*

Remark. *We would like to point out that such definitions are standard in the literature and have been used extensively in prior work (e.g (Zhang et al., 2021)) on meta-learning.*

We require the following Hoeffding inequality for means of i.i.d zero mean sub-Gaussians with parameter σ. It is proved in Section B.6

Lemma 1 (Hoeffding's inequality the mean of sub-Gaussians)**.** *Let $X_1, \ldots, X_n$ be i.i.d mean zero sub-Gaussian random variables with parameter σ. Then,*

$$\mathbb{P}\left(\left|\bar{X}\right| > \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{\sigma^2}\right)$$

Problem Formulation. In this paper, we focus on estimating the support of the PC matrix of a multivariate sub-Gaussian distribution. We first estimate a superset of the support of the PC matrices. Then we solve a novel task utilising the recovered support union.

To be precise, let for $i \in [m]$, Σ be the true underlying covariance matrix and $\Sigma^{(i)}$ be perturbations of Σ, with each $\Sigma^{(i)}$ forming the covariance matrix of an auxiliary task i. Then, we are interested in using those auxiliary tasks to recover support union J of the PC matrix Π corresponding to Σ, i.e., $J = supp(\Pi)$. This is because we are interested in a novel task where the support of its PC matrix is assumed to be a subset of the support of Π. This allows us to achieve a substantial reduction in sample complexity from $O(\log(p))$ to $O(\log|J|)$ while estimating the support in the novel task.

2.1 A Meta-Learning Model

In this section, we define a generative model for the population covariance matrices. Let Σ be a positive semi-definite, real symmetric matrix. For some orthogonal matrix U, and positive diagonal matrix Λ, the spectral decomposition of Σ gives us,

$$\Sigma = U\Lambda U^T := \sum_{l=1}^{p} \lambda_l(\Sigma) U_{*,l} U_{*,l}^T, \tag{2.1}$$

where $U_{*,l}$ are orthonormal vectors and $\lambda_l := \lambda_l(\Sigma)$ is the lth largest eigenvalue of Σ as well as the $(l,l)^{th}$ entry of Λ. If $\lambda_l(\Sigma) > \lambda_{l+1}(\Sigma)$, $\forall l$, then,

$$\Pi = \sum_{l=1}^{k} U_{*,l} U_{*,l}^T \tag{2.2}$$

gives us the unique k-component PC matrix of Σ. $\mathbb{R}^{p\times p}$. For each $i \in [m]$, the covariance matrix $\Sigma^{(i)}$ corresponding to the i-th auxiliary task is generated as,

$$\Sigma^{(i)} = \Sigma + R^{(i)}. \tag{2.3}$$

Referring to Definition 1, we assume that for $j \in [n^{(i)}]$, our data, $X_j^{(i)}$ is a random p-dimensional multivariate sub-Gaussian random variable with randomised covariance matrix $\Sigma^{(i)}$. Let $S^{(i)} := \frac{1}{n^{(i)}} \sum_{j=1}^{n^{(i)}} X_j^{(i)} (X_j^{(i)})^T$ be the sample covariance matrix corresponding to the auxiliary task $\Sigma^{(i)}$. A "pooled" estimator S of Σ is defined below:

$$S := \frac{1}{m} \sum_{i=1}^{m} S^{(i)}. \tag{2.4}$$

2.2 The Predictive Covariance Loss Function

In this subsection, we describe the objective function that we maximise to obtain the improper estimator. Let $\Sigma = \mathrm{Cov}(X)$ and H be a rank k orthogonal matrix.

Table 1: Notations used in the paper

Notation	Description	Notation	Description
$[a]$	The first a integers, i.e., $\{1, 2, \ldots, a\}$	$\lvert J\rvert$	Cardinality of a set J
$A_{i,j}$	The $(i,j)^{th}$ entry of matrix A	U_i	The ith coordinate of a vector u
$supp(A)$	Support of A, i.e., $\{i : A_{i,i} \neq 0\}$	$A_{*,i}$ (or $A_{i,*}$)	The ith column (or row) of matrix A
$sign(x)$	The sign of $x \in \mathbb{R}$	Σ	The true covariance
$\Sigma^{(i)}$	Covariance of auxiliary task i	$n^{(i)}$	Number of samples for auxiliary task i
J^c	Complement of set J, i.e., $[p] - J$	$\lambda_k(A)$	The kth largest eigenvalue of A
$\lVert u\rVert_q$	l_q-norm of the vector $u \in \mathbb{R}^p$	$\lVert A\rVert_{q_1,q_2}$	$\lVert (\lVert A_{1,*}\rVert_{q_1}, \ldots, \lVert A_{p,*}\rVert_{q_1}) \rVert_{q_2}$
$\lVert A\rVert_F$	The Frobenius norm of a matrix A	$\langle A, B\rangle$	The matrix inner product, i.e., $\mathrm{tr}(A^T B)$
$diag(A)$	Diagonal of matrix A	$A \otimes B$	The matrix Kronecker product
$\mathbb{D}(A)$	The diagonal matrix with entries $A_{i,i}$	$\mathbb{I}_p$	The p-dimensional identity matrix
$A_{I,J}$	Submatrix of A with rows in set I and columns in set J		

From Theorem 1 in Overton and Womersley (1992), k-component PCA seeks to maximize $\langle \Sigma, H\rangle$. Accordingly, the k-component sparse PCA seeks the rank-k orthogonal projection matrix H such that for a penalty $\rho > 0$ the penalised predictive covariance

$$\langle \Sigma, H\rangle - \rho\lVert H\rVert_{1,1} \tag{2.5}$$

is maximised. Crucial to getting a consistent estimate of the principal component matrix is its identifiability. Unless the PC matrix is unique, there is no hope to correctly estimate the PC matrix. For instance, if $\Sigma = \mathbb{I}_p$, then every k-dimensional subspace of $\mathbb{R}^p$ is a valid principal component. Slightly abusing notation, by λ_k we denote $\lambda_k(\Sigma)$ and by λ_{k+1} we denote $\lambda_{k+1}(\Sigma)$. We make our first assumption about the uniqueness of the PC matrix, denoted in terms of the spectral gap of the covariance matrix Σ.

Assumption 1 (Uniqueness)**.** *Let λ_k and λ_{k+1} be the k-th and $k+1$-th eigenvalues of Σ. Then, $\lambda_k - \lambda_{k+1} > 0$.*

Remark. *The above condition ensures that the principal subspace is identifiable. Subsequently, we need to define the notion of sparsity of the PC matrix. Recall from eq.* (2.2) *that Π is a positive semi-definite matrix. Consequently, if the i-th row/column of Π is non-zero, then $\Pi_{i,i} \neq 0$. Hence, it is sufficient to recover the set of non-zero entries in the diagonal of Π. Hence, we can make our sparsity assumption.*

Assumption 2 (Sparsity)**.** *Let Π be the k-component PC matrix corresponding to Σ. The support union $J = supp(\Pi)$ fulfills $\lvert J\rvert \ll p$.*

Throughout section 3, we will assume that Σ satisfies both Assumption 1 and Assumption 2.

2.3 The Improper Objective Function

As illustrated in Section 2, we recover the support union by estimating the true covariance matrix. To be specific, we pool together all the samples from m tasks and maximise the l_1-regularised predictive covariance between the true and the estimated covariance matrices, i.e., we solve the following optimisation problem with regularisation constant $\rho > 0$

$$\hat{\Pi} := \arg\sup_H \sum_{i=1}^m T^{(i)} \left\{ \langle S^{(i)}, H\rangle - \rho\lVert H\rVert_{1,1} \right\}$$
$$\text{subject to } H \in \mathcal{F}^k, \text{ where,} \tag{2.6}$$
$$\mathcal{F}^k := \{M \in \mathbb{R}^{p\times p} : 0 \preceq M \preceq \mathbb{I}_p \text{ and } \mathrm{tr}(M) = k\},$$

and $T^{(i)}$'s are weights proportional to the number of samples $n^{(i)}$ for each i. For clarity of exposition, we assume that $n^{(i)} = n$ for all values of i for the reminder of this paper. This allows us to assume $T^{(i)} = \frac{1}{m}$ without losing generality. Consequently, we can rewrite the objective function in Equation (2.6) as

$$\hat{\Pi} = \arg\sup_H \langle S, H\rangle - \rho\lVert H\rVert_{1,1} \text{ subject to } H \in \mathcal{F}^k, \tag{2.7}$$

where S is as defined in eq. (2.4). It is noteworthy that Equation (2.6) is an improper estimation technique since it estimates a single PC matrix with data from different tasks. This enables us to efficiently recover the true support with an optimal number of sample size per task (see Section 3).

Remark. *In contrast with manifold optimization formulations (like over the Stiefel manifold) Fantope formulation yields a convex optimization problem, allowing the use of strong duality and primal–dual witness arguments, and enabling practical implementation with standard convex solvers (e.g., CVX).*

For the next step, suppose that we have successfully recovered the true support union J in the first step. Then for the novel task, ($m+1$-th task), we can assume that the support of its PC matrix is also a subset of the recovered support union. Then, for a noisy estimate $S^{(m+1)}$ of $\Sigma^{(m+1)}$ we propose the following l_1-regularised predictive covariance loss function

$$\langle S^{(m+1)}, H\rangle - \rho\lVert H\rVert_{1,1} \text{ subject to } H \in \mathcal{F}^k \text{ and } supp(H) \subseteq J. \tag{2.8}$$

Due to presence of the extra constraint, the above estimation technique reduces the dimension of the optimisation problem from $p \times p$ to $|J| \times |J|$.

3 SUPPORT UNION RECOVERY FOR THE AUXILIARY TASKS

In this section, we formally show that a sample complexity of $O(\sqrt{m^{-1}\log p})$ for each task, is sufficient for the recovery of the support union via optimising Equation (2.7). Thus, our meta-learning method works on a challenging scenario where there are many tasks $m = O(\log p)$, each with very few samples, i.e., $n = O(1)$. In this same scenario, multi-task learning fails. One key task in achieving the said sample complexity is to carefully upper bound the tail probabilities of $\|S - \Sigma\|_{\infty,\infty}$. To that end, we prove in Section B.1 the following tail probability for a family of multivariate sub-Gaussian vectors with randomised covariance matrices. Let $S^{(i)} = \frac{1}{n}\sum_{t=1}^{n} X_t^{(i)}(X_t^{(i)})^{\mathrm{T}}$.

Lemma 2. *Let $\Sigma^{(i)}$ be a sequence of random covariance matrices such that $\left\|\Sigma^{(i)}\right\|_{\infty,\infty} \leq \eta$. Then, for $\{X_j^{(i)}\}_{1\leq j\leq n, 1\leq i\leq m}$ following a family of random p-dimensional multivariate sub-Gaussian distributions with parameter σ and random covariance matrices $\Sigma^{(i)}$,*

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^{m}\left(S^{(i)} - \Sigma^{(i)}\right)\right\|_{\infty,\infty} \geq \varepsilon\right) \leq \frac{p(p+1)}{2}e^{-\frac{nm\varepsilon^2}{512\sigma^4\eta}},$$

whenever $0 < \varepsilon < 32\eta\sigma^2$.

Recall from eq. (2.3) that we defined each auxiliary task $\Sigma^{(i)}$ in terms of random matrices $R^{(i)}$. Key to the process of improper estimation is the control of the maximal loss between the auxiliary tasks $\Sigma^{(i)}$ and the true underlying task Σ. In other words, we need to find theoretical guarantees for $\left\|\Sigma^{(i)} - \Sigma\right\|_{\infty,\infty}$, which is derived in eq. (B.7) in the supplements under the following assumption.

Assumption 3. *Assume that $\|\Sigma\|_{\infty,\infty} \leq \eta$ for some $\eta > 0$. Assume that for each $1 \leq i \leq m$, $R^{(i)}$ in Equation* (2.3) *are independent symmetric random matrices such that $R_{j,k}^{(i)}$ are zero mean subgaussian random variables with parameter σ_a for some constant $\sigma_a^2 > 0$ such that $\mathbb{P}(\Sigma + R^{(i)} \succeq 0) = 1$.*

The perturbation matrices $R^{(i)}$ and $D^{(i)}$ capture task-specific deviations from the underlying covariance structure. In practice, these matrices model natural sources of heterogeneity across tasks, such as scanner- or site-specific noise in neuroimaging studies, sample-handling variability across genomic cohorts, or experimental conditions that introduce mild shifts in covariance structure.

White noise additive models have been used extensively in prior literature to study precision and covariance matrices (Wang and Honorio, 2021; Zhang et al., 2021). We note that Assumption 3 is weak, but still sufficient to produce a minimax optimal sample complexity. Our final technical assumption is given below.

Assumption 4. *Let Σ be a covariance satisfying assumption 2. Then there exists a constant $\alpha \in (0,1)$ such that,*

$$\frac{8|J|}{\lambda_k - \lambda_{k+1}}\|\Sigma_{J^c,J}\|_{2,\infty} =: (1-\alpha) < 1.$$

This condition previously appears in Lei and Vu (2015) and recovers the conditions by Amini and Wainwright (2008) when $\Sigma_{J^c,J} = 0$. It is also similar to the irrepresentability condition in the seminal papers (Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006; Meinshausen and Yu, 2009) for l_1 penalised sparse regression and lasso type recovery. As far as we know, this is the most general condition for the support recovery of PC matrices and encompasses the conditions when Σ is block diagonal. For ease of exposition, we define the following notation:

$$\lambda_{diff} := \lambda_k - \lambda_{k+1} \quad \lambda^\dagger := \eta + \sqrt{10\sigma_a^2\log(p+1)/m},$$
$$\rho_1 := \sqrt{8\sigma_a^2\frac{\log(p+1)}{m}} \quad \rho_2 := \sqrt{512\sigma^4\lambda^\dagger\frac{\log(p+1)}{nm}}.$$

The following theorem establishes the $O(|J|^2\log p/m)$ per-task sample complexity for the support union recovery from multiple auxiliary tasks.

Theorem 1 (Support Union Recovery)**.** *Let assumption 3 and assumption 4 hold. Then there exists a large enough constant $c(\alpha)_\pi > 0$ such that, whenever*

$$\alpha\max\{\rho_1,\rho_2\} < \rho$$
$$< \min\left\{\frac{(\lambda_{diff})\min_{j\in J}\Pi_{j,j}}{16|J|}, \frac{(\lambda_{diff})^2}{4|J|\left(\lambda_{diff} + 8\lambda_1(\Sigma)\right)}\right\}$$

then, with probability at least $1 - \frac{2}{p^2} - \frac{1}{2(p+1)^2}$ the solution $\hat{\Pi}$ to the objective function Equation (2.7) *satisfies,*

$$1.\ \ supp(\hat{\Pi}) = J, \qquad 2.\ \left\|\hat{\Pi} - \Pi\right\|_{\infty,\infty} \leq c(\alpha)_\pi\rho.$$

In other words, Theorem 1 gives us our required sample complexity for sign consistency and error bound. Therefore, as long as

$$\alpha\max\left\{4\lambda^\dagger\sqrt{\frac{\log(p)}{m}}, 16\sqrt{2\sigma^4\lambda^\dagger\frac{\log(p+1)}{nm}}\right\}$$
$$< \min\left\{\frac{(\lambda_{diff})\min_{j\in J}\Pi_{j,j}}{16|J|}, \frac{(\lambda_{diff})^2}{4|J|\left(\lambda_{diff} + 8\lambda_1(\Sigma)\right)}\right\}$$

with probability at least $1 - 2/p^2 - 1/(2(p+1)^2)$, we recover the true support. Although the admissible range

for ρ involves several structural constants, its leading dependence on (m, n, p) implies that choosing ρ of order $\sqrt{\log(p)/(mn)}$ is sufficient for support recovery. Therefore, settings such as $n = O(\log p/m)$ or $n = O(1)$ and $m = O(\log p)$ suffice to recover the correct support union. Our next objective is to provide a matching lower bound for the previous per-task sample complexity.

Theorem 2. *In the setup of Theorem 1, there exists an estimation problem such that, for all p large enough any estimator $\hat{J}$ of the support union J, if $n \leq (|J| \log p - |J| \log |J|)/(2m)$, then*

$$\mathbb{P}\left(\hat{J} \neq J\right) \geq 1/4.$$

The proof of the above theorem is provided in Appendix B.3.

Remark. *Observe that, by setting $n = 1$, we also obtain a corresponding error lower bound whenever $m \leq |J| \log |p| - |J| \log |J|/2$ which matches the upper bound.*

3.1 Sketch of Proof for Theorem 1

The complete proof of this theorem may be found at appendix B.2 and we only make the outline here. The proof of Theorem 1 is via primal-dual witness method. It starts by writing down the objective function eq. (2.7) as follows. Define

$$\max_{H \in \mathcal{F}^k} \langle S, H \rangle - \rho \|H\|_{1,1}, \text{ and}$$

$$\mathbb{B}_p := \left\{ Z \in \mathbb{R}^{p \times p} : diag(Z) = 0, Z = Z^T, \|Z\|_{\infty,\infty} \leq 1 \right\}.$$

Using strong-duality of $\|\cdot\|_{1,1}$, we can rewrite the objective function as,

$$\begin{aligned} &\max_{\substack{H \in \mathcal{F}^k \\ Z \in \mathbb{B}}} \langle S, H \rangle - \rho \langle H, Z + \mathbb{I}_p \rangle \equiv \\ &\max_{\substack{H \in \mathcal{F}^k \\ Z \in \mathbb{B}}} \langle S, H \rangle - \rho \langle H, Z \rangle - \rho k \equiv \\ &\max_{\substack{H \in \mathcal{F}^k \\ Z \in \mathbb{B}}} \langle H, S - \rho Z \rangle, \end{aligned} \tag{3.1}$$

where the last step follows from the fact that k is a constant. To maximise Equation (3.1), we need to find solutions $(\hat{H}, \hat{Z})$ such that the following Karush-Kuhn-Tucker (KKT) (Boyd et al., 2004) conditions satisfied:

$$\hat{Z}_{i,j} = sign(\hat{H}_{i,j}), \forall i \neq j \text{ and } \hat{H}_{i,j} \neq 0, \tag{3.2}$$

$$\hat{Z}_{i,j} \in [-1, 1], \forall i \neq j \text{ and } \hat{H}_{i,j} = 0, \tag{3.3}$$

$$\hat{H} = \arg\max \langle S - \rho \hat{Z}, H \rangle. \tag{3.4}$$

Proceeding according to the primal-dual witness method, we construct the constrained optimisation problem where the $supp(H) \subseteq J$.

$$\max_{H \in \mathcal{F}^k, supp(H) \subseteq J} \langle S, H \rangle - \rho \|H\|_{1,1}. \tag{3.5}$$

Let $(\tilde{H}, \tilde{Z})$ be the corresponding primal and dual variables. The proof selects an appropriate $|J| \times |J|$ orthonormal matrix Q according to lemma 5 with the correct rotation such that the Frobenius norm between the true eigenvectors $U_{J,*}$ and the estimated eigenvectors $\hat{U}_{J,*}$ is small. Then we propose the following solution $(\hat{H}, \hat{Z})$ to the dual objective eq. (3.1).

$$\begin{aligned} &\hat{H} = \tilde{H},\ \hat{Z}_{J,J} = \tilde{Z}_{J,J}, \\ &\hat{Z}_{ij} = \frac{1}{\rho}\left\{S_{ij} - \langle Q_{i,*}, \Sigma_{J,j} \rangle\right\}, \forall (i,j) \in J \times J^c, \\ &\hat{Z}_{ij} = \frac{1}{\rho}(S - \Sigma)_{ij}, \forall (i,j) \in J^c \times J^c, i \neq j. \end{aligned}$$

The rest of the proof follows in 4 parts.

1. Finding an appropriate Q matrix and an appropriate tail probability bound under Assumption 3 by using a combination of Matrix Chernoff bound (see appendix A) and Lemma 2.
2. Proving that the proposed solution satisfies the KKT conditions (3.2), (3.3), and (3.4) with high probability for our choices of ρ and is unique.
3. Find an upper bound for estimation error. We achieve this through the following upper bound, $\left\|\hat{H} - \Pi\right\|_{\infty,\infty} < 2\frac{\rho|J|}{\lambda_{diff}}$, which is a mild variation of a similar upper bound from Lemma 2 in Vu et al. (2013). Thus, for sufficiently small enough ρ, we prove that $sign(\hat{\Pi}) = sign(\Pi)$, and our estimate is sign consistent.
4. Showing that the proposed solution is sign consistent using a similar technique to part 3.

On Choice of Regularisation. Theorem 1 characterizes an admissible range for ρ, but as is standard in nonparametric analysis the constants involved (e.g., spectral gap, irrepresentability factor) are typically unknown. We suggest the following recipe for choosing this multiplier:

- Start with $\rho = c\sqrt{\log(p)/(mn)}$.
- **Tuning constant:** adjust c via a small grid or cross-validation.
- **Small per-task sample sizes:** when n is very small, slightly larger c may improve stability.
- **Larger pooled sample sizes:** when mn is large, c close to 1 or even less generally suffices.

4 SUPPORT UNION RECOVERY FOR THE NOVEL TASK

In this section, we show that $O(\log |J|)$ sample complexity is sufficient for the support recovery for a novel task where

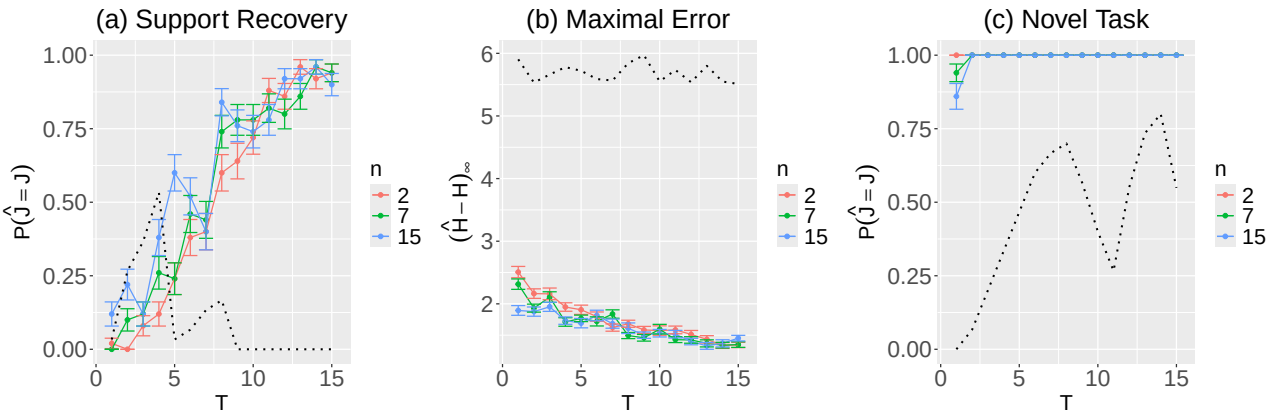

Figure 1: (a), (b): Simulations for Theorem 1 on the probability of support union recovery and maximal error under various settings of n and auxilliary tasks generated by Gaussian random variables. $\rho = \sqrt{\log(p+1)/(mn)}$. The x-axis is the constant T used to set the number of tasks $m = \lceil T \log(p+1)/n \rceil$. We vary T from $\{1, \ldots, 15\}$. (c): (c): Probability of support recovery for novel task. $\rho = \sqrt{\log(|J|+1)/n}$. The x-axis is the constant T used to set the number of samples in the novel task $n^{(m+1)} = \lceil T \log(|J|+1) \rceil$, for $T \in \{1, \ldots, 15\}$. The dotted black line represents the baseline, which is the best result that either single or multitask learning obtained for any n. Our meta learning method is clearly superior.

J is the recovered support union in Section 3. This considerably reduces the original sample complexity of $O(\log p)$ for learning a single task. We also show that for the novel task, the estimation problem can be reduced from estimating a $p \times p$ dimensional matrix to a $|J| \times |J|$ dimensional matrix. Let $n^{(m+1)}$ be the number of samples in the novel task. We begin by establishing the reduction in dimension. Let $\Sigma^{(m+1)}$ be the variance matrix corresponding to the novel task with spectral decomposition

$$U^{(m+1)} \Lambda^{(m+1)} \left(U^{(m+1)} \right)^T,$$

and the k-component PC matrix $\Pi^{(m+1)}$. Throughout the rest of this section, we assume that $\Sigma^{(m+1)}$ satisfies Assumption 1 and Assumption 2. We write $\Sigma^{(m+1)}$ and its estimator $S^{(m+1)}$ as the following block matrices

$$\Sigma^{(m+1)} = \begin{bmatrix} \Sigma^{(m+1)}_{J,J} & \Sigma^{(m+1)}_{J,J^c} \\ \Sigma^{(m+1)}_{J^c,J} & \Sigma^{(m+1)}_{J^c,J^c} \end{bmatrix} \text{ and}$$
$$S^{(m+1)} = \begin{bmatrix} S^{(m+1)}_{J,J} & S^{(m+1)}_{J,J^c} \\ S^{(m+1)}_{J^c,J} & S^{(m+1)}_{J^c,J^c} \end{bmatrix}. \tag{4.1}$$

We obtain the k-component PC matrix of the novel task $\Pi^{(m+1)}$ by optimising the objective function,

$$\max_{H \in \mathcal{F}^k} \langle S^{(m+1)}, H \rangle - \rho \|H\|_{1,1}. \tag{4.2}$$

However, we have already recovered the support union J. Hence, we can add another constraint on H: $supp(H) = J$. Moreover, since the row and column i of the principal component $\Pi^{(m+1)}$ is 0 if and on $\Pi^{(m+1)}_{i,i} = 0$, we can rewrite this additional constraint as,

$$H = \begin{bmatrix} H_{J,J} & 0 \\ 0 & 0 \end{bmatrix}.$$

Since $H \in \mathcal{F}^k$, the maximisation problem only makes sense if $|J| > k$. This also automatically implies that $H \in \mathcal{F}^k_J$, where $\mathcal{F}^k_J := \left\{ M \in \mathbb{R}^{|J| \times |J|} : 0 \preceq M \preceq \mathbb{I}_{|J|} \text{ and } \operatorname{tr}(M) = k \right\}$. Therefore, the objective corresponding to the novel task maximises $\langle S^{(m+1)}, H \rangle - \rho \|H\|_{1,1}$ subject to $H \in \mathcal{F}^k$ and $supp(H) \subseteq J$. This is equivalent to maximising

$$\hat{\Pi}^{(m+1)} := \arg\sup_{H_{J,J}} \langle S^{(m+1)}_{J,J}, H_{J,J} \rangle - \rho \|H_{J,J}\|_{1,1}, \quad \text{for } H_{J,J} \in \mathcal{F}^k_J \tag{4.3}$$

where $supp(\hat{\Pi}) = J$ from theorem 1 is the recovered support union, and $H_{J,J}, S_{J,J} \in \mathbb{R}^{|J| \times |J|}$. The following proposition on the eigenvalues and eigenvectors of $\Sigma^{(m+1)}_{J,J}$ shows that the estimation problem can be reduced from estimating a $p \times p$ dimensional matrix to a $|J| \times |J|$ dimensional matrix.

Proposition 3. *For $1 \le l \le k$, $U^{(m+1)}_{l,J}$ are the first k eigenvectors of $\Sigma^{(m+1)}_{J,J}$ with corresponding eigenvalues $\lambda^{(m+1)}_l$.*

The proof of this proposition is given in Appendix B.4. Following from the fact that $\Sigma^{(m)}$ satisfies Assumption 1 and Assumption 2, proposition 3 establishes as an immediate consequence that

$$\lambda^{(m+1)}_k - \lambda^{(m+1)}_{k+1} > 0, \text{ and, } J^{(m+1)} = supp(\Pi^{(m+1)}).$$

This establishes that the k-component PC matrix of $\Sigma^{(m+1)}_{J,J}$ is sparse and unique, as well as the solution to a lower dimensional objective function.

Overloading notation, define $\lambda_{diff}^{(m+1)} := \lambda_k^{(m+1)} - \lambda_{k+1}^{(m+1)}$. We make the following technical assumption on the entries of $\Sigma_{J,J}^{(m+1)}$.

Assumption 5. *Let $\Sigma^{(m+1)}$ be the covariance matrix corresponding to the novel task and J be the recovered support union. Then,*

$$\frac{8|J^{(m+1)}|}{\lambda_{diff}^{(m+1)}} \|\Sigma^{(m+1)}_{(J^{(m+1)})^c, J^{(m+1)}}\|_{2,\infty} =: \left(1 - \alpha^{(m+1)}\right) < 1.$$

This condition is similar to Assumption 4, and as we mention before, previously appears in (Lei and Vu, 2015) and recovers the conditions by Amini and Wainwright (2008) when $\Sigma_{J^c,J} = 0$. With this, we can now proceed to state our theorem regarding support recovery in novel task.

The following theorem shows the $O(\log |J|)$ sample complexity for support recovery of the novel task.

Theorem 3 (Novel Task Support Recovery)**.** *Let $\Sigma^{(m+1)}$ be the covariance matrix corresponding to the novel task satisfying Assumption 5 and J be the correctly recovered support union. Furthermore, assume that, for some $\varepsilon, \delta > 0$, we have*

$$n \geq \frac{512\sigma^2 \lambda_1\left(\Sigma^{(m+1)}\right)}{\varepsilon^2} \log\left(\frac{|J|(|J|+1)}{\delta}\right),$$

$$\text{and } \frac{\varepsilon^2}{\alpha^{(m+1)}} < \rho < \frac{\lambda_k^{(m+1)} - \lambda_{k+1}^{(m+1)}}{4|J^{(m+1)}|\left(1 + \frac{8\lambda_1^{(m+1)}}{\lambda_k^{(m+1)} - \lambda_{k+1}^{(m+1)}}\right)}.$$

Then, with probability at least $1 - \delta$, we have $supp(\Pi^{(m+1)}) = J^{(m+1)}$.

The proof of this theorem is deferred to Appendix B.5. Observe that $\log(|J|(|J|+1)) \leq 2\log(|J|+1)$. Therefore, we have shown the $O(\log |J|)$ sample complexity for support recovery of the novel task.

5 EXPERIMENTS

In this section, we empirically validate our theoretical claims in Sections 3 and 4 in one numerical example. Further numerical examples are deferred to Appendices C, C.1, and C.2 for a lack of space.

5.1 Synthetic Experiments

To address the challenging regime with many tasks (m) but few samples per task (n), we fix $p = 50$, $k = 10$, $|J| = 10$ and run 50 trials. Sample sizes are $n \in \{6, 7, 8\}$ with $\rho = \sqrt{\log(p+1)/(mn)}$. The number of tasks is summarized by $T = mn/\log(p+1)$. For each $j \in [n]$, generate $X_j^{(i)} \sim \mathcal{N}(0, \Sigma^{(i)})$. Figure 1 shows that for every n, the curves for $\mathbb{P}(\hat{J} = J)$ and $\|\hat{\Pi} - \Pi\|_{\infty,\infty}$ coincide. Using the recovered support union, we evaluate a novel task with $n^{(m+1)} \propto T\log(|J|+1)$. With a single active coordinate ($k_{\text{nov}} = 1$), the probability of correctly identifying the extra zeros exceeds 95%. In Appendix C we provide more details and show more examples, including the performance in high dimensions. Then, we show the superiority of meta-learning over multi-task learning and single task learning in Appendices C.2, and C.1.

5.2 Real World Experiments

In Appendix C.3, we show that our method $\sim 85 - 95\%$ accuracy in a real world brain-imaging dataset and cancer datasets. We also found consistent findings with brain imaging (Rubia et al., 2000; Yan et al., 2005; Baron-Cohen et al., 2006) and cancer genetics (Huang et al., 2014; Sun et al., 2020; Feber et al., 2004; Hu et al., 2009) studies. These findings demonstrate the applicability and interpretability of meta-learning in practical applications.

6 CONCLUDING REMARKS

We believe we are taking a meaningful and important first step that will potentially motivate further developments in the community. Dependencies in the data, such as when the data is from a time series makes an interesting scenario. Similarly, it is also unclear how estimation should proceed when there is missing data. Finally, we would like to point out that extending our results to kernel methods would also be an interesting future endeavor.

Acknowledgments

The first author acknowledges the IEMS Alumni Fellowship at Northwestern University and the Ross-Lynn Fellowship at Purdue University for financial support during the conduct of this research. The first author acknowledges Ksheera Sagar for providing feedback on an earlier draft of the manuscript. The authors also thank the four anonymous reviewers for their useful comments and suggestions, which significantly improved the readability of the paper.

References

Amini, A. A. and Wainwright, M. J. (2008). High-dimensional analysis of semidefinite relaxations for sparse principal components. In *2008 IEEE international symposium on information theory*, pages 2454–2458. IEEE.

Amit, R. and Meir, R. (2018). Meta-learning by adjusting priors based on extended PAC-Bayes theory. *International Conference on Machine Learning*, 35:205–214.

Bai, Z., Miao, B., and Pan, G. (2007). On asymptotics of eigenvectors of large sample covariance matrix. *Annals of Probability*, 35(4):1532–1572.

Baron-Cohen, S., Ring, H., Chitnis, X., Wheelwright, S., Gregory, L., Williams, S., Brammer, M., and Bullmore, E. (2006). fMRI of parents of children with asperger syndrome: a pilot study. *Brain and cognition*, 61(1):122–130.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227.

Binkofski, F., Buccino, G., Posse, S., Seitz, R. J., Rizzolatti, G., and Freund, H.-J. (1999). A fronto-parietal circuit for object manipulation in man: evidence from an fMRI-study. *European Journal of Neuroscience*, 11(9):3276–3286.

Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Chen, L. and Chen, T. (2022). Is Bayesian model-agnostic meta learning better than model-agnostic meta learning, provably? In *Artificial Intelligence and Statistics*, volume 151, pages 1733–1774. PMLR.

Cooper, C. S., Nicholson, A. G., Foster, C., Dodson, A., Edwards, S., Fletcher, A., Roe, T., Clark, J., Joshi, A., Norman, A., et al. (2006). Nuclear overexpression of the E2F3 transcription factor in human lung cancer. *Lung Cancer*, 54(2):155–162.

Dai, W., Wang, Y., Yang, T., Wang, J., Wu, W., and Gu, J. (2019). Downregulation of exosomal CLEC3B in hepatocellular carcinoma promotes metastasis and angiogenesis via AMPK and VEGF signals. *Cell communication and signaling*, 17(1):1–17.

Den Hollander, P., Rawls, K., Tsimelzon, A., Shepherd, J., Mazumdar, A., Hill, J., Fuqua, S. A., Chang, J. C., Osborne, C. K., Hilsenbeck, S. G., et al. (2016). Phosphatase PTP4A3 promotes triple-negative breast cancer growth and predicts poor patient survival. *Cancer research*, 76(7):1942–1953.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. (2018). Learning to learn around a common mean. In *Neural Information Processing Systems*, volume 31.

Deng, Y. (2019). *Group Sparsity in Regression and PCA*. PhD thesis, The University of Michigan.

Deshpande, Y. and Montanari, A. (2014). Information-theoretically optimal sparse PCA. In *2014 IEEE International Symposium on Information Theory*, pages 2197–2201.

Du, H., Huang, X., Wang, S., Wu, Y., Xu, W., and Li, M. (2009). PSMA7, a potential biomarker of diseases. *Protein and peptide letters*, 16(5):486–489.

Fallah, A., Mokhtari, A., and Ozdaglar, A. (2020). On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *Artificial Intelligence and Statistics*, pages 1082–1092.

Fano, R. M. (1949). *The transmission of information*, volume 65. Massachusetts Institute of Technology, Research Laboratory of Electronics ....

Farid, A. and Majumdar, A. (2021). Generalization bounds for meta-learning via PAC-Bayes and uniform stability. In *Neural Information Processing Systems*, volume 34, pages 2173–2186.

Feber, A., Clark, J., Goodwin, G., Dodson, A. R., Smith, P. H., Fletcher, A., Edwards, S., Flohr, P., Falconer, A., Roe, T., et al. (2004). Amplification and overexpression of E2F3 in human bladder cancer. *Oncogene*, 23(8):1627–1630.

Fincham, J. M., Carter, C. S., Van Veen, V., Stenger, V. A., and Anderson, J. R. (2002). Neural mechanisms of planning: a computational analysis using event-related fMRI. *Proceedings of the National Academy of Sciences*, 99(5):3346–3351.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. (2019). Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930.

Friederici, A. D., Opitz, B., and Von Cramon, D. Y. (2000). Segregating semantic and syntactic aspects of processing in the human brain: an fMRI investigation of different word types. *Cerebral cortex*, 10(7):698–705.

Fulford, J., Milton, F., Salas, D., Smith, A., Simler, A., Winlove, C., and Zeman, A. (2018). The neural correlates of visual imagery vividness–an fMRI study and literature review. *Cortex*, 105:26–40.

Gao, K. and Sener, O. (2020). Modeling and optimization trade-off in meta-learning. In *Neural Information Processing Systems*, volume 33, pages 11154–11165.

Gao, Y., Feng, B., Lu, L., Han, S., Chu, X., Chen, L., and Wang, R. (2017). MiRNAs and E2F3: a complex network of reciprocal regulations in human cancers. *Oncotarget*, 8(36):60624.

Grant, E., Finn, C., Levine, S., Darrell, T., and Griffiths, T. (2018). Recasting gradient-based meta-learning as hierarchical Bayes. *International Conference on Learning Representations*.

Guan, J., Liu, Y., and Lu, Z. (2022). Fine-grained analysis of stability and generalization for modern meta learning algorithms. In *Neural Information Processing Systems*, volume 35, pages 18487–18500.

Honorio, J. and Jaakkola, T. (2014). Tight bounds for the expected risk of linear classifiers and PAC-Bayes finite-sample guarantees. In *Artificial Intelligence and Statistics*, pages 384–392. PMLR.

Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. (2022). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5149–5169.

Hu, X.-T., Chen, W., Wang, D., Shi, Q.-L., Zhang, F.-B., Liao, Y.-Q., Jin, M., and He, C. (2008a). High expression of proteasome subunit PSMA7 in colorectal cancer is significantly correlated with liver metastasis. *Zhonghua Zhong liu za zhi [Chinese Journal of Oncology]*, 30(7):515–518.

Hu, X.-T., Chen, W., Wang, D., Shi, Q.-L., Zhang, F.-B., Liao, Y.-Q., Jin, M., and He, C. (2008b). The proteasome subunit PSMA7 located on the 20q13 amplicon is overexpressed and associated with liver metastasis in colorectal cancer. *Oncology reports*, 19(2):441–446.

Hu, X.-T., Chen, W., Zhang, F.-B., Shi, Q.-L., Hu, J.-B., Geng, S.-M., and He, C. (2009). Depletion of the proteasome subunit PSMA7 inhibits colorectal cancer cell tumorigenicity and migration. *Oncology reports*, 22(5):1247–1252.

Huang, Y., Huang, W., Li, L., and Li, Z. (2020). Meta-learning PAC-Bayes priors in model averaging. *Association for the Advancement of Artificial Intelligence Conference*, pages 4198–4205.

Huang, Y.-H., Al-Aidaroos, A. Q. O., Yuen, H.-F., Zhang, S.-D., Shen, H.-M., Rozycka, E., McCrudden, C. M., Tergaonkar, V., Gupta, A., Lin, Y. B., et al. (2014). A role of autophagy in PTP4A3-driven cancer progression. *Autophagy*, 10(10):1787–1800.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547.

Kato, T. (1949). On the convergence of the perturbation method. I. *Progress of Theoretical Physics*, 4(4):514–523.

Khodak, M., Balcan, M.-F., and Talwalkar, A. (2019a). Provable guarantees for gradient-based meta-learning. *International Conference on Machine Learning*, 36:424–433.

Khodak, M., Balcan, M.-F. F., and Talwalkar, A. S. (2019b). Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, 32.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Lei, J. and Vu, V. Q. (2015). Sparsistency and agnostic inference in sparse PCA. *The Annals of Statistics*, 43(1):299–322.

Lemke, C., Budka, M., and Gabrys, B. (2015). Metalearning: a survey of trends and technologies. *Artificial intelligence review*, 44(1):117–130.

Lin, C.-T., Chen, Y.-C., Huang, T.-Y., Chiu, T.-T., Ko, L.-W., Liang, S.-F., Hsieh, H.-Y., Hsu, S.-H., and Duann, J.-R. (2008). Development of wireless brain computer interface with embedded multitask scheduling and its application on real-time driver's drowsiness detection and warning. *IEEE Transactions on Biomedical Engineering*, 55(5):1582–1591.

Maurer, A. (2005). Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6(Jun):967–994.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The annals of statistics*, 37(1):246–270.

Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817.

Olsson, A., Feber, A., Edwards, S., Te Poele, R., Giddings, I., Merson, S., and Cooper, C. (2007). Role of E2F3 expression in modulating cellular proliferation rate in human bladder and prostate cancer cells. *Oncogene*, 26(7):1028–1037.

Overton, M. L. and Womersley, R. S. (1992). On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):41–45.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642.

Pentina, A. and Lampert, C. H. (2014). A PAC-Bayesian bound for lifelong learning. *International Conference on Machine Learning*, 31:991–999.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, volume 2. NSF-CBMS Regional Conference Series in Probability and Statistics.

Ravi, S. and Larochelle, H. (2017). Optimization as a model for few-shot learning. *International Conference on Learning Representations*.

Rubia, K., Overmeyer, S., Taylor, E., Brammer, M., Williams, S., Simmons, A., Andrew, C., and Bullmore, E. (2000). Functional frontalisation with age: mapping neurodevelopmental trajectories with fMRI. *Neuroscience & Biobehavioral Reviews*, 24(1):13–19.

Shuster, L. I. and Lemieux, S. K. (2005). An fMRI investigation of covertly and overtly produced mono-and multisyllabic words. *Brain and language*, 93(1):20–31.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. In *Neural Information Processing Systems*, pages 4077–4087.

Sun, J., Xie, T., Jamal, M., Tu, Z., Li, X., Wu, Y., Li, J., Zhang, Q., and Huang, X. (2020). CLEC3B as a potential diagnostic and prognostic biomarker in lung cancer and association with the immune microenvironment. *Cancer cell international*, 20(1):1–15.

Sun, Y., Li, L., Zheng, L., Hu, J., Li, W., Jiang, Y., and Yan, C. (2019). Image classification base on PCA of multi-view deep representation. *Journal of Visual Communication and Image Representation*, 62:253–258.

Sylvester, C.-Y. C., Wager, T. D., Lacey, S. C., Hernandez, L., Nichols, T. E., Smith, E. E., and Jonides, J. (2003). Switching attention and resolving interference: fMRI measures of executive functions. *Neuropsychologia*, 41(3):357–370.

Tan, L. H., Feng, C.-M., Fox, P. T., and Gao, J.-H. (2001). An fMRI study with written chinese. *Neuroreport*, 12(1):83–88.

Tiomoko, M., Couillet, R., and Pascal, F. (2023). PCA-based multi task learning: a random matrix approach. *International Conference on Machine Learning*, pages 34280–34300.

Tripuraneni, N., Jordan, M. I., and Jin, C. (2020). On the theory of transfer learning: The importance of task diversity. *Neural Information Processing Systems*, 33.

Tropp, J. A. (2015). *An introduction to matrix concentration inequalities*. Now Foundations and Trends.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*, pages 210–268. Cambridge University Press.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In *Neural Information Processing Systems*, pages 3630–3638.

Vu, V. Q., Cho, J., Lei, J., and Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. *Advances in neural information processing systems*, 26.

Wang, H., Zhao, H., and Li, B. (2021). Bridging multi-task learning and meta-learning: Towards efficient training and effective adaptation. In *International Conference on Machine Learning*, pages 10991–11002. PMLR.

Wang, L., Yu, T., Li, W., Li, M., Zuo, Q., Zou, Q., and Xiao, B. (2019). The miR-29c-KIAA1199 axis regulates gastric cancer migration by binding with WBP11 and PTP4A3. *Oncogene*, 38(17):3134–3150.

Wang, Z. and Honorio, J. (2021). The sample complexity of meta sparse regression. In *International Conference on Artificial Intelligence and Statistics*, pages 2323–2331. PMLR.

Wohlschläger, A. M., Specht, K., Lie, C., Mohlberg, H., Wohlschläger, A., Bente, K., Pietrzyk, U., Stöcker, T., Zilles, K., Amunts, K., et al. (2005). Linking retinotopic fMRI mapping and anatomical probability maps of human occipital areas V1 and V2. *Neuroimage*, 26(1):73–82.

Wu, J., Xiong, W., and Wang, W. Y. (2019). Learning to learn and predict: A meta-learning approach for multi-label classification. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4354–4364.

Xie, H. and Honorio, J. (2022). Meta learning for high-dimensional Ising model selection using l_1-regularized logistic regression. *arXiv preprint arXiv:2208.09539*.

Yamane, I., Yger, F., Berar, M., and Sugiyama, M. (2016). Multitask principal component analysis. In *Asian Conference on Machine Learning*, pages 302–317. PMLR.

Yan, B., Li, K., Xu, J., Wang, W., Li, K., Liu, H., Shan, B., and Tang, X. (2005). Acupoint-specific fMRI patterns in human brain. *Neuroscience letters*, 383(3):236–240.

Yoon, J., Kim, T., Dia, O., Kim, S., Bengio, Y., and Ahn, S. (2018). Bayesian model-agnostic meta-learning. In *Neural Information Processing Systems*, volume 31.

Yu, S., Yu, K., Tresp, V., Kriegel, H.-P., and Wu, M. (2006). Supervised probabilistic principal component analysis. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 464–473.

Zhang, Q., Zheng, Y., and Honorio, J. (2021). Meta learning for support recovery in high-dimensional precision matrix estimation. In *International Conference on Machine Learning*, pages 12642–12652. PMLR.

Zhang, X., Wan, J.-X., Ke, Z.-P., Wang, F., Chai, H.-X., and Liu, J.-Q. (2017). TMEM88, CCL14 and CLEC3B as prognostic biomarkers for prognosis and palindromia of human hepatocellular carcinoma. *Tumor Biology*, 39(7):1010428317708900.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zheng, H., Yang, Z., Qiao, L., Li, J., and Liu, W. (2015). Attribute knowledge integration for speech recognition based on multi-task learning neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Zhu, H.-F., Zhang, X.-H., Gu, C.-S., Zhong, Y., Long, T., Ma, Y.-D., Hu, Z.-Y., Li, Z.-G., and Wang, X.-Y. (2019).

Cancer-associated fibroblasts promote colorectal cancer progression by secreting CLEC3B. *Cancer Biology & Therapy*, 20(7):967–978.

Zimmerman, M. W., Homanics, G. E., and Lazo, J. S. (2013). Targeted deletion of the metastasis-associated phosphatase Ptp4a3 (PRL-3) suppresses murine colon cancer. *PloS one*, 8(3):e58300.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No, but we provide all the experimental details to reproduce the main results]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable, no special computing infrastructure was used.]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix to "Meta Sparse Principal Component Analysis"

A Technical Desiderata

In this section of the appendix, we state some technical definitions and lemmas. We begin with the definition of a real valued sub-Gaussian random variable (Zhang et al., 2021).

Definition 2. *A random variable X is called sub-Gaussian with parameter $\sigma \geq 0$ if, $\forall \rho \in \mathbb{R}$*

$$\mathbb{E}[e^{\rho X}] \leq e^{\frac{\rho^2 \sigma^2}{2}}. \tag{A.1}$$

We can extend the definition of univariate sub-Gaussian distributions to define a general class of multivariate sub-Gaussian distribution on $\mathbb{R}^p$. Let $\Sigma \in \mathbb{R}^{p \times p}$ be a positive semi-definite, real symmetric matrix.

Definition 3. *We say that a random vector $X \in \mathbb{R}^p$ follows a multivariate sub-Gaussian distribution with covariance matrix Σ and parameter σ if*

1. $\mathbb{E}[X] = 0$, *and* $\mathrm{Cov}(X) = \Sigma$.
2. *For* $1 \leq i \leq p$, $\frac{X_i}{\sqrt{\Sigma_{i,i}}}$ *is a sub-Gaussian random variable with parameter* σ.

Our last definition of this section is the sub-exponential distribution (Pollard, 1990).

Definition 4. *A random variable X with mean 0 is said to follow a sub-exponential distribution with parameters (η, b) if,*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2 \eta^2}{2}} \text{ for } |\lambda| < \frac{1}{b}.$$

Remark. *It is well established that (see for instance Equation (6) in Honorio and Jaakkola (2014)) the square of a sub-Gaussian random variable with parameter σ is sub-exponential with parameters $(4\sqrt{2}\sigma^2, 4\sigma^2)$.*

Next, we recall the matrix Chernoff bound (Theorem 5.1.1 in Tropp (2015)).

Lemma 4. *Let A_i be random matrices such that $\mathbb{E}[A_i] = 0$. If $\|A_i\|_{\infty,\infty} \leq \gamma$ almost surely, then for every $\varepsilon > 0$,*

$$\mathbb{P}\left(\left\| \frac{1}{m} \sum A_i \right\|_{\infty,\infty} \geq \varepsilon \right) \leq 2pe^{-\frac{3\varepsilon^2 m}{8\gamma^2}}.$$

We then recall the following lemma from Lei and Vu (2015).

Lemma 5. *Under the assumptions in Theorem 1, let $\tilde{H}$ be the solution to the further constrained problem* (3.5)*. Then $\tilde{H}$ is rank k and unique. Furthermore, there exist $|J| \times k$ orthonormal matrices U_J, $\hat{U}_J$ such that:*

1. $\begin{bmatrix} U_J \\ 0 \end{bmatrix}$ *and* $\begin{bmatrix} \hat{U}_J \\ 0 \end{bmatrix}$ *spans the k-dimensional principal subspaces of Σ and $S - \rho\tilde{Z}$, respectively.*
2. *There exists a $|J| \times |J|$ orthonormal matrix Q such that*

$$\begin{aligned} \hat{U}_J &= QU_J, \\ \|Q - \mathbb{I}_{|J|}\|_F &\leq \frac{8\rho|J|}{\lambda_{diff}}. \end{aligned}$$

We also recall the following lemma about the principal subspace projection of a matrix. Theorem 1 in Overton and Womersley (1992).

Lemma 6. *Let A be a symmetric matrix with eigenvalues $\gamma_1 \geq \cdots \geq \gamma_p$ and orthonormal eigenvectors $v_1, \ldots, v_p$. $\max_{H \in \mathcal{F}^k} \langle A, H \rangle = \gamma_1 + \cdots + \gamma_k$ and the maximum is achieved by the projector of a k-dimensional principal subspace of A. Moreover, the maximiser is unique if and only if $\gamma_k > \gamma_{k+1}$.*

We also produce the following lemma along with the proof on the upper bound of the infinite norm of the product of two matrices.

Lemma 7. *Let A and B be two matrices with conformable dimensions. Then,*

$$\|AB\|_{\infty,\infty} \leq \|A\|_{\infty,\infty} \|B\|_{1,\infty}$$

Proof.

$$\|AB\|_{\infty,\infty} = \max_{i,j \in [p]} |\sum_{k \in [p]} A_{i,k} B_{k,j}| \leq \max_{i,j,k_0 \in [p]} |A_{i,k_0}| |\sum_{k \in [p]} B_{k,j}| \leq \|A\|_{\infty,\infty} \|B\|_{1,\infty}.$$

□

B Proofs of Main Theorems

B.1 Proof of Lemma 2

Proof. We note that $S^{(i)} - \Sigma^{(i)}$ are i.i.d with $\mathbb{E}\left[S^{(i)} - \Sigma^{(i)}\right] = 0$. We begin by finding an entrywise bound for $\frac{1}{m} \sum \left(S^{(i)} - \Sigma^{(i)}\right)$. Since $S^{(i)} = \frac{1}{p} \sum X_p^{(i)} \left(X_p^{(i)}\right)^T$,

$$\mathbb{P}\left(\left|\frac{1}{mn} \sum_{i,p} X_{k,p}^{(i)} X_{l,p}^{(i)} - \sigma_{k,l}^{(i)}\right| \geq \varepsilon\right).$$

Let $\sigma_k := \max_{1 \leq i \leq K} \sigma_{k,k}^{(i)}$, $\tilde{X}_{k,p}^{(i)} := \frac{X_{k,p}^{(i)}}{\sqrt{\sigma_k}}$, $\tilde{\rho}_{ij}^{(k)} := \frac{\sigma_{k,l}^{(i)}}{\sqrt{\sigma_k \sigma_l}}$. We have

$$\mathbb{P}\left(\left|\frac{1}{mn} \sum_{i,p} X_{k,p}^{(i)} X_{l,p}^{(i)} - \sigma_{k,l}^{(i)}\right| \geq \varepsilon\right) = \mathbb{P}\left[4\left|\sum_{i,p}\left(\tilde{X}_{k,p}^{(i)} \tilde{X}_{l,p}^{(i)} - \tilde{\rho}_{k,l}^{(i)}\right)\right| > \frac{4nm\varepsilon}{\sqrt{\sigma_k \sigma_l}}\right]$$

Define $U_{p,k,l}^{(i)} := \tilde{X}_{p,k}^{(i)} + \tilde{X}_{p,l}^{(i)}$, $V_{p,k,l}^{(i)} := \tilde{X}_{p,k}^{(i)} - \tilde{X}_{p,l}^{(i)}$. Then for any $r \in \mathbb{R}$,

$$4\sum_{i,p}\left(\tilde{X}_{k,p}^{(i)} \tilde{X}_{l,p}^{(i)} - \tilde{\rho}_{k,l}^{(i)}\right) = \sum_{k,t}\left\{\left(U_{p,k,l}^{(i)}\right)^2 - 2\left(r + \tilde{\rho}_{k,l}^{(i)}\right)\right\} - \sum_{k,t}\left\{\left(V_{p,k,l}^{(i)}\right)^2 - 2\left(r - \tilde{\rho}_{k,l}^{(i)}\right)\right\} \tag{B.1}$$

Thus,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{mn} \sum_{i,p} X_{k,p}^{(i)} X_{l,p}^{(i)} - \sigma_{k,l}^{(i)}\right| \geq \varepsilon\right) \leq & \mathbb{P}\left[\left|\sum_{k,t}\left\{\left(U_{t,ij}^{(k)}\right)^2 - 2\left(r + \tilde{\rho}_{ij}^{(k)}\right)\right\}\right| > \frac{2nm\varepsilon}{\sqrt{\sigma_k \sigma_l}}\right] \\ & + \mathbb{P}\left[\left|\sum_{k,t}\left\{\left(V_{t,ij}^{(k)}\right)^2 - 2\left(r - \tilde{\rho}_{ij}^{(k)}\right)\right\}\right| > \frac{2nm\varepsilon}{\sqrt{\sigma_k \sigma_l}}\right].\end{aligned} \tag{B.2}$$

Consider the first term

$$\mathbb{P}\left[\left|\sum_{i,p}\left\{\left(U_{p,k,l}^{(i)}\right)^2 - 2\left(r + \tilde{\rho}_{ij}^{(k)}\right)\right\}\right| > \frac{2nm\varepsilon}{\sqrt{\sigma_k \sigma_l}}\right],$$

and let $r = \frac{\sigma^{(i)}_{k,k}}{\sigma_k} + \frac{\sigma^{(i)}_{l,l}}{\sigma_l}$. Then,

$$\mathbb{E}\left[\left(U^{(i)}_{p,k,l}\right)^2 - 2\left(r + \tilde{\rho}^{(k)}_{ij}\right)\right] = 0.$$

Let $Z^{(i)}_{p,k,l} := \left(U^{(i)}_{p,k,l}\right)^2 - 2\left(r + \tilde{\rho}^{(k)}_{ij}\right)$. Consequently,

$$\mathbb{P}\left(\Big|\sum_{i,p}\left\{\left(U^{(i)}_{p,k,l}\right)^2 - 2\left(r + \tilde{\rho}^{(k)}_{ij}\right)\right\}\Big| > \frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}\right) = \mathbb{P}\left(\sum_{i,p} Z^{(i)}_{p,k,l} > \frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}\right) + \mathbb{P}\left(\sum_{i,p} Z^{(i)}_{p,k,l} < -\frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}\right) \tag{B.3}$$

To see that $U^{(i)}_{p,k,l}$ is sub-Gaussian, we note that,

$$\begin{aligned}\mathbb{E}(e^{\lambda U^{(i)}_{p,k,l}}) &= \mathbb{E}\left(e^{\lambda\left(X^{(i)}_{p,k} + X^{(i)}_{p,l}\right)}\right)\\ &\leq \left[\mathbb{E}e^{2\lambda\tilde{X}^{(i)}_{p,k}}\right]^{0.5}\left[\mathbb{E}e^{2\lambda\tilde{X}^{(i)}_{p,l}}\right]^{0.5}\\ &\leq e^{\lambda^2\left(\frac{\sigma^{(i)}_{k,k}}{\sigma_k}\sigma\right)^2} e^{\lambda^2\left(\frac{\sigma^{(i)}_{l,l}}{\sigma_l}\sigma\right)^2}.\end{aligned}$$

Since $\frac{\sigma^{(i)}_{k,k}}{\sigma_k}$ and $\frac{\sigma^{(i)}_{l,l}}{\sigma_l}$ are upper bounded by 1, we consequently have,

$$\mathbb{E}(e^{\lambda U^{(i)}_{p,k,l}}) \leq e^{2\lambda^2\sigma^2},$$

which shows that $U^{(i)}_{p,k,l}$ is sub-Gaussian with parameter 2σ. We also know, from Section 5.2.4 in Vershynin (2012), that if $U^{(i)}_{p,k,l}$ is a sub-Gaussian random variable with then $\left(U^{(i)}_{p,k,l}\right)^2$ is a sub-exponential random variable. As noted in Section 2 if X is a sub-Gaussian with parameter ω, then, X^2 is a sub-exponential distribution with parameters $(4\sqrt{2}\omega^2, 4\omega^2)$. Consequently, if $|t| < \frac{1}{4\omega^2}$,

$$\mathbb{E}[e^{tX^2}] \leq e^{16\omega^4 t^2}.$$

Returning to eq. (B.3), we observe that each $Z^{(i)}_{p,k,l}$ is i.i.d. subexponential with parameter $(16\sqrt{2}\sigma^2, 16\sigma^2)$. Hence for any $\frac{1}{16\sigma^2} > \theta > 0$,

$$\begin{aligned}\mathbb{P}\left(\sum_{i,p} Z^{(i)}_{p,k,l} > \frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}\right) &= \mathbb{P}\left(\theta\sum_{i,p} Z^{(i)}_{p,k,l} > \theta\frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}\right)\\ &= \mathbb{P}\left(e^{\theta\sum_{i,p} Z^{(i)}_{p,k,l}} > e^{\theta\frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}}\right).\end{aligned}$$

Thus, by Markov's inequality, it follows that,

$$\begin{aligned}\mathbb{P}\left(\sum_{i,p} Z^{(i)}_{p,k,l} > \frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}\right) &\leq \mathbb{E}\left[e^{\theta\sum_{i,p} Z^{(i)}_{p,k,l}}\right] e^{-\theta\frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}} \\ &= \prod_{i,p}\mathbb{E}\left[e^{\theta Z^{(i)}_{p,k,l}}\right] e^{-\theta\frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}} \\ &\leq \prod_{i,p} e^{512\theta^2\sigma^4} e^{-\theta\frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}} \\ &= e^{512nm\theta^2\sigma^4} e^{-\theta\frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}} \\ &= e^{512mn\theta^2\sigma^4-\theta\frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}}\end{aligned}$$

Minimising over $\theta > 0$, we get the derivative of the exponent with respect to θ to be equal to 0. In other words,

$$512mn\theta\sigma^4 - \frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}} = 0,$$

which in turn implies that whenever

$$\frac{\varepsilon}{512\sigma^4\sqrt{\sigma_k\sigma_l}} < \frac{1}{16\sigma^2},$$

the minimum is achieved when $\theta = \frac{\varepsilon}{512\sigma^4\sqrt{\sigma_k\sigma_l}}$. In other words, if

$$\varepsilon < 32\sigma^2\sqrt{\sigma_k\sigma_l},$$

the minimum is, $e^{-\frac{nm\varepsilon^2}{512\sigma^4\sqrt{\sigma_k\sigma_l}}}$. Consequently, our upper bound can be further refined as,

$$\mathbb{P}\left(\sum_{i,p} Z^{(i)}_{p,k,l} > \frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}\right) \leq e^{-\frac{nm\varepsilon^2}{512\sqrt{\sigma_k\sigma_l}\sigma^4}}.$$

Since for each $k \in [m]$, $\sigma_k < \max_{i\in[m]} \left\|\Sigma^{(i)}\right\|_{\infty,\infty} < \eta$, this in turn implies that the upper bound is independent of the value of $\Sigma^{(i)}$'s. In particular, this upper bound is distribution free. Hence, for $\varepsilon < 32\eta\sigma^2$,

$$\mathbb{P}\left(\sum_{i,p} Z^{(i)}_{p,k,l} > \frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}\right) \leq e^{-\frac{nm\varepsilon^2}{512\eta\sigma^4}}.$$

Similarly, the upper bound to $\mathbb{P}\left(\sum_{i,p} Z^{(i)}_{p,k,l} < -\frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}\right)$, can be achieved; and,

$$\mathbb{P}\left(\sum_{i,p} Z^{(i)}_{p,k,l} < -\frac{2nm\varepsilon}{\sqrt{\sigma_k\sigma_l}}\right) \leq e^{-\frac{nm\varepsilon^2}{512\eta\sigma^4}}.$$

Hence, we have found the entrywise error bound to $\frac{1}{m}\sum\left(S^{(i)} - \Sigma^{(i)}\right)$. To find an upper bound to $\left\|\frac{1}{m}\sum\left(S^{(i)} - \Sigma^{(i)}\right)\right\|_{\infty,\infty}$, we use a union bound.

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{m}\sum\left(S^{(i)} - \Sigma^{(i)}\right)\right\|_{\infty,\infty} > \varepsilon\right) &\leq \sum_{k,l}\mathbb{P}\left(\left|\frac{1}{mn}\sum_{i,p} X^{(i)}_{k,p}X^{(i)}_{l,p} - \sigma^{(i)}_{k,l}\right| \geq \varepsilon\right). \\ &\leq \frac{p(p+1)}{2} e^{-\frac{nm\varepsilon^2}{512\eta\sigma^4}}.\end{aligned}$$

This completes the proof. □

B.2 Proof of Theorem 1

Proof. Continue from the setup in Section 3.1. We now have to complete the proof by verifying the remaining 4 steps.

STEP 1. We begin by finding a probabilistic upper bound on $\|S-\Sigma\|_{\infty,\infty}$. As seen in eq. (2.4), $S=\frac{1}{m}\sum_{i=1}^{m}S^{(i)}$, where $S^{(i)}=\frac{1}{n}\sum_{j=1}^{n}X_j^{(i)}(X_j^{(i)})^T$ are estimated variance matrix from auxiliary tasks.

$$\|S-\Sigma\|_{\infty,\infty}\leq\underbrace{\left\|\frac{1}{m}\sum_{i=1}^{m}\left(S^{(i)}-\Sigma^{(i)}\right)\right\|_{\infty,\infty}}_{\text{Term 1}}+\underbrace{\left\|\frac{1}{m}\sum_{i=1}^{m}\Sigma^{(i)}-\Sigma\right\|_{\infty,\infty}}_{\text{Term 2}} \tag{B.4}$$

TERM 2.We begin by analysing the second term in eq. (B.4). Observe that,

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^{m}\Sigma^{(i)}-\Sigma\right\|_{\infty,\infty}>\varepsilon\right)&=\mathbb{P}\left(\bigcup_{k\leq j}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\Sigma_{k,j}^{(i)}-\Sigma_{k,j}\right|>\varepsilon\right\}\right)\\&\overset{(i)}{\leq}\sum_{k\leq j}\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\underbrace{\Sigma_{k,j}^{(i)}-\Sigma_{k,j}}_{=R_{k,j}^{(i)}}\right|>\varepsilon\right)\\&=\sum_{k\leq j}\mathbb{P}\left(\left|\bar{R}_{k,j}\right|>\varepsilon\right)\\&\overset{(ii)}{\leq}\sum_{k\leq j}2e^{-\frac{m\varepsilon^2}{\sigma_a^2}}\\&=p(p+1)e^{-\frac{m\varepsilon^2}{2\sigma_a^2}}\end{aligned} \tag{B.5}$$

where (i) follows from an union bound, and (ii) follows from a combination of Assumption 3 and Lemma 1.

Now, substituting $\varepsilon=\sqrt{8\sigma_a^2\log(p+1)/m}$ in eq. (B.5), we have

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^{m}\Sigma^{(i)}-\Sigma\right\|_{\infty,\infty}>\sqrt{\frac{8\sigma_a^2\log(p+1)}{m}}\right)&\leq\frac{p(p+1)}{(p+1)^4}\\&\leq\frac{1}{p^2}.\end{aligned}$$

We move to Term 1.

TERM 1. In this part of the proof, we provide an upper bound for the first term from eq. (B.4). We begin by finding a

high-probability upper bound to $\left\|\Sigma^{(i)}\right\|_{\infty,\infty}$. Observe that,

$$\begin{aligned}
\mathbb{P}\left(\sup_{1\le i\le m}\left\|\Sigma^{(i)}-\Sigma\right\|_{\infty,\infty}>\varepsilon\right) &= \mathbb{P}\left(\bigcup_{1\le i\le m}\bigcup_{k\le j}\left\{\left|\Sigma^{(i)}_{k,j}-\Sigma_{k,j}\right|>\varepsilon\right\}\right)\\
&\overset{(i)}{\le}\sum_{i=1}^{m}\sum_{k\le j}\mathbb{P}\left(\left|\underbrace{\Sigma^{(i)}_{k,j}-\Sigma_{k,j}}_{=R^{(i)}_{k,j}}\right|>\varepsilon\right)\\
&=\sum_{i=1}^{m}\sum_{k\le j}\mathbb{P}\left(\left|R^{(i)}_{k,j}\right|>\varepsilon\right)\\
&\overset{(ii)}{\le}\sum_{i=1}^{m}\sum_{k\le j}2e^{-\frac{m\varepsilon^2}{\sigma_a^2}}\\
&=mp(p+1)e^{-\frac{m\varepsilon^2}{2\sigma_a^2}}
\end{aligned}\tag{B.6}$$

where (i) follows from an union bound, and (ii) follows from a combination of Assumption 3 and Lemma 1. Now, substituting $\varepsilon=\sqrt{8\sigma_a^2\log(p+1)/m}$ in eq. (B.6), we have

$$\begin{aligned}
\mathbb{P}\left(\sup_{1\le i\le m}\left\|\Sigma^{(i)}-\Sigma\right\|_{\infty,\infty}>\sqrt{\frac{10\sigma_a^2\log(p+1)}{m}}\right) &\le \frac{mp(p+1)}{(p+1)^5}\\
&\overset{(i)}{\le}\frac{1}{p^2}.
\end{aligned}\tag{B.7}$$

(i) follows because $m\le p$. Therefore, it follows with probability at least $1-1/p^2$ that, $\sup_{i,j,k}\Sigma^{(i)}_{j,k}\le\eta+\sqrt{10\sigma_a^2\log(p+1)/m}$. We define $\lambda^\dagger:=\eta+\sqrt{10\sigma_a^2\log(p+1)/m}$

We define $A:=\left\{\sup_{1\le i\le m}\left\|\Sigma^{(i)}-\Sigma\right\|_{\infty,\infty}\le\sqrt{10\sigma_a^2\log(p+1)/m}\right\}$. Observe that,

$$\begin{aligned}
\mathbb{P}\left(\left\|\frac{1}{m}\sum\left(S^{(i)}-\Sigma^{(i)}\right)\right\|_{\infty,\infty}\ge\varepsilon\right) &= \mathbb{P}\left(\left\|\frac{1}{m}\sum\left(S^{(i)}-\Sigma^{(i)}\right)\right\|_{\infty,\infty}\ge\varepsilon\bigcap A\right)\\
&\quad+\mathbb{P}\left(\left\|\frac{1}{m}\sum\left(S^{(i)}-\Sigma^{(i)}\right)\right\|_{\infty,\infty}\ge\varepsilon\bigcap A^c\right)\\
&\overset{(i)}{=}\mathbb{P}\left(\left\|\frac{1}{m}\sum\left(S^{(i)}-\Sigma^{(i)}\right)\right\|_{\infty,\infty}\ge\varepsilon\bigcap A\right)+\frac{1}{p^2}\\
&\overset{(ii)}{\le}\mathbb{P}\left(\left\|\frac{1}{m}\sum\left(S^{(i)}-\Sigma^{(i)}\right)\right\|_{\infty,\infty}\ge\varepsilon\mid A\right)+\frac{1}{p^2}
\end{aligned}$$

By using Lemma 2 we get,

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum\left(S^{(i)}-\Sigma^{(i)}\right)\right\|_{\infty,\infty}\ge\varepsilon\mid A\right)\le\frac{p(p+1)}{2}e^{-\frac{nm\varepsilon^2}{512\sigma^4\lambda^\dagger}}.$$

putting $\varepsilon=16\sqrt{2\sigma^4\lambda^\dagger\frac{\log(p+1)}{nm}}$, we get with probability $1-\frac{1}{2(p+1)^2}$,

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum\left(S^{(i)}-\Sigma^{(i)}\right)\right\|_{\infty,\infty}\le\sqrt{512\sigma^4\lambda^\dagger\frac{\log(p+1)}{nm}}\right).$$

Thus,

$$\mathbb{P}\left(\|S-\Sigma\|_{\infty,\infty} > \max\left\{\sqrt{\frac{8\sigma_a^2\log(p+1)}{m}}, \sqrt{512\sigma^4\lambda^\dagger\frac{\log(p+1)}{nm}}\right\}\right) \leq \frac{1}{p^2} + \frac{1}{2(p+1)^2}$$
$$\leq \frac{3}{p^2}.$$

STEP 2. In this section of the proof, we show that our proposed estimator is feasible, unique, and optimal, i.e., satisfies the KKT conditions (3.2,3.3,3.4).

FEASIBLITY. The feasibility of $\hat{H}$ is obvious from construction. To check the feasibility of $\hat{Z}$, we need to verify that for a given choice of ρ,

$$\hat{Z}_{ij} \in [-1,1],\ \forall (i,j) \in (J\times J)^c.$$

Therefore, it is sufficient to verify

$$\frac{1}{\rho}\left[|S_{ij}-\Sigma_{ij}| + \left|\Sigma_{ij} - \langle Q_{i*}, \Sigma_{J,j}\rangle\right|\right] \leq 1.$$

By an application of Cauchy-Schwarz inequality, we get

$$\frac{1}{\rho}\left[|S_{ij}-\Sigma_{ij}| + \left|\Sigma_{ij} - \langle Q_{i*}, \Sigma_{J,j}\rangle\right|\right] \leq \frac{1}{\rho}\left[|\ \|S-\Sigma\|_{\infty,\infty} + \|Q-\mathbb{I}_{|J|}\|_F\|\Sigma\|_{2,\infty}\right]$$
$$\leq \frac{1}{\rho}\left[\|S-\Sigma\|_{\infty,\infty} + \frac{8\rho|J|}{\lambda_{diff}}\|\Sigma\|_{2,\infty}\right].$$

Thus, with probability at least $1-\frac{2}{p^2}-\frac{1}{2(p+1)^2}$,

$$\rho^{-1}\|S-\Sigma\|_{\infty,\infty} + \frac{8|J|}{\lambda_{diff}}\|\Sigma_{J^c,J}\|_{2,\infty} \leq 1. \tag{B.8}$$

OPTIMALITY. $\hat{H}$ only has non-zero entries in the sub-matrix $\hat{H}_{J,J}$. By construction $(\hat{H}_{i,j}, \hat{Z}_{i,j}) = (\tilde{H}_{i,j}, \tilde{Z}_{i,j}), \forall (i,j) \in J\times J$. From the optimality of $(\tilde{H}, \tilde{Z})$, Equation (3.2) is satisfied.

The fact that Equation (3.3) is satisfied follows from the optimality of $(\tilde{H}, \tilde{Z})$ when $(i,j) \in J\times J$ and feasibility of $\tilde{Z}$ otherwise.

Note that $\hat{U}_J$ spans the k-dimensional principal subspace of $S-\rho\tilde{Z}$. We now show that $\hat{U}$ spans one (not necessarily principal) k-dimensional subspace of $S-\rho\hat{Z}$. In other words, it is enough to show that, for some diagonal matrix D_0,

$$(S-\rho\hat{Z})\begin{bmatrix}\hat{U}_J\\0\end{bmatrix} = D\begin{bmatrix}\hat{U}_J\\0\end{bmatrix}. \tag{B.9}$$

Analysing the term on the left-hand side of the previous equation, we get,

$$(S-\rho\widehat{Z})\begin{bmatrix}\hat{U}_J\\0\end{bmatrix} = \begin{bmatrix}S_{J,J}-\rho\tilde{Z}_{J,J} & Q\Sigma_{J,J^c}\\ \Sigma_{J^c,J}Q^T & \Sigma_{J^c,J^c}+\mathbb{D}(S_{J^c,J^c}-\Sigma_{J^c,J^c})\end{bmatrix}\begin{bmatrix}\hat{U}_J\\0\end{bmatrix}$$
$$= \begin{bmatrix}(S_{J,J}-\rho\tilde{Z}_{J,J})\hat{U}_J\\ \Sigma_{J^cJ}Q^T\hat{U}_J\end{bmatrix}$$
$$= \begin{bmatrix}(S_{J,J}-\rho\tilde{Z}_{J,J})\hat{U}_J\\ \Sigma_{J^cJ}Q^TQU_J\end{bmatrix}$$

Recall from Lemma 5 that there exists a diagonal matrix D_1 such that $(S_{J,J}-\rho\tilde{Z}_{J,J})\hat{U}_J = D_1\hat{U}_J$. Also, the fact that $\begin{bmatrix}U_J\\0\end{bmatrix}$

spans the k-dimensional principal subspace of Σ implies that $\Sigma_{J^cJ}U_J = 0$. Therefore,

$$\begin{aligned}\begin{bmatrix}(S_{J,J} - \rho\tilde{Z}_{J,J})\hat{U}_J \\ \Sigma_{J^cJ}Q^TQU_J\end{bmatrix} &= \begin{bmatrix}D_1\hat{U}_J \\ \Sigma_{J^cJ}U_J\end{bmatrix} \\ &= \begin{bmatrix}D_1\hat{U}_J \\ 0\end{bmatrix} \\ &= \begin{bmatrix}D_1\hat{U}_J \\ 0\end{bmatrix} \\ &= \begin{bmatrix}D_1 & 0 \\ 0 & 0\end{bmatrix}\begin{bmatrix}\hat{U}_J \\ 0\end{bmatrix}\end{aligned}$$

Setting

$$D_0 = \begin{bmatrix}D_1 & 0 \\ 0 & 0\end{bmatrix}$$

we successfully show that Equation (B.9) holds. We will further prove that $diag(D_1)$ are the first k eigenvalues of $(S - \rho\widehat{Z})$, and $\lambda_k(S - \rho\widehat{Z}) - \lambda_{k+1}(S - \rho\widehat{Z}) > 0$. Then uniqueness will follow by lemma 5.

$$\begin{aligned}(S - \rho\hat{Z}) &= \begin{bmatrix}S_{JJ} - \rho\tilde{Z}_{JJ} - Q\Sigma_{JJ}Q^T & 0 \\ 0 & \Sigma_{J^c,J^c} + \mathbb{D}(S_{J^c,J^c} - \Sigma_{J^c,J^c})\end{bmatrix} + \begin{bmatrix}Q\Sigma_{JJ}Q^T & Q\Sigma_{JJ^c} \\ \Sigma_{J^cJ}Q^T & \Sigma_{J^cJ^c}\end{bmatrix} \\ &= \begin{bmatrix}S_{JJ} - \rho\tilde{Z}_{JJ} - Q\Sigma_{JJ}Q^T & 0 \\ 0 & \Sigma_{J^c,J^c} + \mathbb{D}(S_{J^c,J^c} - \Sigma_{J^c,J^c})\end{bmatrix} + \begin{bmatrix}Q & 0 \\ 0 & I\end{bmatrix} \times \Sigma \times \begin{bmatrix}Q^T & 0 \\ 0 & I\end{bmatrix}\end{aligned}$$

However, since we just proved that $(S - \rho\widehat{Z})$ satisfies the conditions of Proposition 3, the problem reduces to finding an upper bound to $\lambda_l(S_{J,J} - \rho\tilde{Z}_{J,J} - Q\Sigma_{J,J}Q^T)$ for each $l \in [|J|]$. Note that,

$$\begin{aligned}\lambda_l(S_{J,J} - \rho\tilde{Z}_{J,J} - Q\Sigma_{J,J}Q^T) &\leq \lambda_l(S_{J,J} - \rho\tilde{Z}_{J,J} - \Sigma_{J,J}) + \lambda_l(\Sigma_{J,J} - Q\Sigma_{J,J}Q^T) \\ &\leq \|S_{J,J} - \rho\tilde{Z}_{J,J} - \Sigma_{J,J}\|_F + \|\Sigma_{J,J} - Q\Sigma_{J,J}Q^T\|_F \\ &\leq \|S_{J,J} - \rho\tilde{Z}_{J,J} - \Sigma_{J,J}\|_F + 2\|\Sigma_{J,J}\|_F \times \|Q - I\|_F \\ &\leq 2\rho|J| + 2\lambda_1(\Sigma) \times 8\rho|J|/(\lambda_k - \lambda_{k+1}).\end{aligned}$$

Where the last inequality follows from the combination of part 1 and Lemma 5. The rest of the proof follows since for a large enough ρ we have,

$$4\rho|J| + \frac{32\lambda_1(\Sigma)\rho|J|}{(\lambda_k - \lambda_{k+1})} \leq \lambda_k - \lambda_{k+1}.$$

This completes the proof of step 2.

STEP 3. In this section of the proof, we prove that $\left\|\hat{H} - \Pi\right\|_{\infty,\infty} < 2\frac{\rho|J|}{\lambda_{diff}}$. By step 2, we have,

$$\begin{aligned}\left\|\hat{H} - \Pi\right\|_{\infty,\infty} &= \left\|\begin{bmatrix}\hat{U}_J\hat{U}_J^T & 0 \\ 0 & 0\end{bmatrix} - \begin{bmatrix}U_JU_J^T & 0 \\ 0 & 0\end{bmatrix}\right\|_{\infty,\infty} \\ &= \left\|\begin{bmatrix}QU_JU_J^TQ^T & 0 \\ 0 & 0\end{bmatrix} - \begin{bmatrix}U_JU_J^T & 0 \\ 0 & 0\end{bmatrix}\right\|_{\infty,\infty} \\ &= \left\|QU_JU_J^TQ^T - U_JU_J^T\right\|_{\infty,\infty} \\ &\leq \lambda_1(U_JU_J^T)\|\mathbb{I} - Q\|_F \\ &\leq \frac{8|J|\rho}{\lambda_{diff}}.\end{aligned}$$

Setting $c(\alpha)_\pi = \frac{8|J|}{\lambda_{diff}}$, completes the proof of the upper bound for $\left\|\hat{H} - \Pi\right\|_{\infty,\infty}$.

STEP 4. Set

$$\rho < \min_{i \in J} \frac{\Pi_{i,i}\lambda_{diff}}{16|J|}.$$

The rest of the proof now follows from the step 3. □

B.3 Proof of Theorem 2

Proof. To show this part, let p be an odd number, let g_p be $2p$. Set $U = \mathbb{I}_p$, and let the true support be J_*. It is easy to see that $|J_*| = k$. Observe that there are $\binom{p}{k}$ many choices for J_* and assume that nature chooses one uniformly. Fix one such choice J. For that J, define the diagonal matrix $\Sigma(J)$ as follows:

$$\Sigma_{jj}(J) = \begin{cases} \frac{k+1}{k};\ j \in J \\ 1;\ j \notin J. \end{cases}$$

Next, recall that there are $\binom{p}{k}$ many choices for J. Define $R^{(i)}(J)$ as follows: Let s be the smallest element of i modulo $k+1$, $i \in J$ and t be the smallest element of i modulo $p-k+1$, $i \in J^c$. Then,

$$R^{(i)}_{jj}(J) = \begin{cases} -\frac{1}{k};\ j = s \\ \frac{1}{k};\ j = t \\ 0;\ otherwise \end{cases}$$

Note that, $\Sigma(J)$ and $\Sigma(J) + R^{(i)}(J)$ has k values $(k+1)/k$ and $p-k$ values 1. Observe that,

$$\Sigma^{(i)}(J) = \Sigma(J) + R^{(i)}(J) \tag{B.10}$$

It follows that, $|\Sigma^{(i)}(J)| = \left(\frac{k+1}{k}\right)^k$. Let $X^{(i)}_j(J) \sim N(0, \Sigma^{(i)}(J))$. Observe that the KL divergence between two multivariate Gaussian random variables of dimension p can be written as

$$KL(N(0,\Sigma_1)||N(0,\Sigma_2)) = \frac{1}{2}\left(\log\frac{|\Sigma_1|}{|\Sigma_2|} + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - p\right)$$

Substituting, $\Sigma^{(i)}(J)$ in place of Σ_1 and $\Sigma^{(i)}(J')$ in place of Σ_2, and assume that nature chose $J \in J^{(1)}$ and $J' \in J^{(2)}$. It follows that

$$\begin{aligned} KL\left(N(0,\Sigma^{(i)}(J))||N(0,\Sigma^{(i)}(J'))\right) = \frac{1}{2}\Bigg(&\log\frac{|\Sigma(J') + R^{(i)}(J')|}{|\Sigma(J) + R^{(i)}(J)|} \\ &+ \mathrm{tr}\left(\left(\Sigma(J') + R^{(i)}(J')\right)^{-1}\left(\left(\Sigma(J) + R^{(i)}\right)(J)\right)\right) - p\Bigg) \end{aligned}$$

Observe that $|\Sigma(J') + R^{(i)}(J')| = |\Sigma(J) + R^{(i)}(J)| = \left(\frac{k+1}{k}\right)^k$. Thus,

$$\log\frac{|\Sigma(J') + R^{(i)}(J')|}{|\Sigma(J) + R^{(i)}(J)|} = \log 1 = 0$$

Furthermore, observe that $\left(\Sigma(J') + R^{(i)}(J')\right)$ is a diagonal matrix with entries at least 1. Therefore,

$$\mathrm{tr}\left(\left(\Sigma(J') + R^{(i)}(J')\right)^{-1}\left(\Sigma(J) + R^{(i)}(J)\right)\right) \leq \mathrm{tr}\left(\Sigma(J) + R^{(i)}(J)\right)$$

We now evaluate the right hand side of the previous expression to get

$$\mathrm{tr}\left(\Sigma(J) + R^{(i)}(J)\right) = k\left(\frac{k+1}{k}\right) + p - k.$$

Therefore,

$$KL\left(N(0,\Sigma^{(i)}(J))||N(0,\Sigma^{(i)}(J'))\right) \leq k\left(\frac{k+1}{k}\right) + p - k - p = 1$$

where the inequality follows from the fact that k many of $\Sigma(J) + R^{(i)}(J)$'s entries were $(k+1)/k$ and the rest were 1. We now have that,

$$KL\left(N(0,\Sigma^{(i)}(J))||N(0,\Sigma^{(i)}(J'))\right) \leq 1,$$

Using Fano's inequality combined with data processing inequality (Fano, 1949) we now have that,

$$\begin{aligned}
\mathbb{P}\left(\hat{J} \neq J_*\right) &\geq 1 - \frac{1}{2}\frac{1}{\binom{p}{k}^2}\sum_{J,J'}\sum_{m,n}\frac{KL\left(N(0,\Sigma^{(i)}(J))||N(0,\Sigma^{(i)}(J'))\right)}{\log\binom{p}{k}} - \frac{1}{2}\frac{\log 2}{\log\binom{p}{k}} \\
&\geq 1 - \frac{1}{2}\frac{1}{\binom{p}{k}^2}\sum_{\substack{J\in J^{(1)}\\ J'\in J^{(2)}}}\sum_{m,n}\frac{KL\left(N(0,\Sigma^{(i)}(J))||N(0,\Sigma^{(i)}(J'))\right)}{\log\binom{p}{k}} - \frac{1}{2}\frac{\log 2}{\log\binom{p}{k}} \\
&\geq 1 - \frac{1}{4}\frac{mn}{\log\binom{p}{k}} - \frac{1}{2}\frac{\log 2}{\log\binom{p}{k}} \\
&\geq 1 - \frac{1}{4}\frac{mn}{k\log p - k\log k} - \frac{1}{2}\frac{\log 2}{k\log p - k\log k},
\end{aligned}$$

where the last inequality follows since $(p/k)^k \leq \binom{p}{k}$. Let p be large enough such that

$$\frac{1}{2}\frac{\log 2}{k\log p - k\log k} \leq \frac{1}{2}.$$

Then,

$$\mathbb{P}\left(\hat{J} \neq J_*\right) \geq \frac{1}{4}$$

whenever,

$$n \leq \frac{k\log p - k\log k}{m}.$$

Recall that g_p was set to be $2p$. This implies that,

$$\mathbb{P}\left(\hat{J} \neq J_*\right) \geq \frac{1}{4}$$

whenever,

$$n \leq \frac{k\log p - k\log k}{2m}.$$

This completes the proof. □

B.4 Proof of Proposition 3

Proof. Let $\Pi^{(m+1)}$ be the k-component PC matrix corresponding to $\Sigma^{(m+1)}$. Hence,

$$\Pi^{(m+1)} = \sum_{i=1}^{k} U_{*,i}^{(m+1)}\left(U_{*,i}^{(m+1)}\right)^T.$$

We also know from the recovered support union that $\Pi_{i,i}^{(m+1)} = 0$ whenever $i \notin J$. However,

$$\Pi_{i,i}^{(m+1)} = \sum_{j=1}^{k}\left(U_{j,i}^{(m+1)}\right)^2.$$

Hence, $\forall j \in J^c,\ U_{i,j}^{(m+1)} = 0$. Therefore,

$$U_i^{(m+1)} = \begin{pmatrix} U_{J,i}^{(m+1)} \\ 0 \end{pmatrix}.$$

Moreover, we also know that

$$\Sigma^{(m+1)}U_i^{(m+1)} = \lambda_i U_i^{(m+1)}.$$

In other words,

$$\begin{bmatrix} \Sigma_{J,J}^{(m+1)} & \Sigma_{J,J^c}^{(m+1)} \\ \Sigma_{J^c,J}^{(m+1)} & \Sigma_{J^c,J^c}^{(m+1)} \end{bmatrix} \begin{pmatrix} U_{J,i}^{(m+1)} \\ 0 \end{pmatrix} = \lambda_i \begin{pmatrix} U_{J,i}^{(m+1)} \\ 0 \end{pmatrix}.$$

Which in turn implies,

$$\Sigma_{J,J}^{(m+1)} U_{J,i}^{(m+1)} = \lambda_i U_{J,i}^{(m+1)}.$$

This completes the proof. □

B.5 Proof of Theorem 3

Proof. As given, $\Sigma^{(m+1)}$ is the novel covariance matrix with the following block structure,

$$\Sigma^{(m+1)} = \begin{bmatrix} \Sigma_{J,J}^{(m+1)} & \Sigma_{J,J^c}^{(m+1)} \\ \Sigma_{J^c,J}^{(m+1)} & \Sigma_{J^c,J^c}^{(m+1)} \end{bmatrix}.$$

For $1 \leq i \leq k$, let $U_i^{(m+1)}$ be the eigenvectors corresponding to $\Sigma^{(m+1)}$, and $\lambda_i^{(m+1)}$ be the corresponding eigenvalues. Let J be the recovered support union. $U_{J,i}^{(m+1)}$ denotes those coordinates of $U_i^{(m+1)}$ that belong to set J. Similarly to the proof of Theorem 1, if the penalty parameter ρ satisfies,

$$\rho^{-1}\|S_{J,J}^{(m+1)} - \Sigma_{J,J}^{(m+1)}\|_{\infty,\infty} + \frac{8|J|^{(m+1)}}{\lambda_k^{(m+1)} - \lambda_{k+1}^{(m+1)}}\|\Sigma_{(J^{(m+1)})^c J^{(m+1)}}\|_{2,\infty} \leq 1 \quad \text{and}$$

$$4\rho|J^{(m+1)}|\left(1 + \frac{8\lambda_1^{(m+1)}}{\lambda_k^{(m+1)} - \lambda_{k+1}^{(m+1)}}\right) < \lambda_k^{(m+1)} - \lambda_{k+1}^{(m+1)},$$

then the solution $\widehat{H}^{(m+1)}$ of Equation (4.2) is unique and satisfies $supp(\widehat{H}^{(m+1)}) \subseteq J^{(m+1)}$. It is noteworthy that if for each $j \in [n^{(m+1)}]$, $X_j^{(m+1)}$ is a p-dimensional multivariate sub-Gaussian distribution with covariance matrix Σ and parameter σ, then for any set of sub-indices J, $X_{j,J}^{(m+1)}$ is a $|J|$-dimensional multivariate sub-Gaussian distribution with covariance matrix $\Sigma_{J,J}$ and parameter σ. Following from eq. (4.1), we can also see that by construction

$$S_{J,J}^{(m+1)} = \frac{1}{n^{(m+1)}} \sum_{l=1}^{n^{(m+1)}} X_{l,J}^{(m+1)} \left(X_{l,J}^{(m+1)}\right)^T.$$

By using Lemma 2, we know that as long as $\varepsilon \leq 32\lambda_1\left(\Sigma^{(m+1)}\right)\sigma^2$,

$$\mathbb{P}\left(\left\|\left(S^{(m+1)} - \Sigma^{(m+1)}\right)\right\|_{\infty,\infty} \geq \varepsilon\right) \leq |J|(|J|+1)e^{-\frac{n\varepsilon^2}{512\sigma^4\lambda_1\left(\Sigma^{(m+1)}\right)}}.$$

As before, let $\alpha^{(m+1)} = 1 - \frac{8|J|^{(m+1)}}{\lambda_k^{(m+1)} - \lambda_{k+1}^{(m+1)}}\|\Sigma_{(J^{(m+1)})^c J^{(m+1)}}\|_{2,\infty}$. By Assumption 5, $\alpha^{(m+1)} \in (0,1)$.

Therefore, as long as, for some $\varepsilon, \delta > 0$,

$$n \geq \frac{512\sigma^2\lambda_1\left(\Sigma^{(m+1)}\right)}{\varepsilon^2} \log\left(\frac{|J|(|J|+1)}{\delta}\right)$$

and

$$\frac{\varepsilon^2}{\alpha^{(m+1)}} < \rho < \frac{\lambda_k^{(m+1)} - \lambda_{k+1}^{(m+1)}}{4|J^{(m+1)}|\left(1 + \frac{8\lambda_1^{(m+1)}}{\lambda_k^{(m+1)} - \lambda_{k+1}^{(m+1)}}\right)}$$

we have, with probability at least $1 - \delta$, we have $supp(\Pi^{(m+1)}) = J^{(m+1)}$.

□

B.6 Proof of Lemma 1

Proof. **Step I:** We first prove that if $X_1, \ldots, X_n$ are i.i.d zero mean sub-Gaussians with parameter σ, then $\bar{X}$ is a sub-Gaussian with parameter $\sigma/\sqrt{n}$. For any $\rho \in \mathbb{R}$

$$\begin{aligned}
\mathbb{E}[e^{\rho \bar{X}}] &= \mathbb{E}[e^{\frac{\rho}{n}\sum_{i=1}^{n} X_i}] \\
&= \mathbb{E}\left[\prod_{i=1}^{n} e^{\frac{\rho}{n} X_i}\right] \\
&\overset{(i)}{=} \prod_{i=1}^{n} \mathbb{E}\left[e^{\frac{\rho}{n} X_i}\right] \\
&\overset{(ii)}{=} \mathbb{E}\left[e^{\frac{\rho}{n} X_1}\right]^n \\
&\overset{(iii)}{\leq} \left(e^{\frac{\rho^2\sigma^2}{n^2}}\right)^n \\
&= e^{\rho^2 \frac{\sigma^2}{n}}.
\end{aligned}$$

Where (i) follows by independence of X_i's, (ii) follows because they are identically distributed, (iii) follows from the definition of sub-Gaussianity of X_1. It now follows from Definition 1 that $\bar{X}$ is sub-Gaussian with parameter $\sigma/\sqrt{n}$.

Step II: We now prove that if X is a mean zero sub-Gaussian random variable with parameter σ, then

$$\mathbb{P}\left(|X| > \varepsilon\right) \leq e^{-\frac{\varepsilon^2}{2\sigma^2}}$$

Pick some arbitrary $t > 0$. By Markov's inequality we have

$$\mathbb{P}[X \geq \varepsilon] \;=\; \mathbb{P}\big(e^{tX} \geq e^{t\varepsilon}\big) \;\leq\; e^{-t\varepsilon}\,\mathbb{E}\big[e^{tX}\big] \;\leq\; e^{-t\varepsilon}\, e^{\frac{\sigma^2 t^2}{2}} \;=\; e^{-t\varepsilon + \frac{\sigma^2 t^2}{2}}.$$

To minimize the quadratic function $f(t) = -t\varepsilon + \frac{\sigma^2 t^2}{2}$, set its derivative to zero:

$$0 \;=\; \frac{\partial f(t)}{\partial t} \;=\; -\varepsilon + \sigma^2 t.$$

Hence $t = \varepsilon/\sigma^2$. Substituting this value of t into the bound above we get

$$\mathbb{P}\left(X > \varepsilon\right) \leq \mathbb{E}^{-\frac{\varepsilon^2}{2\sigma^2}}.$$

We can similarly show

$$\mathbb{P}\left(X < -\varepsilon\right) \leq \mathbb{E}^{-\frac{\varepsilon^2}{2\sigma^2}}.$$

Combining, we have

$$\begin{aligned}
\mathbb{P}\left(|X| > \varepsilon\right) &= \mathbb{P}\left(X > \varepsilon \bigcup X < -\varepsilon\right) \\
&\leq \mathbb{P}\left(X > \varepsilon\right) + \mathbb{P}\left(X < -\varepsilon\right) \\
&\leq 2e^{-\frac{\varepsilon^2}{2\sigma^2}}.
\end{aligned}$$

From **Step I**, it follows that $\bar{X}$ is sub-Gaussian with parameter $\sigma/\sqrt{n}$. Combining, this with **Step II**, our lemma is proved. □

C Additional Experiments

To emulate the challenging "small-n/large-m" regime highlighted in the introduction, we adopt the following configuration. For all experiments we fix $p = 50$, $k = 10$ and $|J| = 10$, and perform 50 independent repetitions. We sweep the sample size over $\{6, 7, 8\}$, keeping all other dimensions unchanged. The regularization parameter is set to

$$\rho = 3\sqrt{\frac{\log(p+1)}{mn}},$$

and the number of auxiliary tasks scales as

$$m = \lfloor T \log(p+1)/n \rfloor + 1,$$

so the rescaled horizon $T \in \{1, \ldots, 15\}$ provides a convenient horizontal axis in all figures.

For each task, the rows $X_j^{(i)} \in \mathbb{R}^p$ are drawn from one of three distributions—Gaussian, uniform on $[-\sqrt{3}, \sqrt{3}]$, or a 50-50 mixture of the two—while the underlying covariances follow the rank-k signal-plus-noise model described in the code listing.

Figures 1, 2, and 3 displays, respectively for Gaussian, Uniform, and sub-Gaussian distributions, (i) the probability of exact support recovery, (ii) the relative estimation error $\|\hat{\Pi} - \Pi\|_{\infty,\infty}$ for the main tasks, and (iii) the support recovery for the novel task. The three sample-size curves overlap almost perfectly, confirming that once the factor $\sqrt{\log(p+1)/(mn)}$ is accounted for, the performance depends essentially only on T.

The recovered support union is then transferred to a novel task whose sample size is chosen as $n^{(m+1)} = \lceil T \log(|J|+1) \rceil$. With a single new active coordinate ($k_{\text{nov}} = 1$), the probability of correctly identifying that coordinate consistently exceeds 95% in every configuration we tested.

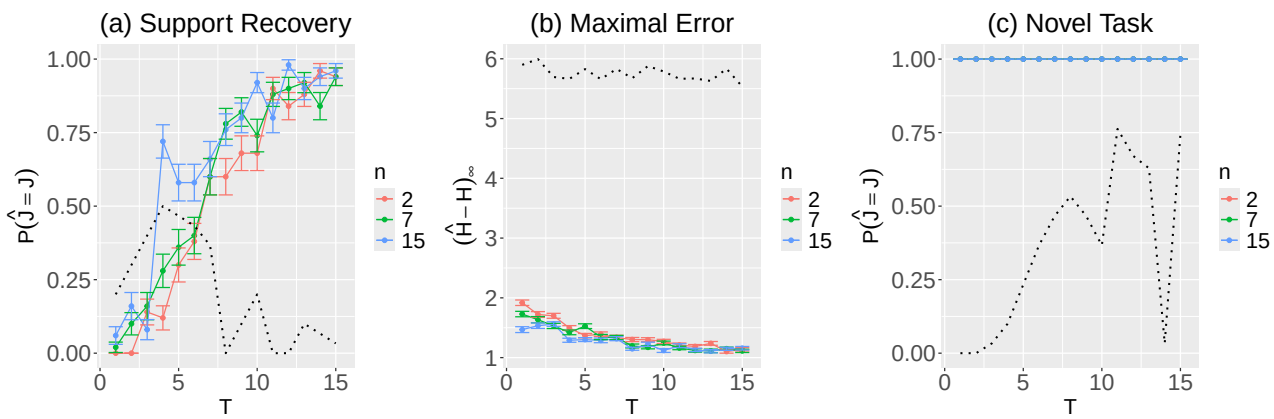

Figure 2: (a), (b): Simulations for Theorem 1 on the probability of support union recovery and maximal error under various settings of n and auxilliary tasks generated by uniform random variables. Let $\rho = \sqrt{\frac{\log(p+1)}{mn}}$. The x-axis is the constant T used to set the number of tasks $m = \lceil \frac{T \log(p+1)}{n} \rceil$. We vary T from $\{1, \ldots, 15\}$. (c): Probability of support recovery for novel task. We let $\rho = \sqrt{\frac{\log(|J|+1)}{n}}$. The x-axis is the constant T used to set the number of samples in the novel task $n^{(m+1)} = \lceil T \log(|J|+1) \rceil$. We vary T from $\{1, \ldots, 15\}$. The dotted black line represents the baseline, which is the best result that either single or multitask learning obtained for any n. Our meta learning method is clearly superior.

Dimension and Rank Dependence. To assess robustness of our method against changes in *dimension* and *rank*, we run two more complementary experiments.

For this section, all data are centred Gaussian and every configuration is repeated 50 times. The common parameters are $n = 5$ (samples per task) and the task budget

$$m = \lfloor T \log(p+1)/n \rfloor + 1, \qquad \rho = 3\sqrt{\frac{\log(p+1)}{mn}}.$$

For the ambient–dimension sweep we fix $k = 10$ and vary $p \in \{50, 100, 200\}$; for the rank sweep we fix $p = 50$ and vary $k \in \{5, 10, 15\}$. In both cases the novel task receives $n^{(m+1)} = \lceil T \log(k+1) \rceil$ samples and contains a single additional active coordinate ($k_{\text{nov}} = 1$).

Figures 4 and 5 overlay the main-task hit-rate, subspace error, and novel-task recovery probability for all settings:

- Dimension (p). After normalising by $\log(p+1)$, the curves for $p = 50, 100, 200$ are virtually identical. Exact support recovery exceeds 95% once $T \geq 6$, and the relative subspace error shrinks. The novel task is solved with $> 95\%$ accuracy even for small values of T.

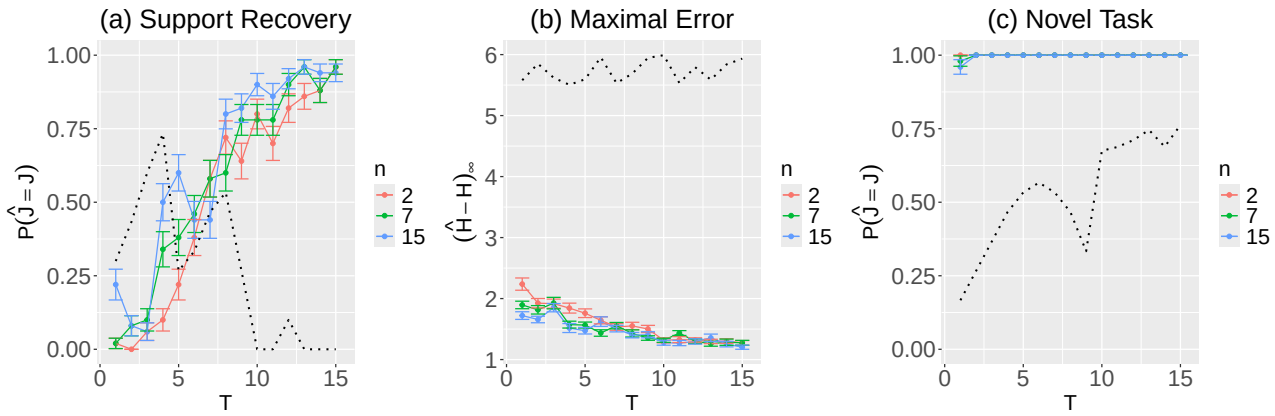

Figure 3: (a), (b): Simulations for Theorem 1 on the probability of support union recovery and maximal error under various settings of n and auxilliary tasks generated by sub-Gaussian random variables. We let $\rho = \sqrt{\frac{\log(p+1)}{mn}}$. The x-axis is the constant T used to set the number of tasks $m = \lceil \frac{T \log(p+1)}{n} \rceil$. We vary T from $\{1, \ldots, 15\}$. (c): Probability of support recovery for novel task. We take $\rho = \sqrt{\frac{\log(|J|+1)}{n}}$. The x-axis is the constant T used to set the number of samples in the novel task $n^{(m+1)} = \lceil T \log(|J| + 1) \rceil$. We vary T from $\{1, \ldots, 15\}$. The dotted black line represents the baseline, which is the best result that either single or multitask learning obtained for any n. Our meta learning method is clearly superior.

- Rank (k). Increasing the intrinsic rank from 5 to 15 induces only a mild degradation: the main-task hit-rate remains above 95% for $T \geq 6$, and the error curves retain the same downward trend. Thanks to the adaptive choice $n^{(m+1)} \propto \log(k + 1)$, the novel-task success rate again stays above 95% across the board.

Conclusion. Once the theoretical $\sqrt{\log(p + 1)/(mn)}$ factor is incorporated, performance is essentially invariant to the dimension, and scales gracefully with moderate increases in model rank. The proposed estimator therefore remains reliable even as the problem dimensionality and latent complexity evolve.

C.1 Comparison With Multi-Task Sparse PCA

In this section, we compare the meta-learning method against a multi-task learning method when applied to support union recovery under a Gaussian distribution setting. The method of generating the dataset and meta-learning is as described in Section 5.1. As before, we let $p = 30$, $|J| = 5$, and fix $n = 3$. The values we choose for m are chosen from the set $\mathcal{M} \in \{2, 3, 4, 5, 6, 7, 9, 10\}$. We perform 24 replications and report the mean probability and the 95% confidence interval as the error bars around the mean. The results are plotted in fig. 6 (a). The penalty ρ is always applied to be $\sqrt{\log(p + 1)/mn}$. As previously noted, as the number of tasks increase, meta-learning achieves a perfect support recovery, while multi-task learning gets progressively worse.

Multi-Task Learning. For the multi-task learning method, we maximise the following objective function:

$$\sum_{i=1}^{m} \langle S^{(i)}, H^{(i)} \rangle - \rho \|H^{(1)}, \ldots, H^{(m)}\|_1 \text{ subject to } H^{(i)} \in \mathcal{F}^k, \ \forall i \in [m],$$

where $S^{(i)}$ is the covariance matrix corresponding to task i and $\|H^{(1)}, \ldots, H^{(m)}\|_1 = \sum_{j,k} \sup_{i \in [m]} |H^{(i)}_{j,k}|$. We apply $\sqrt{\frac{\log(p+1)}{mn}}$ as the penalty for the tasks. We recover the support union by taking the union of the recovered support.

Meta-Learning. For the meta-learning setup, we proceed according to Section 5.1 and recover the support union from the pooled covariance estimator. We apply $\sqrt{\frac{\log(p+1)}{mn}}$ as the penalty when m is the total number of tasks.

As we can see, meta-learning outperforms multi-task learning with 100% accuracy for large values of m.

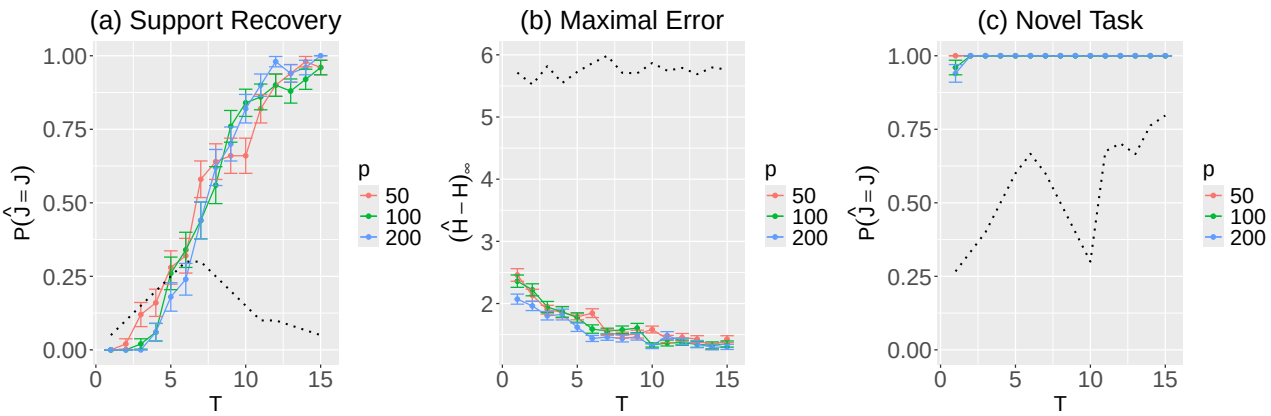

Figure 4: (a), (b): Simulations for Theorem 1 on the probability of support union recovery and maximal error under various settings of p and auxilliary tasks generated by Gaussian random variables. Let $\rho = \sqrt{\frac{\log(p+1)}{mn}}$. The x-axis is set by $T := \frac{mn}{\log(p+1)}$ varying from $\{1, \ldots, 15\}$ for support recovery. (c): Probability of support recovery for novel task. We let $\rho = \sqrt{\frac{\log(|J|+1)}{n}}$. The x-axis is set by $T := \frac{n}{\log(|J|+1)}$ varying from $\{10, \ldots, 25\}$. The dotted black line represents the baseline, which is the best result that either single or multitask learning obtained for any n. Our meta learning method is clearly superior.

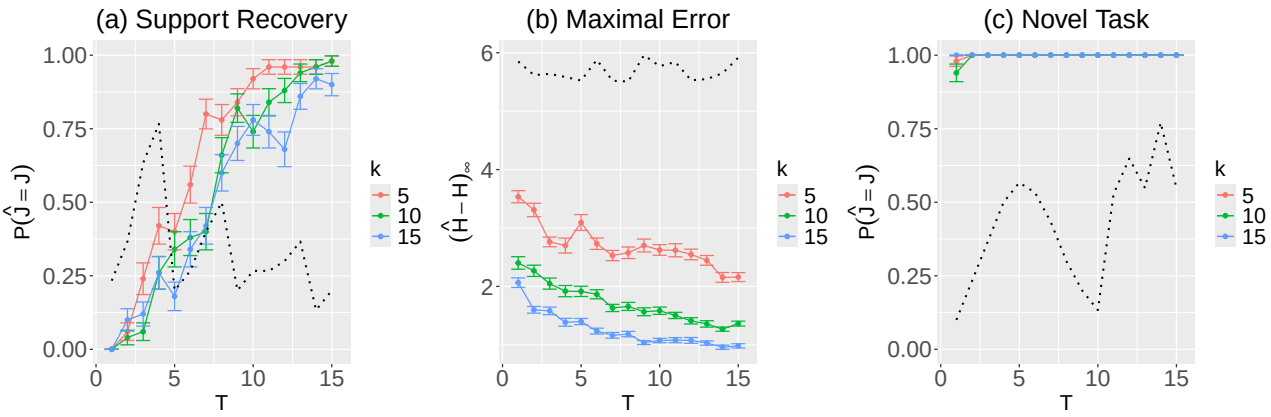

Figure 5: (a), (b): Simulations for Theorem 1 on the probability of support union recovery and maximal error under various settings of k and auxilliary tasks generated by Gaussian random variables. Let $\rho = \sqrt{\frac{\log(p+1)}{mn}}$. The x-axis is set by $T := \frac{mn}{\log(p+1)}$ varying from $\{1, \ldots, 15\}$ for support recovery. (c): Probability of support recovery for novel task. We let $\rho = \sqrt{\frac{\log(|J|+1)}{n}}$. The x-axis is set by $T := \frac{n}{\log(|J|+1)}$ varying from $\{10, \ldots, 25\}$. The dotted black line represents the baseline, which is the best result that either single or multitask learning obtained for any k. Our meta learning method is clearly superior.

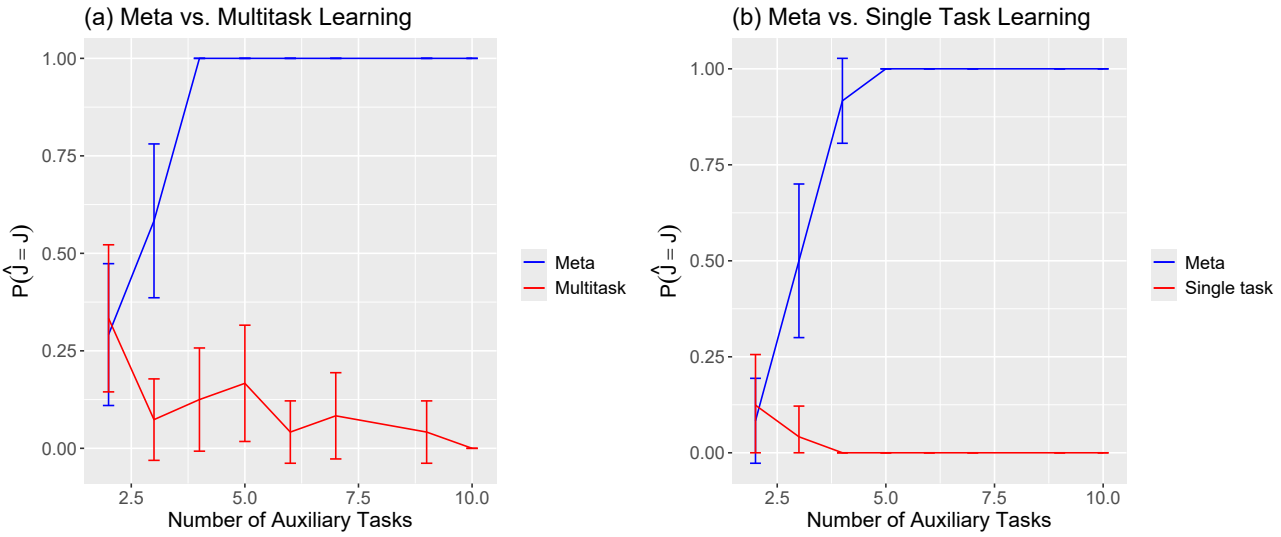

Figure 6: (a) Comparison showing the difference in the probability of recovering the correct support union of meta-learning versus multi-task learning. (b) Comparison showing the difference in the probability of recovering the correct support union of meta-learning versus single task sparse PCA. Y-axis : Probability of support union recovery. X-axis : Number of auxiliary tasks.

C.2 Comparison with Single Task Sparse PCA

In this section, we compare the meta-learning method against a single task learning method when applied to support union recovery under a Gaussian distribution setting. The method of generating the dataset and meta-learning is as described in Section 5.1. We let $p = 30$, $|J| = 5$, and fix $n = 3$. The values we choose for m are chosen from the set $\mathcal{M} \in \{2, 3, 4, 5, 6, 7, 9, 10\}$. We perform 24 replications and report the mean probability and the 95% confidence interval as the error bars around the mean. The results are plotted in fig. 6 (b).

Single Task Learning. For the single task learning setup there are m many objective functions being maximised simultaneously. Let $S^{(i)}$ be the covariance matrix associated with the task i. Then the i-th objective is to maximise,

$$\langle S^{(i)}, H\rangle - \rho\|H\|_{1,1} \text{ subject to } H \in \mathcal{F}^k. \tag{C.1}$$

Let $\hat{H}^{(i)}$ be the maximiser of this objective function. Let $\hat{J}^{(i)} = supp(H^{(i)})$ and take $\hat{J} = \bigcup_{i\in\mathcal{M}} \hat{J}^{(i)}$ to be the natural estimator for the support union. We apply $\sqrt{\frac{\log(p+1)}{n}}$ as the penalty for each of the m tasks.

Meta-Learning. For the meta-learning setup, we proceed according to Section 5.1 and recover the support union from the pooled covariance estimator. We apply $\sqrt{\frac{\log(p+1)}{mn}}$ as the penalty when m is the total number of tasks.

Single task learning of sparse PCA can be seen to be completely useless whenever the number of tasks get large and fails to recover the correct support union after a while. On the other hand, meta-learning improves steadily with the increasing number of tasks and the results coincide with our experiments in Section 5.1.

C.3 Real World Datasets

Brain Imaging Data. The 1000 functional connectomes dataset is from nitrc.org and contains resting-state functional magnetic resonance signals from 1128 subjects, 41 sites around the world, and $p = 157$ brain regions. This dataset consists of labs around the world that collected MRI from different subsets of subjects/persons. Each task is a different research lab with different magnetic resonance imaging scanner/equipment with different physical properties, and each brain region forms a covariate. Exploratory research like PCA is often performed to unveil important brain regions, which is the support. Although different labs might not have the same important brain regions (support) it is reasonable to believe

that there are some similarities across labs. This dataset is therefore a prime candidate for meta-learning. We used a set of independent samples to construct the true covariance matrix, and thus the true support union. We then constructed the pooled estimator using m auxiliary tasks, where $m \in \{1, 2, 3, 5, 8, 10, 12, 15, 18, 20, 25, 30, 35, 41\}$. Then we compared the recovered support with the true support to calculate the mean percentage and standard deviation of error in the recovered support. The error bars were obtained by creating a 95% confidence interval around the mean. As we clearly see in fig. 7(a) the probability of incurring an error decreases as we increase the number of tasks.

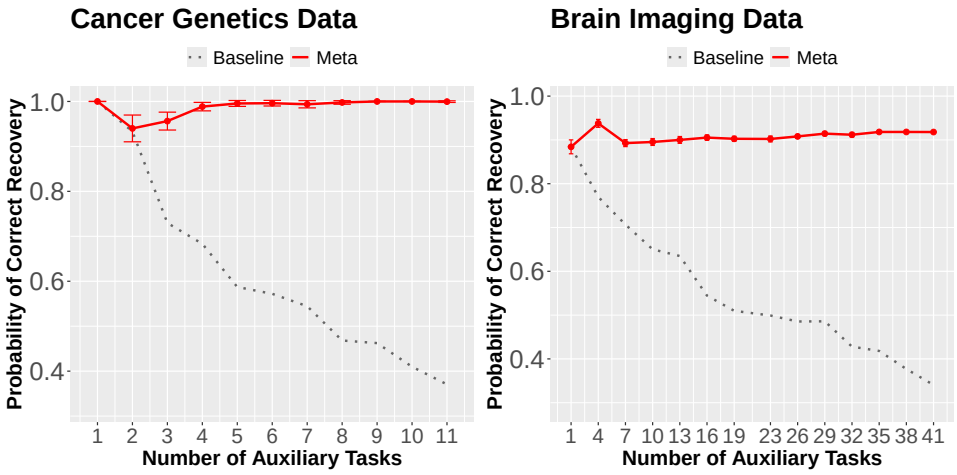

Figure 7: (a) Probability of error of support recovery for brain imaging dataset. We took an increasing number of auxiliary tasks $m \in \{1, 2, 3, 5, 8, \ldots, 30, 35, 41\}$. (b) Probability of error of support recovery for cancer genetics dataset. We took an increasing number of auxiliary tasks $m \in \{1, 2, \ldots, 10\}$. In both cases, the probability of error decreases as the number of auxiliary tasks increase. The dotted black line represents the baseline, which is the best result that either single or multitask learning obtained. Our meta learning method is clearly superior.

Recovered Broadmann areas in the support union. At a penalty of $\rho = 10\sqrt{\frac{\log(p+1)}{mn}}$, the recovered significant brain regions were: Broadmann areas 6, 17, and 44. Interestingly, these are widely reported in other fMRI studies, as shown in table 2.

Area	Found significant in
6	Rubia et al. (2000); Shuster and Lemieux (2005); Sylvester et al. (2003); Fincham et al. (2002)
17	Yan et al. (2005); Fulford et al. (2018); Wohlschläger et al. (2005); Tan et al. (2001)
44	Baron-Cohen et al. (2006); Friederici et al. (2000); Fincham et al. (2002); Binkofski et al. (1999)

Table 2: Recovered Broadmann areas in the support union

Cancer Genetics Data. The cancer genome atlas program dataset is available at gdc.cancer.gov and contains the genetic sequencing of 1576 patients. In cancer research, exploratory research is often performed to discover the important genes, which forms the support of PCA studies. In this dataset, we have 11 types of cancer, each one being an auxiliary task, and $p = 187$ genes, each one being a covariate. Different cancer types might not elicit the same important genes (support) but it is reasonable to believe that there are some similarities across cancer types. Therefore, this dataset is a prime candidate for meta-learning. We use a set of independent samples to construct the true covariance matrix and the true support union. We then constructed the pooled estimator using m auxiliary tasks where $m \in \{1, \ldots 11\}$. Then we compared the recovered support with the true support to calculate the mean percentage and the standard deviation of the error in recovered support. The error bars were obtained by empirically calculating the 95% confidence interval around the mean. As we can see in fig. 7(b) the probability of incurring an error decreases as we increase the number of tasks.

Recovered genes in the support union. At a penalty of $\rho = 10\sqrt{\frac{\log(p+1)}{mn}}$, the recovered significant genes were: PTP4A3, CLEC3B, E2F3, PSMA7. Interestingly, these genes are widely reported in other studies of cancer research, as shown in table 3.

Gene	Found significant in
PTP4A3	Huang et al. (2014); Den Hollander et al. (2016); Wang et al. (2019); Zimmerman et al. (2013)
CLEC3B	Sun et al. (2020); Zhu et al. (2019); Dai et al. (2019); Zhang et al. (2017)
E2F3	Feber et al. (2004); Cooper et al. (2006); Gao et al. (2017); Olsson et al. (2007)
PSMA7	Hu et al. (2009, 2008b,a); Du et al. (2009)

Table 3: Recovered genes in the support union