Meta Sparse Principal Component Analysis

Anonymous authors
Paper under double-blind review

Abstract

We study the meta-learning for support recovery (i.e., non-zero entries 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 & 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 & Lu, 2009; Nadler, 2008; Paul, 2007). In particular, it has been shown that the eigenvalues of the covariance matrix in high-dimensional data can exhibit unexpected behaviour (Kato, 1949; Bickel & 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 & 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:

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 PCA. 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 & 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 & Lampert, 2014; Amit & Meir, 2018; Denevi et al., 2018; Khodak et al., 2019a; Huang et al., 2020; Tripuraneni et al., 2020; Farid & Majumdar, 2021; Chen & Chen, 2022; Guan et al., 2022) and convergence rates in optimisation (Fallah et al., 2020; Finn et al., 2019; Khodak et al., 2019a; Gao & 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. 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 & Honorio, 2021), logistic regression (Xie & 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 $\Sigma^{(1)} \dots \Sigma^{(m)}$ be m many covariance matrices with their corresponding PC matrices $\Pi^{(1)} \dots, \Pi^{(m)}$, 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 nonzero entries in 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 3.2 and 3.3). Thus, our a 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 4.2). 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.

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.

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} cell 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 $x, x \in \mathbb{R}$	Σ	The true covariance A
$\Sigma^{(i)}$	Covariance of auxiliary task i	$\lambda_k(A)$	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 I and columns J		

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.

Definition 2.1. *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 $\left\{\Sigma^{(i)}\right\}$ distribution of size m and parameter σ if*

1. *$X_j^{(i)}$ are independent random variables.*
2. *For each i and all j, $X_j^{(i)}$ is distributed according to a distribution P_i such that $\mathbb{E}[X_j^{(i)}] = 0$, $\mathrm{Cov}\left(X_j^{(i)}\right) = \Sigma^{(i)}$.*
3. *$\frac{X_{l,j}^{(i)}}{\Sigma_{l,l}^{(i)}}$ is sub-Gaussian with parameter σ, $\forall i \in [m]$ and $j \in [n^{(i)}]$ and $l \in [p]$.*

Remark 2.2. *We would like to point out that such definitions are standard in the literature and have been used extensively in prior work (see for instance Zhang et al. (2021)) for modelling sub-Gaussian random variables with randomised covariance matrices under the meta-learning setup. We have deferred the definition of sub-Gaussian and sub-exponential distributions to Appendix A.*

Problem Formulation. In this paper, we have focused 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 λ_i are the diagonal entries 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. With random matrices $R^{(i)}$ and $D^{(i)}$, we generate a sequence of random matrices $\Sigma^{(1)}, \ldots, \Sigma^{(m)} \in \mathbb{R}^p$. For each $i \in [m]$, the covariance matrix $\Sigma^{(i)}$ corresponding to the i-th auxiliary task is generated as,

$$\Sigma^{(i)} = R^{(i)} U(\Lambda + D^{(i)}) U^T \left(R^{(i)}\right)^T. \tag{2.3}$$

Referring to Definition 2.1, 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. Recall from Theorem 1 in (Overton & Womersley, 1992) that PCA can be interpreted as a covariance maximisation technique maximising the predictive covariance $\operatorname{tr}(\operatorname{Cov}(X, HX|H))$. Observe that,

$$\operatorname{tr}(\operatorname{Cov}(X, HX|H)) = \langle \Sigma, H \rangle,$$

for $\operatorname{Cov}(X) = \Sigma$ and a rank k orthogonal matrix H. 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

$$\langle \Sigma, H \rangle - \rho \|H\|_{1,1} \tag{2.5}$$

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 2.3. *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 Equation (2.2) that Π is a positive semi-definite matrix. Consequently, the i-th row/column of Π is 0 is non-zero if and only if $\Pi_{i,i} \neq 0$. 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 $|J| \ll p$.*

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 $\rho > 0$

$$\begin{aligned} \hat{\Pi} := \arg\sup_{H} \sum_{i=1}^{m} T_i \left\{ \langle S^{(i)}, H \rangle - \rho \|H\|_{1,1} \right\} \\ \text{subject to } H \in \mathcal{F}^k, \text{ where,} \\ \mathcal{F}^k := \{ M \in \mathbb{R}^{p \times p} : 0 \preceq M \preceq \mathbb{I}_p \text{ and } \operatorname{tr}(M) = k \}, \end{aligned} \tag{2.6}$$

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 (2.6) as

$$\hat{\pi} = \arg\sup_{H} \langle S, H \rangle - \rho \|H\|_{1,1} \text{ subject to } H \in \mathcal{F}^k, \tag{2.7}$$

where S is as defined in Equation (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).

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

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

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 a 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 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 3.1. *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 $\Sigma^{(i)}$,*

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

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

The proof of this lemma is presented in Appendix B.1. Recall from Equation (2.3) that we defined each auxiliary task $\Sigma^{(i)}$ in terms of random matrices $R^{(i)}$ and random matrices $D^{(i)}$. 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 $\left\|\Sigma^{(i)}-\Sigma\right\|_{\infty,\infty}$. We achieve it by controlling the variance and support of the perturbations $R^{(i)}$ and $D^{(i)}$. To that end, we make the following assumption.

Assumption 3. *Assume that for each $1\leq i\leq m$, $R^{(i)}$ and $D^{(i)}$ in Equation (2.3) are random matrices that fulfill the following conditions:*

1. *Each $R^{(i)}$ is i.i.d. with,*

$$\mathbb{E}\left[R^{(i)}\otimes R^{(i)}\right]=\mathbb{I}_{p^2} \tag{3.1}$$

Furthermore, we assume that

$$\|R^{(i)}-\mathbb{I}_p\|_{\infty,1}\leq c\sqrt{g_p}-1=:\mathbb{C}_R \tag{3.2}$$

almost surely for some g_p bounded from above by $2p$, and c is a positive universal constant. Without losing generality, we put $c=1$. The proofs would proceed very similarly in all other cases.

2. *Each $D^{(i)}$ is i.i.d. with $\mathbb{E}[D^{(i)}]=0$. Moreover, for some constant $L>0$,*

$$\lambda_1(D^{(i)})<L \text{ almost surely.} \tag{3.3}$$

White noise additive models have been used extensively in prior literature to study precision and covariance matrices (Wang & Honorio, 2021; Zhang et al., 2021). It is noteworthy that under our assumptions, we can recover additive models from Equation (2.3) when $R^{(i)}=\mathbb{I}_p$ for every i. However, $R^{(i)}$ suffers from curse of dimensionality, and consequently Equation (3.2) is required to control the rate of growth of $R^{(i)}$. We note that Assumption 3.2 is weak, but still sufficient to produce a minimax optimal sample complexity. However, it is not a necessary assumption. In this light, a necessary condition on the matrices $R^{(i)}$ is by itself an interesting question and is part of a future work. Our final technical assumption is given below.

Assumption 4. *If Σ is a covariance satisfying Assumption 2. Then,*

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

This condition previously appears in (Lei & Vu, 2015) and recovers the conditions by Amini & Wainwright (2008) when $\Sigma_{J^c,J}=0$. It is also similar to the irrepresentability condition in the seminal papers (Zhao & Yu, 2006; Meinshausen & Bühlmann, 2006; Meinshausen & 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:

$$\lambda_{diff}:=\lambda_k-\lambda_{k+1}\qquad\lambda^\dagger:=2\left(\mathbb{C}_R+1\right)^2\left(\lambda_1(\Sigma)+L\right),$$
$$\rho_1:=4\lambda^\dagger\sqrt{\frac{\log(p)}{m}}\qquad\rho_2:=16\sqrt{2\sigma^4\lambda^\dagger\frac{\log\left(p+1\right)}{nm}}.$$

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 3.2 (Support Union Recovery)**.** *Let Assumption 3 and Assumption 4 hold. Then there exists a large enough constant $\mathbb{C}_\pi > 0$ such that, whenever*

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

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. $supp(\hat{\Pi}) = J$,
2. $\left\|\hat{\Pi} - \Pi\right\|_{\infty,\infty} \leq \mathbb{C}_\pi \rho$.

In other words, Theorem 3.2 gives us our required sample complexity for sign consistency and error bound. Recalling $\mathbb{C}_R$ from Equation (3.2) we find that $\lambda^\dagger = (\lambda(\Sigma) + L) \times g_p$. 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 $1 - \frac{2}{p^2} - \frac{1}{2(p+1)^2}$, we recover the true support. Our next objective is to provide a matching lower bound for the previous per-task sample complexity.

Theorem 3.3. *In the setup of Theorem 3.2, there exists an estimation problem such that, for all p large enough any estimator $\hat{J}$ of the support union J, if $n \leq \frac{g_p|J|\log p - |J|\log|J|}{2m}$., then*

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

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

Remark 3.4. *Observe that, by setting $n = 1$, we also obtain a corresponding error lower bound whenever $m \leq \frac{g_p|J|\log|p| - |J|\log|J|}{2}$ which matches the upper bound up to a factor of g_p.*

3.1 Sketch of Proof for Theorem 3.2

The complete proof of this theorem may be found at Appendix B.2 and we only make the outline here. The proof of Theorem 3.2 is via primal-dual witness method. It starts by writing down the objective function Equation (2.7).

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

Define the subspace

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

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

$$\begin{aligned}\max_{\substack{H\in\mathcal{F}^k\\ Z\in\mathbb{B}}} \langle S, H\rangle - \rho\langle H, Z + \mathbb{I}_p\rangle &\equiv \max_{\substack{H\in\mathcal{F}^k\\ Z\in\mathbb{B}}} \langle S, H\rangle - \rho\langle H, Z\rangle - \rho k \\ &\equiv \max_{\substack{H\in\mathcal{F}^k\\ Z\in\mathbb{B}}} \langle H, S - \rho Z\rangle, \end{aligned} \tag{3.4}$$

where the last step follows from the fact that k is a constant. To maximise Equation (3.4), we need to find solutions $(\hat{H}, \hat{Z})$ such that Karush-Kuhn-Tucker (KKT) (Boyd et al., 2004) conditions

$$\hat{Z}_{i,j} = sign(\hat{H}_{i,j}), \forall i \neq j \text{ and } \hat{H}_{i,j} \neq 0, \tag{3.5}$$
$$\hat{Z}_{i,j} \in [-1, 1], \forall i \neq j \text{ and } \hat{H}_{i,j} = 0, \tag{3.6}$$
$$\hat{H} = \arg\max\langle S - \rho\hat{Z}, H\rangle, \tag{3.7}$$

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

$$\max_{H \in \mathcal{F}^k, supp(H) \subseteq J} \langle S, H \rangle - \rho \|H\|_{1,1}. \tag{3.8}$$

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 A.6 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 Equation (3.4). It is noteworthy that this construction is similar to that in Theorem 1 in (Lei & Vu, 2015).

$$\hat{H} = \tilde{H},$$

$$\hat{Z}_{J,J} = \tilde{Z}_{J,J}, \tag{3.9}$$

$$\hat{Z}_{ij} = \frac{1}{\rho}\left\{S_{ij} - \langle Q_{i,*}, \Sigma_{J,j} \rangle\right\}, (i,j) \in J \times J^c, \tag{3.10}$$

$$\hat{Z}_{ij} = \frac{1}{\rho}(S - \Sigma)_{ij}, (i,j) \in J^c \times J^c, i \neq j. \tag{3.11}$$

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 3.1.
2. Proving that the proposed solution satisfies the KKT conditions (3.5), (3.6), and (3.7) 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.

4 Support 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 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

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

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

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

where $supp(\hat{\Pi}) = J$ from Theorem 3.2 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 4.1. *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 4.1 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 $\lambda^{(m+1)}_{diff} := \lambda^{(m+1)}_k - \lambda^{(m+1)}_{k+1}$. We make the following technical assumption on the entries of $\Sigma^{(m+1)}_{J,J}$.

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

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

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

$$\begin{aligned} \mathbb{C}_{n1} &:= \frac{1}{32\lambda_1\left(\Sigma^{(m+1)}\sigma^2\right)}, \\ \mathbb{C}_{n2} &:= \frac{4|J^{(m+1)}|\left(1 + \frac{8\lambda^{(m+1)}_1}{\lambda^{(m+1)}_k - \lambda^{(m+1)}_{k+1}}\right)}{\lambda^{(m+1)}_k - \lambda^{(m+1)}_{k+1}}, \\ \mathbb{C}_{n3} &:= \frac{\left(\lambda^{(m+1)}_k - \lambda^{(m+1)}_{k+1}\right)\min_{j\in\left[|J^{(m+1)}|\right]} \Pi^{(m+1)}_{j,j}}{4|J^{(m+1)}|}. \end{aligned}$$

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

Theorem 4.2 (Novel Task Support Recovery)**.** *Let $\Sigma^{(m+1)}$ be the covariance matrix corresponding to the novel task satisfying Assumption 5 and J be the recovered support union. Furthermore, assume that,*

$$n^{\frac{1}{3}} > \max\{\mathbb{C}_{n1}, \mathbb{C}_{n2}, \mathbb{C}_{n3}\} \text{ and,}$$
$$\frac{1}{n^{1/3}\alpha^{(m+1)}} < \rho < \frac{\lambda_k^{(m+1)} - \lambda_{k+1}^{(m+1)}}{4|J^{(m+1)}|\left(1 + \frac{8\lambda_1^{(m+1)}}{\lambda_k^{(m+1)} - \lambda_{k+1}^{(m+1)}}\right)},$$

then, with probability at least $1 - \frac{|J|(|J|+1)}{2} e^{-\frac{n^{1/3}}{8\mathbb{C}\sigma^2\left(2\lambda_1(\Sigma^{(m+1)})\right)}}$, the maximiser $\hat{\Pi}^{(m+1)}$ to the objective function Equation (2.8) satisfies, supp$(\hat{\Pi}^{(m+1)}) = J^{(m+1)}$.

The proof of this theorem is deferred to Appendix B.5.

5 Experiments

In this section, we empirically validate our theoretical claims in Sections 3 and 4.

Synthetic Experiments. For all the experiments in this subsection, we let $p = 50$, $k = 5$ and $|J| = 5$ and perform 100 repetitions of each setting. We first consider the setting of different sample sizes. We choose $n \in \{3, 5, 7, 9\}$ and use $\rho = \sqrt{\frac{\log(p+1)}{mn}}$ for all pairs of (m, n). The number of tasks m is rescaled to T defined by $\frac{mn}{\log(p+1)}$. For each $j \in [n]$, $X_j^{(i)}$ is generated from a Gaussian distribution with mean a p-dimensional 0 vector and covariance matrix $\Sigma^{(i)}$. Figure 1 depicts the outcome of our experiments. For different choices of n, the graphs overlap each other perfectly (both $\mathbb{P}(\hat{J} = J)$ and $\|\hat{\Pi} - \Pi\|_{\infty,\infty}$). Then, we took the recovered support union and derived the probability of support recovery for a novel task. We perform the experiments when there are $2, 3$, or 4 extra zeros and all of those have greater than 95% probability of accurately identifying the extra zeros.

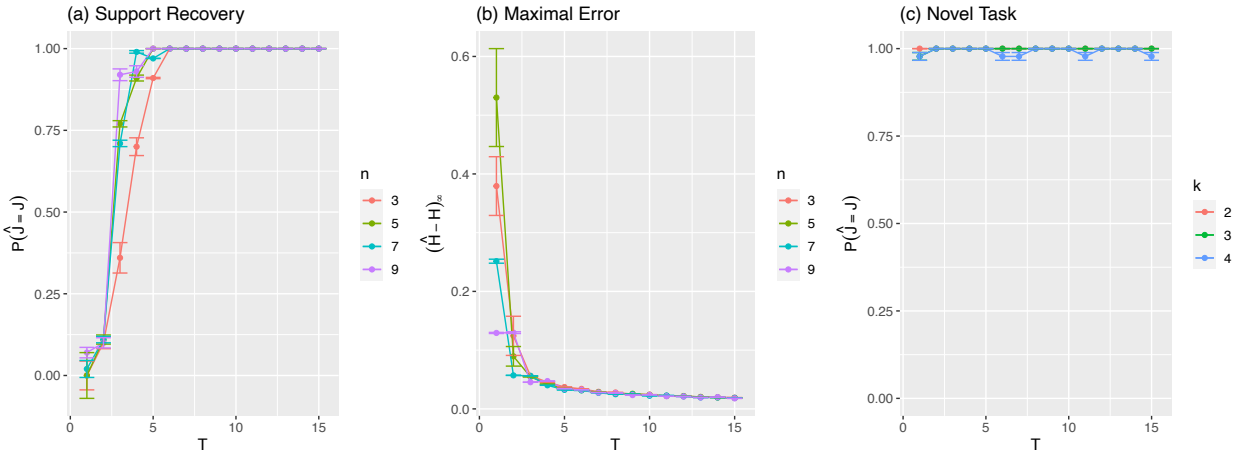

Figure 1: (a), (b): Simulations for Theorem 3.2 on the probability of support union recovery and maximal error under various settings of n. $\rho = \sqrt{\frac{\log(p+1)}{mn}}$. The x-axis is set by $T := \frac{mn}{\log(p+1)}$ varying from $\{1, \ldots, 15\}$ for support union recovery. (c): Probability of support recovery for novel task. $\rho = \sqrt{\frac{\log(|J|+1)}{n}}$. The x-axis is set by $T := \frac{n}{\log(|J|+1)}$ varying from $\{10, \ldots, 25\}$

In Appendix C.1 we provide more details and show extra results for the uniform distribution as well as a mixture of sub-Gaussian distributions. Furthermore, we conduct comparisons to show the superiority of meta-learning over multi-task learning Appendix C.2 and over single task learning in Appendix C.3.

Real World Experiments. In Appendix C.4, we show that our method obtains anywhere between $85 - 95\%$ accuracy in a real world brain-imaging dataset and similar results in a cancer dataset. We also found consistent results 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.

References

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

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

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

Simon Baron-Cohen, Howard Ring, Xavier Chitnis, Sally Wheelwright, Lloyd Gregory, Steve Williams, Mick Brammer, and Ed Bullmore. fMRI of parents of children with asperger syndrome: a pilot study. *Brain and cognition*, 61(1):122–130, 2006.

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

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

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Lisha Chen and Tianyi Chen. Is Bayesian model-agnostic meta learning better than model-agnostic meta learning, provably? In *Artificial Intelligence and Statistics*, volume 151, pp. 1733–1774. PMLR, 2022.

Colin S Cooper, Andrew G Nicholson, Christopher Foster, Andrew Dodson, Sandra Edwards, Anne Fletcher, Toby Roe, Jeremy Clark, Anupam Joshi, Andrew Norman, et al. Nuclear overexpression of the e2f3 transcription factor in human lung cancer. *Lung Cancer*, 54(2):155–162, 2006.

Wenjuan Dai, Yilin Wang, Tianxiao Yang, Jing Wang, Weicheng Wu, and Jianxin Gu. Downregulation of exosomal clec3b in hepatocellular carcinoma promotes metastasis and angiogenesis via ampk and vegf signals. *Cell communication and signaling*, 17(1):1–17, 2019.

Petra Den Hollander, Kathryn Rawls, Anna Tsimelzon, Jonathan Shepherd, Abhijit Mazumdar, Jamal Hill, Suzanne AW Fuqua, Jenny C Chang, C Kent Osborne, Susan G Hilsenbeck, et al. Phosphatase ptp4a3 promotes triple-negative breast cancer growth and predicts poor patient survivalptp4a3 is critical for triple-negative breast cancer growth. *Cancer research*, 76(7):1942–1953, 2016.

Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Learning to learn around a common mean. In *Neural Information Processing Systems*, volume 31, 2018.

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

Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse PCA. In *2014 IEEE International Symposium on Information Theory*, pp. 2197–2201, 2014. doi: 10.1109/ISIT.2014.6875223.

Hongyan Du, Xiang Huang, Suihai Wang, Yingsong Wu, Weiwen Xu, and Ming Li. Psma7, a potential biomarker of diseases. *Protein and peptide letters*, 16(5):486–489, 2009.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *Artificial Intelligence and Statistics*, pp. 1082–1092, 2020.

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

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

Andrew Feber, Jeremy Clark, Graham Goodwin, Andrew R Dodson, Paul H Smith, Anne Fletcher, Sandra Edwards, Penny Flohr, Alison Falconer, Toby Roe, et al. Amplification and overexpression of e2f3 in human bladder cancer. *Oncogene*, 23(8):1627–1630, 2004.

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

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International Conference on Machine Learning*, pp. 1920–1930, 2019.

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

Jon Fulford, Fraser Milton, David Salas, Alicia Smith, Amber Simler, Crawford Winlove, and Adam Zeman. The neural correlates of visual imagery vividness–an fMRI study and literature review. *Cortex*, 105:26–40, 2018.

Katelyn Gao and Ozan Sener. Modeling and optimization trade-off in meta-learning. In *Neural Information Processing Systems*, volume 33, pp. 11154–11165, 2020.

Yanping Gao, Bing Feng, Lu Lu, Siqi Han, Xiaoyuan Chu, Longbang Chen, and Rui Wang. Mirnas and e2f3: a complex network of reciprocal regulations in human cancers. *Oncotarget*, 8(36):60624, 2017.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical Bayes. *International Conference on Learning Representations*, 2018.

Jiechao Guan, Yong Liu, and Zhiwu Lu. Fine-grained analysis of stability and generalization for modern meta learning algorithms. In *Neural Information Processing Systems*, volume 35, pp. 18487–18500, 2022.

Jean Honorio and Tommi Jaakkola. Tight bounds for the expected risk of linear classifiers and PAC-Bayes finite-sample guarantees. In *Artificial Intelligence and Statistics*, pp. 384–392. PMLR, 2014.

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

Xiao-Tong Hu, Wei Chen, Da Wang, Qing-Lan Shi, Fu-Biao Zhang, Yong-Qiang Liao, Mei Jin, and Chao He. 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, 2008a.

Xiao-Tong Hu, Wei Chen, Da Wang, Qing-Lan Shi, Fu-Biao Zhang, Yong-Qiang Liao, Mei Jin, and Chao He. 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, 2008b.

Xiao-Tong Hu, Wei Chen, Fu-Biao Zhang, Qing-Lan Shi, Jian-Bin Hu, Shu-Min Geng, and Chao He. Depletion of the proteasome subunit psma7 inhibits colorectal cancer cell tumorigenicity and migration. *Oncology reports*, 22(5):1247–1252, 2009.

Yimin Huang, Weiran Huang, Liang Li, and Zhenguo Li. Meta-learning PAC-Bayes priors in model averaging. *Association for the Advancement of Artificial Intelligence Conference*, pp. 4198–4205, 2020.

Yu-Han Huang, Abdul Qader O Al-Aidaroos, Hiu-Fung Yuen, Shu-Dong Zhang, Han-Ming Shen, Ewelina Rozycka, Cian M McCrudden, Vinay Tergaonkar, Abhishek Gupta, You Bin Lin, et al. A role of autophagy in ptp4a3-driven cancer progression. *Autophagy*, 10(10):1787–1800, 2014.

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

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

Mikhail Khodak, Maria-Florina Balcan, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. *International Conference on Machine Learning*, 36:424–433, 2019a.

Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, 32, 2019b.

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

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

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

Chin-Teng Lin, Yu-Chieh Chen, Teng-Yi Huang, Tien-Ting Chiu, Li-Wei Ko, Sheng-Fu Liang, Hung-Yi Hsieh, Shang-Hwa Hsu, and Jeng-Ren Duann. 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, 2008.

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

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

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

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

AY Olsson, A Feber, S Edwards, R Te Poele, I Giddings, S Merson, and CS Cooper. Role of e2f3 expression in modulating cellular proliferation rate in human bladder and prostate cancer cells. *Oncogene*, 26(7): 1028–1037, 2007.

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

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

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pp. 1617–1642, 2007.

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

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

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. *International Conference on Learning Representations*, 2017.

Katya Rubia, Stephan Overmeyer, El Taylor, Michael Brammer, SCR Williams, Andy Simmons, C Andrew, and ET Bullmore. Functional frontalisation with age: mapping neurodevelopmental trajectories with fMRI. *Neuroscience & Biobehavioral Reviews*, 24(1):13–19, 2000.

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

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Neural Information Processing Systems*, pp. 4077–4087, 2017.

Jiaxing Sun, Tian Xie, Muhammad Jamal, Zhenbo Tu, Xinran Li, Yingjie Wu, Jingyuan Li, Qiuping Zhang, and Xiaoxing Huang. Clec3b as a potential diagnostic and prognostic biomarker in lung cancer and association with the immune microenvironment. *Cancer cell international*, 20(1):1–15, 2020.

Yaoqi Sun, Liang Li, Liang Zheng, Ji Hu, Wenchao Li, Yatong Jiang, and Chenggang Yan. Image classification base on PCA of multi-view deep representation. *Journal of Visual Communication and Image Representation*, 62:253–258, 2019.

Ching-Yune C Sylvester, Tor D Wager, Steven C Lacey, Luis Hernandez, Thomas E Nichols, Edward E Smith, and John Jonides. Switching attention and resolving interference: fMRI measures of executive functions. *Neuropsychologia*, 41(3):357–370, 2003.

Li Hai Tan, Ching-Mei Feng, Peter T Fox, and Jia-Hong Gao. An fMRI study with written chinese. *Neuroreport*, 12(1):83–88, 2001.

Malik Tiomoko, Romain Couillet, and Frédéric Pascal. PCA-based multi task learning: a random matrix approach. *International Conference on Machine Learning*, pp. 34280–34300, 2023.

Nilesh Tripuraneni, Michael I. Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Neural Information Processing Systems*, 33, 2020.

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

Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, pp. 210–268. Cambridge University Press, 2012.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Neural Information Processing Systems*, pp. 3630–3638, 2016.

Vincent Q Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. *Advances in neural information processing systems*, 26, 2013.

Haoxiang Wang, Han Zhao, and Bo Li. Bridging multi-task learning and meta-learning: Towards efficient training and effective adaptation. In *International Conference on Machine Learning*, pp. 10991–11002. PMLR, 2021.

Lina Wang, Ting Yu, Wei Li, Mengmeng Li, Qianfei Zuo, Quanming Zou, and Bin Xiao. The mir-29c-kiaa1199 axis regulates gastric cancer migration by binding with wbp11 and ptp4a3. *Oncogene*, 38(17):3134–3150, 2019.

Zhanyu Wang and Jean Honorio. The sample complexity of meta sparse regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 2323–2331. PMLR, 2021.

Afra M Wohlschläger, Karsten Specht, C Lie, Hartmut Mohlberg, Andreas Wohlschläger, Kay Bente, Uwe Pietrzyk, Tony Stöcker, Karl Zilles, Katrin Amunts, et al. Linking retinotopic fMRI mapping and anatomical probability maps of human occipital areas v1 and v2. *Neuroimage*, 26(1):73–82, 2005.

Jiawei Wu, Wenhan Xiong, and William Yang Wang. 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*, pp. 4354–4364, 2019.

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

Ikko Yamane, Florian Yger, Maxime Berar, and Masashi Sugiyama. Multitask principal component analysis. In *Asian Conference on Machine Learning*, pp. 302–317. PMLR, 2016.

Bin Yan, Ke Li, Jianyang Xu, Wei Wang, Kuncheng Li, Hua Liu, Baoci Shan, and Xiaowei Tang. Acupoint-specific fMRI patterns in human brain. *Neuroscience letters*, 383(3):236–240, 2005.

Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In *Neural Information Processing Systems*, volume 31, 2018.

Shipeng Yu, Kai Yu, Volker Tresp, Hans-Peter Kriegel, and Mingrui Wu. Supervised probabilistic principal component analysis. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 464–473, 2006.

Qian Zhang, Yilin Zheng, and Jean Honorio. Meta learning for support recovery in high-dimensional precision matrix estimation. In *International Conference on Machine Learning*, pp. 12642–12652. PMLR, 2021.

Xin Zhang, Jin-Xiang Wan, Zun-Ping Ke, Feng Wang, Hai-Xia Chai, and Jia-Qiang Liu. Tmem88, ccl14 and clec3b as prognostic biomarkers for prognosis and palindromia of human hepatocellular carcinoma. *Tumor Biology*, 39(7):1010428317708900, 2017.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Hao Zheng, Zhanlei Yang, Liwei Qiao, Jianping Li, and Wenju Liu. Attribute knowledge integration for speech recognition based on multi-task learning neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Hui-Fang Zhu, Xu-Hui Zhang, Chuan-Sha Gu, Yan Zhong, Ting Long, Yi-Dan Ma, Zhi-Yan Hu, Zu-Guo Li, and Xiao-Yan Wang. Cancer-associated fibroblasts promote colorectal cancer progression by secreting clec3b. *Cancer Biology & Therapy*, 20(7):967–978, 2019.

Mark W Zimmerman, Gregg E Homanics, and John S Lazo. Targeted deletion of the metastasis-associated phosphatase ptp4a3 (prl-3) suppresses murine colon cancer. *PloS one*, 8(3):e58300, 2013.