Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation, and Minimax Optimality

Sai Li

Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

T. Tony Cai

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

Hongzhe Li

Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

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.

1. Introduction

Modern scientific research is characterized by massive and diverse data sets. It is of significant interest to integrate different data sets to make a more accurate prediction and statistical inference. Given a target problem to solve, transfer learning (Torrey and 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 labeled data for cars but some labeled 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 and 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

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

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

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

where $w^{(k)} \in \mathbb{R}^p$ is the regression vector for the k-th 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 k-th 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 are those whose contrasts are sufficiently sparse:

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

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 favorable. 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 etiology, including various autoimmune diseases (Li et al., 2015; Zhernakova et al., 2009) and psychiatric disorders (Lee et al., 2013; Cross-Disorder Group of the Psychiatric Genomics Consortium, 2019). The similarity characterization we proposed captures the sparse nature of genome-wide association data and shared genetic etiology 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 (Pierson et al., 2015; Fagny et al., 2017). 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 (Ando and Zhang, 2005; Lounici et al., 2009; Agarwal et al., 2012), where the goal is to estimate multiple mod-

els 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 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 and 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.

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 (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 k-th 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 such that its j-th 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 $|a_n/b_n| \leq c$ 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 $\mathbb{P}(|a_n/b_n| \leq c) \to 1$ 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 general. We then correct its bias using the primary data in the second step. Algorithm 1 formally presents our proposed Oracle Trans-Lasso algorithm.

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: $\hat{\beta}$

transfer step

Step 1. Compute

$$\hat{w}^{\mathcal{A}} = \arg\min_{w\in\mathbb{R}^p} \left\{ \frac{1}{2n_{\mathcal{A}}} \sum_{k\in\mathcal{A}} \|y^{(k)} - X^{(k)}w\|_2^2 + \lambda_w \|w\|_1 \right\} \tag{4}$$

for $\lambda_w = c_1\sqrt{\log p/n_{\mathcal{A}}}$ with some constant c_1.

debiasing step?

Step 2. Let

$$\hat{\beta} = \hat{w}^{\mathcal{A}} + \hat{\delta}^{\mathcal{A}}, \tag{5}$$

where

$$\hat{\delta}^{\mathcal{A}} = \arg\min_{\delta\in\mathbb{R}^p} \left\{ \frac{1}{2n_0} \|y^{(0)} - X^{(0)}(\hat{w}^{\mathcal{A}} + \delta)\|_2^2 + \lambda_\delta \|\delta\|_1 \right\} \tag{6}$$

for $\lambda_\delta = c_2\sqrt{\log p/n_0}$ with some constant c_2.

In Step 1, $\hat{w}^{\mathcal{A}}$ 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 $\hat{w}^{\mathcal{A}}$, $w^{\mathcal{A}}$, 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 and van de Geer, 2015) and moment estimators. The estimator $\hat{w}^{\mathcal{A}}$ 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.

2.2. Theoretical Properties of Oracle Trans-Lasso

Formally, the parameter space we consider can be written as

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

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 $\max_{k\in\mathcal{A}\cup\{0\}} \mathbb{E}[\exp\{t\epsilon_i^{(k)}\}] \le \exp\{t^2 C_0\}$ 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 Condition 1 and Condition 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*

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

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, i.e., $s\log p = o(n_{\mathcal{A}})$ 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 (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 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 $\hat{w}^{\mathcal{A}}$ and that of $\hat{\delta}^{\mathcal{A}}$. 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 (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 and 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 Condition 1 and Condition 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 $k\in\mathcal{A}$, i.e., 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 centered 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.

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 (Rigollet and Tsybakov, 2011; Dai et al., 2012, 2018) 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 $\{1, \ldots, n_0\}$. 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, $\hat{\beta}(\widehat{G}_l)$. 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 and 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 $\{1, \ldots, n_0\}$ such that $|\mathcal{I}| \approx c_0 n_0$ with some constant $0 < c_0 < 1$. Let $\mathcal{I}^c = \{1, \ldots, n_0\} \setminus \mathcal{I}$.

Step 2. Construct $L+1$ candidate sets of $\mathcal{A}$, $\{\widehat{G}_0, \widehat{G}_1, \ldots, \widehat{G}_L\}$ such that $\widehat{G}_0 = \emptyset$ and $\widehat{G}_1, \ldots, \widehat{G}_L$ are based on (14) using $\left(X^{(0)}_{\mathcal{I},\cdot}, y^{(0)}_{\mathcal{I}}\right)$ and $\{X^{(k)}, y^{(k)}\}_{k=1}^{K}$.

Step 3. For each $0 \le l \le L$, run the Oracle Trans-Lasso algorithm with primary sample $(X^{(0)}_{\mathcal{I},\cdot}, y^{(0)}_{\mathcal{I}})$ and auxiliary samples $\{X^{(k)}, y^{(k)}\}_{k \in \widehat{G}_l}$. Denote the output as $\hat{\beta}(\widehat{G}_l)$ for $0 \le l \le L$.

Step 4. Compute

$$\hat{\theta} = \tag{10}$$
$$\underset{\theta \in \Lambda^{L+1}}{\arg\min} \left\{ \widehat{Q}\big(\mathcal{I}^c, \sum_{l=0}^{L} \hat{\beta}(\widehat{G}_l)\theta_l\big) + \sum_{l=0}^{L} \theta_l \widehat{Q}(\mathcal{I}^c, \hat{\beta}(\widehat{G}_l)) + \frac{2\lambda_\theta}{n_0} \sum_{l=0}^{L} \theta_l \log(\theta_l) \right\}$$

for some $\lambda_\theta > 0$. Output

$$\hat{\beta}^{\hat{\theta}} = \sum_{l=0}^{L} \hat{\theta}_l \hat{\beta}(\widehat{G}_l). \tag{11}$$

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 $\hat{\beta}(\widehat{G}_0)$ 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 $\hat{\beta}^{\hat{\theta}}$ 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

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

i.e., $\widehat{G}_l$ is a non-empty subset of the informative set $\mathcal{A}$, then the model selection aggregation property implies that the performance of $\hat{\beta}^{\hat{\theta}}$ 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 $1 \leq l \leq L$ 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 $\{1, \ldots, K\}$, 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 (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 $1 \leq k \leq K$ 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 (13), we can define the candidate sets as

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

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 (13) and (14) is that $\mathbb{P}(\widehat{G}_{|\mathcal{A}^o|} = \mathcal{A}^o) \to 1$ and 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 (13) holds for $\sum_{k\in\mathcal{A}^o} n_k$ as large as possible. Notice that $\widehat{G}_{K+1} = \{1, \ldots, K\}$ is always included as candidates according to (14). Hence, in the special cases where all the auxiliary samples are informative or none of the auxiliary samples are informative, it holds that $\widehat{G}_{|\mathcal{A}|} = \mathcal{A}$ 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 $\|\hat{\beta}(\widehat{G}_0) - \hat{w}^{(k)}\|_1$, where $\hat{w}^{(k)}$ is the Lasso estimate of $w^{(k)}$ based on the k-th 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 k-th 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 ν_k^2 for each k, then it follows from the properties of Laplacian distribution (Liu and 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, $\widehat{R}^{(k)}$, 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 and 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 $n_* = \min_{0\le k\le K} n_k$.

Algorithm 3: Step 2 of the Trans-Lasso Algorithm

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

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

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

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

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

$$\widehat{R}^{(k)} = \Big\|\widehat{\Delta}^{(k)}_{\widehat{T}_k}\Big\|_2^2. \tag{16}$$

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

One can see that $\widehat{\Delta}^{(k)}$ are empirical marginal statistics such that $\mathbb{E}[\widehat{\Delta}^{(k)}] = \Sigma\delta^{(k)}$ for $k \in \mathcal{A}$. The set $\widehat{T}_k$ is the set of first t_* largest marginal statistics for the k-th sample. The purpose of screening the marginal statistics is to reduce the magnitude of noise. Notice that the un-screened version $\|\widehat{\Delta}^{(k)}\|_2^2$ 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 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 $\widehat{T}_k$ 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 $\widehat{R}^{(k)}$ based on the marginal statistics which are selected by SURE screening. In practice, different choices of t_* may lead to different realizations of $\widehat{G}_l$. One can compute multiple sets of $\{\widehat{R}^{(k)}\}_{k=1}^K$ with different t_* which give multiple sets of $\{\widehat{G}_l\}_{l=1}^K$. 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 Condition 1 and Condition 2 hold true. Let $\hat{\theta}$ be computed via (10) with $\lambda_\theta \geq 4\sigma_0^2$. With probability at least $1-t$, it holds that*

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

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

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

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, (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 (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 (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 (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 (18) that the estimation error is of the same order if the cardinality of the dictionary is $L \leq cn_0$ for some small enough c. For our constructed dictionary, it suffices to require $K \leq cn_0$. In many practical applications, K is relatively small compared to the sample sizes and hence this assumption is not very restrictive.

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

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

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 $\Sigma_{j,\cdot}^{(k)} w^{(k)} - \Sigma_{j,\cdot}^{(0)} \beta = \Sigma_{j,\cdot} \delta^{(k)}$ if $\Sigma^{(k)} = \Sigma^{(0)}$ for $k \in \mathcal{A}^c$. The definition of $\mathcal{H}_k$ in (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 $\log p \vee \log K \leq c_1 \sqrt{n_*}$ for a small enough constant c_1. Moreover,

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

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 $\tau \geq 0$ 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 and 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 $p \vee K \leq n_*^\xi$

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 $\min_{k\in\mathcal{A}^c} \|\mathbb{E}[\widehat{\Delta}^{(k)}]\|_\infty \geq c_0 > 0$, then (20) holds with any $\alpha \leq 1/2$. In words, a sufficient condition for (20) is that at least one marginal statistic in the k-th 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

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

for some small constant $c_1 < 1$ and H_k defined in (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 fact that $\max_{k\in\mathcal{A}} \|\Sigma^{(0)}\delta^{(k)}\|_2^2 \leq \|\Sigma^{(0)}\|_2^2 h^2$ implies $\mathcal{A}^o = \mathcal{A}$ when h is sufficiently small. We will show (13) for such $\mathcal{A}^o$ with $\widehat{R}^{(k)}$ defined in (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 $\hat{\beta}^{\hat{\theta}}$ be computed using the Trans-Lasso algorithm with $\lambda_\theta \geq 4\sigma_0^2$. 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 $K \leq cn_0$ for a sufficiently small constant $c > 0$, then

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

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

REMARK 1. Under the conditions of Theorem 3, if

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

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

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

Theorem 3 establishes the convergence rate of the Trans-Lasso when $\mathcal{A}$ is unknown. The result in (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 (21) is (12) for the candidate sets $\widehat{G}_l$ defined in (14). Together with Theorem 1 and Lemma 1, we arrive at (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 (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.

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

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

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 $\Sigma^{(k)}$ are 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 Condition 2 and Condition 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*

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

The right-hand side of (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 favorable. 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

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

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

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 $s\log p/(n_{\tilde{\mathcal{A}}^o}+n_0)+\{C_\Sigma h(\log p/n_0)^{1/2}\}\wedge(s\log p/n_0) = o(1)$ and $K \leq cn_0$ for a small enough constant c, then*

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

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

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 $\widehat{R}^{(k)} = \|\hat{\beta}^L - \hat{w}^{(k)}\|_1$, $k = 1,\ldots,K$, 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 $\widehat{G}_l$ 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 $\widehat{R}^{(k)}$ based on SURE screening and Q-aggregation. In the Supplementary Materials, we report the performance of the estimated sparse indices $\widehat{\tilde{R}}^{(k)}$ 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

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

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 $k = 1, \ldots, K$.

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

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

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

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

We set $w_1^{(k)} = -0.3$ for $k = 1, \ldots, K$. 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 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

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

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.

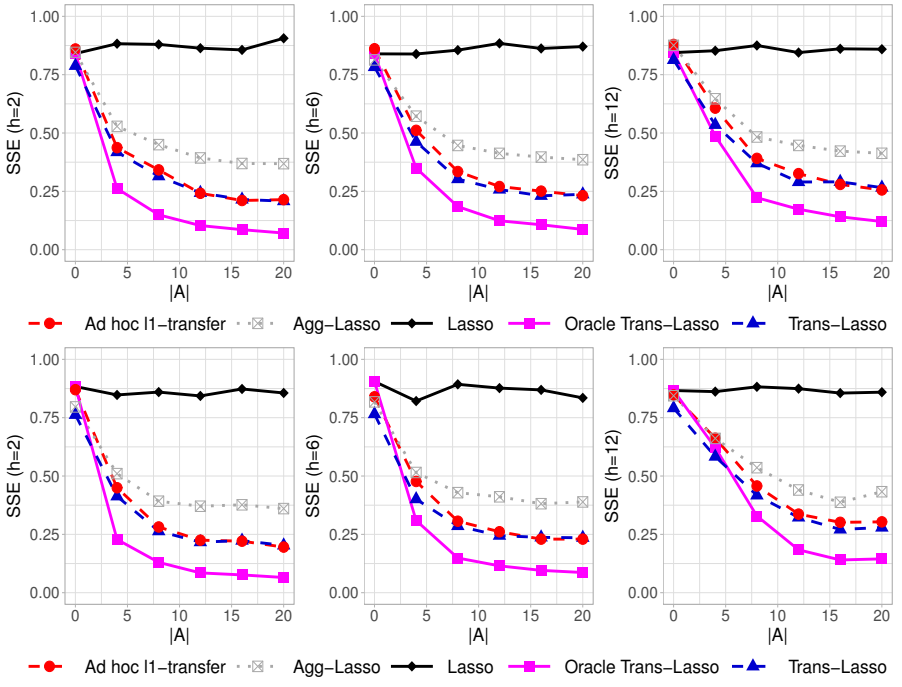

Figure 1. Estimation errors of the Ad hoc ℓ_1-transfer, Agg-Lasso, Lasso, Oracle Trans-Lasso, and Trans-Lasso with identity covariance matrices of the predictors. The two rows correspond to configurations (i) and (ii), respectively. The y-axis corresponds to $\|b - \beta\|_2^2$ for some estimator b.

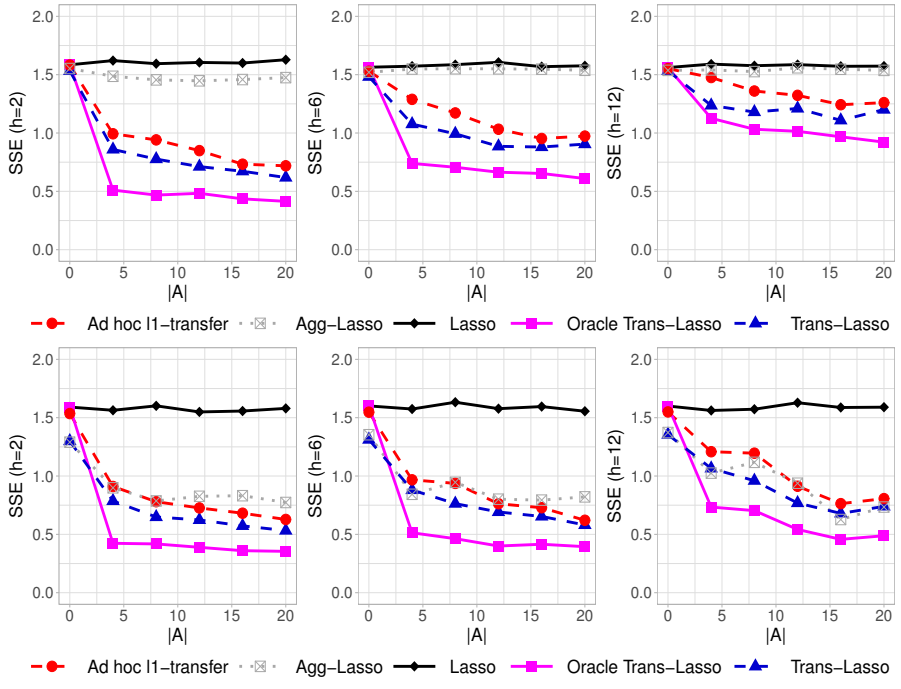

Figure 2. Estimation errors of the Ad hoc ℓ_1-transfer, Agg-Lasso, Lasso, Oracle Trans-Lasso, and Trans-Lasso with homogeneous covariance matrices. The two rows correspond to configurations (i) and (ii), respectively. The y-axis corresponds to $\|b - \beta\|_2^2$ for some estimator b.

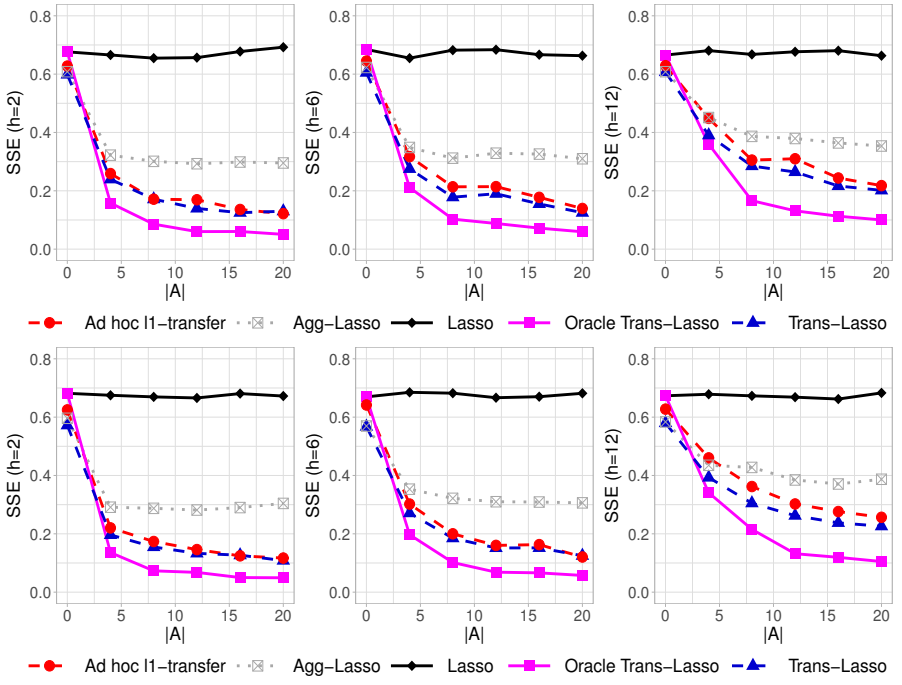

Figure 3. Estimation errors of the Ad hoc ℓ_1-transfer, Agg-Lasso, Lasso, Oracle Trans-Lasso, and Trans-Lasso with heterogeneous covariance matrices. The two rows correspond to configurations (i) and (ii), respectively. The y-axis corresponds to $\|b - \beta\|_2^2$ for some estimator b.

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 (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.

6. Application to Genotype-Tissue Expression Data

In this section, we demonstrate the performance of our proposed transfer learning algorithm in analyzing the Genotype-Tissue Expression (GTEx) data (`https://gtexportal.org/`). Overall, the data sets 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 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` has 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` ex-

presses 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 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 1,079 genes. The data is 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 five folds and use four folds 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 8-fold 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 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.

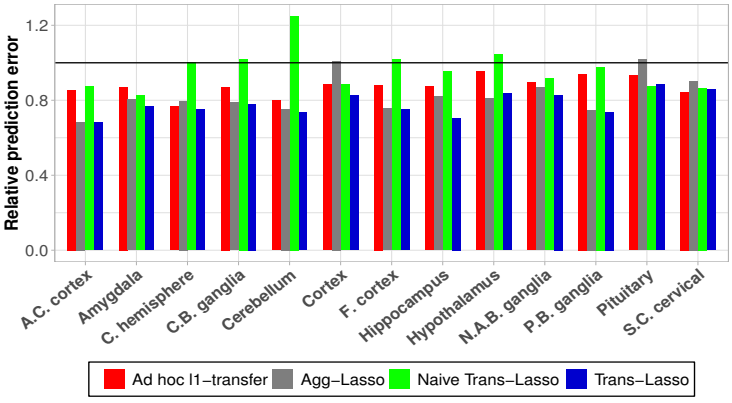

Figure 4. Prediction errors of Agg-Lasso, Naive Trans-Lasso,Trans-Lasso, and Ad hoc ℓ_1-transfer relative to the Lasso evaluated via 5-fold cross validation for gene JAM2 in multiple tissues.

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.

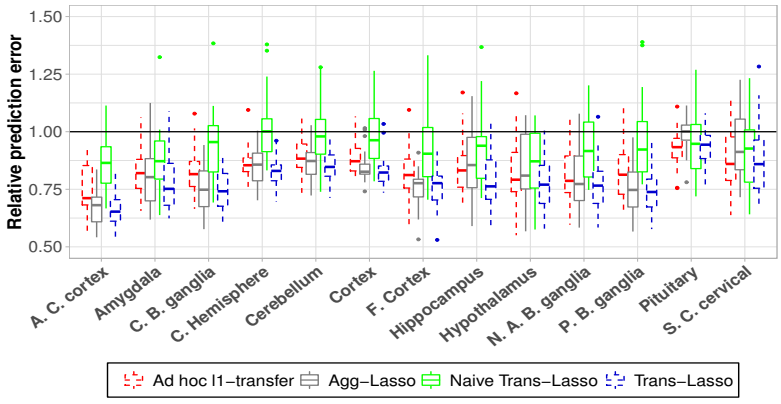

Figure 5. Prediction errors of Ad hoc ℓ_1-transfer, Agg-Lasso, Naive Trans-Lasso*, and Trans-Lasso relative to the Lasso for the 25 genes on chromosome 21 and in Module_137, in multiple target tissues. The Naive Trans-Lasso has two outliers for the tissue Cerebellum not showing in the figure with values 1.61 and 1.95.

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. 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.

Acknowledgments

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

References

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

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

Ando, R. K. and T. Zhang (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. and S. van de Geer (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics 9*(1), 1449–1473.

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

Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics 35*(6), 2313–2351.

Cen, Z., Y. Chen, S. Chen, et al. (2020). Biallelic loss-of-function mutations in jam2 cause primary familial brain calcification. *Brain 143*(2), 491–502.

Chen, X., S. Kim, Q. Lin, J. G. Carbonell, and E. P. Xing (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., L. Han, T. Yang, and T. Zhang (2018). Bayesian Model Averaging with Exponentiated Least Squares Loss. *IEEE Transactions on Information Theory 64*(5), 3331–3345.

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

Danaher, P., P. Wang, and D. M. Witten (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., S. Mukherjee, and A. D. N. Initiative (2020). The joint lasso: high-dimensional regression for group structured data. *Biostatistics 21*(2), 219–235.

Fagny, M., J. N. Paulson, M. L. Kuijjer, et al. (2017). Exploring regulation in tissues with eqtl networks. *Proceedings of the National Academy of Sciences 114*(37), E7841–E7850.

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

Fan, J. and J. Lv (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., M. Li, Q. Lu, et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature genetics 51*(3), 568–576.

Johnson-Léger, C. A., M. Aurrand-Lions, N. Beltraminelli, et al. (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., S. Ripke, B. M. Neale, et al. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide snps. *Nature genetics 45*, 984–994.

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

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

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

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

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

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

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

Pierson, E., D. Koller, A. Battle, et al. (2015). Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Comput Biol 11*(5), e1004220.

Raskutti, G., M. J. Wainwright, and B. Yu (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. and A. Tsybakov (2011). Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics 39*(2), 731–771.

Schottlaender, L. V., R. Abeti, Z. Jaunmuktane, 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., H. R. Roth, M. Gao, 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, T. and C.-H. Zhang (2012). Scaled sparse linear regression. *Biometrika 99*(4), 879–898.

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

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. and J. Shavlik (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242–264. IGI Global.

Tsybakov, A. B. (2014). Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians*, Volume 3, pp. 225–246.

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

Weiss, K., T. M. Khoshgoftaar, and D. Wang (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., C. C. Van Diemen, and C. Wijmenga (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.