Received: 14 May 2020 | Accepted: 12 July 2021

DOI: 10.1111/rssb.12479

ORIGINAL ARTICLE

Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality

Sai Li[1] | T. Tony Cai[2] | Hongzhe Li[1]

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[2]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Correspondence
Hongzhe Li, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.
Email: hongzhe@upenn.edu

Funding information
NIH, Grant/Award Number: GM129781 and GM123056; NSF, Grant/Award Number: DMS-1712735

Abstract

This paper considers estimation and prediction of a high-dimensional linear regression in the setting of transfer learning where, in addition to observations from the target model, auxiliary samples from different but possibly related regression models are available. When the set of informative auxiliary studies is known, an estimator and a predictor are proposed and their optimality is established. The optimal rates of convergence for prediction and estimation are faster than the corresponding rates without using the auxiliary samples. This implies that knowledge from the informative auxiliary samples can be transferred to improve the learning performance of the target problem. When the set of informative auxiliary samples is unknown, we propose a data-driven procedure for transfer learning, called Trans-Lasso, and show its robustness to non-informative auxiliary samples and its efficiency in knowledge transfer. The proposed procedures are demonstrated in numerical studies and are applied to a dataset concerning the associations among gene expressions. It is shown that Trans-Lasso leads to improved performance in gene expression prediction in a target tissue by incorporating data from multiple different tissues as auxiliary samples.

KEYWORDS

auxiliary studies, data aggregation, domain adaptation, GTEx data, multitask learning, Q-aggregation

接收：2021年5月14日 2020 | 接受：2021年7月12日

DOI: 10.1111/rssb.12479

ORIGINAL ARTICLE

高维线性回归的迁移学习：预测、估计和极大极小最优性

Sai Li[1] | T. Tony Cai[2] | Hongzhe Li1

1生物统计学、流行病学与信息学系，佩雷尔曼医学院，宾夕法尼亚大学，费城，宾夕法尼亚州，美国 2统计系，沃顿商学院，宾夕法尼亚大学，费城，宾夕法尼亚州，美国

通讯Hongzhe Li，生物统计学、流行病学与信息学系，佩雷尔曼医学院，宾夕法尼亚大学，费城，宾夕法尼亚州19104，美国。邮箱：hongzhe@upenn.edu

资助信息NIH，资助编号：GM129781 和 GM123056；NSF，资助编号：DMS- 1712735

摘要

本文考虑了迁移学习设置下高维线性回归的估计和预测，其中除了目标模型的观测值外，还有来自不同但可能相关的回归模型的辅助样本。当信息辅助研究集已知时，提出了一个估计器和预测器，并证明了其最优性。预测和估计的最优收敛率比不使用辅助样本的相应速率更快。这意味着信息辅助样本中的知识可以迁移以提高目标问题的学习性能。当信息辅助样本集未知时，我们提出了一种数据驱动的迁移学习过程，称为 Trans- Lasso，并展示了其对非信息辅助样本的鲁棒性和知识迁移效率。所提出的方法在数值研究中得到了验证，并应用于一个关于基因表达关联的数据集。结果表明，通过将来自多个不同组织的辅助样本数据结合起来，Trans- Lasso 能够在目标组织中提高基因表达预测的性能。

KEYWORDS

辅助研究, 数据聚合, 领域自适应, GTEx 数据, 多任务学习, Q- 聚合

1 | INTRODUCTION

Modern scientific research is characterized by massive and diverse datasets. It is of significant interest to integrate different datasets to make a more accurate prediction and statistical inference. Given a target problem to solve, transfer learning (Torrey & Shavlik, 2010) aims at transferring the knowledge from different but related samples to improve the learning performance of the target problem. A typical example of transfer learning is that one can improve the accuracy of recognizing cars by using not only the labelled data for cars but some labelled data for trucks (Weiss et al., 2016). Besides classification, another important transfer learning problem is linear regressions with auxiliary samples. In biomedical studies, some clinical or biological outcomes are hard to obtain due to ethical or cost issues, in which case transfer learning can be leveraged to boost the prediction and estimation performance by effectively utilizing information from related studies.

Transfer learning has been applied to problems in medical and biological studies, including predictions of protein localization (Mei et al., 2011), biological imaging diagnosis (Shin et al., 2016), drug sensitivity prediction (Turki et al., 2017) and integrative analysis of ‘multi-omics’ data, see, for instance, Sun and Hu (2016), Hu et al. (2019), and Wang et al. (2019). It has also been applied to natural language processing (Daumé III, 2007) and recommendation systems (Pan & Yang, 2013) in machine learning. The application that motivated the present paper is the integration of the gene expression measurements in different issues for understanding the gene regulations using the Genotype-Tissue Expression (GTEx) data (https://gtexportal.org/). These datasets are always high dimensional with relatively small sample sizes. When studying the gene regulation relationships of a specific tissue or cell type, it is possible to incorporate information from other tissues to enhance the learning accuracy. This motivates us to consider transfer learning in high-dimensional linear regression.

1.1 | Transfer learning in high-dimensional linear regression

Regression analysis is one of the most widely used statistical methods to understand the association of an outcome with a set of covariates. In many modern applications, the dimension of the covariates is usually very high as compared to the sample size. Typical examples include genome-wide association and gene expression studies. In this paper, we consider transfer learning in high-dimensional linear models. Formally, the target model can be written as

$$y_i^{(0)} = (x_i^{(0)})^\intercal \beta + \epsilon_i^{(0)}, \quad i = 1, \ldots, n_0, \tag{1}$$

where $((x_i^{(0)})^\intercal, y_i^{(0)})$, $i = 1, \ldots, n_0$, are independent samples, $\beta \in \mathbb{R}^p$ is the coefficient vector of interest, and $\epsilon_i^{(0)}$, $i = 1, \ldots, n_0$ are independently distributed random noises with $\mathbb{E}[\epsilon_i^{(0)} | x_i^{(0)}] = 0$. In the high-dimensional regime, where p can be larger and much larger than n_0, β is often assumed to be sparse such that the number of nonzero elements of β, denoted by s, is much smaller than p.

In the context of transfer learning, we observe additional samples from K auxiliary studies, That is, we observe $((x_i^{(k)})^\intercal, y_i^{(k)})$ generated from the auxiliary model

$$y_i^{(k)} = (x_i^{(k)})^\intercal w^{(k)} + \epsilon_i^{(k)}, \quad i = 1, \ldots, n_k, \ k = 1, \ldots, K, \tag{2}$$

1 | 引言

现代科学研究的特点是海量且多样的数据集。整合不同数据集以进行更准确的预测和统计推断具有重要意义。给定一个要解决的问题，迁移学习（Torrey & Shavlik，2010）旨在将知识从不同但相关的样本中迁移过来，以提高目标问题的学习性能。迁移学习的一个典型例子是，通过使用不仅限于汽车标签数据，还包括一些卡车标签数据（Weiss 等人，2016），来提高识别汽车准确率。除了分类，另一个重要的迁移学习问题是辅助样本的线性回归。在生物医学研究中，由于伦理或成本问题，一些临床或生物结果难以获得，在这种情况下，迁移学习可以通过有效利用相关研究的信息来提高预测和估计性能。

迁移学习已应用于医学和生物医学研究中的问题，包括蛋白质定位预测（Mei 等人，2011）、生物成像诊断（Shin 等人，2016）、药物敏感性预测（Turki 等人，2017）以及‘多组学’数据的整合分析，例如 Sun 和 Hu（2016）、Hu 等人（2019）和 Wang 等人（2019）。它还应用于机器学习中的自然语言处理（Daumé III，2007）和推荐系统（ Pan 和 Yang，2013）。本文所研究的应用是利用基因型-组织表达（GTEx）数据（ https://gtexp ortal.org/）整合不同问题中的基因表达测量值，以理解基因调控。这些数据集通常是高维的，且样本量相对较小。在研究特定组织或细胞类型的基因调控关系时，可以结合其他组织的信息来提高学习准确率。这促使我们考虑在高维线性回归中应用迁移学习。

1.1 | 高维线性回归中的迁移学习

回归分析是最广泛使用的统计方法之一，用于理解结果与协变量集之间的关联。在许多现代应用中，协变量的维度通常远高于样本量。典型例子包括全基因组关联分析和基因表达研究。在本文中，我们考虑高维线性模型中的迁移学习。形式上，目标模型可以写为

$$y_i^{(0)} = (x_i^{(0)})^\intercal \beta + \epsilon_i^{(0)}, \quad i = 1, \ldots, n_0, \tag{1}$$

其中 $((x^{(0)}{}_i)^\intercal, y^{(0)}{}_i), i = 1, \ldots, n_0$, 是独立样本， $\beta \in \mathbb{R}^p$ 是感兴趣的系数向量，而 $\epsilon^{(0)}{}_i, i = 1, \ldots, n_0$ 是独立分布的随机噪声，有 $\mathbb{E}[\epsilon_i^{(0)} | x_i^{(0)}] = 0$. 在高维情况下，其中 p 可以大于且远大于 n_0 β，通常假设是稀疏的，使得 β 的非零元素数量，记为 s，远小于 p。

在迁移学习的背景下，我们从 K 项辅助研究中观察到额外的样本，也就是说，我们观察到由辅助模型生成的 $((x_i^{(k)})^\intercal, y_i^{(k)})$

$$y_i^{(k)} = (x_i^{(k)})^\intercal w^{(k)} + \epsilon_i^{(k)}, \quad i = 1, \ldots, n_k, \ k = 1, \ldots, K, \tag{2}$$

where $w^{(k)} \in \mathbb{R}^p$ is the regression vector for the kth study, and $\epsilon_i^{(k)}$ is the random noise such that $\mathbb{E}[\epsilon_i^{(k)}|x_i^{(k)}] = 0$. The regression coefficients $w^{(k)}$ are unknown and different from our target β in general. The number of auxiliary studies, K, is allowed to grow but practically K may not be too large. We will study the estimation and prediction of target model (1) utilizing the primary data $((x_i^{(0)})^\intercal, y_i^{(0)})$, $i = 1, \ldots, n_0$, as well as the data from K auxiliary studies $((x_i^{(k)})^\intercal, y_i^{(k)})$, $i = 1, \ldots, n_k$, $k = 1, \ldots, K$.

If an auxiliary model is 'similar' to the target model, we say that this auxiliary sample/study is informative. In this work, we characterize the informative level of the kth auxiliary study using the sparsity of the difference between $w^{(k)}$ and β. Let $\delta^{(k)} = \beta - w^{(k)}$ denote the contrast between $w^{(k)}$ and β. The set of informative auxiliary samples is those whose contrasts are sufficiently sparse:

$$\mathcal{A}_q = \{1 \le k \le K : \|\delta^{(k)}\|_q \le h\}, \tag{3}$$

for some $q \in [0, 1]$. The set $\mathcal{A}_q$ contains the auxiliary studies whose contrast vectors have ℓ_q-sparsity at most h and is called the *informative set*. It will be seen later that as long as h is relatively small compared to the sparsity of β, the studies in $\mathcal{A}_q$ can be useful in improving the prediction and estimation of β. In the case of $q = 0$, the set $\mathcal{A}_q$ corresponds to the auxiliary samples whose contrast vectors have at most h nonzero elements. We also consider approximate sparsity constraints ($q \in (0, 1]$), which allows all of the coefficients to be nonzero but their magnitude decays at a relatively rapid rate. For any $q \in [0, 1]$, smaller h implies that the auxiliary samples in $\mathcal{A}_q$ are more informative; larger cardinality of $\mathcal{A}_q$ ($|\mathcal{A}_q|$) implies that a larger number of informative auxiliary samples. Therefore, smaller h and larger $|\mathcal{A}_q|$ should be favourable. We allow $\mathcal{A}_q$ to be empty in which case none of the auxiliary samples is informative. For the auxiliary samples outside of $\mathcal{A}_q$, we do not assume sparse $\delta^{(k)}$ and hence $w^{(k)}$ can be very different from β for $k \notin \mathcal{A}_q$.

In polygenic risk score (PRS) prediction and gene expression partial-correlation analysis, this similarity characterization of two different high-dimensional regression models is motivated by commonly adopted assumptions. In PRS prediction, for example, high-dimensional sparse regression models are commonly assumed (Mak et al., 2017). In addition, it has been observed that many complex traits have a shared genetic aetiology, including various autoimmune diseases (Li et al., 2015; Zhernakova et al., 2009) and psychiatric disorders (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2019; Lee et al., 2013). The similarity characterization we proposed captures the sparse nature of genome-wide association data and shared genetic aetiology of multiple genetically related traits. In the gene expression data analysis, one is interested in understanding how a set of genes regulate another gene based on data measured in different tissues. Such an analysis provides useful insights into gene regulatory networks, which are often sparse. In addition, many tissues have shared regulatory relationships among the genes (Fagny et al., 2017; Pierson et al., 2015). In such applications, we also expect sparse and similar regression coefficients for the models assumed for different tissues.

There is a paucity of methods and fundamental theoretical results for high-dimensional linear regression in the transfer learning setting. In the case where the set of informative auxiliary samples $\mathcal{A}_q$ is known, there is a lack of rate optimal estimation and prediction methods. A closely related topic is multi-task learning (Agarwal et al., 2012; Ando & Zhang, 2005; Lounici et al., 2009), where the goal is to estimate multiple models simultaneously. The multi-task learning considered in Lounici et al. (2009) estimates multiple high-dimensional sparse linear models under the assumption that the supports of all the regression coefficients are the

其中 $w^{(k)} \in \mathbb{R}^p$ 是第 k 项研究的回归向量，而 $\epsilon_i^{(k)}$ 是随机噪声，使得 $\mathbb{E}[\epsilon_i^{(k)}|x_i^{(k)}] = 0$。回归系数 $w^{(k)}$ β

通常是未知的，并且与我们的目标不同。辅助研究的数量 K 允许增长，但实际上 K 可能不会太大。我们将研究目标模型（1）的估计和预测，利用原始数据 ((x$^{(0)}$i)ᵀ, y$^{(0)}$i), i = 1, …, n0，以及来自 K 项辅助研究的数据 ((x$^{(k)}$i)ᵀ, y$^{(k)}$i), i = 1, …, n^k, k = 1, …, K。

如果辅助模型与目标模型“相似”，我们就说这个辅助样本/研究是信息量大的。在这项工作中，我们使用 $w^{(k)}$ 和 β 之间的稀疏性来表征第 k 项辅助研究的 informative 级别。令 $\delta^{(k)} = \beta - w^{(k)}$ 表示 $w^{(k)}$ 和 β 之间的对比。信息辅助样本的集合是那些对比足够稀疏的样本：

$$\mathcal{A}_q = \{1 \le k \le K : \|\delta^{(k)}\|_q \le h\}, \tag{3}$$

对于某些 $q \in [0, 1]$。集合 $\mathcal{A}_q$ 包含对比向量具有最多 h - 稀疏性的辅助研究，并称为 信息集。稍后将会看到，只要 h 相对于 β 的稀疏性相对较小，那么 $\mathcal{A}_q$ 中的研究就可以用于提高 β 的预测和估计。在 $q = 0$ 的情况下，集合 $\mathcal{A}_q$ 对应于对比向量最多具有 h 个非零元素的辅助样本。我们还考虑近似稀疏性约束 ($q \in (0, 1]$)，它允许所有系数都不为零，但它们的幅度以相对较快的速率衰减。对于任何 $q \in [0, 1]$，较小的 h 意味着 $\mathcal{A}_q$ 中的辅助样本更具信息量；$\mathcal{A}_q$ ($|\mathcal{A}_q|$) 的基数较大意味着有更多具有信息量的辅助样本。因此，较小的 h 和较大的

$|\mathcal{A}_q|$ 应该更有利。我们允许 $\mathcal{A}_q$ 为空，在这种情况下，没有任何辅助样本具有信息量。对于 $\mathcal{A}_q$ 外部的辅助样

本，我们不假设稀疏性 $\delta^{(k)}$，因此 $w^{(k)}$ 可以与 β 对于 $k \notin \mathcal{A}_q$. 非常不同

在多基因风险评分 (PRS) 预测和基因表达偏相关分析中，这
两种不同高维回归模型的相似性表征是由通常采用的假设驱动的。例如，在 PRS 预测中，通常假设高维稀疏回归模型 (Mak 等人，2017)。此外，已经观察到许多复杂性状具有共享遗传病因，包括各种自身免疫性疾病 (Li 等人，2015；Zhernakova 等人，2009) 和精神疾病 (精神基因组联盟的跨疾病组，2019；Lee 等人，2013)。我们提出的相似性表征捕捉了全基因组关联数据的稀疏性以及多种遗传相关性状的共享遗传病因。在基因表达数据分析中，人们感兴趣的是基于在不同组织中测量的数据，了解一组基因如何调控另一基因。这种分析为基因调控网络提供了有用的见解，这些网络通常是稀疏的。此外，许多组织在基因之间具有共享的调控关系 (Fagny 等人，2017；Pierson 等人，2015)。在这些应用中，我们也期望不同组织假设的模型具有稀疏和相似的回归系数。

在迁移学习设置下，高维线性回归的方法和基本理论结果非常匮乏。当信息辅助样本集 $\mathcal{A}_q$ 已知时，缺乏速率最优的估计和预测方法。一个密切相关的话题是多任务学习（Agarwal 等人，2012；Ando & Zhang，2005；Lounici 等人，2009），其目标是在同一时间估计多个模型。Lounici 等人（2009）考虑的多任务学习在所有回归系数的支持集都为…的假设下估计多个高维稀疏线性模型。

same. In multi-task learning, different regularization formats have been considered to model the similarity among different studies (Chen et al., 2010; Danaher et al., 2014; Dondelinger et al., 2020).

The goal of transfer learning is, however, different, as one is only interested in estimating the target model and this remains to be a largely unsolved problem. Cai and Wei (2021) studied the minimax and adaptive methods for nonparametric classification in the transfer learning setting under the assumption that all the auxiliary samples are similar to the target distribution (Cai & Wei, 2021, Definition 5). In a more challenging setting where the set $\mathcal{A}_q$ is unknown as is typical in real applications, it is unclear how to avoid the effects of adversarial auxiliary samples. Bastani (2020) studied estimation and prediction in high-dimensional linear models with one informative auxiliary study and $q = 1$, where the sample size of the auxiliary study is larger than the number of covariates. The current work considers more general scenarios under weaker assumptions. Specifically, the sample size of auxiliary samples can be smaller than the number of covariates and some auxiliary studies can be non-informative, which is more practical in applications. Additional challenges include the heterogeneity among the design matrices, which does not arise in the conventional high-dimensional regression problems and hence requires novel proposals.

The problem we study here is certainly related to the high-dimensional prediction and estimation in the conventional settings where only samples from the target model are available. Several penalized or constrained minimization methods have been proposed for prediction and estimation for high-dimensional linear regression; see, for example, Tibshirani (1996); Fan and Li (2001); Zou (2006); Candes and Tao (2007); Zhang (2010). The minimax optimal rates for estimation and prediction are studied in Raskutti et al. (2011) and Verzelen (2012).

1.2 | Our contributions

In the setting where the informative set $\mathcal{A}_q$ is known, we propose a transfer learning algorithm, called Oracle Trans-Lasso, for estimation of the target regression vector and prediction and prove its minimax optimality under mild conditions. The results demonstrate a faster rate of convergence when $\mathcal{A}_q$ is non-empty and h is sufficiently smaller than s, in which case the knowledge from the informative auxiliary samples can be optimally transferred to substantially improve estimation and prediction of the regression problem under the target model.

In the more challenging setting where $\mathcal{A}_q$ is unknown a priori, we introduce a data-driven algorithm, called Trans-Lasso, to adapt to the unknown $\mathcal{A}_q$. The adaption is achieved by aggregating a number of candidate estimators. The desirable properties of the aggregation methods guarantee that the Trans-Lasso does not perform much worse than the best one among the candidate estimators. We construct the candidate estimators and demonstrate the robustness and the efficiency of Trans-Lasso under mild conditions. In terms of robustness, the Trans-Lasso is guaranteed to be not much worse than the Lasso estimator using only the primary samples no matter how adversarial the auxiliary samples are. In terms of efficiency, the knowledge from a subset of the informative auxiliary samples can be transferred to the target problem under proper conditions. Furthermore, If the contrast vectors in the informative samples are sufficiently sparse, the Trans-Lasso estimator performs as if the informative set $\mathcal{A}_q$ is known.

相同。在多任务学习中，已经考虑了不同的正则化格式来建模不同研究之间的相似性（Chen 等人，2010；Danaher 等人，2014；Dondelinger 等人，2020）。

然而，迁移学习的目标不同，因为人们只对估计目标模型感兴趣，而这仍然是一个尚未完全解决的问题。Cai 和 Wei (2021) 在所有辅助样本都与目标分布相似的假设下（Cai & Wei, 2021, 定义 5），研究了迁移学习设置下非参数分类的 minimax 和自适应方法。在一个更具挑战性的设置中，其中集合 $\mathcal{A}_q$ 未知，这在实际应用中很典型，不清楚如何避免对抗性辅助样本的影响。Bastani (2020) 研究了具有一个信息辅助研究和 $q = 1$ 的高维线性模型中的估计和预测，其中辅助研究的样本量大于协变量的数量。当前工作考虑了在更弱的假设下的更一般的场景。具体来说，辅助样本的样本量可以小于协变量的数量，并且一些辅助研究可以是非信息的，这在实际应用中更实用。额外的挑战包括设计矩阵之间的异质性，这在传统的多维回归问题中不会出现，因此需要新的提案。

我们在此研究的问题当然与传统的仅能获取目标模型样本的高维预测和估计问题相关。已经提出了几种用于高维线性回归预测和估计的惩罚或约束最小化方法；例如，参见 Tibshirani (1996)；Fan 和 Li (2001)；Zou (2006)；Candes 和 Tao (2007)；Zhang (2010)。估计和预测的极大极小最优率在 Raskutti 等人 (2011) 和 Verzelen (2012) 中进行了研究。

1.2 | 我们的贡献

在信息集 $\mathcal{A}_q$ 已知的设置中，我们提出了一种迁移学习算法，称为 Oracle Trans- Lasso，用于目标回归向量的估计以及预测，并在温和条件下证明了其极大极小最优性。结果表明，当 $\mathcal{A}_q$ 非空且 h 足够小于 s 时，收敛速度更快，在这种情况下，来自信息辅助样本的知识可以被最优地转移，从而显著提高目标模型下回归问题的估计和预测。

在$\mathcal{A}_q$ 未知且需要先验信息的更具挑战性的设置中，我们引入了一种数据驱动的算法，称为Trans- Lasso，以适应未知的 $\mathcal{A}_q$。这种适应是通过聚合多个候选估计量来实现的。聚合方法的理想特性保证了Trans- Lasso的表现不会比候选估计量中最好的那个差很多。我们构建了候选估计量，并在温和条件下展示了Trans- Lasso的鲁棒性和效率。在鲁棒性方面，无论辅助样本多么对抗性，Trans- Lasso都保证不会比仅使用主要样本的Lasso估计量差很多。在效率方面，在适当条件下，来自信息辅助样本子集的知识可以迁移到目标问题。此外，如果信息样本中的对比向量足够稀疏，那么Trans- Lasso估计量的表现就好像信息集$\mathcal{A}_q$ 是已知的。

When the distributions of the design matrices are distinct in different samples, the effect of heterogeneous designs in transfer learning is studied. The performance of the proposed algorithm is investigated theoretically and numerically in various settings.

1.3 | Organization and notation

The rest of this paper is organized as follows. Section 2 focuses on the setting where the informative set $\mathcal{A}_q$ is known and with the sparsity in Equation (3) measured in ℓ_1-norm. A transfer learning algorithm is proposed for estimation and prediction of the target parameter and its minimax optimality is established. In Section 3, we study the estimation and prediction of the target model when $\mathcal{A}_q$ is unknown for $q = 1$. In Section 4, we justify the theoretical performance of our proposals under heterogeneous designs. In Section 5, the numerical performance of the proposed methods is studied in various settings. In Section 6, the proposed algorithms are applied to the GTEx data to investigate the association of one gene with other genes in a target tissue by leveraging data measured on other related tissues or cell types. The proofs and results for ℓ_q-sparse contrasts with $q \in [0, 1)$ are provided in the supplementary materials (Li et al., 2020).

We finish this section with notation. Let $X^{(0)} \in \mathbb{R}^{n_0 \times p}$ and $y^{(0)} \in \mathbb{R}^{n_0}$ denote the design matrix and the response vector for the primary data respectively. Let $X^{(k)} \in \mathbb{R}^{n_k \times p}$ and $y^{(k)} \in \mathbb{R}^{n_k}$ denote the design matrix and the response vector for the kth auxiliary data respectively. For a class of matrices $R_l \in \mathbb{R}^{n_l \times p_0}$, $l \in \mathcal{L}$, we use $\{R_l\}_{l \in \mathcal{L}}$ to denote R_l, $l \in \mathcal{L}$. Let $n_{\mathcal{A}_q} = \sum_{k \in \mathcal{A}_q} n_k$. For a generic semi-positive definite matrix $\Sigma \in \mathbb{R}^{m \times m}$, let $\Lambda_{\max}(\Sigma)$ and $\Lambda_{\min}(\Sigma)$ denote the largest and smallest eigenvalues of Σ respectively. Let $\mathrm{Tr}(\Sigma)$ denote the trace of Σ. Let e_j be a vector such that its jth element is 1 and all other elements are zero. Let $a \vee b$ denote $\max\{a, b\}$ and $a \wedge b$ denote $\min\{a, b\}$. We use $c, c_0, c_1, \ldots$ to denote generic constants which can be different in different statements. Let $a_n = O(b_n)$ and $a_n \lesssim b_n$ denote $|a_n/b_n| \le c$ for some constant c when n is large enough. Let $a_n \asymp b_n$ denote $|a_n/b_n| \to c$ for some constant c as $n \to \infty$. Let $a_n = O_P(b_n)$ and $a_n \lesssim_{\mathbb{P}} b_n$ denote $\mathbb{P}(|a_n/b_n| \le c) \to 1$ for some constant $c < \infty$. Let $a_n = o_P(b_n)$ denote $\mathbb{P}(|a_n/b_n| > c) \to 0$ for any constant $c > 0$.

2 | ESTIMATION WITH KNOWN INFORMATIVE AUXILIARY SAMPLES

We consider in this section transfer learning for high-dimensional linear regression when the informative set $\mathcal{A}_q$ is known. The focus is on the ℓ_1-sparse characterization of the contrast vectors. The notation $\mathcal{A}_1$ will be abbreviated as $\mathcal{A}$ in the sequel without special emphasis. Section C in the supplementary materials generalizes the sparse contrasts from ℓ_1-constraint to ℓ_q-constraint for $q \in [0, 1)$ and presents a rate-optimal estimator in this setting.

2.1 | Oracle Trans-Lasso algorithm

We propose a transfer learning algorithm, called *Oracle Trans-Lasso*, for estimation and prediction when $\mathcal{A}$ is known. As an overview, we first compute an initial estimator using all the informative auxiliary samples. However, its probabilistic limit is biased from β as $w^{(k)} \neq \beta$ in

当设计矩阵在不同样本中的分布不同时，研究了迁移学习中异构设计的影响。在各种设置下，从理论上和数值上研究了所提出算法的性能。

1.3 | 组织与符号

本文的其余部分组织如下。第 2 节重点讨论信息集 $\mathcal{A}_q$ 已知且方程 (3) 中的稀疏性以 ℓ_1- 范数度量的设置。提出了一种迁移学习算法，用于估计和预测目标参数及其极大极小最优性。在第 3 节中，我们研究了当 $\mathcal{A}_q$ 未知时，$q = 1$ 的目标模型的估计和预测。在第 4 节中，我们证明了在异构设计下我们提案的理论性能。在第 5 节中，在各种设置下研究了所提出方法的数值性能。在第 6 节中，所提出的算法被应用于 GTEx 数据，通过利用在其它相关组织或细胞类型上测量的数据，研究了目标组织中一个基因与其他基因的关联。ℓ_q- 稀疏对比与 $q \in [0, 1)$ 的证明和结果在补充材料 (Li 等人，2020) 中提供。

我们用符号结束这一节。令 X $^{(0)} \in \mathbb{R}^{n_0 \times}$p 和 $y^{(0)} \in \mathbb{R}^{n_0}$ 分别表示原始数据的设计矩阵和响应向量。令 $X^{(k)} \in \mathbb{R}^{n_k \times p}$ 和 $y^{(k)} \in \mathbb{R}^{n_k}$ 分别表示第 k 个辅助数据的设计矩阵和响应向量。对于一类矩阵 R$_l \in \mathbb{R}^{n_l \times}$p0，l $\in \mathcal{L}$，我们用 $\{R_l\}_{l \in \mathcal{L}}$ 表示 Rl，l $\in \mathcal{L}$。令 $n_{\mathcal{A}_q} = \sum_{k \in \mathcal{A}_q} n_k$. 对于一个通用的半正定矩阵 Σ ∈ℝ $^{m \times m}$，令 $\Lambda_{\max}(\Sigma)$ 和

$\Lambda_{\min}(\Sigma)$ 分别表示 Σ 的最大和最小特征值。令 Tr(Σ) 表示 Σ 的迹。令 e$_j$ 是一个向量，其第 j 个元素为 1，其余元素为 0。令 a ∨ b 表示 max{a, b}，a ∧ b 表示 min{a, b}。我们用 c, c0, c1, … 表示在不同语句中可能不同的通用常数。令 $a_n = O(b_n)$ 和 $a_n \lesssim b_n$ 表示当 n 足够大时，$|a_n/b_n| \le c$ 对于某个常数 c。令 $a_n \asymp b_n$ 表示 $|a_n/b_n| \to c$ 对

于某个常数 c，当 n → ∞。令 $a_n = O_P(b_n)$ 和 $a_n \lesssim_{\mathbb{P}} b_n$ 表示 ℙ(a_n/b_n $\le c) \to 1$ 对于某个常数 c < ∞。令 $a_n = o_P(b_n)$ 表示 ℙ(

| |

$|a_n/b_n|$ $> c) \to 0$ 对于任何常数$c > 0$。

已知信息辅助样本的估计

在本节中，我们考虑当信息集$\mathcal{A}_q$ 已知时的高维线性回归的迁移学习。重点是对比向量的ℓ_1-稀疏特征。符号$\mathcal{A}_1$ 将在后续中缩写为$\mathcal{A}$，而无需特别强调。补充材料中的第C节将稀疏对比从ℓ_1-约束推广到ℓ_q-约束，并针对$q \in [0, 1)$ 在此设置中提出一个速率最优估计器。

2.1 | Oracle Trans- Lasso算法

我们提出了一种迁移学习算法，称为<code>Oracle Trans- Lasso</code>，用于在$\mathcal{A}$ 已知时的估计和预测。概述如下，我们首先使用所有信息辅助样本计算一个初始估计器。然而，其概率极限从β 偏差，如$w^{(k)} \neq \beta$ 中

```
Algorithm 1: Oracle Trans-Lasso algorithm
Input   : Primary data $(X^{(0)}, y^{(0)})$ and informative auxiliary
          samples $\{X^{(k)}, y^{(k)}\}_{k\in\mathcal{A}}$
Output: $\hat{\beta}$
Step 1. Compute
    $\hat{w}^{\mathcal{A}} = \arg\min_{w\in\mathbb{R}^p} \left\{ \frac{1}{2n_{\mathcal{A}}} \sum_{k\in\mathcal{A}} \|y^{(k)} - X^{(k)}w\|_2^2 + \lambda_w\|w\|_1 \right\}$    (4)
    for $\lambda_w = c_1\sqrt{\log p/n_{\mathcal{A}}}$ with some constant $c_1$.
Step 2. Let
    $\hat{\beta} = \hat{w}^{\mathcal{A}} + \hat{\delta}^{\mathcal{A}}$,    (5)
    where
    $\hat{\delta}^{\mathcal{A}} = \arg\min_{\delta\in\mathbb{R}^p} \left\{ \frac{1}{2n_0} \|y^{(0)} - X^{(0)}(\hat{w}^{\mathcal{A}} + \delta)\|_2^2 + \lambda_\delta\|\delta\|_1 \right\}$    (6)
    for $\lambda_\delta = c_2\sqrt{\log p/n_0}$ with some constant $c_2$.
```

general. We then correct its bias using the primary data in the second step. Algorithm 1 formally presents our proposed Oracle Trans-Lasso algorithm.

In Step 1, $\widehat{w}^{\mathcal{A}}$ is realized based on the Lasso (Tibshirani, 1996) using all the informative auxiliary samples. Its probabilistic limit is $w^{\mathcal{A}}$, which can be defined via the following moment condition

$$\mathbb{E}\left[\sum_{k\in\mathcal{A}} (X^{(k)})^\intercal (y^{(k)} - X^{(k)}w^{\mathcal{A}})\right] = 0.$$

Denoting $\mathbb{E}[x_i^{(k)}(x_i^{(k)})^\intercal] = \Sigma^{(k)}$, $w^{\mathcal{A}}$ has the following explicit form:

$$w^{\mathcal{A}} = \beta + \delta^{\mathcal{A}} \tag{7}$$

for $\delta^{\mathcal{A}} = \sum_{k\in\mathcal{A}} \alpha_k \delta^{(k)}$ and $\alpha_k = n_k/n_{\mathcal{A}}$ given that $\Sigma^{(k)} = \Sigma^{(0)}$ for all $k \in \mathcal{A}$. That is, the probabilistic limit of $\widehat{w}^{\mathcal{A}}$, $w^{\mathcal{A}}$, has bias $\delta^{\mathcal{A}}$, which is a weighted average of $\delta^{(k)}$. Step 1 is related to the approach for high-dimensional misspecified models (Bühlmann & van de Geer, 2015) and moment estimators. The estimator $\widehat{w}^{\mathcal{A}}$ converges relatively fast as the sample size used in Step 1 is relatively large. Step 2 corrects the bias, $\delta^{\mathcal{A}}$, using the primary samples. In fact, $\delta^{\mathcal{A}}$ is a sparse high-dimensional vector whose ℓ_1-norm is no larger than h. Hence, the error of step 2 is under control for a relatively small h. The choice of the tuning parameters λ_w and λ_δ will be further specified in Theorem 1.

We compare the proposed Oracle Trans-Lasso method to the multi-task regression methods, say section 3.4.3 of Agarwal et al. (2012) and Danaher et al. (2014). The Oracle Trans-Lasso does not penalize the differences among the regression coefficients in the auxiliary studies. This is again because the focus of transfer learning is only the target study. Theoretically, extra penalization terms and the joint analysis of multiple estimators may not help improve the estimation accuracy of the parameter of interest.

```
Algorithm 1: Oracle Trans-Lasso algorithm
Input   : Primary data $(X^{(0)}, y^{(0)})$ and informative auxiliary
          samples $\{X^{(k)}, y^{(k)}\}_{k\in\mathcal{A}}$
Output: $\hat{\beta}$
Step 1. Compute
    $\hat{w}^{\mathcal{A}} = \arg\min_{w\in\mathbb{R}^p} \left\{ \frac{1}{2n_{\mathcal{A}}} \sum_{k\in\mathcal{A}} \|y^{(k)} - X^{(k)}w\|_2^2 + \lambda_w\|w\|_1 \right\}$    (4)
    for $\lambda_w = c_1\sqrt{\log p/n_{\mathcal{A}}}$ with some constant $c_1$.
Step 2. Let
    $\hat{\beta} = \hat{w}^{\mathcal{A}} + \hat{\delta}^{\mathcal{A}}$,    (5)
    where
    $\hat{\delta}^{\mathcal{A}} = \arg\min_{\delta\in\mathbb{R}^p} \left\{ \frac{1}{2n_0} \|y^{(0)} - X^{(0)}(\hat{w}^{\mathcal{A}} + \delta)\|_2^2 + \lambda_\delta\|\delta\|_1 \right\}$    (6)
    for $\lambda_\delta = c_2\sqrt{\log p/n_0}$ with some constant $c_2$.
```

一般而言。然后我们使用第二步中的原始数据来纠正其偏差。算法1正式提出了我们提出的Oracle Trans- Lasso算法。

在步骤1, $\widehat{w}^{\mathcal{A}}$ 中基于Lasso（Tibshirani，1996）使用所有信息量大的辅助样本实现。其概率极限是$w^{\mathcal{A}}$，可以通过以下矩条件定义

$$\mathbb{E}\left[\sum_{k\in\mathcal{A}} (X^{(k)})^\intercal (y^{(k)} - X^{(k)}w^{\mathcal{A}})\right] = 0.$$

记$\mathbb{E}[x_i^{(k)}(x_i^{(k)})^\intercal] = \Sigma^{(k)}$，$w^{\mathcal{A}}$ 具有以下显式形式：

$$w^{\mathcal{A}} = \beta + \delta^{\mathcal{A}} \tag{7}$$

对于$\delta^{\mathcal{A}} = \sum_{k\in\mathcal{A}} \underset{\wedge\mathcal{A}}{\alpha_k} \underset{\mathcal{A}}{\delta^{(k)}}$ 和$\alpha_k = n_k/n_{\mathcal{A}}$，给定$\Sigma^{(k)} = \Sigma^{(0)}$ 对于所有$k \in \mathcal{A}$. 这意味着，w，w的

概率极限存在偏差$\delta^{\mathcal{A}}$，这是一个$\delta^{(k)}$的加权平均。步骤1与高维错定模型（Bühlmann 和 van de Geer，2015）和矩估计量相关。估计器$\widehat{w}^{\mathcal{A}}$ 随着步骤1中使用的样本量相对较大而相对快速收敛。步骤2使用主要样本纠正偏差$\delta^{\mathcal{A}}$。事实上，$\delta^{\mathcal{A}}$ 是一个稀疏的高维向量，其ℓ_1-范数不超过<样式id='15'>h</样式>。因此，对于相对较小的<样式id='17'>h</样式>，步骤2的误差在控制范围内。调整参数λ_w 和λ_δ 的选择将在定理1中进一步指定。

我们将所提出的 Oracle Trans- Lasso 方法与多任务回归方法进行比较，例如 Agarwal 等人 (2012) 的第 3.4.3 节和 Danaher 等人 (2014) 的研究。Oracle Trans-Lasso 不会惩罚辅助研究中回归系数之间的差异。这再次是因为迁移学习的重点仅是目标研究。理论上，额外的惩罚项和多个估计量的联合分析可能无法提高目标参数的估计精度。

2.2 | Theoretical properties of Oracle Trans-Lasso

Formally, the parameter space we consider can be written as

$$\Theta_q(s,h)=\left\{B=(\beta,\delta^{(1)},\ldots,\delta^{(K)}):\|\beta\|_0\le s,\ \max_{k\in\mathcal{A}_q}\|\delta^{(k)}\|_q\le h\right\} \tag{8}$$

for $\mathcal{A}_q\subseteq\{1,\ldots,K\}$ and $q\in[0,1]$. We study the rate of convergence for the Oracle Trans-Lasso algorithm under the following two conditions.

Condition 1 *For each $k\in\mathcal{A}\cup\{0\}$, each row of $X^{(k)}$ is i.i.d. Gaussian distributed with mean zero and covariance matrix Σ. The smallest and largest eigenvalues of Σ are bounded away from zero and infinity respectively.*

Condition 2 *For each $k\in\mathcal{A}\cup\{0\}$, $\mathbb{E}[(y_i^{(k)})^2]$ is finite and the random noises $\epsilon_i^{(k)}$ are i.i.d. sub-Gaussian with mean zero and variance σ_k^2. For some constant C_0, it holds that $\max_{k\in\mathcal{A}\cup\{0\}}\mathbb{E}[\exp\{t\epsilon_i^{(k)}\}]\le\exp\{t^2C_0\}$ for all $t\in\mathbb{R}$.*

Condition 1 assumes Gaussian designs, which provides convenience for bounding the restricted eigenvalues of sample covariance matrices. Moreover, the designs are identically distributed for $k\in\mathcal{A}\cup\{0\}$. This assumption simplifies some technical conditions and will be relaxed in Section 4. We mention that the conditions on the eigenvalues of Σ can be replaced with some eigenvalue conditions restricted to a convex cone. Condition 2 assumes sub-Gaussian random noises for primary and informative auxiliary samples and the second moment of the response vector is finite. Conditions 1 and 2 make no assumptions on the non-informative auxiliary samples as they are not used in the Oracle Trans-Lasso algorithm. In the next theorem, we prove the convergence rate of the Oracle Trans-Lasso. Let $\eta_h=h\sqrt{\log p/n_0}\wedge h^2$.

Theorem 1 (Convergence rate of Oracle Trans-Lasso). *Assume that Conditions 1 and 2 hold true. Suppose that $\mathcal{A}$ is known with $h\lesssim s\sqrt{\log p/n_0}$ and $n_0\lesssim n_{\mathcal{A}}$. We take $\lambda_w=\max_{k\in\mathcal{A}}c_1\sqrt{\mathbb{E}[(y_i^{(k)})^2]\log p/n_{\mathcal{A}}}$ and $\lambda_\delta=c_2\sqrt{\log p/n_0}$ for some sufficiently large constants c_1 and c_2. If $s\log p/n_{\mathcal{A}}+h(\log p/n_0)^{1/2}=o(1)$, then there exists some positive constant c_1 such that*

$$\inf_{B\in\Theta_1(s,h)}\mathbb{P}\left(\frac{1}{n_0}\|X^{(0)}(\hat\beta-\beta)\|_2^2\vee\|\hat\beta-\beta\|_2^2\lesssim\frac{s\log p}{n_{\mathcal{A}}+n_0}+\frac{s\log p}{n_0}\wedge\eta_h\right)\ge1-\exp(-c_1\log p), \tag{9}$$

where $B=\{\beta,w^{(1)},\ldots,w^{(k)}\}$ denotes all the unknown parameters. Theorem 1 provides the convergence rate of $\hat\beta$ for any true parameters in $\Theta_1(s,h)$ when an informative set $\mathcal{A}$ is known. We illustrate Theorem 1 by contrasting to the estimation results of the Lasso. First, the results of Theorem 1 hold under a weaker condition on s, that is, $s\log p=o(n_{\mathcal{A}})$ when $n_{\mathcal{A}}\gtrsim n_0$, while $s\log p=o(n_0)$ is always assumed in the single-task regression. Hence, the Oracle Trans-Lasso can deal with more challenging scenarios with less sparse target parameter. Second, the right-hand side of Equation (9) is sharper than the convergence rate of Lasso, $s\log p/n_0$, if $h\ll s\sqrt{\log p/n_0}$ and $n_{\mathcal{A}}\gg n_0$. That is, if the informative auxiliary samples have contrast vectors sufficiently

2.2 | Oracle Trans- Lasso 的理论性质

形式上，我们考虑的参数空间可以写为

$$\Theta_q(s,h)=\left\{B=(\beta,\delta^{(1)},\ldots,\delta^{(K)}):\|\beta\|_0\le s,\ \max_{k\in\mathcal{A}_q}\|\delta^{(k)}\|_q\le h\right\} \tag{8}$$

对于 $\mathcal{A}_q\subseteq\{1,\ldots,K\}$ 和 $q\in[0,1]$。我们研究 Oracle Trans- Lasso 算法在以下两个条件下的收敛速度。

条件 1 对于每个 $k\in\mathcal{A}\cup\{0\}$，$X^{(k)}$ 的每一行是独立同分布的高斯分布，均值为零，协方差矩阵为 Σ。Σ 的最小和最大特征值分别有界远离零和无穷大。

条件 2 对于每个 $k\in\mathcal{A}\cup\{0\}$ $\mathbb{E}[(y_i^{(k)})^2]$,是有限的，随机噪声 $\epsilon_i^{(k)}$ 是独立同分布的次高斯分布，均值为零，方差为 σ_k^2. 。对于某个常数 C_0，它满足 $\max_{k\in\mathcal{A}\cup\{0\}}\mathbb{E}[\exp\{t\epsilon_i^{(k)}\}]\le\exp\{t^2C_0\}$ *for all* $t\in\mathbb{R}$。

条件 1 假设高斯设计，这为界定样本协方差矩阵的受限特征值提供了便利。此外，设计对于 $k\in\mathcal{A}\cup\{0\}$ 是同分布的。这一假设简化了一些技术条件，并在第 4 节中将得到放宽。我们指出，关于 Σ 的特征值的条件可以替换为一些限制在凸锥上的特征值条件。条件 2 假设主要样本和信息辅助样本具有次高斯随机噪声，并且响应向量的二阶矩是有限的。条件 1 和 2 对非信息辅助样本不做假设，因为它们在 Oracle Trans- Lasso 算法中不被使用。$\surd$在下一个定理中，我们证明了 Oracle Trans- Lasso 的收敛速度。令$\eta_h=h\quad\log p/n_0\wedge h^2$.

定理1 (Oracle Trans- Lasso 的收敛速度)。 假设条件 *1* 和 *2* 成立。假设 $\mathcal{A}$ 已知，并且 $h\lesssim s\quad\log p/n_0$ 和 $n_0\lesssim n_{\mathcal{A}}.$ 。我们取 $\surd$ $\lambda_w=\surd\quad {}_{k\in\mathcal{A}}\ {}_1\sqrt{\mathbb{E}[\ {}^{(k)}\ {}^2]}\quad/\ {}_{\mathcal{A}}\quad \lambda_\delta=c_2\quad\log p/n_0$ *for* max$c(y_i)\log pn$以及一些足够大的常数 c_1 和 c_2。如果 $s\log p/n_{\mathcal{A}}+h(\log p/n_0)^{1/2}=o(1)$，则存在某个正常数 c_1 使得

$$\inf_{B\in\Theta_1(s,h)}\mathbb{P}\left(\frac{1}{n_0}\|X^{(0)}(\hat\beta-\beta)\|_2^2\vee\|\hat\beta-\beta\|_2^2\lesssim\frac{s\log p}{n_{\mathcal{A}}+n_0}+\frac{s\log p}{n_0}\wedge\eta_h\right)\ge1-\exp(-c_1\log p), \tag{9}$$

其中 $B=\{\beta,w^{(1)},\ldots,w^{(k)}\}$ 表示所有未知参数。定理1提供了 $\hat\beta$ 在 $\Theta_1(s,h)$ 中对任何真实参数的收敛速度，当已知信息集 $\mathcal{A}$ 时。我们通过对比Lasso的估计结果来说明定理1。首先，定理1在 s 的较弱条件下成立，即 $s\log p=o(n_{\mathcal{A}})$ 当 $n_{\mathcal{A}}\gtrsim n_0$，而 $s\log p=o(n_0)$ 总是在单任务回归中假设。因此，Oracle Trans- Lasso 可以处理具有更稀疏目标参数的更具挑战性的场景。其次，方程 (9) 的右侧比Lasso的收敛速度更严格，$s\log p/n_0$，如果 $h\ll s\sqrt{\log p/n_0}$ 和 $n_{\mathcal{A}}\gg n_0.$ 。也就是说，如果信息辅助样本的对向量具有足够的对比性

sparser than β and the total sample size is significantly larger than the primary sample size, then the knowledge from the auxiliary samples can significantly improve the learning performance of the target model. The condition for improvement, $h \ll s\sqrt{\log p/n_0}$, allows a wide range of h. For example, the typical regime for single-task regression is $s\log p/n_0 = O(1)$ and it implies that $s\sqrt{\log p/n_0}$ can be as large as $\sqrt{n_0/\log p}$. Hence, the condition for improvement of Theorem 1 allows h to be as large as $\sqrt{n_0/\log p}$. Larger the s, weaker the condition for improvement.

The sample size requirement in Theorem 1 guarantees the lower restricted eigenvalues of the sample covariance matrices in use are bounded away from zero with high probability. The proof of Theorem 1 involves an error analysis of $\widehat{w}^{\mathcal{A}}$ and that of $\widehat{\delta}^{\mathcal{A}}$. While $w^{\mathcal{A}}$ may be neither ℓ_0- nor ℓ_1-sparse, it can be decomposed into an ℓ_0-sparse component plus an ℓ_1 -sparse component as illustrated in Equation (7). Exploiting this sparse structure is a key step in proving Theorem 1. Regarding the choice of tuning parameters, λ_w depends on the second moment of $y_i^{(k)}$, which can be consistently estimated by $\|y^{(k)}\|_2^2/n_k$. The other tuning parameter λ_δ depends on the noise levels, which can be estimated by the scaled Lasso (Sun & Zhang, 2012). In practice, cross-validation can be performed for selecting tuning parameters.

We now establish the minimax lower bound for estimating β in the transfer learning setup, which shows the minimax optimality of the Oracle Trans-Lasso algorithm in $\Theta_1(s, h)$.

Theorem 2 (Minimax lower bound for q = 1). *Assume Conditions* 1 *and* 2. *If* $\max\{s\log p/(n_{\mathcal{A}} + n_0),\ h(\log p/n_0)^{1/2}\} = o(1)$, *then*

$$\inf_{\hat\beta}\sup_{B\in\Theta_1(s,h)} \mathbb{P}\left(\|\hat\beta-\beta\|_2^2 \geq c_1\frac{s\log p}{n_{\mathcal{A}}+n_0} + c_2\frac{s\log p}{n_0}\wedge\eta_h\right) \geq \frac{1}{2}$$

for some positive constants c_1 *and* c_2.

Theorem 2 implies that $\hat\beta$ obtained by the Oracle Trans-Lasso algorithm is minimax rate optimal in $\Theta_1(s, h)$ under the conditions of Theorem 1. To understand the lower bound, the term $s\log p/(n_{\mathcal{A}} + n_0)$ is the optimal convergence rate when $w^{(k)} = \beta$ for all $k \in \mathcal{A}$. This is an extremely ideal case where we have $n_{\mathcal{A}} + n_0$ *i.i.d.* samples from the target model. The second term in the lower bound is the optimal convergence rate when $w^{(k)} = 0$ for all $k \in \mathcal{A}$, that is, the auxiliary samples are not helpful at all. Let $\mathcal{B}_q(r) = \{u \in \mathbb{R}^p : \|u\|_q \leq r\}$ denote the ℓ_q-ball with radius r centred at zero. In this case, the definition of $\Theta_1(s, h)$ implies that $\beta \in \mathcal{B}_0(s) \cap \mathcal{B}_1(h)$ and the second term in the lower bound is indeed the minimax optimal rate for estimation when $\beta \in \mathcal{B}_0(s) \cap \mathcal{B}_1(h)$ with n_0 *i.i.d.* samples (Tsybakov, 2014).

3 | UNKNOWN SET OF INFORMATIVE AUXILIARY SAMPLES

The Oracle Trans-Lasso algorithm is based on the knowledge of the informative set $\mathcal{A}$. In some applications, the informative set $\mathcal{A}$ is not given, which makes the transfer learning problem more challenging. In this section, we propose a data-driven method for estimation and prediction when $\mathcal{A}$ is unknown. The proposed algorithm is described in detail in Sections 3.1 and 3.2. Its theoretical properties are studied in Section 3.3.

稀疏于 β 并且总样本量明显大于原始样本量，那么来自辅助样本的知识可以显著提高目标模型的学习性能。改进的条件， $h \ll s\sqrt{\log p/n_0}$，允许广泛的 h。例如，单任务回归的典型情况是 $s\log p/n_0 = O(1)$ 并且它意味着 $s\sqrt{\log p/n_0}$ 可以像 $\sqrt{n_0/\log p}$那么大。因此，定理1的改进~~条件允许~~ h 大到 $n_0/\log p$。~~越大的 s~~，改进的条件越弱.

定理1中的样本量要求保证了在使用中的样本协方差矩阵的受限特征值以高概率有界远离零。定理1的证明涉及 $\widehat{w}^{\mathcal{A}}$ 和 $\widehat{\delta}^{\mathcal{A}}$ 的误差分析。虽然 $w^{\mathcal{A}}$ 可能既不是 ℓ_0-稀疏也不是 ℓ_1-稀疏，但它可以分解为一个 ℓ_0-稀疏分量加上一个 ℓ_1-稀疏分量，如公式(7)所示。利用这种稀疏结构是证明定理1的关键步骤。关于调整参数的选择， λ_w 取决于 $y_i^{(k)}$ 的二阶矩，这可以通过 $\|y^{(k)}\|_2^2/n_k$ 一致估计。另一个调整参数 λ_δ 取决于噪声水平，可以通过缩放Lasso估计 (Sun & Zhang, 2012). 在实践中，交叉验证可用于选择调整参数。

我们现在建立估计 β 在迁移学习设置中的极小极大下界，这显示了Oracle Trans-Lasso算法在 $\Theta_1(s, h)$.

定理2 (估计的极小极大下界。 q = 1)。 假设条件 1 和 2。 如果$\max\{s\log p/(n_{\mathcal{A}} + n_0),\ h(\log p/n_0)^{1/2}\} = o(1)$，则

$$\inf_{\hat\beta}\sup_{B\in\Theta_1(s,h)} \mathbb{P}\left(\|\hat\beta-\beta\|_2^2 \geq c_1\frac{s\log p}{n_{\mathcal{A}}+n_0} + c_2\frac{s\log p}{n_0}\wedge\eta_h\right) \geq \frac{1}{2}$$

对于某些正常数 c_1 和 c_2.

定理2意味着Oracle Trans- Lasso算法得到的 $\hat\beta$ 在定理1的条件下的 $\Theta_1(s, h)$ 中是极小极大最优的。为了理解下界，项 $s\log p/(n_{\mathcal{A}} + n_0)$ 是当 $w^{(k)} = \beta$ 对所有 $k \in \mathcal{A}$ 时的最优收敛速度。这是一个极其理想的情况，其中我们有 $n_{\mathcal{A}} + n_0$ *i.i.d* 来自目标模型的样本。下界的第二项是当 $w^{(k)} = 0$ 对所有 $k \in \mathcal{A}$ 时的最优收敛速度，即辅助样本完全无帮助。令 $\mathcal{B}_q(r) = \{u \in \mathbb{R}^p : \|u\|_q \leq r\}$ 表示以零为中心、半径为 Θ_1

r 的 ℓ_q- 球。在这种情况下，(s, h) 的定义意味着 $\beta \in \mathcal{B}_0(s) \cap \mathcal{B}_1(h)$，并且下界的第二项确实是当 $\beta \in \mathcal{B}_0(s) \cap \mathcal{B}_1(h)$ 且 n_0 *i.i.d* 样本时的估计极小极大最优速度 (Tsybakov, 2014)。

3 | 未知集OF INFORMATIVE AUXILIARY SAMPLES

Oracle Trans- Lasso算法基于对信息集$\mathcal{A}$的知识。在一些应用中，信息集$\mathcal{A}$ 没有给出，这使得迁移学习问题更具挑战性。在本节中，我们提出了一种数据驱动的方法，用于估计和预测当$\mathcal{A}$ 未知时。所提出的算法在3.1节和3.2节中详细描述。其理论性质在3.3节中研究。

3.1 | The Trans-Lasso algorithm

Our proposed algorithm, called Trans-Lasso, consists of two main steps. First, we construct a collection of candidate estimators, each of which is based on an estimate of $\mathcal{A}$. Second, we perform an aggregation step (Dai et al., 2012, 2018; Rigollet & Tsybakov, 2011) on these candidate estimators. Under proper conditions, the aggregated estimator is guaranteed to be not much worse than the best candidate estimator under consideration in terms of prediction. For technical reasons, we need the candidate estimators and the samples for aggregation to be independent. Hence, we start with sample splitting. We need some more notation. For a generic estimate of β, b, denote its sum of squared prediction error as

$$\widehat{Q}(\mathcal{I}, b) = \sum_{i \in \mathcal{I}} \|y_i^{(0)} - (x_i^{(0)})^{\intercal} b\|_2^2,$$

where $\mathcal{I}$ is a subset of $\{1, \dots, n_0\}$. Let $\Lambda^{L+1} = \{\nu \in \mathbb{R}^{L+1} : \nu_l \geq 0, \sum_{l=0}^{L} \nu_l = 1\}$ denote an L-dimensional simplex. The Trans-Lasso algorithm is presented in Algorithm 2.

As an illustration, steps 2 and 3 of the Trans-Lasso algorithm construct some initial estimates of β, $\widehat{\beta}(\widehat{G}_l)$. They are computed using the Oracle Trans-Lasso algorithm by treating each $\widehat{G}_l$ as the set of informative auxiliary samples. We construct $\widehat{G}_l$ to be some estimates of $\mathcal{A}$ using the procedure provided in Section 3.2. Step 4 is based on the Q-aggregation proposed in Dai et al. (2012) with a uniform prior, a Kullback–Leibler penalty, and a simplified tuning parameter. The Q-aggregation can be viewed as a weighted version of least square aggregation and exponential aggregation (Rigollet & Tsybakov, 2011 and it has been shown to be rate optimal both in expectation and with high probability for model selection aggregation problems.

```
Algorithm 2: Trans-Lasso Algorithm
Input  : Primary data $(X^{(0)}, y^{(0)})$ and samples from $K$
         auxiliary studies $\{X^{(k)}, y^{(k)}\}_{k=1}^{K}$.
Output: $\hat{\beta}^{\hat{\theta}}$.
Step 1. Let $\mathcal{I}$ be a random subset of $\{1, \dots, n_0\}$ such that
    $|\mathcal{I}| \approx c_0 n_0$ with some constant $0 < c_0 < 1$. Let
    $\mathcal{I}^c = \{1, \dots, n_0\} \setminus \mathcal{I}$.
Step 2. Construct $L+1$ candidate sets of $\mathcal{A}$, $\{\widehat{G}_0, \widehat{G}_1, \dots, \widehat{G}_L\}$
    such that $\widehat{G}_0 = \emptyset$ and $\widehat{G}_1, \dots, \widehat{G}_L$ are based on (14) using
    $\left(X_{\mathcal{I},\cdot}^{(0)}, y_{\mathcal{I}}^{(0)}\right)$ and $\{X^{(k)}, y^{(k)}\}_{k=1}^{K}$.
Step 3. For each $0 \leq l \leq L$, run the Oracle Trans-Lasso algorithm
    with primary sample $(X_{\mathcal{I},\cdot}^{(0)}, y_{\mathcal{I}}^{(0)})$ and auxiliary samples
    $\{X^{(k)}, y^{(k)}\}_{k \in \widehat{G}_l}$. Denote the output as $\hat{\beta}(\widehat{G}_l)$ for $0 \leq l \leq L$.
Step 4. Compute
    $\hat{\theta} =$                                                        (10)
    $\arg\min_{\theta \in \Lambda^{L+1}} \left\{ \widehat{Q}\left(\mathcal{I}^c, \sum_{l=0}^{L} \hat{\beta}(\widehat{G}_l)\theta_l\right) + \sum_{l=0}^{L} \theta_l \widehat{Q}(\mathcal{I}^c, \hat{\beta}(\widehat{G}_l)) + \frac{2\lambda_\theta}{n_0} \sum_{l=0}^{L} \theta_l \log(\theta_l) \right\}$
    for some $\lambda_\theta > 0$. Output
    $\hat{\beta}^{\hat{\theta}} = \sum_{l=0}^{L} \hat{\theta}_l \hat{\beta}(\widehat{G}_l)$.          (11)
```

3.1 | Trans- Lasso算法

我们提出的算法，称为Trans- Lasso，包含两个主要步骤。首先，我们构建一组候选估计器，每个估计器都基于对$\mathcal{A}$的估计。其次，我们对这些候选估计器执行聚合步骤（Dai 等人，2012年，2018年；Rigollet & Tsybakov，2011年）。在适当条件下，聚合估计器在预测方面保证不会比所考虑的最佳候选估计器差很多。出于技术原因，我们需要候选估计器和聚合样本独立。因此，我们从样本拆分开始。我们需要一些额外的符号。对于一般的估计量β，b，将其预测误差的平方和表示为

$$\widehat{Q}(\mathcal{I}, b) = \sum_{i \in \mathcal{I}} \|y_i^{(0)} - (x_i^{(0)})^{\intercal} b\|_2^2,$$

其中$\mathcal{I}$是$\{1, \dots, n_0\}$的子集。令$\Lambda^{L+1} = \{\nu \in \mathbb{R}^{L+1} : \nu_l \geq 0, \sum_{l=0}^{L} \nu_l = 1\}$表示一个$L$-维单纯形。Trans- Lasso算法在算法2中提出。

以示例说明，Trans- Lasso算法的步骤2和3构建一些β, $\widehat{\beta}(\hat{}Gl)$的初始估计。它们通过将每个$\hat{}G_l$视为信息辅助样本集，使用Oracle Trans- Lasso算法计算。我们使用第3.2节中提供的程序构建$\hat{}G_l$作为$\mathcal{A}$的一些估计。步骤4基于Dai等人（2012年）提出的Q-聚合，具有均匀先验、Kullback– Leibler惩罚和简化的调整参数。Q-聚合可以看作是最小二乘聚合和指数聚合的加权版本（Rigollet & Tsybakov，2011年），并且已经证明它在期望和高概率下对模型选择聚合问题都是速率最优的。

```
Algorithm 2: Trans-Lasso Algorithm
Input  : Primary data $(X^{(0)}, y^{(0)})$ and samples from $K$
         auxiliary studies $\{X^{(k)}, y^{(k)}\}_{k=1}^{K}$.
Output: $\hat{\beta}^{\hat{\theta}}$.
Step 1. Let $\mathcal{I}$ be a random subset of $\{1, \dots, n_0\}$ such that
    $|\mathcal{I}| \approx c_0 n_0$ with some constant $0 < c_0 < 1$. Let
    $\mathcal{I}^c = \{1, \dots, n_0\} \setminus \mathcal{I}$.
Step 2. Construct $L+1$ candidate sets of $\mathcal{A}$, $\{\widehat{G}_0, \widehat{G}_1, \dots, \widehat{G}_L\}$
    such that $\widehat{G}_0 = \emptyset$ and $\widehat{G}_1, \dots, \widehat{G}_L$ are based on (14) using
    $\left(X_{\mathcal{I},\cdot}^{(0)}, y_{\mathcal{I}}^{(0)}\right)$ and $\{X^{(k)}, y^{(k)}\}_{k=1}^{K}$.
Step 3. For each $0 \leq l \leq L$, run the Oracle Trans-Lasso algorithm
    with primary sample $(X_{\mathcal{I},\cdot}^{(0)}, y_{\mathcal{I}}^{(0)})$ and auxiliary samples
    $\{X^{(k)}, y^{(k)}\}_{k \in \widehat{G}_l}$. Denote the output as $\hat{\beta}(\widehat{G}_l)$ for $0 \leq l \leq L$.
Step 4. Compute
    $\hat{\theta} =$                                                        (10)
    $\arg\min_{\theta \in \Lambda^{L+1}} \left\{ \widehat{Q}\left(\mathcal{I}^c, \sum_{l=0}^{L} \hat{\beta}(\widehat{G}_l)\theta_l\right) + \sum_{l=0}^{L} \theta_l \widehat{Q}(\mathcal{I}^c, \hat{\beta}(\widehat{G}_l)) + \frac{2\lambda_\theta}{n_0} \sum_{l=0}^{L} \theta_l \log(\theta_l) \right\}$
    for some $\lambda_\theta > 0$. Output
    $\hat{\beta}^{\hat{\theta}} = \sum_{l=0}^{L} \hat{\theta}_l \hat{\beta}(\widehat{G}_l)$.          (11)
```

Model selection aggregation is an effective method for the transfer learning task under consideration. On one hand, it guarantees the robustness of Trans-Lasso in the following sense. Notice that $\widehat{\beta}(\widehat{G}_0)$ corresponds to the single-task Lasso estimator and it is always included in our dictionary. The purpose is that, invoking the property of model selection aggregation, the performance of $\widehat{\beta}^{\hat{\theta}}$ is guaranteed to be not much worse than the performance of the original Lasso estimator under mild conditions. This shows that the performance of Trans-Lasso will not be ruined by adversarial auxiliary samples. Formal statements are provided in Section 3.3. On the other hand, the gain of Trans-Lasso relates to the qualities of $\widehat{G}_1, \ldots, \widehat{G}_L$. If

$$\mathbb{P}\left(\widehat{G}_l \subseteq \mathcal{A}, \text{ for some } 1 \le l \le L\right) \to 1, \tag{12}$$

that is, $\widehat{G}_l$ is a non-empty subset of the informative set $\mathcal{A}$, then the model selection aggregation property implies that the performance of $\widehat{\beta}^{\hat{\theta}}$ is not much worse than the performance of the Oracle Trans-Lasso with $\sum_{k \in \widehat{G}_l} n_k$ informative auxiliary samples. Ideally, one would like to achieve $\widehat{G}_l = \mathcal{A}$ for some $1 \le l \le L$ with high probability. However, it can rely on strong assumptions that may not be guaranteed in practical situations.

To motivate our constructions of $\widehat{G}_l$, let us first point out a naive construction of candidate sets, which consists of 2^K candidates. These candidates are all different combinations of $\{1, ..., K\}$, denoted by $\widehat{G}_1, \ldots, \widehat{G}_{2^K}$. It is obvious that $\mathcal{A}$ is an element of these candidate sets. However, the number of candidates is too large and it can be computationally burdensome. Furthermore, the cost of aggregation can be significantly high, which is of order K/n_0 as will be seen in Lemma 1. In contrast, we would like to pursue a much smaller number of candidate sets such that the cost of aggregation is almost negligible and Equation (12) can be achieved under mild conditions. We introduce our proposed construction of candidate sets in the next subsection.

3.2 | Constructing the candidate sets for aggregation

As illustrated in Section 3.1, the goal of Step 2 is to have a class of candidate sets, $\{\widehat{G}_0, \ldots, \widehat{G}_L\}$, that satisfy (12) under certain conditions. Our idea is to exploit the sparsity patterns of the contrast vectors. Recall that the definition of $\mathcal{A}$ implies that $\{\delta^{(k)}\}_{k \in \mathcal{A}}$ are sparser than $\{\delta^{(k)}\}_{k \in \mathcal{A}^c}$, where $\mathcal{A}^c = \{1, \ldots, K\} \setminus \mathcal{A}$. This property motivates us to find a sparsity index $R^{(k)}$ and its estimator $\widehat{R}^{(k)}$ for each $1 \le k \le K$ such that

$$\max_{k \in \mathcal{A}^o} R^{(k)} < \min_{k \in \mathcal{A}^c} R^{(k)} \quad \text{and} \quad \mathbb{P}\left(\max_{k \in \mathcal{A}^o} \widehat{R}^{(k)} < \min_{k \in \mathcal{A}^c} \widehat{R}^{(k)}\right) \to 1, \tag{13}$$

where $\mathcal{A}^o$ is some subset of $\mathcal{A}$. In words, the sparsity indices in $\mathcal{A}^o$ are no larger than the sparsity indices in $\mathcal{A}^c$ and so are their estimators with high probability. To utilize Equation (13), we can define the candidate sets as

$$\widehat{G}_l = \left\{1 \le k \le K : \widehat{R}^{(k)} \text{ is among the first } l \text{ smallest of all}\right\} \tag{14}$$

for $1 \le l \le K$. That is, $\widehat{G}_l$ is the set of auxiliary samples whose estimated sparsity indices are among the first l smallest. A direct consequence of Equations (13) and (14) is that $\mathbb{P}(\widehat{G}_{|\mathcal{A}^o|} = \mathcal{A}^o) \to 1$ and

模型选择聚合是迁移学习任务的一种有效方法。一方面，它保证了Trans- Lasso的鲁棒性，具体表现为：注意到$\widehat{\beta}(\widehat{G}_0)$对应于单任务Lasso估计器，并且它始终包含在我们的字典中。目的是利用模型选择聚合的性质，保证$\widehat{\beta}^{\hat{\theta}}$在温和条件下表现不会比原始Lasso估计器差很多。这表明Trans- Lasso不会因对抗性辅助样本而性能受损。正式陈述见第3.3节。另一方面，Trans- Lasso的增益与^G1, …, ^GL的质量有关。

$$\mathbb{P}\left(\widehat{G}_l \subseteq \mathcal{A}, \text{ for some } 1 \le l \le L\right) \to 1, \tag{12}$$

也∑是说，如果$\widehat{G}_l$是信息集$\mathcal{A}$的非空子集，那么模型选择聚合性质意味着$\widehat{\beta}^{\hat{\theta}}$的表现不会比带有 $_{k \in \widehat{G}_l} n_k$ 信息辅助样本的Oracle Trans- Lasso差很多。理想情况下，人们希望以高概率实现$\widehat{G}_l = \mathcal{A}$，但实际情况下可能依赖于一些不一定能保证的强假设。

为了激发我们构建$\widehat{G}_l$的思路，让我们首先指出一种简单的候选集构建方法，它由2^K个候选集组成。这些候选集是$\{1, \ldots, K\}$的所有不同组合，表示为^G1, …, ^G2ᴷ。显然，$\mathcal{A}$是这些候选集的一个元素。然而，候选集的数量太大，并且可能计算负担沉重。此外，聚合的成本可能非常高，其阶数为K/n_0，正如引理1中将要看到的。相比之下，我们希望追求数量远小的候选集，使得聚合成本几乎可以忽略不计，并且在温和条件下可以实现方程(12)。我们在下一节中介绍我们提出的候选集构建方法。

3.2 | 构建用于聚合的候选集

如第3.1节所述，步骤2的目标是找到一个候选集类$\{\widehat{G}_0, \ldots, \widehat{G}_L\}$，在特定条件下满足 (12)。我们的想法是利用对比向量的稀疏模式。回想一下$\mathcal{A}$的定义意味着$\{\delta^{(k)}\}_{k \in \mathcal{A}}$比$\{\delta^{(k)}\}_{k \in \mathcal{A}^c}$更稀疏，其中$\mathcal{A}^c = \{1, \ldots, K\} \setminus \mathcal{A}$。这一特性促使我们寻找一个稀疏指数$R^{(k)}$及其估计器^R(k)，对于每个$1 \le k \le K$，使得

$$\max_{k \in \mathcal{A}^o} R^{(k)} < \min_{k \in \mathcal{A}^c} R^{(k)} \quad \text{and} \quad \mathbb{P}\left(\max_{k \in \mathcal{A}^o} \widehat{R}^{(k)} < \min_{k \in \mathcal{A}^c} \widehat{R}^{(k)}\right) \to 1, \tag{13}$$

其中$\mathcal{A}^o$是$\mathcal{A}$的某个子集。用文字来说，$\mathcal{A}^o$中的稀疏指标不大于$\mathcal{A}^c$中的稀疏指标，并且它们的高概率估计量也是如此。为了利用公式(13)，我们可以将候选集定义为

$$\widehat{G}_l = \left\{1 \le k \le K : \widehat{R}^{(k)} \text{ is among the first } l \text{ smallest of all}\right\} \tag{14}$$

对于$1 \le l \le K$.。也就是说，$\widehat{G}_l$是估计稀疏指标在前l个最小的辅助样本的集合。公式(13)和(14)的一个直接推论是$\mathbb{P}(\widehat{G}_{\mathcal{A}^o} = \mathcal{A}^o) \to 1$和 | |

hence the desirable property (12) is satisfied. To achieve the largest gain with transfer learning, we would like to find proper sparsity indices such that Equation (13) holds for $\sum_{k\in\mathcal{A}^o} n_k$ as large as possible. Notice that $\hat{G}_{K+1} = \{1, \dots, K\}$ is always included as candidates according to Equation (14). Hence, in the special cases where all the auxiliary samples are informative or none of the auxiliary samples are informative, it holds that $\hat{G}_{|\mathcal{A}|} = \mathcal{A}$ and the Trans-Lasso is not much worse than the Oracle Trans-Lasso. The more challenging cases are $0 < |\mathcal{A}| < K$.

As $\{\delta^{(k)}\}_{k\in\mathcal{A}^c}$ are not necessarily sparse, the estimation of $\delta^{(k)}$ or functions of $\delta^{(k)}$, $1 \le k \le K$, is not trivial. As an example, an intuitive sparsity index can be $\|\delta^{(k)}\|_1$ and its estimate is $\|\hat{\beta}(\hat{G}_0) - \hat{w}^{(k)}\|_1$, where $\hat{w}^{(k)}$ is the Lasso estimate of $w^{(k)}$ based on the kth study. However, such a Lasso-based estimate is not guaranteed to converge to the oracle $\|\delta^{(k)}\|_1$ when $\delta^{(k)}$ is non-sparse. Therefore, we consider using $R^{(k)} = \|\Sigma\delta^{(k)}\|_2^2$, which is a function of the population-level marginal statistics, as the oracle sparsity index for kth auxiliary sample. The advantage of $R^{(k)}$ is that it has a natural unbiased estimate even when $\delta^{(k)}$ is non-sparse. Let us relate $R^{(k)}$ to the sparsity of $\delta^{(k)}$ using a Bayesian characterization of sparse vectors assuming $\Sigma^{(k)} = \Sigma$ for all $0 \le k \le K$. If $\delta_j^{(k)}$ are *i.i.d.* Laplacian distributed with mean zero and variance v_k^2 for each k, then it follows from the properties of Laplacian distribution (Liu & Kozubowski, 2015) that $\mathbb{E}[\|\delta^{(k)}\|_1] \asymp \mathbb{E}^{1/2}[\|\Sigma\delta^{(k)}\|_2^2]$. Hence, the rank of $\mathbb{E}[\|\Sigma\delta^{(k)}\|_2^2]$ is the same as the rank of $\mathbb{E}[\|\delta^{(k)}\|_1]$. As $\max_{k\in\mathcal{A}} \|\delta^{(k)}\|_1 < \min_{k\in\mathcal{A}^c} \|\delta^{(k)}\|_1$, it is reasonable to expect $\max_{k\in\mathcal{A}} \|\Sigma\delta^{(k)}\|_2^2 < \min_{k\in\mathcal{A}^c} \|\Sigma\delta^{(k)}\|_2^2$. The above derivation holds for many other zero mean prior distributions besides Laplacian. This illustrates our motivation for considering $R^{(k)}$ as the oracle sparsity index.

We next introduce the estimated version, $\hat{R}^{(k)}$, based on the primary data $\{(x_i^{(0)})^\intercal, y_i^{(0)}\}_{i\in\mathcal{I}}$ (after sample splitting) and auxiliary samples $\{X^{(k)}, y^{(k)}\}_{k=1}^K$. We first perform a SURE screening (Fan & Lv, 2008) on the marginal statistics to reduce the effects of random noises. We summarize our proposal for Step 2 of the Trans-Lasso as follows (Algorithm 3). Let $n_* = \min_{0\le k\le K} n_k$.

One can see that $\hat{\Delta}^{(k)}$ are empirical marginal statistics such that $\mathbb{E}[\hat{\Delta}^{(k)}] = \Sigma\delta^{(k)}$ for $k \in \mathcal{A}$. The set $\hat{T}_k$ is the set of first t_* largest marginal statistics for the kth sample. The purpose of screening the marginal statistics is to reduce the magnitude of noise. Notice that the un-screened version $\|\hat{\Delta}^{(k)}\|_2^2$ is a sum of p random variables and it contains noise of order $p/(n_k \wedge n_0)$, which diverges fast as p is much larger than the sample sizes. By screening with t_* of order n_*^α, $\alpha < 1$, the errors

Algorithm 3: Step 2 of the Trans-Lasso Algorithm

Step 2.1. For $1 \le k \le K$, compute the marginal statistics

$$\hat{\Delta}^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_i^{(k)} y_i^{(k)} - \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} x_i^{(0)} y_i^{(0)}. \qquad (15)$$

For each $k \in \{1, \dots, K\}$, let $\hat{T}_k$ be obtained by SURE screening such that

$$\hat{T}_k = \left\{1 \le j \le p : |\hat{\Delta}_j^{(k)}| \text{ is among the first } t_* \text{ largest of all}\right\}$$

for a fixed $t_* = n_*^\alpha$, $0 \le \alpha < 1$.

Step 2.2. Define the estimated sparse index for the k-th auxiliary sample as

$$\hat{R}^{(k)} = \left\|\hat{\Delta}_{\hat{T}_k}^{(k)}\right\|_2^2. \qquad (16)$$

Step 2.3. Compute $\hat{G}_l$ as in (14) for $l = 1, \dots, L$.

因此，理想的属性 (12) 得以满足。为了通过迁移学习实现最大增益，我们希望找到合适的稀疏指数，使得公式 (13) 对于 $\sum_{k\in\mathcal{A}^o} n_k$ 尽可能大。注意到 $\hat{G}_{K+1} = \{1, \dots, K\}$根据方程 (14) 总是作为候选者。因此，在所有辅助样本都是信息量大的或没有辅助样本是信息量大的这些特殊情况下，成立的是 $\hat{G}$

$= \mathcal{A}$ 并且 Trans- Lasso 并不比 Oracle Trans- Lasso 差很多。更具挑战性的情况是 $0 < |\mathcal{A}| < K$.

由于 $\{\delta^{(k)}\}_{k\in\mathcal{A}^c}$ 不一定是稀疏的，因此 $\delta^{(k)}$ 或 $\delta^{(k)}, 1 \le k \le K$的函数的估计并不简单。作为一个例子，一个直观的稀疏指数可以是 $\|\delta^{(k)}\|_1$，其估计是 $\|\hat{\beta}(\hat{G}_0) - \hat{w}^{(k)}\|_1$，其中 $\hat{w}^{(k)}$ 是基于第 k个研究的 Lasso 估计。然而，这种基于 Lasso 的估计并不保证在 $\delta^{(k)}$ 不是稀疏时收敛到 oracle $\|\delta^{(k)}\|_1$ 。因此，我们考虑使用 $R^{(k)} = \|\Sigma\delta^{(k)}\|_2^2$，它是总体层面的边缘统计量的函数，作为第 k个辅助样本的 oracle 稀疏指数。$R^{(k)}$ 的优点是，即使 $\delta^{(k)}$ 不是稀疏的，它也有一个自然的无偏估计。让我们使用稀疏向量的贝叶斯特征将 $R^{(k)}$ 与 $\delta^{(k)}$ 的稀疏性联系起来，假设对所有 $0 \le k \le K$都有 $\Sigma^{(k)} = \Sigma$ 。如果 $\delta_j^{(k)}$ 是 *i.i.d* 拉普拉斯分布的，均值为零，方差为 $v^2{}_k$ 对于每个 k，则根据拉普拉斯分布的性质 (Liu & Kozubowski, 2015)，得到 $\mathbb{E}[\|\delta^{(k)}\|_1] \asymp \mathbb{E}^{1/2}[\|\Sigma\delta^{(k)}\|_2^2]$。因此，$\mathbb{E}[\|\Sigma\delta^{(k)}\|_2^2]$ 的秩与 $\mathbb{E}[\|\delta^{(k)}\|_1]$的秩相同。 As $\max_{k\in\mathcal{A}} \|\delta^{(k)}\|_1 < \min_{k\in\mathcal{A}^c} \|\delta^{(k)}\|_1$，可以合理地期望 $\max_{k\in\mathcal{A}} \|\Sigma\delta^{(k)}\|_2^2 < \min_{k\in\mathcal{A}^c} \|\Sigma\delta^{(k)}\|_2^2$。上述推导对许多其他零均值先验分布也成立，除了拉普拉斯分布。这说明了我们考虑 $R^{(k)}$ 作为 oracle 稀疏指数的动机。

我们接下来介绍基于原始数据$\{(x_i^{(0)})^\intercal, y_i^{(0)}\}_{i\in\mathcal{I}}$ (样本拆分后)和辅助样本$\{X^{(k)}, y^{(k)}\}_{k=1}^K$估计的版本 $\hat{R}^{(k)}$。我们首先对边缘统计量进行SURE筛选(Fan & Lv, 2008)，以减少随机噪声的影响。我们将Trans- Lasso第2步的提案总结如下(算法3)。令$n_* = \min_{0\le k\le K} n_k$。

可以看出$\hat{\Delta}^{(k)}$ 是经验边缘统计量，使得$\mathbb{E}[\hat{\Delta}^{(k)}] = \Sigma\delta^{(k)}$ 对于$k \in \mathcal{A}$。集合$\hat{T}_k$ 是第t_* 个最大的边缘统计量的集合，对应于第k个样本。筛选边缘统计量的目的是减少噪声的幅度。注意未筛选的版本$\|\hat{\Delta}^{(k)}\|_2^2$ 是p个随机变量的和，它包含阶数为$p/(n_k \wedge n_0)$的噪声，当p远大于样本量时，该噪声会快速发散。通过使用t_* 阶数的n_*^α筛选， $\alpha < 1$，误差

Algorithm 3: Step 2 of the Trans-Lasso Algorithm

Step 2.1. For $1 \le k \le K$, compute the marginal statistics

$$\hat{\Delta}^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_i^{(k)} y_i^{(k)} - \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} x_i^{(0)} y_i^{(0)}. \qquad (15)$$

For each $k \in \{1, \dots, K\}$, let $\hat{T}_k$ be obtained by SURE screening such that

$$\hat{T}_k = \left\{1 \le j \le p : |\hat{\Delta}_j^{(k)}| \text{ is among the first } t_* \text{ largest of all}\right\}$$

for a fixed $t_* = n_*^\alpha$, $0 \le \alpha < 1$.

Step 2.2. Define the estimated sparse index for the k-th auxiliary sample as

$$\hat{R}^{(k)} = \left\|\hat{\Delta}_{\hat{T}_k}^{(k)}\right\|_2^2. \qquad (16)$$

Step 2.3. Compute $\hat{G}_l$ as in (14) for $l = 1, \dots, L$.

induced by the random noises is under control. In practice, the auxiliary samples with very small sample sizes can be removed from the analysis as their contributions to the target problem is mild. Desirable choices of $\hat{T}_k$ should keep the variation of $\Sigma\delta^{(k)}$ as much as possible. Under proper conditions, SURE screening can consistently select a set of strong marginal statistics and hence is appropriate for the current purpose. In Step 2.2, we compute $\hat{R}^{(k)}$ based on the marginal statistics which are selected by SURE screening. In practice, different choices of t_* may lead to different realizations of $\hat{G}_l$. One can compute multiple sets of $\{\hat{R}^{(k)}\}_{k=1}^K$ with different t_* which give multiple sets of $\{\hat{G}_l\}_{l=1}^K$. It will be seen from Lemma 1 that a finite number of choices on t_* does not affect the rate of convergence.

3.3 | Theoretical properties of Trans-Lasso

In this subsection, we derive the theoretical guarantees for the Trans-Lasso algorithm. We first establish the model selection aggregation type of results for the Trans-Lasso estimator $\hat{\beta}^{\hat{\theta}}$.

Lemma 1 (Q-aggregation for Trans-Lasso). *Assume that Conditions 1 and 2 hold true. Let $\hat{\theta}$ be computed via Equation (10) with $\lambda_\theta \ge 4\sigma_0^2$. With probability at least* $1-\mathrm{t}$, *it holds that*

$$\frac{1}{|\mathcal{I}^c|}\left\|X^{(0)}_{\mathcal{I}^c,.}(\hat{\beta}^{\hat{\theta}}-\beta)\right\|_2^2 \le \min\nolimits_{0\le l\le L}\frac{1}{|\mathcal{I}^c|}\left\|X^{(0)}_{\mathcal{I}^c,.}(\hat{\beta}(\hat{G}_l)-\beta)\right\|_2^2+\frac{\lambda_\theta\log(L/t)}{n_0}. \qquad (17)$$

If $L \le c_1 n_0$ for some small enough constant c_1, then

$$\left\|\hat{\beta}^{\hat{\theta}}-\beta\right\|_2^2 \lesssim_{\mathbb{P}} \min_{0\le l\le L}\|\hat{\beta}(\hat{G}_l)-\beta\|_2^2+\frac{\log L}{n_0}. \qquad (18)$$

Lemma 1 implies that the performance of $\hat{\beta}^{\hat{\theta}}$ only depends on the best candidate regardless of the performance of other candidates under mild conditions. As commented before, this result guarantees the robustness and efficiency of Trans-Lasso, which can be formally stated as follows. As the original Lasso is always in our dictionary, Equations (17) and (18) imply that $\hat{\beta}^{\hat{\theta}}$ is not much worse than the Lasso in prediction and estimation. Formally, 'not much worse' refers to the last term in Equation (17), which can be viewed as the cost of 'searching' for the best candidate model within the dictionary which is of order $\log L/n_0$. This term is almost negligible, say, when $L = O(K)$, which corresponds to our constructed candidate estimators. This demonstrates the robustness of $\hat{\beta}^{\hat{\theta}}$ to adversarial auxiliary samples. Furthermore, if Equation (12) holds, then the prediction and estimation errors of Trans-Lasso are comparable to the Oracle Trans-Lasso using the auxiliary samples in $\mathcal{A}^o$.

The prediction error bound in Equation (17) follows from Corollary 3.1 in Dai et al. (2012). However, the aggregation methods do not have theoretical guarantees in estimation errors in general. Indeed, an estimator with ℓ_2-error guarantee is crucial for more challenging tasks, such as out-of-sample prediction and inference. For our transfer learning task, we show in Equation (18) that the estimation error is of the same order if the cardinality of the dictionary is $L \le cn_0$ for some small enough c. For our constructed dictionary, it suffices to require $K \le cn_0$. In many practical applications, K is relatively small compared to the sample sizes and hence this assumption is not very restrictive.

由随机噪声引起的误差在可控范围内。在实际应用中，样本量非常小的辅助样本可以从分析中移除，因为它们对目标问题的贡献较小。理想的选项是f $\hat{T}_k$ 应尽可能保留f $\Sigma\delta^{(k)}$ 的变化。在适当条件下，SURE筛选可以一致地选择一组强边缘统计量，因此适用于当前目的。在步骤2.2中，我们基于SURE筛选选择的边缘统计量计算$\hat{R}^{(k)}$。在实际应用中，不同的t_* 选择可能导致$\hat{G}_l$. 的不同实现。可以计算多个具有不同t_* 的$\{\hat{R}^{(k)}\}_{k=1}^K$，从而得到多个$\{\hat{G}_l\}_{l=1}^K$。从引理1可以看出，对$t_*$ 的有限选择不会影响收敛速度。

3.3 |Trans- Lasso的理论性质

在本小节中，我们推导了Trans- Lasso算法的理论保证。我们首先建立了Trans- Lasso估计量$\hat{\beta}^{\hat{\theta}}$的模型选择聚合类型的结果。

引理1 (Q- 聚合 for Trans- Lasso). 假设条件 *1* 和 *2* 成立。令 $\hat{\theta}$ 通过公式 *(10)* 计算，使用 $\lambda_\theta \ge 4\sigma_0^2$，以至少 *1* − t 的概率，它成立。

$$\frac{1}{|\mathcal{I}^c|}\left\|X^{(0)}_{\mathcal{I}^c,.}(\hat{\beta}^{\hat{\theta}}-\beta)\right\|_2^2 \le \min\nolimits_{0\le l\le L}\frac{1}{|\mathcal{I}^c|}\left\|X^{(0)}_{\mathcal{I}^c,.}(\hat{\beta}(\hat{G}_l)-\beta)\right\|_2^2+\frac{\lambda_\theta\log(L/t)}{n_0}. \qquad (17)$$

如果 L $\le c_1 n_0$ 对于某个足够小的常数 c_1，则

$$\left\|\hat{\beta}^{\hat{\theta}}-\beta\right\|_2^2 \lesssim_{\mathbb{P}} \min_{0\le l\le L}\|\hat{\beta}(\hat{G}_l)-\beta\|_2^2+\frac{\log L}{n_0}. \qquad (18)$$

引理1意味着 $\hat{\beta}^{\hat{\theta}}$ 的性能仅取决于最佳候选，而与其他候选的性能无关，在温和条件下。正如之前评论的那样，这个结果保证了Trans- Lasso的鲁棒性和效率，这可以正式表述如下。由于原始Lasso始终在我们的字典中，公式 (17) 和 (18) 意味着 $\hat{\beta}^{\hat{\theta}}$ 在预测和估计方面并不比Lasso差很多。正式地说，“不差很多”指的是公式 (17) 中的最后一项，这可以看作是在字典中搜索最佳候选模型的成本，其数量级为 $\log L/n_0$。这项成本几乎可以忽略不计，比如当 $L = O(K)$，这对应于我们构造的候选估计量。这表明 $\hat{\beta}^{\hat{\theta}}$ 对抗性辅助样本的鲁棒性。此外，如果公式 (12) 成立，那么Trans- Lasso的预测和估计误差与使用 $\mathcal{A}^o$ 中的辅助样本的Oracle Trans- Lasso相当。

方程（17）中的预测误差界限源自 Dai 等人（2012）中的推论 3.1。然而，聚合方法在估计误差方面通常没有理论保证。实际上，对于更具挑战性的任务（如样本外预测和推断），具有 ℓ_2-误差保证的估计器至关重要。对于我们的迁移学习任务，我们在方程（18）中表明，如果字典的基数为 $L \le cn_0$ 且对于足够小的 c，估计误差具有相同的量级。对于我们的构建字典，只需要求 $K \le cn_0$.。在许多实际应用中，K 与样本量相比相对较小，因此这一假设并不十分严格。

In the following, we provide sufficient conditions such that the desirable property (13) holds with $\widehat{R}^{(k)}$ defined in Equation (16) and hence Equation (12) is satisfied. For each $k \in \mathcal{A}^c$, define a set

$$H_k = \left\{1 \le j \le p: \mid \Sigma^{(k)}_{j,.} w^{(k)} - \Sigma^{(0)}_{j,.}\beta \mid > n_*^{-\kappa},\ \kappa < \alpha/2\right\}. \tag{19}$$

Recall that $\alpha < 1$ is defined such that $t_* = n^\alpha$. In fact, H_k is the set of 'strong' marginal statistics that can be consistently selected into $\widehat{T}_k$ for each $k \in \mathcal{A}^c$. We see that $\Sigma^{(k)}_{j,.}w^{(k)} - \Sigma^{(0)}_{j,.}\beta = \Sigma_{j,.}\delta^{(k)}$ if $\Sigma^{(k)} = \Sigma^{(0)}$ for $k \in \mathcal{A}^c$. The definition of $\mathcal{H}_k$ in Equation (19) allows for heterogeneous designs among non-informative auxiliary samples.

Condition 3

(a) *For each $k \in \mathcal{A}^c$, each row of $X^{(k)}$ is i.i.d. Gaussian with mean zero and covariance matrix $\Sigma^{(k)}$ and $\max_{k\in\mathcal{A}^c} \Lambda_{\max}(\Sigma^{(k)})$ is finite. For each $k \in \mathcal{A}^c$, the random noises $\epsilon_i^{(k)}$ are i.i.d. Gaussian with mean zero and variance σ_k^2 and $\mathbb{E}[(y_i^{(k)})^2]$ is finite.*

(b) *It holds that $\log p \vee \log K \le c_1\sqrt{n_*}$ for a small enough constant c_1. Moreover,*

$$\min_{k\in\mathcal{A}^c} \sum_{j\in H_k} |\Sigma^{(k)}_{j,.}w^{(k)} - \Sigma^{(0)}_{j,.}\beta|^2 \ge \frac{c_2 \log p}{n_*^{1-\alpha}} \tag{20}$$

for some constant $c_2 > 0$.

The Gaussian assumptions in Condition 3(a) guarantee the desirable properties of SURE screening for the non-informative auxiliary studies. In fact, the largest eigenvalue of $\Sigma^{(k)}$, $k \in \mathcal{A}^c$ can grow as $O(n_*^\tau)$ for some $\tau \ge 0$ and $\tau + \alpha < 1$ following the proof in Fan and Lv (2008). The Gaussian assumption can be relaxed to be sub-Gaussian random variables according to some recent studies (Ahmed & Bajwa, 2019). For the conciseness of the proof, we consider Gaussian distributed random variables with bounded eigenvalues. Condition 3(b) puts a constraint on the relative dimensions. It is trivial in the regime that $p \vee K \le n_*^\xi$ for any finite $\xi > 0$. The expression (20) requires that for each $k \in \mathcal{A}^c$, there exists a subset of strong marginal statistics with not-so-small cardinality. This condition is mild by choosing α such that $\log p \ll n_*^{1-\alpha}$ and $\alpha = 1/2$ is an obvious choice revoking the first part of Condition 3(b). For instance, if $\min_{k\in\mathcal{A}^c} \|\mathbb{E}[\widehat{\Delta}^{(k)}]\|_\infty \ge c_0 > 0$, then Equation (20) holds with any $\alpha \le 1/2$. In words, a sufficient condition for (20) is that at least one marginal statistic in the kth study is of constant order for $k \in \mathcal{A}^c$. We see that larger n_* makes Condition 3 weaker. As mentioned before, it is helpful to remove the auxiliary samples with very small sample sizes from the analysis.

In the next theorem, we demonstrate the theoretical properties of $\widehat{R}^{(k)}$ and provide a complete analysis of the Trans-Lasso algorithm. Let $\mathcal{A}^o$ be a subset of $\mathcal{A}$ such that

$$\mathcal{A}^o = \left\{k \in \mathcal{A} : \|\Sigma^{(0)}\delta^{(k)}\|_2^2 \le c_1 \min_{k\in\mathcal{A}^c} \sum_{j\in H_k} |\Sigma^{(k)}_{j,.}w^{(k)} - \Sigma^{(0)}_{j,.}\beta|^2\right\}$$

for some small constant $c_1 < 1$ and H_k defined in Equation (19). In general, one can see that the informative auxiliary samples with sparser $\delta^{(k)}$ are more likely to be included into $\mathcal{A}^o$. Specially, the

在以下内容中，我们提供充分条件，使得理想的属性（13）在方程(16)中定义的$\widehat{R}^{(k)}$下成立，从而满足方程(12)。对于每个$k \in \mathcal{A}^c$，定义一个集合

$$H_k = \left\{1 \le j \le p: \mid \Sigma^{(k)}_{j,.} w^{(k)} - \Sigma^{(0)}_{j,.}\beta \mid > n_*^{-\kappa},\ \kappa < \alpha/2\right\}. \tag{19}$$

回想一下 $\alpha < 1$ 被定义使得 $t_* = n^\alpha$。事实上， H_k 是可以在 $\widehat{T}_k$ 中一致选择的“强”边缘统计量的集合，对于每个 $k \in \mathcal{A}^c$。我们看到 $\Sigma^{(k)}_{j\,.}w^{(k)} - \Sigma^{(0)}_{j\,.}\beta = \Sigma_{j\,.}\delta^{(k)}$，，，如果 $\Sigma^{(k)} = \Sigma^{(0)}$ 对于 $k \in \mathcal{A}^c$。方程(19)中$\mathcal{H}_k$ 的定义允许非信息辅助样本之间存在异构设计。

条件3

(a) 对于每个$k \in \mathcal{A}^c$，每个 $X^{(k)}$ 的行是 独立同分布的高斯分布，均值为零，协方差矩阵 $\Sigma^{(k)}$ ，并且 $\max_{k\in\mathcal{A}^c} \Lambda_{\max}(\Sigma^{(k)})$ 是有限的。对于每个 $k \in \mathcal{A}^c$，随机噪声 $\epsilon_i^{(k)}$ 是 独立同分布的高斯分布，均值为零，方差 k，并且 $\mathbb{E}[(y_i^{(k)})^2]$ 是有限的。

(b) 它成立 log$p$ ∨log $K \le c_1$ n_* 对$\sqrt{}$于足够小的常数 c₁。此外，

$$\min_{k\in\mathcal{A}^c} \sum_{j\in H_k} |\Sigma^{(k)}_{j,.}w^{(k)} - \Sigma^{(0)}_{j,.}\beta|^2 \ge \frac{c_2 \log p}{n_*^{1-\alpha}} \tag{20}$$

对于某个常数 $c_2 > 0$。

条件3(a)中的高斯假设保证了SURE筛选对于非信息辅助研究的理想特性。事实上，$\Sigma^{(k)}$，$k \in \mathcal{A}^c$ 的最大特征值可以随着 $O(n_*^\tau)$ 增长，对于某些 $\tau \ge 0$ 和 $\tau + \alpha < 1$，根据Fan 和 Lv (2008)中的证明。根据一些最近的研究 (Ahmed 和 Bajwa, 2019)，高斯假设可以放宽为次高斯随机变量。为了证明的简洁性，我们考虑具有有界特征值的高斯分布随机变量。条件3(b)对相对维度施加了约束。在 $p \vee K \le n_*^\xi$ 对于任何有限 $\xi > 0$的范围内，这是平凡的。表达式 (20) 要求对于每个 $k \in \mathcal{A}^c$，存在一个具有不太小基数的强边缘统计量的子集。通过选择 α 使得 $\log p \ll n_*^{1-\alpha}$，$\alpha = 1/2$ 是一个明显的选择，撤回了条件3(b)的第一部分。例如，如果 $\min_{k\in\mathcal{A}^c}$ $\mathbb{E}[\widehat{\Delta}^{(k)}]$ $_\infty \ge c_0 > 0$，则方程 (20) 对任何 $\alpha \le 1/2$都成立。换句话说，(20) 的一个充分条件是第 k个研究中至少有一个边缘统计量是

对$k \in \mathcal{A}^c$具有恒定阶数。我们看到更大的 n_* 使条件3变弱。如前所述，从分析中移除样本量非常小的辅助样本是有帮助的。

在下一个定理中，我们展示了$\widehat{R}^{(k)}$ 的理论性质，并提供了Trans- Lasso算法的完整分析。设$\mathcal{A}^o$ 是$\mathcal{A}$ 的一个子集，使得

$$\mathcal{A}^o = \left\{k \in \mathcal{A} : \|\Sigma^{(0)}\delta^{(k)}\|_2^2 \le c_1 \min_{k\in\mathcal{A}^c} \sum_{j\in H_k} |\Sigma^{(k)}_{j,.}w^{(k)} - \Sigma^{(0)}_{j,.}\beta|^2\right\}$$

对于某个小的常数$c_1 < 1$和公式(19)中定义的H_k。一般来说，可以看到信息量大的辅助样本具有更稀疏的$\delta^{(k)}$更有可能被包含到$\mathcal{A}^o$中。特别地，

fact that $\max_{k\in\mathcal{A}}\|\Sigma^{(0)}\delta^{(k)}\|_2^2 \le \|\Sigma^{(0)}\|_2^2h^2$ implies $\mathcal{A}^o = \mathcal{A}$ when h is sufficiently small. We will show Equation (13) for such $\mathcal{A}^o$ with $\widehat{R}^{(k)}$ defined in Equation (16). Let $n_{\mathcal{A}^o} = \sum_{k\in\mathcal{A}^o} n_k$.

Theorem 3 (Convergence rate of the Trans-Lasso). *Assume Conditions 1, 2, and 3. Then*

$$\mathbb{P}\left(\max_{k\in\mathcal{A}^o}\widehat{R}^{(k)} < \min_{k\in\mathcal{A}^c}\widehat{R}^{(k)}\right)\to 1. \tag{21}$$

Let $\widehat{\beta}^{\hat{\theta}}$ be computed using the Trans-Lasso algorithm with $\lambda_\theta \ge 4\sigma_0^2$. If $s\log p/(n_{\mathcal{A}^o}+n_0) + \{h(\log p/n_0)^{1/2}\}\wedge(s\log p/n_0) = o(1)$ and $K \le cn_0$ for a sufficiently small constant $c>0$, then

$$\inf_{B\in\Theta_1(s,h)}\mathbb{P}\left(\frac{1}{|\mathcal{I}^c|}\left\|X^{(0)}_{\mathcal{I}^c,\cdot}(\widehat{\beta}^{\hat{\theta}}-\beta)\right\|_2^2 \vee \left\|\widehat{\beta}^{\hat{\theta}}-\beta\right\|_2^2 \lesssim \frac{s\log p}{n_{\mathcal{A}^o}+n_0} + \frac{s\log p}{n_0}\wedge\eta_h + \frac{\log K}{n_0}\right)\to 1 \tag{22}$$

as $(n_0, n_{\mathcal{A}^o}, p)\to\infty$.

Remark 1 Under the conditions of Theorem 3, if

$$\|\Sigma^{(0)}\|_2^2h^2 \le c\min_{k\in\mathcal{A}^c}\sum_{j\in H_k}|\Sigma^{(k)}_{j,\cdot}w^{(k)}-\Sigma^{(0)}_{j,\cdot}\beta|^2 \text{ for some } c<1,$$

then $\mathbb{P}(\max_{k\in\mathcal{A}}\widehat{R}^{(k)} < \min_{k\in\mathcal{A}^c}\widehat{R}^{(k)}) \to 1$ and as $(n_0, n_{\mathcal{A}}, p)\to\infty$,

$$\inf_{B\in\Theta_1(s,h)}\mathbb{P}\left(\frac{1}{|\mathcal{I}^c|}\left\|X^{(0)}_{\mathcal{I}^c,\cdot}(\widehat{\beta}^{\hat{\theta}}-\beta)\right\|_2^2 \vee \left\|\widehat{\beta}^{\hat{\theta}}-\beta\right\|_2^2 \lesssim \frac{s\log p}{n_{\mathcal{A}}+n_0} + \frac{s\log p}{n_0}\wedge\eta_h + \frac{\log K}{n_0}\right)\to 1.$$

Theorem 3 establishes the convergence rate of the Trans-Lasso when $\mathcal{A}$ is unknown. The result in Equation (21) implies the estimated sparse indices in $\mathcal{A}^o$ and in $\mathcal{A}^c$ are separated with high probability. As illustrated before, a consequence of Equation (21) is (12) for the candidate sets $\widehat{G}_l$ defined in Equation (14). Together with Theorem 1 and Lemma 1, we arrive at Equation (22).

It is worth mentioning that Condition 3 is only employed to show the gain of Trans-Lasso. The robustness property of Trans-Lasso holds without any conditions on the non-informative samples (Lemma 1). In practice, missing a few informative auxiliary samples may not be a grave concern. One can see that when $n_{\mathcal{A}^o}$ is large enough such that the first term on the right-hand side of Equation (22) no longer dominates, increasing the number of auxiliary samples will not improve the convergence rate. In contrast, it is more important to guarantee that the estimator is not affected by the adversarial auxiliary samples. The empirical performance of Trans-Lasso is carefully studied in Section 5.

事实是 $\max_{k\in\mathcal{A}}\|\Sigma^{(0)}\delta^{(k)}\|_2^2 \le \|\Sigma^{(0)}\|_2^2h^2$ 意味着 $\mathcal{A}^o = \mathcal{A}$ 当 h 足够小时。我们将展示公式 ()，其中定义在公式 (16) 中。Let$n_{\mathcal{A}^o} =$
13 $\mathcal{A}^o$ $\widehat{R}^{(k)}$
$\sum_{k\in\mathcal{A}^o} n_k$.

定理3 (Trans- Lasso的收敛速度)。假设条件*1*、*2*和*3*。则

$$\mathbb{P}\left(\max_{k\in\mathcal{A}^o}\widehat{R}^{(k)} < \min_{k\in\mathcal{A}^c}\widehat{R}^{(k)}\right)\to 1. \tag{21}$$

令 $\widehat{\beta}^{\hat{\theta}}$ 使用*Trans- Lasso*算法和 $\lambda_\theta \ge 4\sigma_0^2$. 计算。如果 $s\log p/(n_{\mathcal{A}^o} + n0) + \{h(\log p/n0)^{1/2}\}\wedge (s\log p/n0) = o(1)$ 并且 $K \le cn_0$ 对于足够小的常数 *c > 0*，则

$$\inf_{B\in\Theta_1(s,h)}\mathbb{P}\left(\frac{1}{|\mathcal{I}^c|}\left\|X^{(0)}_{\mathcal{I}^c,\cdot}(\widehat{\beta}^{\hat{\theta}}-\beta)\right\|_2^2 \vee \left\|\widehat{\beta}^{\hat{\theta}}-\beta\right\|_2^2 \lesssim \frac{s\log p}{n_{\mathcal{A}^o}+n_0} + \frac{s\log p}{n_0}\wedge\eta_h + \frac{\log K}{n_0}\right)\to 1 \tag{22}$$

as $(n_0, n_{\mathcal{A}^o}, p)\to\infty$.

备注 1 在定理3的条件下，如果

$$\|\Sigma^{(0)}\|_2^2h^2 \le c\min_{k\in\mathcal{A}^c}\sum_{j\in H_k}|\Sigma^{(k)}_{j,\cdot}w^{(k)}-\Sigma^{(0)}_{j,\cdot}\beta|^2 \text{ for some } c<1,$$

then $\mathbb{P}(\max_{k\in\mathcal{A}}\widehat{R}^{(k)} < \min_{k\in\mathcal{A}^c}\widehat{R}^{(k)}) \to 1$ and as $(n_0, n_{\mathcal{A}}, p)\to\infty$,

$$\inf_{B\in\Theta_1(s,h)}\mathbb{P}\left(\frac{1}{|\mathcal{I}^c|}\left\|X^{(0)}_{\mathcal{I}^c,\cdot}(\widehat{\beta}^{\hat{\theta}}-\beta)\right\|_2^2 \vee \left\|\widehat{\beta}^{\hat{\theta}}-\beta\right\|_2^2 \lesssim \frac{s\log p}{n_{\mathcal{A}}+n_0} + \frac{s\log p}{n_0}\wedge\eta_h + \frac{\log K}{n_0}\right)\to 1.$$

定理3建立了Trans- Lasso在$\mathcal{A}$ 未知时的收敛速度。公式(21)中的结果表明，$\mathcal{A}^o$ 和 $\mathcal{A}^c$ 中的估计稀疏指标以高概率是分离的。如前所述，公式(21)的一个推论是对于方程(14)中定义的候选集$\widehat{G}_l$ 的(12)。结合定理1和引理1，我们得到了公式(22)。

值得提到的是，条件3仅用于展示Trans- Lasso的优势。Trans- Lasso的鲁棒性属性在没有关于非信息样本的任何条件下成立（引理1）。在实践中，缺少一些信息辅助样本可能不是一个大问题。可以看到，当$n_{\mathcal{A}^o}$ 足够大，使得公式(22)右侧的第一项不再占主导地位时，增加辅助样本的数量将不会提高收敛速度。相反，更重要的是保证估计器不受对抗性辅助样本的影响。Trans- Lasso的经验性能在第五节中进行了仔细研究。

4 | EXTENSIONS TO HETEROGENEOUS DESIGNS

In this section, we extend the algorithms and theoretical results developed in Sections 2 and 3 to the case where the covariates have different covariance structures in different studies.

The Oracle Trans-Lasso algorithm proposed in Section 2 can be directly applied to the setting where the design matrices are moderately heterogeneous. Formally, we first introduce a relaxed version of Condition 1 as follows. Define

$$C_\Sigma = 1 + \max_{j\le p}\max_{k\in\mathcal{A}} \| e_j^{\intercal}(\Sigma^{(k)} - \Sigma^{(0)}) \left(\sum_{k\in\mathcal{A}} \alpha_k \Sigma^{(k)}\right)^{-1} \|_1,$$

which characterizes the differences between $\Sigma^{(k)}$ and $\Sigma^{(0)}$ for $k \in \mathcal{A}$. Notice that C_Σ is a constant if $\max_{1\le j\le p} \|e_j^{\intercal}(\Sigma^{(k)} - \Sigma^{(0)})\|_0 \le C < \infty$ for all $k \in \mathcal{A}$, where examples include block diagonal $\Sigma^{(k)}$ with constant block sizes or banded $\Sigma^{(k)}$ with constant bandwidths for $k \in \mathcal{A}$.

Condition 4 *For each $k \in \mathcal{A} \cup \{0\}$, each row of $X^{(k)}$ is i.i.d. Gaussian with mean zero and covariance matrix $\Sigma^{(k)}$. The smallest eigenvalue of $\Sigma^{(k)}$ is bounded away from zero for all $k \in \mathcal{A} \cup \{0\}$. The largest eigenvalue of $\Sigma^{(0)}$ is bounded away from infinity.*

The following theorem characterizes the rate of convergence of the Oracle Trans-Lasso estimator in terms of C_Σ. Let $\eta_{h,\Sigma} = (C_\Sigma h\sqrt{\log p/n_0}) \wedge (C_\Sigma^2 h^2)$.

Theorem 4 (Oracle Trans-Lasso with heterogeneous designs). *Assume that Conditions 2 and 4 hold true. Suppose $\mathcal{A}$ is known with $C_\Sigma h \lesssim s\sqrt{\log p/n_0}$ and $n_0 \lesssim n_{\mathcal{A}}$. We take λ_w and λ_δ as in Theorem 1. If $s\log p/n_{\mathcal{A}} + C_\Sigma h(\log p/n_0)^{1/2} = o(1)$, then*

$$\begin{aligned}\inf_{B\in\Theta_1(s,h)} \mathbb{P}\left(\frac{1}{n_0}\|X^{(0)}(\hat\beta-\beta)\|_2^2 \vee \|\hat\beta-\beta\|_2^2 \lesssim \frac{s\log p}{n_{\mathcal{A}}+n_0} + \frac{s\log p}{n_0}\wedge\eta_{h,\Sigma}\right)\\ \ge 1-\exp(-c_1\log p).\end{aligned} \tag{23}$$

The right-hand side of Equation (9) is sharper than $s\log p/n_0$ if $n_{\mathcal{A}} \gg n_0$ and $C_\Sigma h\sqrt{\log p/n_0} \ll s$. We see that small C_Σ is favourable. This implies that the Oracle Trans-Lasso is guaranteed to perform well with sparse contrasts and similar covariance matrices to the primary one.

We now provide theoretical guarantees for the Trans-Lasso with heterogeneous designs when $\mathcal{A}$ is unknown. In this case, the sparsity index $R^{(k)}$ takes the format $\|\Sigma^{(k)}w^{(k)} - \Sigma^{(0)}\beta\|_2^2$. It measures the sparsity of $\delta^{(k)}$ but also the covariance heterogeneity. We consider $\tilde{\mathcal{A}}^o$, a subset of $\mathcal{A}$ such that

$$\tilde{\mathcal{A}}^o = \left\{k\in\mathcal{A}\colon \|\Sigma^{(k)}w^{(k)} - \Sigma^{(0)}\beta\|_2^2 < c_1 \min_{k\in\mathcal{A}^c}\sum_{j\in H_k}|\Sigma^{(k)}_{j,.}w^{(k)} - \Sigma^{(0)}_{j,.}\beta|^2\right\}$$

for some $c_1 < 1$ and H_k defined in Equation (19). This is a generalization of $\mathcal{A}^o$ to the case of heterogeneous designs.

4 | 异构设计的扩展

在本节中，我们将第2节和第3节中开发的算法和理论结果扩展到协变量在不同研究中具有不同协方差结构的情形。

第2节中提出的Oracle Trans- Lasso算法可以直接应用于设计矩阵适度异构的情形。形式上，我们首先引入条件1的放宽版本，如下定义

$$C_\Sigma = 1 + \max_{j\le p}\max_{k\in\mathcal{A}} \| e_j^{\intercal}(\Sigma^{(k)} - \Sigma^{(0)}) \left(\sum_{k\in\mathcal{A}} \alpha_k \Sigma^{(k)}\right)^{-1} \|_1,$$

该式描述了 $\Sigma^{(k)}$ 和 $\Sigma^{(0)}$ 之间的差异$k \in \mathcal{A}$。请注意 C_Σ 是一个常数，如果 $\max_{1\le j\le p}$ $e_j^{\intercal}(\Sigma^{(k)} - \Sigma^{(0)})$ $_0 \le C < \infty$ 对所有 $k \in \mathcal{A}$，其中示例包括具有恒定块大小的块对角 $\Sigma^{(k)}$ 或具有恒定带宽的banded$\Sigma^{(k)}$ 对于 r$k \in \mathcal{A}$.

条件 4 对于每个 $k \in \mathcal{A} \cup \{0\}$， $X^{(k)}$ 的每一行是独立同分布的高斯变量，均值为零，协方差矩阵为 $\Sigma^{(k)}$. 。$\Sigma^{(k)}$ 的最小特征值对所有 $k \in \mathcal{A} \cup \{0\}$ 都有界远离零。最大特征值 *of*$\Sigma^{(0)}$ 有界远离无穷大。

以下定理以 $C\Sigma$ 的形式刻画了 Oracle Trans- Lasso 估计量的收敛速度。设 $\eta_{h\,\Sigma} = (C_\Sigma h\ \ \log p/n_0) \wedge (C_\Sigma^2 h^2)$,.

定理 4 (具有异构设计的 Oracle Trans- Lasso)。假设条件 *2* 和条件 *4* 成立。假设 $\mathcal{A}$ 已知，且 $C_\Sigma h \lesssim s\sqrt{\log p/n_0}$ 和 $n_0 \lesssim n_{\mathcal{A}}$。我们取 λ_w 和 λ_δ 如定理 *1* 中所示。如果 $s\log p/n_{\mathcal{A}} + C_\Sigma h(\log p/n_0)^{1/2} = o(1)$，则

$$\begin{aligned}\inf_{B\in\Theta_1(s,h)} \mathbb{P}\left(\frac{1}{n_0}\|X^{(0)}(\hat\beta-\beta)\|_2^2 \vee \|\hat\beta-\beta\|_2^2 \lesssim \frac{s\log p}{n_{\mathcal{A}}+n_0} + \frac{s\log p}{n_0}\wedge\eta_{h,\Sigma}\right)\\ \ge 1-\exp(-c_1\log p).\end{aligned} \tag{23}$$

方程 (9) 的右侧比 $s\log p/n_0$ if$n_{\mathcal{A}} \gg n_0$and$C_\Sigma h\sqrt{\log p/n_0} \ll s$ 更严格。我们看到小的 C_Σ 是有利的。这意味着 Oracle Trans- Lasso 在稀疏对比和与主要矩阵相似的协方差矩阵的情况下保证表现良好。

我们现在为具有异构设计的 Trans- Lasso 提供理论保证，当$\mathcal{A}$ 未知时。在这种情况下，稀疏指数 $R^{(k)}$ 采用格式 $\Sigma^{(k)}w^{(k)} - \Sigma^{(0)}\beta$ $_2^2$。它衡量了 $\delta^{(k)}$ 的稀疏性，但也衡量了协方差异质性。我们考虑 $\tilde{\mathcal{A}}^o$，$\mathcal{A}$ 的一个子集，

那

$$\tilde{\mathcal{A}}^o = \left\{k\in\mathcal{A}\colon \|\Sigma^{(k)}w^{(k)} - \Sigma^{(0)}\beta\|_2^2 < c_1 \min_{k\in\mathcal{A}^c}\sum_{j\in H_k}|\Sigma^{(k)}_{j,.}w^{(k)} - \Sigma^{(0)}_{j,.}\beta|^2\right\}$$

对于某些 $c_1 < 1$ 和 H_k，它们在公式(19)中定义。这是对 $\mathcal{A}^o$ 在异构设计情况下的推广。

Corollary 1 (Trans-Lasso with heterogeneous designs). *Assume Conditions* 2, 3, *and* 4. *Let* $\hat{\beta}^{\hat{\theta}}$ *be computed via the Trans-Lasso algorithm with* $\lambda_\theta \geq 4\sigma_0^2$. *If* $s\log p/(n_{\hat{\mathcal{A}}^o}+n_0)+\{C_\Sigma h(\log p/n_0)^{1/2}\}\wedge(s\log p/n_0)=o(1)$ *and* $K\leq cn_0$ *for a small enough constant c, then*

$$\inf_{B\in\Theta_1(s,h)}\mathbb{P}\left(\frac{1}{|\mathcal{I}^c|}\left\|X^{(0)}_{\mathcal{I}^c,.}(\hat{\beta}^{\hat{\theta}}-\beta)\right\|_2^2\vee\|\hat{\beta}^{\hat{\theta}}-\beta\|_2^2\lesssim\frac{s\log p}{n_{\hat{\mathcal{A}}^o}+n_0}+\frac{s\log p}{n_0}\wedge\eta_{h,\Sigma}+\frac{\log K}{n_0}\right)\to 1$$

as $(n_0, n_{\hat{\mathcal{A}}^o}, p)\to\infty$.

Corollary 1 provides an upper bound for the Trans-Lasso with heterogeneous designs. The numerical experiments for this setting are studied in Section 5.

5 | SIMULATION STUDIES

In this section, we evaluate the empirical performance of the proposed methods and some other comparable methods in various numerical experiments. Specifically, we evaluate the performance of five methods, including *Lasso*, *Oracle Trans-Lasso* proposed in Section 2.1, *Trans-Lasso* proposed in Section 3.1, and two other ad hoc transfer learning methods related to ours. The first one implements Trans-Lasso except that the bias-correction step (Step 2) of the Oracle Trans-Lasso is omitted. We call this method the '*aggregated Lasso*' (*Agg-Lasso*), as it implements our proposed adaptive aggregation step and applies Lasso to each candidate set. The purpose is to understand the necessity of the bias-correction step in Oracle Trans-Lasso. The second one follows the steps of Trans-Lasso but uses a different aggregation step. Specifically, we consider $\hat{R}^{(k)}=\|\hat{\beta}^L-\hat{w}^{(k)}\|_1$, $k=1,\ldots,K$, where $\hat{\beta}^L$ and $\hat{w}^{(k)}$ are the Lasso estimators based on each of the corresponding studies. Moreover, the Q-aggregation step is replaced with the cross-validation, where we select the set $\hat{G}_l$ that minimizes the out-of-sample prediction errors. We call this algorithm '*ad hoc* ℓ_1*-transfer*'. The purpose of including this method is to understand the performance of our proposed $\hat{R}^{(k)}$ based on SURE screening and Q-aggregation. In the Supplementary Materials, we report the performance of the estimated sparse indices $\hat{R}^{(k)}$ based on Trans-Lasso and ad hoc ℓ_1-transfer. The R code for all the methods are available at https://github.com/saili0103/TransLasso.

5.1 | Identity covariance matrix for the designs

We consider $p=500$, $n_0=150$, and $n_1,\ldots,n_K=100$ for $K=20$. The covariates $x_i^{(k)}$ are *i.i.d.* Gaussian with mean zero and identity covariance matrix for all $0\leq k\leq K$ and $\epsilon_i^{(k)}$ are *i.i.d.* Gaussian with mean zero and variance one for all $0\leq k\leq K$. For the target parameter β, we set $s=16$, $\beta_j=0.3$ for $j\in\{1,\ldots,s\}$, and $\beta_j=0$ otherwise. For the regression coefficients in auxiliary samples, we consider two configurations.

(i) For a given $\mathcal{A}$, if $k\in\mathcal{A}$, let

$$w_j^{(k)}=\beta_j-0.3\mathbb{1}(j\in H_k),$$

推论 1 2, (Trans- Lasso with 3, 4 $\hat{\beta}^{\hat{\theta}}$ heterogeneous designs). *Assume Conditions and . Let be computed via the Trans- Lasso algorithm with* $\lambda_\theta \geq 4\sigma_0^2$. *If* $s\log p/(n_{\hat{\mathcal{A}}^o}+n_0)+\{C\Sigma h(\log p/n_0)^{1/2}\}\wedge(s\log p/n_0)=o(1)$ *and* $K\leq cn_0$ *for a small enough constant c, then*

$$\inf_{B\in\Theta_1(s,h)}\mathbb{P}\left(\frac{1}{|\mathcal{I}^c|}\left\|X^{(0)}_{\mathcal{I}^c,.}(\hat{\beta}^{\hat{\theta}}-\beta)\right\|_2^2\vee\|\hat{\beta}^{\hat{\theta}}-\beta\|_2^2\lesssim\frac{s\log p}{n_{\hat{\mathcal{A}}^o}+n_0}+\frac{s\log p}{n_0}\wedge\eta_{h,\Sigma}+\frac{\log K}{n_0}\right)\to 1$$

as $(n_0, n_{\hat{\mathcal{A}}^o}, p)\to\infty$.

推论 1 为具有异构设计的 Trans- Lasso 提供了上界。该设置下的数值实验研究在第5节。

5 | 模拟研究

在本节中，我们通过各种数值实验评估了所提出的方法和一些其他可比较方法的经验性能。具体来说，我们评估了五种方法的性能，包括 *Lasso*、*Oracle Trans- Lasso*（在第2.1节中提出）、*Trans-Lasso*（在第3.1节中提出）以及与我们相关的另外两种特制迁移学习方法。第一种方法实现了 Trans- Lasso，但省略了 Oracle Trans- Lasso 的偏差校正步骤（步骤2）。我们将此方法称为“聚合 *Lasso*”（*Agg-Lasso*），因为它实现了我们提出的自适应聚合步骤，并将 Lasso 应用于每个候选集。目的是了解 Oracle Trans- Lasso 中偏差校正步骤的必要性。第二种方法遵循 Trans- Lasso 的步骤，但使用不同的聚合步骤。具体来说，我们考虑 $\hat{R}^{(k)}=\hat{\beta}^L-\hat{w}^{(k)}{}_1\ k=1\ \ldots\ K,,,,$ 其中 $\hat{\beta}^L$ 和 $\hat{}w^{(k)}$ 是基于每个相应研究的 Lasso 估计量。此外，Q- 聚合步骤被交叉验证替换，

我们选择使样本外预测误差最小化的集合 $\hat{G}_l$。我们称此算法为“*ad hoc* ℓ_1-迁移”。包含此方法的目的是了解我们基于 SURE 筛选和 Q- 聚合提出的 $\hat{}R(k)$ 的性能。在补充材料中，我们报告了基于 Trans- Lasso 和 ad hoc ℓ_1- 迁移估计的稀疏指标 $\hat{R}^{(k)}$ 的性能。所有方法的 R 代码可在 https://github.com/saili0103/TransLasso 获得。

5.1 | 设计的单位协方差矩阵

我们考虑 $p=500$，$n_0=150$，和 $n_1,\ldots,$ $n_K=100$ 对于 $K=20$。协变量 $x_i^{(k)}$ 是 *i.i.d* 高斯分布，均值为零且对于所有 $0\leq k\leq K$ 和 $\epsilon_i^{(k)}$ 是 *i.i.d* 高斯分布，均值为零且方差为一对于所有 $0\leq k\leq K$。对于目标参数 β，我们设置 $s=16$，$\beta_j=0.3$ 对于 $j\in\{1,\ldots,s\}$，否则为 $\beta_j=0$。对于辅助样本中的回归系数，我们考虑两种配置。

(i) 对于给定的 $\mathcal{A}$，如果 $k\in\mathcal{A}$，则让

$$w_j^{(k)}=\beta_j-0.3\mathbb{1}(j\in H_k),$$

where H_k is a random subset of $[p]$ with $|H_k| = h \in \{2, 6, 12\}$. If $k \notin \mathcal{A}$, we set H_k to be a random subset of $[p]$ with $|H_k| = 2s$ and $w_j^{(k)} = \beta_j - 0.5\mathbb{1}(j \in H_k)$. We set $w_1^{(k)} = -0.3$ for $k = 1, ..., K$.

(ii) For a given $\mathcal{A}$, if $k \in \mathcal{A}$, let $H_k = \{1, \ldots, 100\}$ and

$$w_j^{(k)} = \beta_j + \xi_j \mathbb{1}(k \in H_k), \quad \text{where } \xi_j \sim_{i.i.d.} N(0, h/100),$$

where $h \in \{2, 6, 12\}$ and $N(a, b)$ is the normal with mean a and standard deviation b. If $k \notin \mathcal{A}$, we set $H_k = \{1, \ldots, 100\}$ and

$$w_j^{(k)} = \beta_j + \xi_j \mathbb{1}(j \in H_k), \quad \text{where } \xi_j \sim_{i.i.d.} N(0, 2s/100).$$

We set $w_1^{(k)} = -0.3$ for $k = 1, ..., K$. The setting (i) can be treated as either ℓ_0- or ℓ_1-sparse contrasts. In practice, the true parameters are unknown and we use $\mathcal{A}$ to denote the set of auxiliary samples without distinguishing ℓ_0- or ℓ_1-sparsity. We consider $|\mathcal{A}| \in \{0, 4, 8, \ldots, 20\}$.

In Figure 1, we report sum of squared estimation errors (SSE) for each estimator b, $\|b - \beta\|_2^2$. Each point is summarized from 200 independent simulations. As expected, the performance of

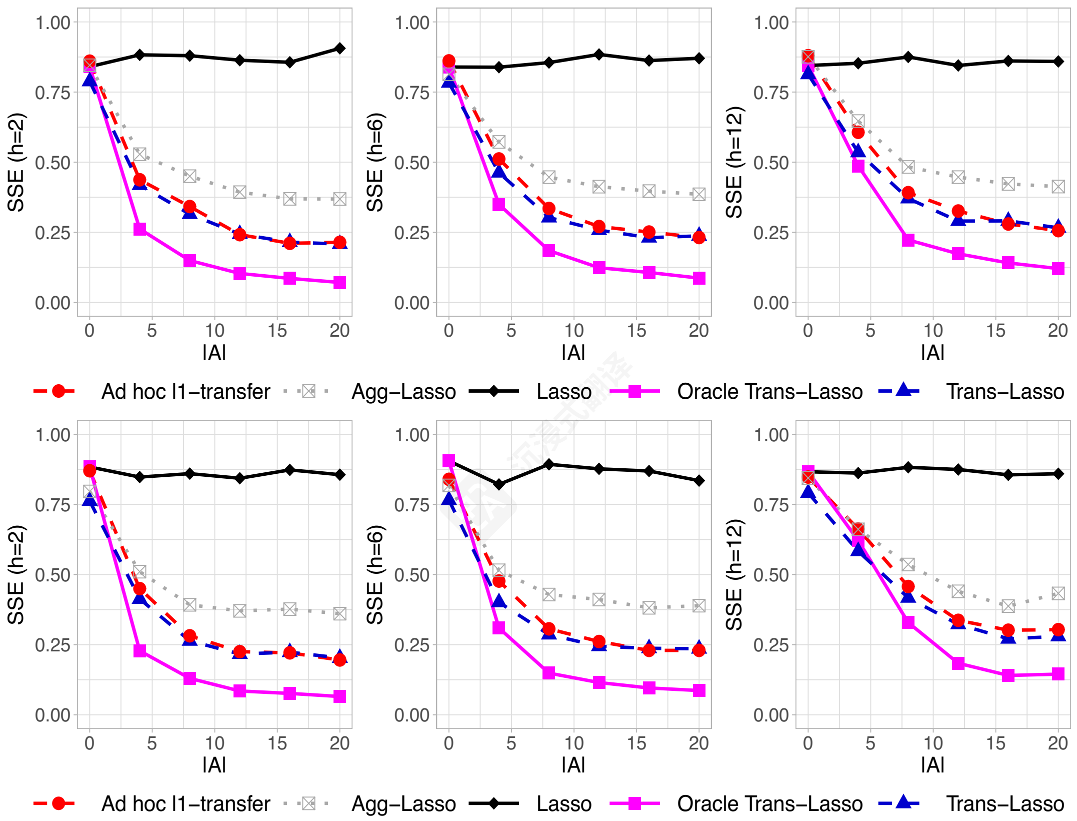

FIGURE 1 Estimation errors of the ad hoc ℓ_1-transfer, Agg-Lasso, Lasso, Oracle Trans-Lasso, and Trans-Lasso with identity covariance matrices of the predictors. The two rows correspond to configurations (i) and (ii) respectively. The y-axis corresponds to $\|b - \beta\|_2^2$ for some estimator b [Colour figure can be viewed at wileyonlinelibrary.com]

在 H_k 是 $[p]$ 的 12 个随机子集 $\mathcal{A}$，我们设置 H_k 为 $[p]$ 的一个随机子集，其中 <style id='22'>H$_k$ = 2 包含 s 和 $w^{(k)}{}_j = \beta_j - 0$。51$(j \in H^k)$。对于 $w_1^{(k)} = -0.3$，我们设置 $k = 1, \ldots, K$。

(ii) 对于给定的 $\mathcal{A}$，如果 $k \in \mathcal{A}$，则令 $H_k = \{1, \ldots, 100\}$

$$w_j^{(k)} = \beta_j + \xi_j \mathbb{1}(k \in H_k), \quad \text{where } \xi_j \sim_{i.i.d.} N(0, h/100),$$

在 $h \in \{2, 6, 12\}$ 和 $N(a, b)$ 是均值为 a 和标准差为 b 的正态分布。如果 $k \notin \mathcal{A}$，我们设置 $H_k = \{1, \ldots, 100\}$ 和

$$w_j^{(k)} = \beta_j + \xi_j \mathbb{1}(j \in H_k), \quad \text{where } \xi_j \sim_{i.i.d.} N(0, 2s/100).$$

我们为 $w_1^{(k)} = -0.3$ 设置 $k = 1, \ldots, K$。设置 (i) 可以被视为 ℓ_0- 或 ℓ_1- 稀疏对比。在实践中，真实参数是未知的，我们使用 $\mathcal{A}$ 来表示辅助样本的集合，而不区分 ℓ_0- 或 ℓ_1- 稀疏性。我们考虑

$|\mathcal{A}| \in \{0, 4, 8, \ldots, 20\}$.

在图1中，我们报告了每个估计器 b $b - \beta$ $^2_2$ 的平方估计误差和 (SSE)。每个点是从200次独立模拟中汇总得到的。正如预期的那样，‖ ‖

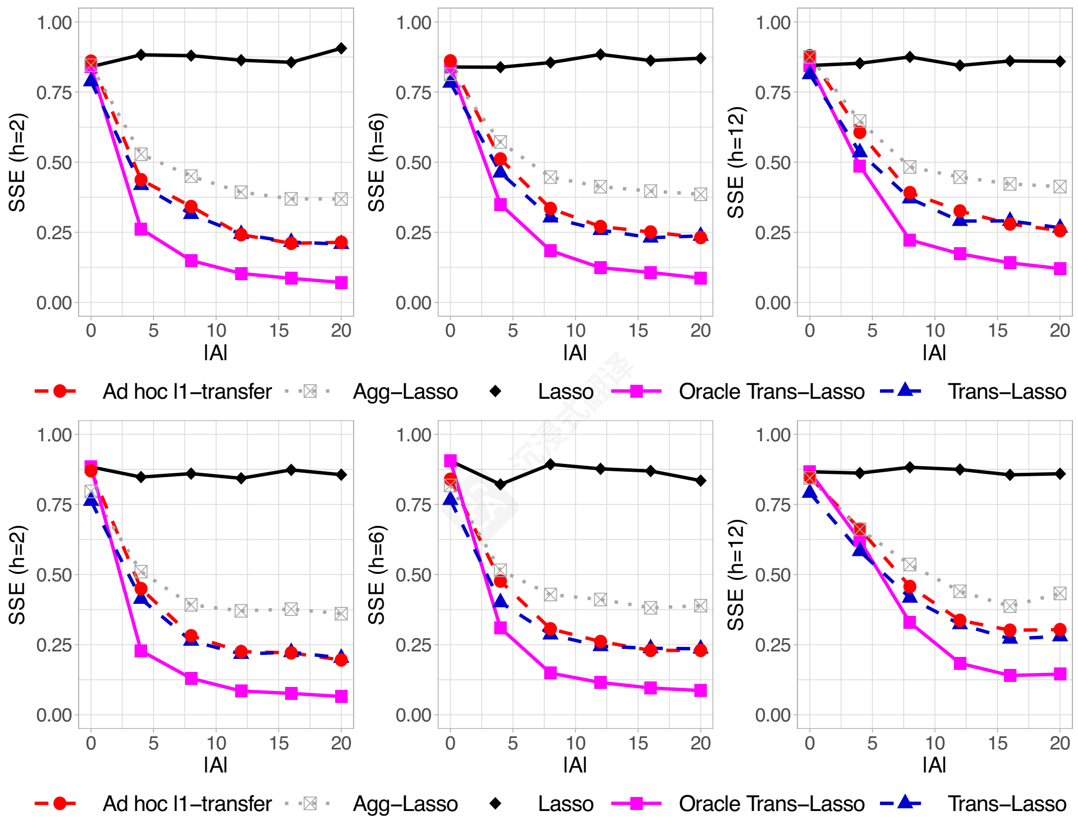

图1 ad hoc ℓ_1-转移、Agg- Lasso、Lasso、Oracle Trans- Lasso 和 Trans- Lasso 的估计误差，其中预测变量的单位协方差矩阵。两行分别对应配置 (i) 和 (ii)。y -轴对应于某些估计器 b [$b - \beta$ $^2_2$。彩色图可查看于 wileyonlinelibrary.com] ‖ ‖

the Lasso does not change as $|\mathcal{A}|$ increases. On the other hand, all four other transfer learning-based algorithms have estimation errors decreasing as $|\mathcal{A}|$ increases. As h increases, the problem gets harder and the estimation errors of all four methods increase. In settings (i) and (ii), the Oracle Trans-Lasso has the smallest estimation errors in most settings. The proposed Trans-Lasso, which is agnostic to $\mathcal{A}$, is always the second best. The gap between the Oracle Trans-Lasso and Trans-Lasso is a result of the uncertainty of aggregation and sample splitting for constructing the initial estimators. We also observe that when $\mathcal{A} = \emptyset$, the Trans-Lasso can have smaller errors than the oracle Trans-Lasso where the latter one does not use auxiliary information. This implies that some auxiliary information can still be borrowed. Due to the randomness of the parameter generation, our definition of $\mathcal{A}$ may not always be the best subset of auxiliary samples that give the smallest estimation errors.

Among the two variants, ad hoc ℓ_1-transfer is also adaptive but has slightly larger estimation errors than Trans-Lasso when h is large. This demonstrates the advantage of Q-aggregation with our proposed sparsity index over the cross-validation type of aggregation with ℓ_1-distance based sparsity index. The Agg-Lasso method has larger estimation errors than Trans-Lasso and ad hoc ℓ_1-transfer, even when h is small. This demonstrates the necessity of the bias-correction step in the Oracle Trans-Lasso.

5.2 | Homogeneous designs among $\mathcal{A} \cup \{0\}$

We now consider $x_i^{(k)}$ as *i.i.d.* Gaussian with mean zero and a equi-correlated covariance matrix, where $\Sigma_{j,j} = 1$ and $\Sigma_{j,k} = 0.8$ if $j \neq k$ for $k \in \mathcal{A} \cup \{0\}$. For $k \notin \mathcal{A} \cup \{0\}$, $x_i^{(k)}$ are *i.i.d.* Gaussian with mean zero and a Toeplitz covariance matrix whose first row is

$$\Sigma^{(k)}_{1,\cdot} = (1, \underbrace{1/(k+1), \ldots, 1/(k+1)}_{2k-1}, 0_{p-2k}). \tag{24}$$

Other true parameters and the dimensions of the samples are set to be the same as in Section 5.1. From the results presented in Figure 2, we see that the Trans-Lasso and Oracle Trans-Lasso have reliable performance in the current setting. The average estimation errors are larger in Figure 2 than those in Section 5.1 as the covariates are highly correlated in the current setting. When h is relatively large, we see that Agg-Lasso and ad hoc ℓ_1-transfer have significantly larger estimation errors than Trans-Lasso. This again demonstrates the advantage of Trans-Lasso over some ad hoc methods.

5.3 | Heterogeneous designs

We next consider a setting where $\Sigma^{(k)}$ are distinct for $k = 0, \ldots, K$. Specifically, for $k = 1, \ldots, K$, let $x_i^{(k)}$ as *i.i.d.* Gaussian with mean zero and a Toeplitz covariance matrix whose first row is Equation (24). Moreover, $\Sigma^{(0)} = I_p$. Other parameters and the dimensions of the samples are set to be the same as in Section 5.1. Figure 3 shows that the general patterns observed under homogeneous designs still hold. Trans-Lasso still gives the best estimation performance under the heterogeneous designs as compared with alternative methods.

Lasso 随 $\mathcal{A}$ 的增加而不会改变。另一方面，所有四种其他基于迁移学习的算法都有估计误差随 $\mathcal{A}$ 的增加而减小。随着 h 的增加，问题变得更难，所有四种方法的估计误差都会增加。在设置 (i) 和 (ii) 中，Oracle Trans- Lasso 在大多数设置中具有最小的估计误差。所提出的 Trans- Lasso，对 $\mathcal{A}$ 无知，始终是第二好的。Oracle Trans- Lasso 和 Trans- Lasso 之间的差距是由于聚合和样本拆分的不确定性，用于构建初始估计器。我们还观察到，当 $\mathcal{A} = \emptyset$ 时，Trans- Lasso 可以比后者（后者不使用辅助信息）具有更小的误差，这表明仍然可以借用一些辅助信息。由于参数生成的随机性，我们的 $\mathcal{A}$ 定义可能不总是给出最小估计误差的辅助样本子集。

在两种变体中，ad hoc ℓ_1-转移也是自适应的，但当 h 较大时，其估计误差略大于 Trans- Lasso。这表明，与基于 ℓ_1-距离的稀疏指数的交叉验证类型聚合相比，我们提出的稀疏指数在 Q-聚合上的优势。Agg- Lasso 方法即使在 h 较小时，其估计误差也大于 Trans- Lasso 和 ad hocℓ_1-转移。这表明 Oracle Trans- Lasso 中的偏差校正步骤的必要性。

5.2 | 同质设计在 $\mathcal{A} \cup \{0\}$

我们现在考虑 $x^{(k)}$ i 作为 *i.i.d.* 高斯分布，均值为零，具有等协相关的协方差矩阵，其中 $\Sigma_{j,j} = 1$ 和 $\Sigma_{j,k} = 0.8$ 如果 $j \neq k$ 对于 $k \in \mathcal{A} \cup \{0\}$。对于 $k \notin \mathcal{A} \cup \{0\}$， $x^{(k)}$ i 是 *i.i.d*。高斯分布，均值为零，具有托普利茨协方差矩阵，其第一行是

$$\Sigma^{(k)}_{1,\cdot} = (1, \underbrace{1/(k+1), \ldots, 1/(k+1)}_{2k-1}, 0_{p-2k}). \tag{24}$$

其他真实参数和样本的维度设置为与第5.1节相同。从图2中展示的结果来看，我们发现Trans- Lasso和Oracle Trans- Lasso在当前设置下具有可靠的性能。图2中的平均估计误差比第5.1节中的更大，因为在当前设置下协变量高度相关。当h相对较大时，我们看到Agg-Lasso和ad hoc ℓ_1- transfer的估计误差明显大于Trans- Lasso。这再次证明了Trans-Lasso相对于某些ad hoc方法的优势。

5.3 | 异构设计

我们接下来考虑一个设置，其中$\Sigma^{(k)}$对于$k = 0, \ldots, K$是不同的。具体来说，对于$k = 1, \ldots, K$，令$x_i^{(k)}$ 作为*i.i.d*高斯分布，均值为零，协方差矩阵为托普利茨协方差矩阵，其第一行是方程(24)。此外，$\Sigma^{(0)} = I_p$. 其他参数和样本的维度设置为与第5.1节相同。图3表明，在异构设计下观察到的通用模式仍然成立。与替代方法相比，Trans- Lasso在异构设计下仍然给出最佳估计性能。

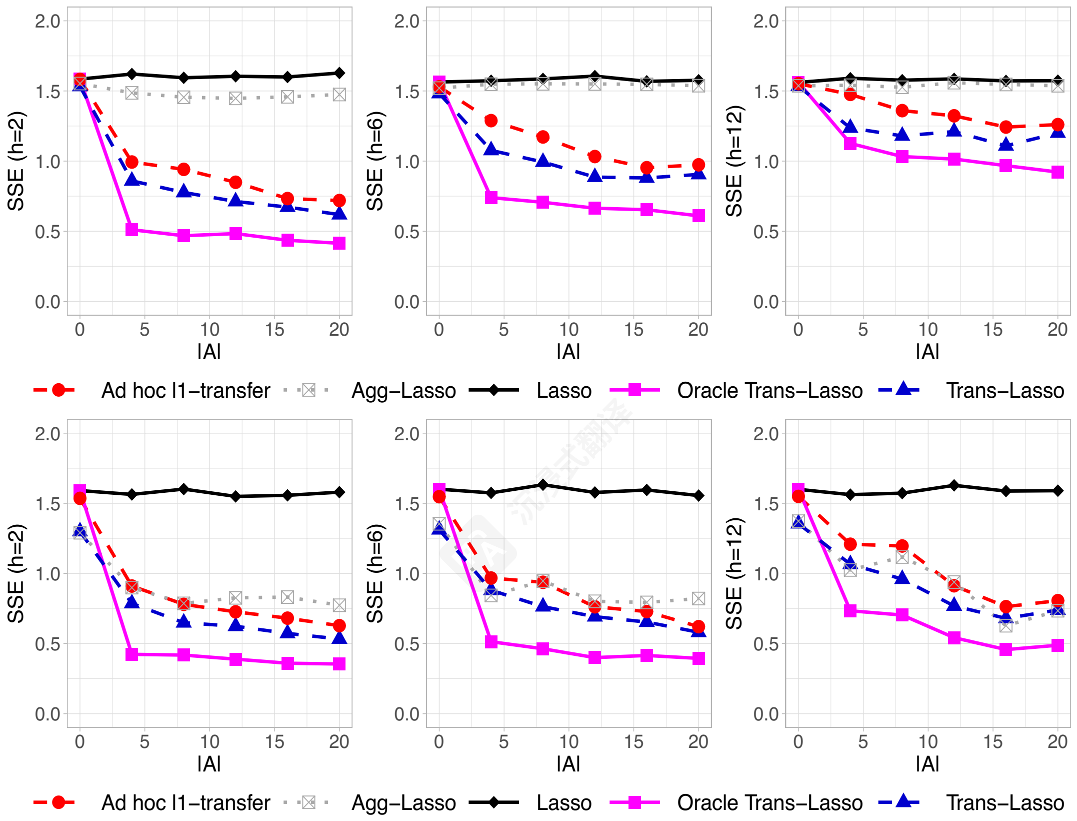

FIGURE 2 Estimation errors of the ad hoc ℓ_1-transfer, Agg-Lasso, Lasso, Oracle Trans-Lasso, and Trans-Lasso with homogeneous covariance matrices. The two rows correspond to configurations (i) and (ii) respectively. The y-axis corresponds to $\|b - \beta\|_2^2$ for some estimator b [Colour figure can be viewed at wileyonlinelibrary.com]

6 | APPLICATION TO GENOTYPE-TISSUE EXPRESSION DATA

In this section, we demonstrate the performance of our proposed transfer learning algorithm in analysing the Genotype-Tissue Expression (GTEx) data (https://gtexportal.org/). Overall, the datasets measure gene expression levels from 49 tissues of 838 human donors, in total comprising 1,207,976 observations of 38,187 genes. In our analysis, we focus on genes that are related to the central nervous system (CNS), which were assembled as MODULE_137 (https://www.gsea-msigdb.org/gsea/msigdb/cards/MODULE_137.html). This module includes a total of 545 genes and additional 1,632 genes that are significantly enriched in the same experiments as the genes of the module. A complete list of genes can be found at http://robotics.stanford.edu/~erans/cancer/modules/module_137.

6.1 | Data analysis method

It is of biological interest to understand the CNS gene regulations in different tissues/cell types. Statistically, we consider predicting the expression levels of a target gene using other

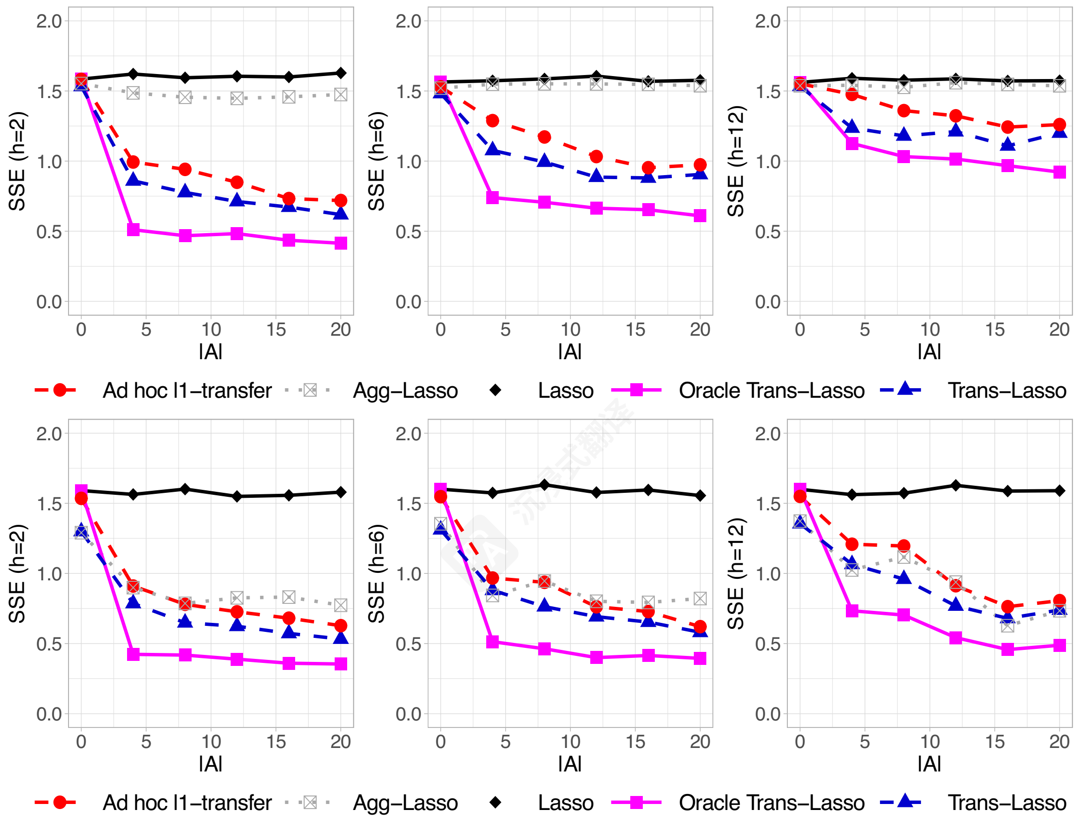

图2特设估计误差ℓ_1-迁移、Agg- Lasso、Lasso、Oracle Trans- Lasso 和 Trans- Lasso 具有同质协方差矩阵的估计误差。两行分别对应配置 (i) 和 (ii)。y-轴对应于$b - \beta$ $^2_2$ 对于某些估计器 b [彩色图可查看于 wileyonlinelibrary.com] ‖ ‖

6 | 一个应用于基因型-组织表达数据的应用

在本节中，我们展示了我们提出的迁移学习算法在分析基因型-组织表达 (GTEx) 数据（https://gtexp ortal.org/）中的性能。总体而言，这些数据集测量了 838 名人类供体中 49 种组织的基因表达水平，总共包含 1,207,976 个 38,187 个基因的观测值。在我们的分析中，我们关注与中枢神经系统 (CNS) 相关的基因，这些基因被组装为 MODULE_137（https://www.gsea- msigdb.org/gsea/msigd b/cards/ MODULE_137.html）。该模块总共包含 545 个基因以及另外 1,632 个基因，这些基因在与模块基因相同的实验中显著富集。基因的完整列表可以在 http://robot ics.stanf ord.edu/~erans/ cance r/ modul es/module_137. 上找到。

6.1 | 数据分析方法

了解不同组织/细胞类型中的中枢神经系统基因调控具有生物学意义。从统计学的角度来看，我们考虑使用其他

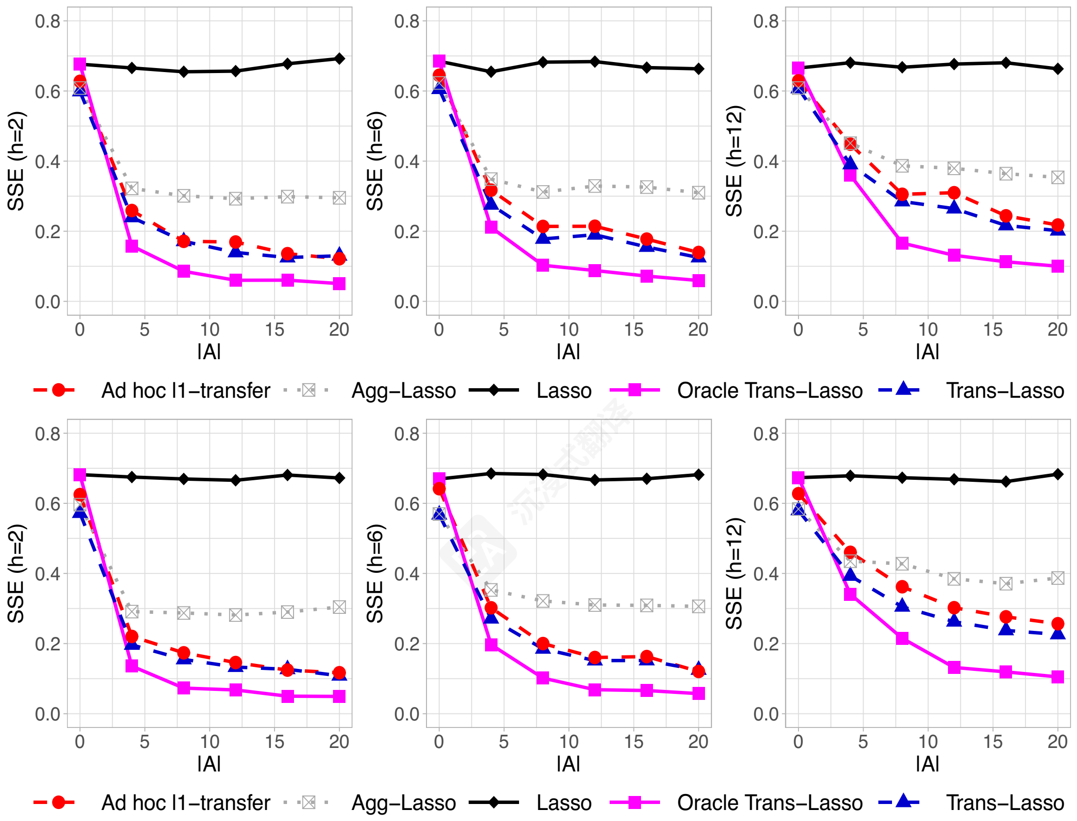

FIGURE 3 Estimation errors of the ad hoc ℓ_1-transfer, Agg-Lasso, Lasso, Oracle Trans-Lasso, and Trans-Lasso with heterogeneous covariance matrices. The two rows correspond to configurations (i) and (ii) respectively. The y-axis corresponds to $\|b-\beta\|_2^2$ for some estimator b [Colour figure can be viewed at wileyonlinelibrary.com]

CNS genes in multiple tissues. Such an analysis provides insights on how other genes regulate the expression of a target gene. To demonstrate the replicability of our proposal, we consider multiple target genes and multiple target tissues and estimate their corresponding models one by one.

For an illustration of the computation process, we consider gene `JAM2` (Junctional adhesion molecule B), as the response variable. `JAM2` is a protein coding gene on chromosome 21 interacting with a variety of immune cell types and may play a role in lymphocyte homing to secondary lymphoid organs Johnson-Léger et al., 2002). Mutations in `JAM2` have been found to cause primary familial brain calcification (Cen et al., 2020; Schottlaender et al., 2020). We consider the association between `JAM2` and other CNS genes in a brain tissue as the target models and the association between `JAM2` and other CNS genes in other tissues as the auxiliary models. As there are multiple brain tissues in the dataset, we treat each of them as the target at each time. The list of target tissues can be found in Figure 4. The min, average, and max of primary sample sizes in these target tissues are 126, 177 and 237 respectively. More information on the target tissues is given in the Supplementary Materials. `JAM2` expresses in 49 tissues in our dataset and we use 47 tissues with more than 120 measurements on `JAM2`. The average number of auxiliary samples for each target model is 14,837 over all the non-target

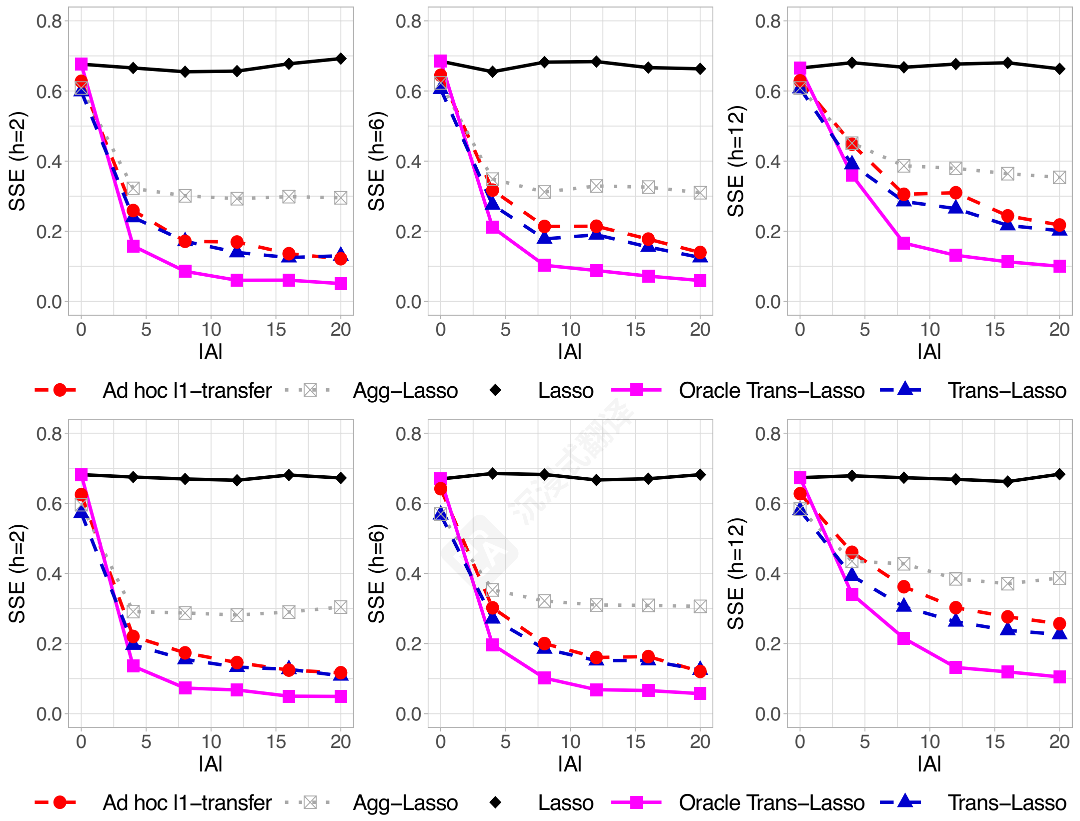

图 3ad hoc 的估计误差ℓ_1- 转移、Agg- Lasso、Lasso、Oracle Trans- Lasso 和 Trans- Lasso 的异质协方差矩阵。两行分别对应配置 (i) 和 (ii)。The y- 轴对应于 $b-\beta$ $^2_2$ 对于某些估计器 b [彩色图可查看于 wileyonlinelibrary.com] ‖ ‖

多种组织中的 CNS 基因。此类分析可提供其他基因如何调控目标基因表达的见解。为证明我们提案的可重复性，我们考虑多个目标基因和多个目标组织，逐一估计其对应模型。

为说明计算过程，我们考虑基因 JAM2（连接粘附分子 B），作为响应变量。JAM2 是位于 21 号染色体上的蛋白质编码基因，与多种免疫细胞类型相互作用，可能参与淋巴细胞归巢至次级淋巴器官（Johnson- Léger 等人，2002）。JAM2 的突变已被发现会导致原发性家族性脑钙化（Cen 等人，2020；Schottlaender 等人，2020）。我们考虑 JAM2 与其他 CNS 基因在脑组织中的关联作为目标模型，JAM2 与其他 CNS 基因在其他组织中的关联作为辅助模型。由于数据集中存在多种脑组织，我们将每种脑组织在每个时间点视为目标。目标组织的列表可在图 4 中找到。这些目标组织中的原始样本量的最小值、平均值和最大值分别为 126、177 和 237。关于目标组织的更多信息，请参见补充材料。在我们的数据集中，JAM2 表达于 49 种组织，我们使用其中 47 种组织，这些组织对 JAM2 的测量值超过 120 个。每个目标模型平均的辅助样本量为 14,837，涵盖所有非目标

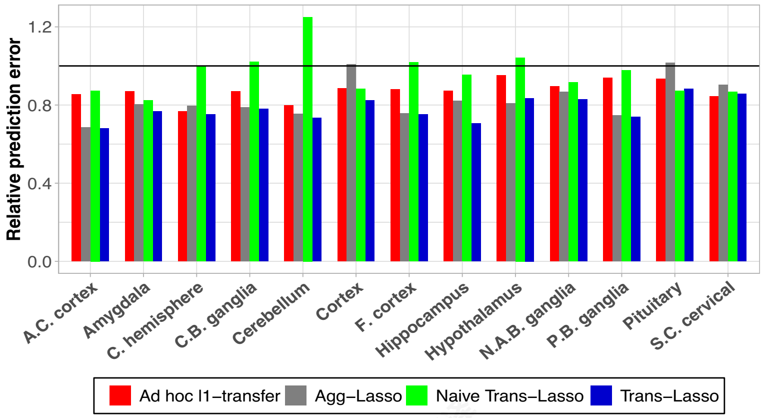

FIGURE 4 Prediction errors of Agg-Lasso, Naive Trans-Lasso, Trans-Lasso and ad hoc ℓ_1-transfer relative to the Lasso evaluated via fivefold cross-validation for gene JAM2 in multiple tissues [Colour figure can be viewed at wileyonlinelibrary.com]

tissues. The covariates in use are the genes that are in the enriched MODULE_137 and do not have missing values in all of the 47 tissues. The final covariates include a total of 1079 genes. The data are standardized before analysis.

We compare the prediction performance of *Trans-Lasso* with *Lasso*, *Agg-Lasso*, *ad hoc*ℓ_1*-transfer*, and *Naive Trans-Lasso*. Implementation of the first four methods is the same as in Section 5. The Naive Trans-Lasso implements the Oracle Trans-Lasso algorithm assuming all the auxiliary studies are informative. Evaluating this method can help us understand the overall informative level of the auxiliary samples. We split the target sample into fivefold and use fourfold to train the algorithms and use the remaining fold to test their prediction performance. We repeat this process five times each with a different fold of test samples. We mention that one individual can provide expression measurements on multiple tissues and these measurements are hard to be independent. While the dependence of the samples can reduce the efficiency of the estimation algorithms, using auxiliary samples may still be beneficial. However, one need to choose proper tuning parameters. The tuning parameter for the Lasso and λ_w are chosen by eightfold cross-validation. The tuning parameter λ_δ is set to be $\lambda_w\sqrt{\sum_{k\in\mathcal{A}} n_k/n_0}$. Other tuning parameters and configurations are the same as for the simulations.

6.2 | Prediction performance of the Trans-Lasso for JAM2 expression

Figure 4 demonstrates the prediction errors of different methods for predicting gene expression JAM2 using other genes. We see that all the transfer learning methods in consideration make improvements over the Lasso in most experiments. The performance of Naive Trans-Lasso implies that there is heterogeneity among tissues and some auxiliary studies can be non-informative. Hence, adaptation to unknown $\mathcal{A}$ is important. Among the adaptive transfer learning methods, Trans-Lasso achieves the smallest prediction errors in almost all the experiments. Its average gain is 22% comparing to the Lasso. This shows that our characterization of the similarity between a target model and a given auxiliary model is suitable for the current problem. Agg-Lasso gives similar prediction errors as Trans-Lasso in most of the tissues but has significantly worse performance for Cortex, Hippocampus, and Pituitary tissues. The average proportion of explained

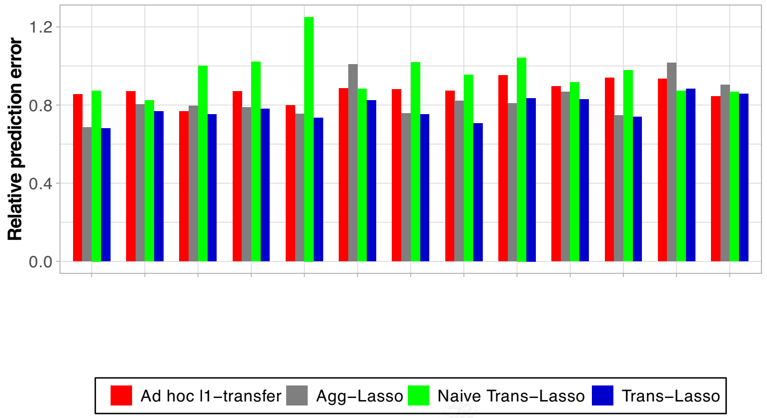

图 URE 4 Agg- Lasso、Naive Trans- Lasso、Trans- Lasso 和 ad hoc ℓ_1- 迁移相对于 Lasso 通过五折交叉验证评估的基因 JAM2 在多种组织中 [彩色图可查看于 wileyonlinelibrary.com]

组织。所使用的协变量是富集的 MODULE_137 中在所有 47 种组织中均无缺失值的基因。最终协变量包括总共 1079 个基因。数据在分析前进行了标准化。

我们比较了 *Trans- Lasso* 与 *Lasso*、 *Agg- Lasso*、 *ad hoc*ℓ_1- 迁移，以及 *Naive Trans-Lasso* 的预测性能。前四种方法的实现与第5节相同。Naive Trans- Lasso 假设所有辅助研究都是信息量大的，实现了 Oracle Trans- Lasso 算法。评估这种方法可以帮助我们了解辅助样本的整体信息量水平。我们将目标样本分成五折，使用四折来训练算法，并使用剩余的折来测试它们的预测性能。我们重复此过程五次，每次使用不同的测试样本折。我们提到一个人可以在多个组织中提供表达测量，这些测量很难是独立的。虽然样本的依赖性会降低估计算法的效率，但使用辅助样本可能仍然有益。然而，需要选择适当的调整参数。Lasso 和 λ_w 的调整参数通过八折交叉验证选择。调整参数 λ_δ 设置为 $\lambda_w\sqrt{\sum_{k\in\mathcal{A}} n_k/n_0}$。 其他调整参数和配置与模拟相同。

6.2 | Trans- Lasso 对 JAM2 表达的预测性能

图 4 展示了使用其他基因预测基因表达 JAM2 的不同方法的预测误差。我们看到，在大多数实验中，所考虑的所有迁移学习方法都比 Lasso 有所改进。Naive Trans- Lasso 的性能表明，组织之间存在异质性，并且一些辅助研究可能没有信息量。因此，适应未知 $\mathcal{A}$ 很重要。在自适应迁移学习方法中，Trans- Lasso 在几乎所有实验中都实现了最小的预测误差。与 Lasso 相比，其平均增益为 22%。这表明我们表征目标模型和给定辅助模型之间相似性的方法适用于当前问题。Agg- Lasso 在大多数组织中给出了与 Trans- Lasso 相似的预测误差，但在皮质、海马体和垂体组织中表现明显较差。解释的平均比例

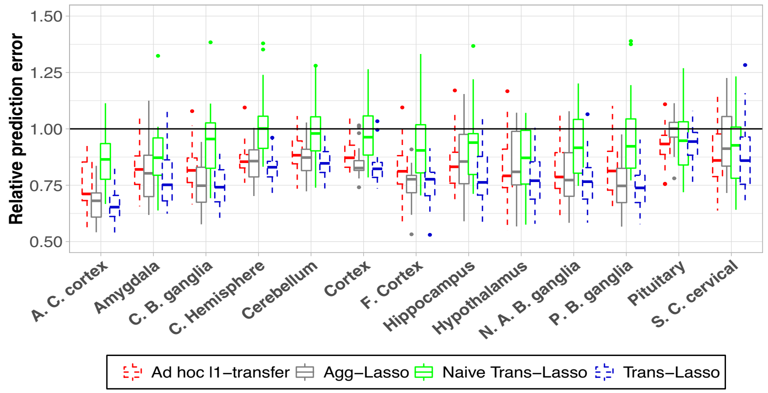

FIGURE 5 Prediction errors of ad hoc ℓ_1-transfer, Agg-Lasso, Naive Trans-Lasso*, and Trans-Lasso relative to the Lasso for the 25 genes on chromosome 21 and in Module_137, in multiple target tissues. The Naive Trans-Lasso has two outliers for the tissue Cerebellum not showing in the figure with values 1.61 and 1.95 [Colour figure can be viewed at wileyonlinelibrary.com]

variance given by the Lasso and that given by the Trans-Lasso are 0.75 and 0.80, respectively, indicating improved fit from transfer learning.

6.3 | Prediction performance of other 25 genes on chromosome 21

To demonstrate the replicability of our proposal, we also consider other genes on chromosome 21 which are in Module_137 as our target genes. We report the overall prediction performance of these 25 genes in Figure 5. A complete list of these genes and some summary information can be found in the Supplementary Materials. Generally speaking, we see that the Trans-Lasso has the best overall performance among all the target tissues when compared to the other two related methods, Agg-Lasso and ad hoc ℓ_1-transfer. The deteriorating performance of the naive Trans-Lasso implies that adaptation to the unknown informative set is crucial for successful knowledge transfer.

7 | DISCUSSION

This paper studies high-dimensional linear regression in the presence of auxiliary samples. The similarity of the target model and a given auxiliary model is characterized by the sparsity of their contrast vectors. Transfer learning algorithms for estimation and prediction are developed that are adaptive to the unknown informative set. Numerical experiments and GTEx data analysis support the theoretical findings and demonstrate its effectiveness in applications.

In the machine learning literature, transfer learning methods have been proposed for different purposes, but few have statistical guarantees. There are several interesting problems related to the present paper for further research. First, transfer learning in nonlinear models can be studied. Using our similarity characterization of the auxiliary studies, transfer learning in high-dimensional generalized linear models (GLMs) can be formulated. GLMs include logistic and Poisson models that are widely used for classification. The main challenge is that the moment equation above (7) is nonlinear and the resulting $\delta^{\mathcal{A}}$ is not necessarily sparse. Hence, transfer learning beyond linear models remain open problems and can be studied under different characterizations for the similarity structure.

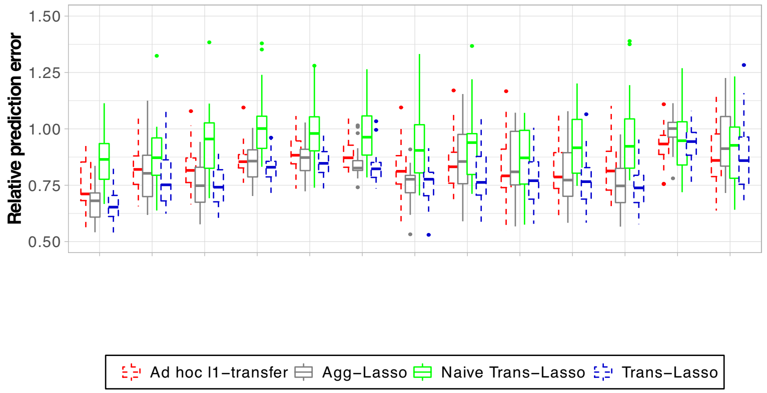

图 5 ad hoc 预测误差ℓ_1- 迁移学习、Agg- Lasso、Naive Trans- Lasso* 和 Trans- Lasso 相对于 Lasso 在 21 号染色体和 Module_137 上的 25 个基因在多个目标组织中的表现。Naive Trans- Lasso 在组织小脑中有两个异常值，未在图中显示，其值为 1.61 和 1。95 [彩色图可查看于 wileyonlinelibrary.com]

Lasso 和 Trans- Lasso 所给出的方差分别为 0.75 和 0.80，这表明通过迁移学习得到了更好的拟合效果。

6.3 | 21 号染色体上其他 25 个基因的预测性能

为了证明我们提案的可重复性，我们还考虑了 21 号染色体上属于 Module_137 的其他基因作为目标基因。我们在图 5 中报告了这些 25 个基因的整体预测性能。这些基因的完整列表和一些总结信息可以在补充材料中找到。总的来说，与其他两种相关方法（Agg- Lasso 和 ad hoc ℓ_1-迁移）相比，Trans- Lasso 在所有目标组织中表现最佳。原始 Trans- Lasso 性能的下降表明，适应未知的信息集对于成功的知识迁移至关重要。

7 | 讨论

本文研究了存在辅助样本时的高维线性回归问题。目标模型与给定辅助模型之间的相似性由其对比向量的稀疏性表征。我们开发了适用于估计和预测的迁移学习算法，这些算法能够适应未知的信息集。数值实验和 GTEx 数据分析支持理论发现，并证明了其在应用中的有效性。

在机器学习文献中，迁移学习方法已被提出用于不同的目的，但很少有统计保证。目前论文有几个有趣的问题值得进一步研究。首先，可以研究非线性模型中的迁移学习。利用我们对辅助研究的相似性表征，可以将高维广义线性模型（GLMs）中的迁移学习进行形式化。GLMs包括广泛用于分类的逻辑回归模型和泊松模型。主要挑战在于上述矩方程（7）是非线性的，并且产生的$\delta^{\mathcal{A}}$不一定是稀疏的。因此，超越线性模型的迁移学习仍然是开放性问题，并且可以在不同的相似性结构表征下进行研究。

Second, it is interesting to study statistical inference, such as constructing confidence intervals and hypothesis testing with auxiliary samples. Given the results derived in this paper, one may expect weaker sample size conditions in the transfer learning setting than those in the single-task setting. It is interesting to provide a precise characterization and to develop a minimax optimal confidence interval in the transfer learning setting.

ACKNOWLEDGEMENT

This research was supported by NIH grants GM129781 and GM123056 and NSF Grant DMS-1712735.

ORCID

Sai Li https://orcid.org/0000-0002-6362-3593
T. Tony Cai https://orcid.org/0000-0002-1673-6296
Hongzhe Li https://orcid.org/0000-0003-3662-3907

REFERENCES

Agarwal, A., Negahban, S. & Wainwright, M.J. (2012) Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. *The Annals of Statistics*, 40(2), 1171–1197.

Ahmed, T. & Bajwa, W.U. (2019) Exsis: extended sure independence screening for ultrahigh-dimensional linear models. *Signal Processing*, 159, 33–48.

Ando, R.K. & Zhang, T. (2005) A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6, 1817–1853.

Bastani, H. (2020). Predicting with proxies: transfer learning in high dimension. *Management Science*, 67(5), 2657–3320.

Bühlmann, P. & van de Geer, S. (2015) High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9(1), 1449–1473.

Cai, T.T. & Wei, H. (2021) Transfer learning for nonparametric classification: minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1), 100–128.

Candes, E. & Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6), 2313–2351.

Cen, Z., Chen, Y., Chen, S., Wang, H., Yang, D., Zhang, H. et al. (2020) Biallelic loss-of-function mutations in JAM2 cause primary familial brain calcification. *Brain*, 143(2), 491–502.

Chen, X., Kim, S., Lin, Q., Carbonell, J.G. & Xing, E.P. (2010) Graph-structured multi-task regression and an efficient optimization method for general fused lasso. *arXiv preprint arXiv:1005.3579*.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2019) Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell*, 179(7), 1469–1482.

Dai, D., Rigollet, P. & Zhang, T. (2012) Deviation optimal learning using greedy *q*-aggregation. *The Annals of Statistics*, 40(3), 1878–1905.

Dai, D., Han, L., Yang, T. & Zhang, T. (2018) Bayesian model averaging with exponentiated least squares loss. *IEEE Transactions on Information Theory*, 64(5), 3331–3345.

Danaher, P., Wang, P. & Witten, D.M. (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2), 373–397.

Daumé III, H. (2007) Frustratingly easy domain adaptation. In: *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 256–263.

Dondelinger, F., Mukherjee, S. & Initiative, A.D.N. (2020) The joint lasso: high-dimensional regression for group structured data. *Biostatistics*, 21(2), 219–235.

Fagny, M., Paulson, J.N., Kuijjer, M.L., Sonawane, A.R., Chen, C.Y., Lopes-Ramos, C.M. et al. (2017) Exploring regulation in tissues with eQTL networks. *Proceedings of the National Academy of Sciences*, 114(37), E7841–E7850.

其次，研究统计推断很有趣，例如使用辅助样本构建置信区间和进行假设检验。根据本文得出的结果，迁移学习设置中的样本量条件可能比单任务设置中的条件更弱。提供精确的特征描述并在迁移学习设置中开发极大极小最优置信区间很有趣。

ACKNOWLEDGEMENT

这项研究得到了 NIH 授款 GM129781 和 GM123056 以及 NSF 授款 DMS- 1712735 的支持。

ORCID

Sai LiT. Tony CaiHongzhe Li

参考文献

Agarwal, A., Negahban, S. & Wainwright, M.J. (2012) Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. *The Annals of Statistics*, 40(2), 1171–1197.

Ahmed, T. & Bajwa, W.U. (2019) Exsis: extended sure independence screening for ultrahigh-dimensional linear models. *Signal Processing*, 159, 33–48.

Ando, R.K. & Zhang, T. (2005) A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6, 1817–1853.

Bastani, H. (2020). Predicting with proxies: transfer learning in high dimension. *Management Science*, 67(5), 2657–3320.

Bühlmann, P. & van de Geer, S. (2015) High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9(1), 1449–1473.

Cai, T.T. & Wei, H. (2021) Transfer learning for nonparametric classification: minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1), 100–128.

Candes, E. & Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6), 2313–2351.

Cen, Z., Chen, Y., Chen, S., Wang, H., Yang, D., Zhang, H. et al. (2020) Biallelic loss-of-function mutations in JAM2 cause primary familial brain calcification. *Brain*, 143(2), 491–502.

Chen, X., Kim, S., Lin, Q., Carbonell, J.G. & Xing, E.P. (2010) Graph-structured multi-task regression and an efficient optimization method for general fused lasso. *arXiv preprint arXiv:1005.3579*.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2019) Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell*, 179(7), 1469–1482.

Dai, D., Rigollet, P. & Zhang, T. (2012) Deviation optimal learning using greedy *q*-aggregation. *The Annals of Statistics*, 40(3), 1878–1905.

Dai, D., Han, L., Yang, T. & Zhang, T. (2018) Bayesian model averaging with exponentiated least squares loss. *IEEE Transactions on Information Theory*, 64(5), 3331–3345.

Danaher, P., Wang, P. & Witten, D.M. (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2), 373–397.

Daumé III, H. (2007) Frustratingly easy domain adaptation. In: *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 256–263.

Dondelinger, F., Mukherjee, S. & Initiative, A.D.N. (2020) The joint lasso: high-dimensional regression for group structured data. *Biostatistics*, 21(2), 219–235.

Fagny, M., Paulson, J.N., Kuijjer, M.L., Sonawane, A.R., Chen, C.Y., Lopes-Ramos, C.M. et al. (2017) Exploring regulation in tissues with eQTL networks. *Proceedings of the National Academy of Sciences*, 114(37), E7841–E7850.

Fan, J. & Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.

Fan, J. & Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849–911.

Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S.M. et al. (2019) A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics*, 51(3), 568–576.

Johnson-Léger, C.A., Aurrand-Lions, M., Beltraminelli, N., Fasel, N. & Imhof, B.A. (2002) Junctional adhesion molecule-2 (JAM-2) promotes lymphocyte transendothelial migration. *Blood, The Journal of the American Society of Hematology*, 100(7), 2479–2486.

Lee, S.H., Ripke, S., Neale, B.M., Faraone, S.V., Purcell, S.M., Perlis, R.H. et al. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, 45, 984–994.

Li, Y.R., Li, J., Zhao, S.D., Bradfield, J.P., Mentch, F.D., Maggadottir, S.M. et al. (2015) Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nature Medicine*, 21, 1018–1027.

Li, S., Cai, T.T. & Li, H. (2020) Supplements to “Transfer learning for high-dimensional linear regression: prediction, estimation, and minimax optimality”.

Liu, Y. & Kozubowski, T. J. (2015) A folded laplace distribution. *Journal of Statistical Distributions and Applications*, 2(1), 1–17.

Lounici, K., Pontil, M. & Tsybakov, A.B. (2009) Taking advantage of sparsity in multi-task learning. *arXiv:0903.1468*.

Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X. & Sham, P.C. (2017) Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, 41(6), 469–480.

Mei, S., Fei, W. & Zhou, S. (2011) Gene ontology based transfer learning for protein subcellular localization. *BMC Bioinformatics*, 12, 44.

Pan, W. & Yang, Q. (2013) Transfer learning in heterogeneous collaborative filtering domains. *Artificial Intelligence*, 197, 39–55.

Pierson, E., GTEx Consortium, Koller, D., Battle, A. & Mostafavi, S. (2015) Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Computational Biology*, 11(5), e1004220.

Raskutti, G., Wainwright, M.J. & Yu, B. (2011) Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10), 6976–6994.

Rigollet, P. & Tsybakov, A. (2011) Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2), 731–771.

Schottlaender, L.V., Abeti, R., Jaunmuktane, Z., Macmillan, C., Chelban, V., O’callaghan, B. et al. (2020) Bi-allelic JAM2 variants lead to early-onset recessive primary familial brain calcification. *The American Journal of Human Genetics*, 106(3), 412–421.

Shin, H.-C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I. et al. (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5), 1285–1298.

Sun, Y.V. & Hu, Y.-J. (2016) Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. In: *Advances in Genetics*, vol. 93, pp. 147–190.

Sun, T. & Zhang, C.-H. (2012) Scaled sparse linear regression. *Biometrika*, 99(4), 879–898.

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

Torrey, L. & Shavlik, J. (2010) Transfer learning. In: Olivas, E.S., Guerrero, J.D.M., Sober, M.M., Benedito, J.R.M. & Lopez, A.J.S. (Eds.), *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI Global, pp. 242–264. Hershey, PA: Information Science Reference - Imprint of: IGI Publishing.

Tsybakov, A.B. (2014) Aggregation and minimax optimality in high-dimensional estimation. In: Proceedings of the international congress of mathematicians, vol. 3, pp. 225–246.

Turki, T., Wei, Z. & Wang, J.T. (2017) Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. *IEEE Access*, 5, 7381–7393.

Verzelen, N. (2012) Minimax risks for sparse regressions: ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6, 38–90.

Wang, S., Shi, X., Wu, M. & Ma, S. (2019) Horizontal and vertical integrative analysis methods for mental disorders omics data. *Scientific Reports*, 9(1), 1–12.

Fan, J. & Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.

Fan, J. & Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849–911.

Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S.M. et al. (2019) A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics*, 51(3), 568–576.

Johnson-Léger, C.A., Aurrand-Lions, M., Beltraminelli, N., Fasel, N. & Imhof, B.A. (2002) Junctional adhesion molecule-2 (JAM-2) promotes lymphocyte transendothelial migration. *Blood, The Journal of the American Society of Hematology*, 100(7), 2479–2486.

Lee, S.H., Ripke, S., Neale, B.M., Faraone, S.V., Purcell, S.M., Perlis, R.H. et al. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, 45, 984–994.

Li, Y.R., Li, J., Zhao, S.D., Bradfield, J.P., Mentch, F.D., Maggadottir, S.M. et al. (2015) Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nature Medicine*, 21, 1018–1027.

Li, S., Cai, T.T. & Li, H. (2020) Supplements to “Transfer learning for high-dimensional linear regression: prediction, estimation, and minimax optimality”.

Liu, Y. & Kozubowski, T. J. (2015) A folded laplace distribution. *Journal of Statistical Distributions and Applications*, 2(1), 1–17.

Lounici, K., Pontil, M. & Tsybakov, A.B. (2009) Taking advantage of sparsity in multi-task learning. *arXiv:0903.1468*.

Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X. & Sham, P.C. (2017) Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, 41(6), 469–480.

Mei, S., Fei, W. & Zhou, S. (2011) Gene ontology based transfer learning for protein subcellular localization. *BMC Bioinformatics*, 12, 44.

Pan, W. & Yang, Q. (2013) Transfer learning in heterogeneous collaborative filtering domains. *Artificial Intelligence*, 197, 39–55.

Pierson, E., GTEx Consortium, Koller, D., Battle, A. & Mostafavi, S. (2015) Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Computational Biology*, 11(5), e1004220.

Raskutti, G., Wainwright, M.J. & Yu, B. (2011) Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10), 6976–6994.

Rigollet, P. & Tsybakov, A. (2011) Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2), 731–771.

Schottlaender, L.V., Abeti, R., Jaunmuktane, Z., Macmillan, C., Chelban, V., O’callaghan, B. et al. (2020) Bi-allelic JAM2 variants lead to early-onset recessive primary familial brain calcification. *The American Journal of Human Genetics*, 106(3), 412–421.

Shin, H.-C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I. et al. (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5), 1285–1298.

Sun, Y.V. & Hu, Y.-J. (2016) Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. In: *Advances in Genetics*, vol. 93, pp. 147–190.

Sun, T. & Zhang, C.-H. (2012) Scaled sparse linear regression. *Biometrika*, 99(4), 879–898.

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

Torrey, L. & Shavlik, J. (2010) Transfer learning. In: Olivas, E.S., Guerrero, J.D.M., Sober, M.M., Benedito, J.R.M. & Lopez, A.J.S. (Eds.), *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI Global, pp. 242–264. Hershey, PA: Information Science Reference - Imprint of: IGI Publishing.

Tsybakov, A.B. (2014) Aggregation and minimax optimality in high-dimensional estimation. In: Proceedings of the international congress of mathematicians, vol. 3, pp. 225–246.

Turki, T., Wei, Z. & Wang, J.T. (2017) Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. *IEEE Access*, 5, 7381–7393.

Verzelen, N. (2012) Minimax risks for sparse regressions: ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6, 38–90.

Wang, S., Shi, X., Wu, M. & Ma, S. (2019) Horizontal and vertical integrative analysis methods for mental disorders omics data. *Scientific Reports*, 9(1), 1–12.

Weiss, K., Khoshgoftaar, T.M. & Wang, D. (2016) A survey of transfer learning. *Journal of Big Data*, 3, 9.

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2), 894–942.

Zhernakova, A., Van Diemen, C.C. & Wijmenga, C. (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nature Reviews Genetics*, 10(1), 43–55.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Li, S., Cai, T.T. & Li, H. (2022) Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 149–173. https://doi.org/10.1111/rssb.12479

Weiss, K., Khoshgoftaar, T.M. & Wang, D. (2016) A survey of transfer learning. *Journal of Big Data*, 3, 9.

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2), 894–942.

Zhernakova, A., Van Diemen, C.C. & Wijmenga, C. (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nature Reviews Genetics*, 10(1), 43–55.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.

支持信息

附加支持信息可能可以在出版商网站的文章在线版本中找到。

如何引用本文： 李, S., 蔡, T.T. & 李, H. (2022) 迁移学习用于高维线性回归：预测、估计和极大极小最优性。 皇家统计学会杂志：*B*辑（统计方法学）, 84, 149– 173. https://doi.org/10.1111/rssb.12479