Guided Transfer Learning for High-Dimensional Linear Regression

Xinhao Qu

(School of Economics, Xiamen University, Xiamen 361005, China)

Abstract This protocol considers transfer learning for high-dimensional linear regression with prior knowledge or data-driven guidance, aiming at enhancing estimation performance under limited target data. After describing the motivation behind, this protocol shows the algorithm design, also the improved estimation and prediction performance in simulation and empirical study.

Keywords Adaptive Lasso; Transfer Learning; High-Dimensional Linear Regression

1 Convergence Rate of Trans-Lasso Could Be Non-Sharp

In model settings as Li *et al.* (2022), target data generating process is formatted within framework of high-dimensional linear regression

$$Y^{(0)} = \mathbf{X}^{(0)}\beta_0 + \varepsilon^{(0)}, \quad \mathbb{E}[\varepsilon^{(0)} \mid \mathbf{X}^{(0)}] = 0 \tag{1}$$

where data $Y^{(0)} \in R^{n_0}$, $\mathbf{X}^{(0)} \in R^{n_0 \times p}$ *i.i.d.*, parameter $\beta_0 \in R^p$.

k-th oracle source model is formatted as

$$Y^{(k)} = \mathbf{X}^{(k)}\beta_0^{(k)} + \varepsilon^{(k)}, \quad \mathbb{E}[\varepsilon^{(k)} \mid \mathbf{X}^{(k)}] = 0, \quad \forall k \in \{1, 2, ..., K\} \tag{2}$$

where K is the total number of transferrable sources, $Y^{(0)} \in R^{n_k}$, $\mathbf{X}^{(k)} \in R^{n_k \times p}$ *i.i.d.*, parameter $\beta_0^{(k)} \in R^p$, $n_0 << n_k,\ \forall k \in \{1, 2, ..., K\}$, $p >> n_0$, while the parameter space focused is

$$\Theta_q(s, h) \equiv \left\{ \left(\beta_0, \delta_0^{(1)}, \ldots, \delta_0^{(K)}\right) : \|\beta_0\|_0 \leq s, \max_{k \in \{1,\ldots,K\}} \left\|\delta_0^{(k)}\right\|_q \leq h \right\} \tag{3}$$

$\delta_0^{(k)} \equiv \left|\beta_0 - \beta_0^{(k)}\right|$, $q \in [0, 1]$. Th.1 of Li *et al.* (2022) derives the convergence rate of oracle Trans-Lasso under $q = 1$ as

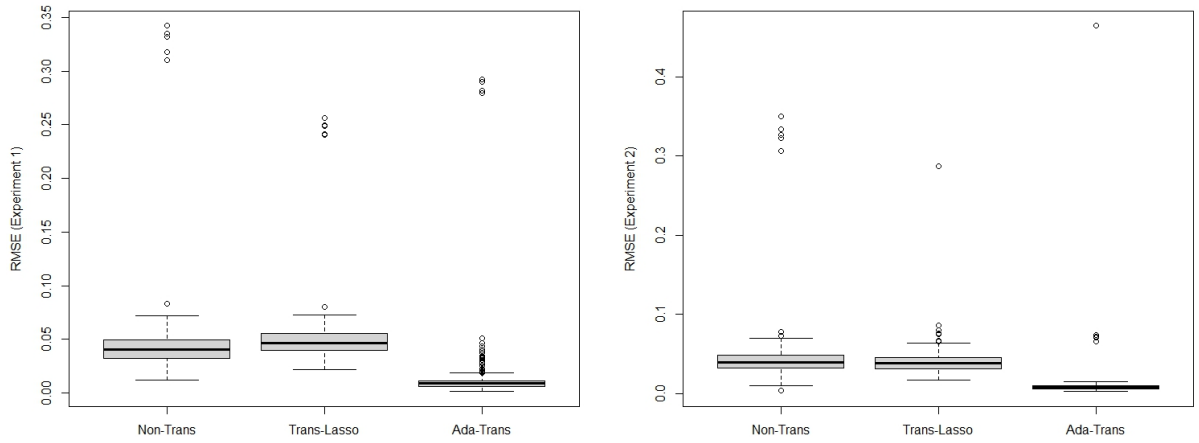
$$CR_1 \equiv O_p\left(\frac{s \log p}{n_0 + n_1 + ... + n_K} + \frac{s \log p}{n_0} \wedge h\sqrt{\log p / n_0} \wedge h^2\right)$$

$$CR_0 \equiv O_p\left(\frac{s \log p}{n_0 + n_1 + ... + n_K} + \frac{s \log p}{n_0} \wedge \frac{h \log p}{n_0}\right)$$

while sharp CR_0 requires exact equivalence with $\delta_{0j}^{(k)} = 0$ for most $j \in \{1, ..., p\}$, which is hardly satisfied as mentioned in Li *et al.* (2022)'s supplement, thus only $p = 1$ is considered in this protocol. However, the sharpness of CR_1 also requires $n_1 + ... + n_K >> n_0$ as well as $h << s$, offering limitations to applications of such methodology, which is illustrated by the following simulations.

(Experiment 1: cumulative dissimilarity) Without loss of generality, consider single source as representative of multiple homogeneous sources. Under 100 DGP of $X_i^{(k)} \overset{i.i.d.}{\sim} N_p(\mathbf{1}, \mathbf{I}_{p \times p})$, $\varepsilon_i^{(k)} \overset{i.i.d.}{\sim} N(0, 1)$ for $k \in \{0, 1\}$, $p = 200$, $n_0 = 50$, $n_1 = 600$. With $\beta_0 = (1 : 5, 0, 0, ...)^\top, \beta_0^{(1)} = (seq(1.4, 5.4, 1), 0, 0, ...)^\top$ with $h = 0.4 \times 5 = 2$. The following box-plot compares Root Mean Square Error (RMSE) between Non-Trans Lasso, Trans-Lasso and Ada-Trans, which would be introduced soon in the next section.

(Experiment 2: dissimilarity point mass) Under same DGP with the first experiment while modifying $\beta_0^{(1)} = (3, 2 : 5, 0, 0, ...)^\top$, still with $h = 2$.



While theoretically speaking, the two aforementioned experiments (*experiment 1* on the left) should enjoy at least similar rate of convergence since all hyper-parameters are identical in CR_1, the formal case reveals slight negative-transfer in average, the latter, however, performs certainly better and whose transfer is non-negative.

Some calculations show $CR_1 = 0.5298 + 0.0407$ while the convergence rate of Non-Trans Lasso is $\frac{s\log(p)}{n_0} = 0.5298$, indicating that the violation of $h << s$ is fatal and would never be compensated by enlarging n_1. Specifically in the case here, whenever $h > 1.6276$, negative transfer occurs.

The second experiment could be further extend to $h = 4$ while still non-negative, however, such extension clearly violates restriction by l_1 similarity. Such paradox could be explained by the ignorance of l_0 norm information, calling for a combined consideration of l_1 and l_0 norm.

2 Guided Trans-Lasso

With prior knowledge of latent dissimilarity point mass when conducting transfer learning, an guided indicator of $I\{\beta_{0j} = \beta_{0j}^{(k)}\}$ could be implemented through adaptive penalty, which utilizes proper l_0 norm information. In particular, algorithm would be designed as the following and is called Ind-Trans.

Algorithm 1: Guided Trans-Lasso with prior knowledge

Input: Target data $\left(\mathbf{X}^{(0)}, Y^{(0)}\right)$ and informative auxiliary samples $\left\{\mathbf{X}^{(k)}, Y^{(k)}\right\}_{k\in\{1,\ldots,K\}}$

Output: $\hat{\beta}$

1 Compute

$$\hat{\beta}^{(k)} = \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2(n_1+\ldots+n_K)}\sum_{k\in\{1,\ldots,K\}}\left\|Y^{(k)} - \mathbf{X}^{(k)}\beta\right\|_2^2 + \lambda_\beta\|\beta\|_1\right\}$$

for λ_β being selected by cross-validation.

2 Let

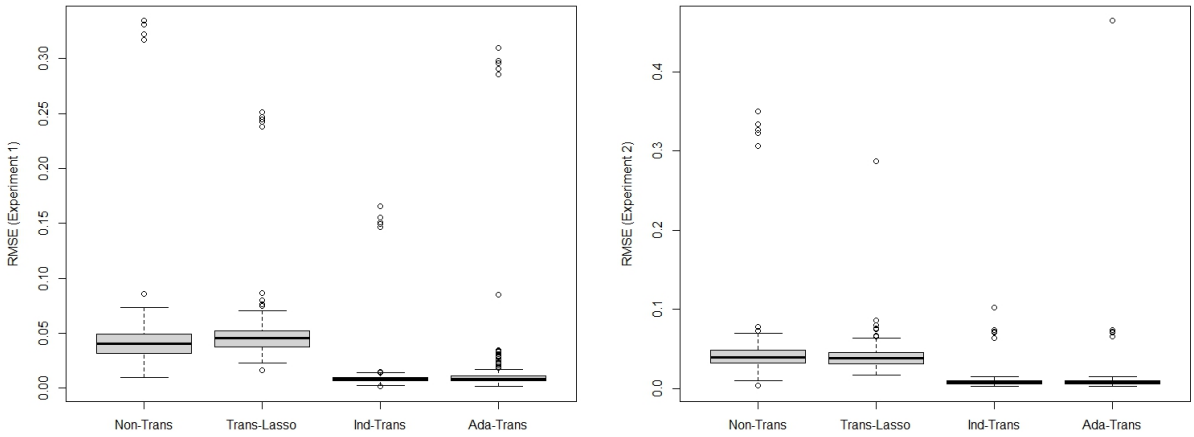
$$\hat{\beta} = \hat{\beta}^{(k)} + \hat{\delta}^{(k)}$$

where

$$\hat{\delta}^{(k)} = \arg\min_{\delta\in\mathbb{R}^p}\left\{\frac{1}{2n_0}\left\|Y^{(0)} - \mathbf{X}^{(0)}\left(\hat{\beta}^{(k)} + \delta\right)\right\|_2^2 + \lambda_\delta\|I_\beta^\top\delta\|_1\right\}$$

for $\lambda_\delta = c_1\sqrt{\log p/n_0}$ with some constant c_1, $I_\beta = I\{\beta_0 = \beta_0^{(k)}\}$ as similarity indicator.

The implementation of **Algorithm 1** in *experiment 1* and *experiment 2* with correct prior information of $I_\delta = (rep(0,5),1,1,...)^\top$ and $(0,1,1,...)^\top$ reveals the following box-plot



Such benchmark is hardly achievable since the underlying DGP is beyond oracle, however, by inheriting the guidance of regularizing for the similar part, a data-driven guidance is introduced as the following algorithm and is noted as Ada-Trans.

Algorithm 2: Guided Trans-Lasso with data-driven adaptiveness

Input: Target data $\left(\mathbf{X}^{(0)}, Y^{(0)}\right)$ and informative auxiliary samples $\left\{\mathbf{X}^{(k)}, Y^{(k)}\right\}_{k\in\{1,...,K\}}$

Output: $\hat{\beta}$

1 Compute

$$\hat{\beta}^{(k)} = \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2(n_1+...+n_K)}\sum_{k\in\{1,...,K\}}\left\|Y^{(k)}-\mathbf{X}^{(k)}\beta\right\|_2^2+\lambda_\beta\|\beta\|_1\right\}$$

for λ_β being selected by cross-validation.

2 Compute

$$\hat{\beta}^{(0k)} = \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2(n_0+n_1+...+n_K)}\sum_{k\in\{0,1,...,K\}}\left\|Y^{(k)}-\mathbf{X}^{(k)}\beta\right\|_2^2+\lambda'_\beta\|\beta\|_1\right\}$$

for λ'_β also being selected by cross-validation.

3 Let

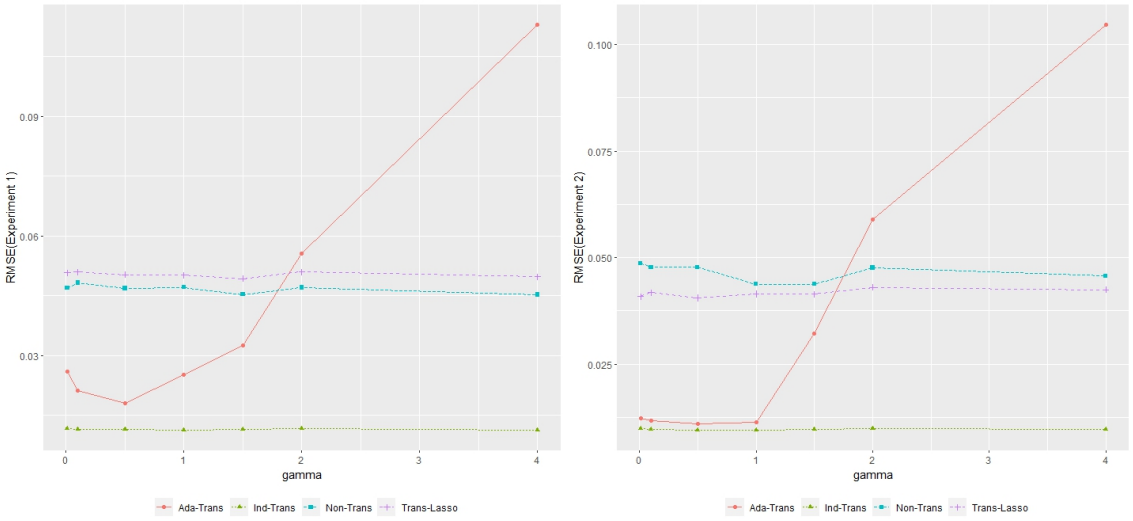
$$\hat{\beta} = \hat{\beta}^{(k)} + \hat{\delta}^{(k)}$$

where

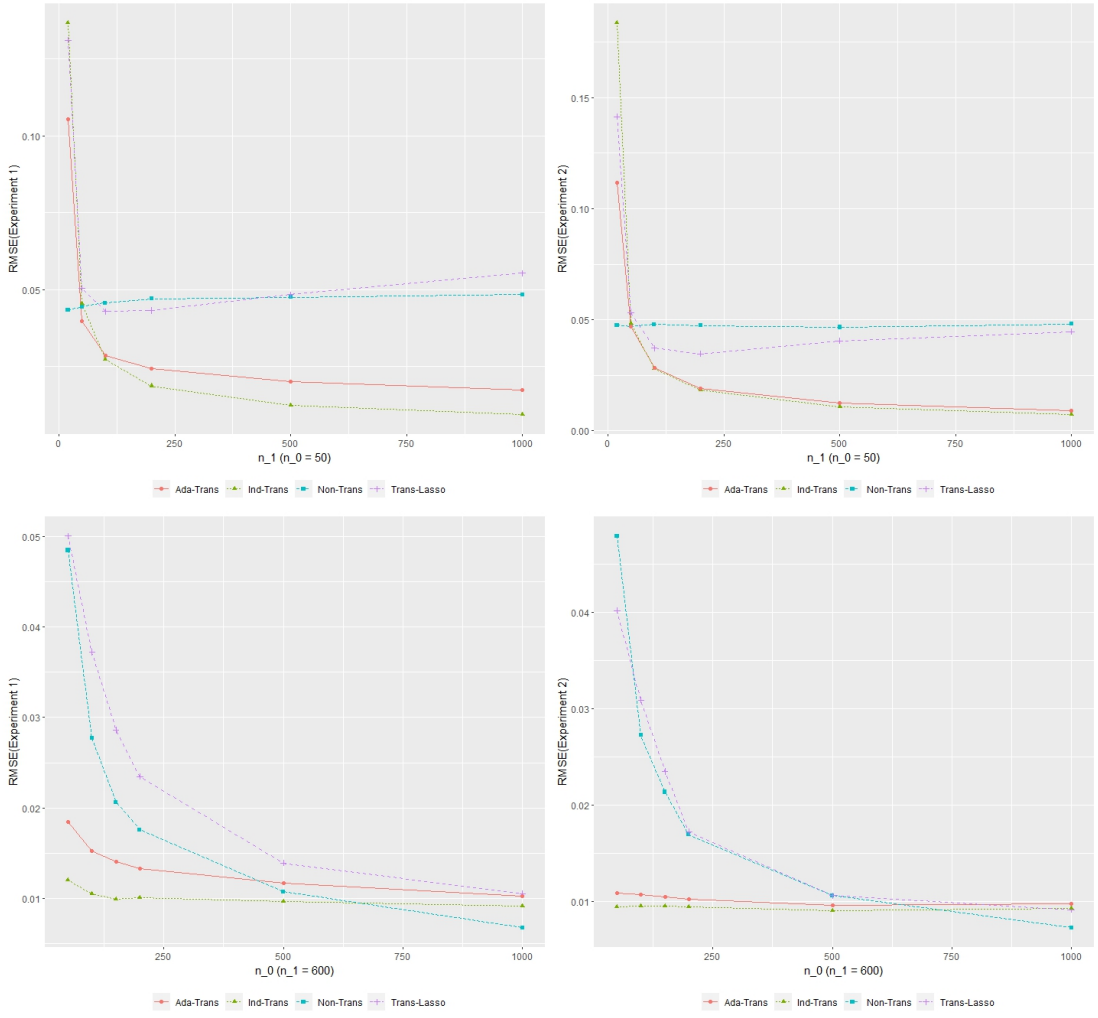
$$\hat{\delta}^{(k)} = \arg\min_{\delta\in\mathbb{R}^p}\left\{\frac{1}{2n_0}\left\|Y^{(0)}-\mathbf{X}^{(0)}\left(\hat{\beta}^{(k)}+\delta\right)\right\|_2^2+\lambda_\delta\|\hat{w}_\beta^\top\delta\|_1\right\}$$

for $\lambda_\delta = c_2\sqrt{\log p/n_0}$ with some constant c_2, $\hat{w}_\beta = \frac{1}{\left|\hat{\beta}^{(k)}-\hat{\beta}^{(0k)}\right|^\gamma}$ as the data-driven guidance through adaptive weighting, where $\gamma>0$ serving as the level of adaptievness.

Step 3 stems from Zou (2006)'s framework of adaptive Lasso, nevertheless, the adaptiveness is formed through similarity guidance, which could also be seen as the estimated version of l_0 norm information. One can easily set extra cutting point for $\hat{w}_\beta$ being either 0 or 1, however, such discretization is largely subjective and is not considered here. For $\gamma = 0.1$ in the aforementioned simulation result, the following graph compares different choices of γ, which is a clear form of scree plot, suggesting cross-validation for selecting γ.



Consider convergence rate through simulation, results for two experiments are shown as the following, where Trans-Lasso fails to converge with $h = 2$ if holding $n_0 = 50$ immutable, nevertheless, Ada-Lasso is able to maintain convergence under the same scenario, and Ind-Lasso serves as the benchmark for comparison. On the other hand, if maintaining $n_1 = 600$ and enlarging n_0, result also shows faster rate of convergence for Ada-Lasso.



3 Empirical Research

The final section intends to demonstrate well-application of Ada-Trans in bioscience especially, and the database refers to Genotype-Tissue Expression (GTEx) (`https://gtexportal.org/`), and a subset of which is picked by choosing MODULE_137 (`https://www.gsea-msigdb.org/gsea/msigdb/cards/MODULE_137.html`) genes and 13 human brains tissues (`https://gtexportal.org/home/samplingSitePage`) samples, the data matrix is then $2,642 \times 546$.

Similar with Li *et al.* (2022)'s transferring strategy, set `JAM2` as the response variable, allocate each brain tissue as target dataset (average sample size of 203) at each time and other 12 tissues as potential sources, the following table compares prediction error between Ada-Lasso and Trans-Lasso, where suffix 'sp' indicates the l_1 norm used in sparsity index $\hat{R}^{(k)}$ and empirical risk minimization. I present the ratio of Trans-Lasso/Trans-Lasso.sp over All-Trans-Lasso where Q-aggregation is ignored, and the ratio of Ada-Lasso over All-Trans-Lasso, in order to compare the level of adaptiveness in these two methods. I also pick up the smallest one between Trans-Lasso and Trans-Lasso.sp for clarity. The comparable time consumption ratio of Ada-Trans over Trans-Lasso/Trans-Lasso.sp is also shown below.

Target Data	Trans-Lasso	Trans-Lasso.sp	Ada-Trans	Time Ratio
Amygdala	—	0.6994	**0.2648**	0.0007
ACC	0.0501	—	**0.0102**	0.0012
Caudate	0.8393	—	**0.6051**	0.0016
Cerebellar Hemisphere	1.0806	—	**0.8764**	0.0024
Cerebellum	—	1.1939	**0.6069**	0.0008
Cortex	—	1.8049	**1.0472**	0.0012
Frontal Cortex	0.6447	—	**0.3369**	0.0008
Hippocampus	0.7819	—	**0.3615**	0.0009
Hippocampus	—	**0.3393**	0.3443	0.0026
Nucleus Accumbens	—	0.9279	**0.3918**	0.0015
Putamen	—	1.2684	**0.5057**	0.0010
Spinal cord	—	**0.8395**	0.8803	0.0024
Substantia nigra	—	1.1822	**0.6989**	0.0007

Result shows generally faster and more precise prediction of Ada-Trans method among 13 transferring sets, indicating efficiency of data-driven guidance for transfer learning.

REFERENCE

[1] Li, S., T. T. Cai, and H. Li (2022). Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation and Minimax Optimality. *Journal of the Royal Statistical Society: Series B*, 84(1), 149-173.

[2] Tian, Y., and Y. Feng (2022). Transfer Learning Under High-Dimensional Generalized Linear Models. *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2022.2071278.

[3] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.