Advances in Applied Mathematics 应用数学进展, 2022, 11(9), 6779-6786
Published Online September 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.119718

Hans汉斯

基于Dantzig Selector的迁移学习

——应用于广义线性模型

孙　飞，梁淑娜*

青岛大学，山东　青岛

收稿日期：2022年8月23日；录用日期：2022年9月18日；发布日期：2022年9月27日

摘　要

小样本高维度的n-p问题一直是统计学家的研究热点，不同于传统的变量选择的参数估计方法，在本篇论文中，我们应用迁移学习的相关知识，借用与需要预测数据相关但分布不同的数据，有效的帮助我们解决目标域数据参数的预测问题。我们提出了一种新颖的基于Dantzig selector的两步迁移学习算法，在数值模拟中，我们验证了提出的迁移学习算法在广义线性模型以及不同的协方差数据设计中的稳健性和有效性，这表明提出的算法具有一定的实际应用价值。

关键词

广义线性模型，Dantzig Selector，迁移学习，两步迁移学习算法

Transfer Learning Based on Dantzig Selector

—Applied to Generalized Linear Models

Fei Sun, Shuna Liang*

Qingdao University, Qingdao Shandong

Received: Aug. 23rd, 2022; accepted: Sep. 18th, 2022; published: Sep. 27th, 2022

Abstract

The problem of small samples and high dimensionality has always been a research hotspot for statisticians. Different from the traditional parameter estimation method of variable selection, in this paper, we apply the relevant knowledge of transfer learning, and borrow data with different distributions from those that need to be predicted. It effectively helps us solve the prediction problem of target domain data. We propose a novel two-step transfer learning algorithm based on

*通讯作者。

文章引用: 孙飞, 梁淑娜. 基于 Dantzig Selector 的迁移学习[J]. 应用数学进展, 2022, 11(9): 6779-6786.
DOI: 10.12677/aam.2022.119718

Dantzig selector. In numerical simulations, we verify the robustness and effectiveness of the proposed transfer learning algorithm in generalized linear models as well as in different covariance data designs, which shows that the proposed algorithm has certain practical application value.

Keywords

Generalized Linear Model, Dantzig Selector, Transfer Learning, Two-Step Transfer Learning Algorithm

1. 引言

大数据时代的背景下，随着数据的规模和维度的增加，统计学习、机器学习算法能够得以成功的应用。但是使得这些算法被成功应用有一基本的假设，即：训练数据和预测数据要求同一分布，这一基本假设在实际生活中是难以满足的。此外，在健康医疗及电商平台中，常会出现用来训练数据标签不足的情形。迁移学习为一个新的机器学习框架，它是一种学习的思想和模式。具体而言，在机器学习范畴，迁移学习通过利用数据、任务或模型之间的相似性，将在旧领域学习过的模型和知识应用于新的领域。它可以很好的解决传统机器学习、统计模型不能适用的场合，利用和预测数据相关的数据，且相关数据的样本量是充足的。在迁移学习中，源域数据就是有知识、有大量数据标注的领域，是我们要迁移的对象；目标数据就是我们最终要赋予标注的对象。知识从源域传递到目标域，就完成了迁移[1]。迁移学习已经成功的应用在各个领域内：推荐系统[2]、生物医疗[3]、计算机视觉[4]等方面。在统计学习方面，文献[5]用二分类源域数据去预测没有标签的目标域数据。另一个应用于回归问题的经典例子，文献[6]中数据集是一个商业飞行数据集，完整的数据集是 1987 年~2008 年，将 1987 到 2007 年的数据作为源域，将 2008 年的数据作为目标，源域和目标域的数据分布是不同的。这使得我们考虑将迁移学习进一步应用到更广泛的统计模型中。

在统计学习中，高维度小样本的 n-p 问题一直是统计学家的研究热点。文献[7]提出了 Lasso 回归用来选择变量，这一高维数据的变量选择方法一经提出，使得研究人员对于高维数据的参数估计问题有了全新的认识。在此基础上，文献[8]提出 dantzig selector，在一般的参数估计问题中，通常会有一个基本的假设，即模型的残差和设计阵是不相关的。但是 dantzig selector 不仅考虑了残差的影响，还考虑了设计阵与残差的相关性。除此之外，还有许多其他的变量选择的方法。包括 group lasso [9]、adaptive lasso [10]、MC+ [11]、Frechet 回归[12]等，这些经典的变量选择方法能够很好的处理训练数据、预测数据服从相同的分布。但是在处理分布不同的数据时，我们可以很好的应用迁移学习的相关思想，即可以借用分布不同但相似的数据来解决目标域数据的预测。在迁移学习中，度量源域数据和目标域数据的相似性是一个十分重要的问题，[13] [14] [15]通过 A-distance 和 Principle Angle 度量相似性。本文中，我们通过 l_1 距离度量相似性。文献[16]利用 $l_q\left(0 \le q \le 1\right)$ 距离度量相似性，研究了基于线性模型的迁移学习，文献[17]同样研究了线性模型的迁移学习。但其仅考虑了源域数据和目标域数据的 l_1 相似性。在非参数模型的假定下，文献[18]研究了基于 KNN 算法的迁移学习分类迁移学习模型。

基于基本统计模型的迁移学习，2021 年[17]研究了基于线性模型的迁移学习，提出了两步联合估计器，并说明了两步联合估计的误差下界较于传统估计方法是较优的，但是其方法局限于单源数据的迁移

学习。2022 年[16]在线性模型的假定下，提出了多源数据的两步迁移学习和数据驱动的迁移学习算法，并在理论方面，证明了迁移学习估计器使用源域数据的收敛速率优于不使用源域数据的传统估计器的参数估计，此外，还证明了迁移学习估计器的极小极大收敛速率。

2022 年，在模型参数稀疏的假定条件下，文献[19]研究了指数分布族的迁移学习，并提出了两步迁移学习算法，即第一步基于 Lasso 算法，利用所有的源域和目标域数据，得到一个较为粗糙的参数估计。第二步再利用 lasso 回归，将第一步的参数估计的偏差进行修正。我们同样基于两步迁移学习算法的思想，研究了广义线性模型的迁移学习，第一步利用 Dantzig selector 得到一个粗糙的参数估计，在假定第一步得到较为粗糙的参数估计和真实参数之间的 l_0 距离较小时。再利用 Lasso 回归去修正第一步的参数估计。通过数值模拟容易发现，我们所提出的两步迁移学习算法是优于经典的 Lasoo 回归、Dantzig selector。

2. 广义线性模型中的迁移学习

2.1. 线性模型

在统计模型中，最为经典的模型便是线性模型，其理论方法也是其他统计模型研究或应用的基础，若给定 $(x,y)\in R^p\times R$，y 来自线性模型，则可被表示为：

$$y=x^{\mathrm{T}}\beta+\varepsilon \tag{1}$$

其中 ε 服从高斯马尔科夫假设：

$$E(\varepsilon)=0$$

$$Var(\varepsilon)=\sigma^2 \tag{2}$$

$$Cov(e_i,e_j)=0, i\neq j; i,j=1,2,3,\cdots,n$$

最小二乘估计是线性模型中最经典的参数估计方法，利用最小二乘估计所获得参数估计值，可以获得很多优良的性质，比如，此参数估计值具有无偏性，并且是一致最小方差无偏估计。但是当设计阵存在着复共线关系时，最小二乘估计的性质不够理想，有时甚至很坏。在这种情况下，我们就需要一些新的估计方法。因此，许多种新估计方法也被统计学家提出，其中在理论上最有影响并且得到广泛应用的就是岭估计，岭估计的本质是对于模型参数的 l_2 惩罚，其不仅很好的处理变量之间的复共线性，而且可以对参数估计进行压缩，但并不能达到变量筛选的目的。

2.2. 广义线性模型

广义线性模型是线性模型的扩展，其可以弥补线性模型的一些缺点，例如，线性模型的取值只能为连续型数据，但是实际生活中，分类数据是十分常见的。广义线性模型的概念最早由文献[20]提出，在这篇论文中详细的介绍了广义线性模型的相关理论知识。但是，相对于线性模型，其主要的特点是，将因变量分布的假设由正态分布放松到任意的指数分布族，故可以很好的处理上述的分类型数据。若给定 $(x,y)\in R^p\times R$，如果 y 来自广义线性模型模型，则其条件密度函数可以表示为：

$$y\mid x\sim P(y\mid x)=\rho(y)\exp\left\{yx^{\mathrm{T}}\beta-\psi\left(x^{\mathrm{T}}\beta\right)\right\} \tag{3}$$

上式可以被改写为：

$$y=f\left(x^{\mathrm{T}}\beta\right)+\varepsilon \tag{4}$$

其中 $\psi(.)$ 被称为链接函数。

易知，$E(y|x)=f(x^{\mathrm{T}}\beta)=\psi'(x^{\mathrm{T}}\beta)$，$Var(y|x)=\psi''(x^{\mathrm{T}}\beta)$且$\psi(.)$的函数形式是已知的。通过不同的函数$\psi(.)$可以区分不同的广义线性模型。例如，当$y$是一个连续变量且$\psi(u)=\frac{1}{2}u^2$时，$y$服从线性模型；当$y$的取值是二分类变量且$\psi(u)=\log(1+e^u)$时，$y$服从逻辑回归；当$y$的取值为非负整数且$\psi(u)=e^u$时，$y$服从泊松回归。广义线性模型较于线性模型，不要求模型服从高斯马尔科夫假定，这放松了线性模型的假定条件，也大大的扩大了线性模型的使用范围。

对于高维的广义线性模型，统计学家也提出了许多经典的参数估计方法，比如：Lasso、Dantzig selector。Lasso 方法的参数估计是一种约束模型参数的最小二乘估计器。假定数据x_{ij}是经过标准化的，即$\sum_j \frac{x_{ij}}{n}=0$，$\sum_j \frac{x_{ij}^2}{n}=1$。Lasso 估计器可以被表示为：

$$\|Y-f(X\beta)\|_2+\lambda\|\beta\|_1 \tag{5}$$

上式可以被等价表示为：

$$\|Y-f(X\beta)\|_2 \quad \text{s.t}\,\|\beta\|_1\le\lambda \tag{6}$$

Lasso 的变量选择是通过阈值估计器实现的，简单描述为：

$$\widehat{\beta}_j^{\,Lasso}=\operatorname{sgn}\left(\widehat{\beta}_j^{\,LS}\right)\left(\left|\widehat{\beta}_j^{\,LS}\right|-\lambda\right)^+ \tag{7}$$

其中，$(a)^+$表示$\max\{0,a\}$，$\hat{\beta}^{LS}$是最小二乘估计计算出的参数估计值。

Lasso 可以避免最小二乘估计的一些缺点，且 Lasso 估计算法是一个凸优化问题，计算复杂度低，计算速度快，无论是在理论上还是在实际的应用中，均可以取得较好的效果。在此基础上，Candes [7]提出了 Dantzig selector。与 Lasso 方法不同，Dantzig selector 不仅考虑了残差的影响，还考虑了设计阵与残差的相关性，在“一致不确定性原则下”，Dantzig selector 可以有效的降低参数估计的误差，其估计器可以被表示为：

$$\beta=\arg\min_{\beta}\|\beta\|_1 \quad \text{s.t.}\left\|X^{\mathrm{T}}(y-f(X\beta))\right\|_\infty\le\lambda \tag{8}$$

虽然这些经典的统计学习方法的许多优良性质已经被统计学家证明出来，但是如前所述，这些方法已经不足以处理情形多变、日益复杂的数据。与传统的变量选择方法不同，我们应用迁移学习的思想来处理高维的广义线性问题。

2.3. 迁移学习

迁移学习，顾名思义，就是要通过知识的迁移进行学习，达到事半功倍的效果。在人工智能和机器学习范畴，迁移学习就是一种学习的思想和模式。其主要的想法是可以类比为找到不同事物之间的相关性，然后进行“举一反三”、“照猫画虎”[1]。迁移学习作为机器学习的一个分支，其大部分研究都是基于计算机领域的应用。近年来，有关迁移学习的应用也越来越成为统计学家的研究热点。

基于迁移学习的统计模型相关应用，最开始是由文献[17]基于线性模型，研究了单源数据的迁移学习。在这篇论文中，我们考虑了多源数据的迁移学习，即存在多个可迁移的数据。单源域数据的迁移学习是多源域数据迁移学习的特例，此外我们的研究是基于广义线性模型，这大大扩展了论文所提出方法实际的应用范围。给定多个源域数据$\{(X_i,Y_i)\}_{i=1,2,\cdots,K}$和目标域数据$(X_0,Y_0)$，源域数据和目标域数据均服从广义线性模型，记源域数据的参数为$\beta_i\,(i=1,2,\cdots,n)$，目标域数据的参数为$\beta_0$。$\beta_i\,(i=1,2,\cdots,n)$与$\beta_0$通常是不相等的。迁移学习的总体思路可以概括为：开发算法来最大限度地利用有标注的领域的知识，来辅

助目标领域的知识获取和学习。其核心是找到源域和目标域之间的相似性，相似性的度量就是描述源域和目标域的距离。在这篇论文中，我们利用 l_1 距离度量源域数据和目标域数据的相似性，定义可迁移数据集合为：$\left\{k:\left\|\beta_k-\beta_0\right\|_1\le h, k=1,2,\cdots,K\right\}$，其中 h 是划分可迁移域和不可迁移域的阈值。

3. 两步迁移学习算法

在本篇论文中，我们提出了一个新颖的两步迁移学习算法(Oracle Trans-DS)。在源域数据和目标域数据均不稀疏的情形下，如果直接使用上述经典的变量选择方法，则得到的参数估计的偏差是较大的。我们遵循文献[18]提出的两步迁移学习算法的思想，在第一步参数估计中(迁移步骤)，利用 Dantzig selector 在源域数据和目标域数据得到一个较为粗糙的参数估计 β^{init} (迁移步骤)，

$$\beta^{init}=\arg\min_{\beta}\|\beta\|_1 \quad \text{s.t.}\left\|\sum_k X^{(k)\mathrm{T}}\left(y^{(k)}-f\left(X^{(k)}\beta^{(k)}\right)\right)\right\|_\infty\le\lambda, k=0,1,2,\cdots,K \tag{9}$$

然后利用 Lasso 在目标域数据上对 β^{init} 进行偏差修正(纠偏步骤)。

$$\hat{\delta}=\left\|Y_{(0)}-f\left(X_0\left(\beta+\delta\right)\right)\right\|_2+\lambda\|\delta\|_1 \tag{10}$$

详细的算法流程可以见图 1。通过这样的迁移学习算法，可以有效的借助相关但分布不同的数据。从而较好的降低参数估计的误差。在第三节的数值模拟中，我们通过与经典的高维数据变量选择对比，可以很好的看出 Oracle Trans-DS 方法的优越性。

算法一： Oracle Trans-DS

输入：目标域数据 (X_0, Y_0) 和源域数据 $(X^{(k)}, Y^{(k)}), k=1,2,...,K$

输出：$\hat{\beta}$

步骤一：计算

$$\beta^{init}=\arg\min_{\beta}\|\beta\|_1 \quad s.t.\|\sum_k X^{(k)T}(y^{(k)}-f(X^{(k)}\beta^{(k)}))\|_\infty\le\lambda, k=0,1,2,...,K$$

步骤二：计算

$$\hat{\delta}=\|Y_{(0)}-f(X_0(\beta+\delta))\|_2+\lambda\|\delta\|_1$$

步骤三：令

$$\hat{\beta}=\beta^{init}+\hat{\delta}$$

Figure 1. Algorithm 1
图 1. 算法 1

4. 数据模拟

在这节中，为了不失一般性，我们分别在线性模型，逻辑回归和泊松回归三种场合下，比较 Oracle Trans-DS 与 Lasso、Dantzig selector 估计高维广义线性模型参数的性能。其中，Naïve-Lasso、naïve-dantzig-selector 表示仅使用目标域数据，分别利用 Lasso、Dantzig selector 估计高维数据的参数，Oracle Trans-DS 表示用算法 1 估计得到的参数。

所有的模拟实验均使用 R 语言实现，其中 Lasso 通过 glmnet 包实现，Dantzig selector 通过 hdme 包实现，其中 Lasso 的调谐参数是通过 10 折交叉验证选择的，Dantzig selector 的调谐参数设置为 $\sqrt{C\frac{\log p}{n}}$。

我们设置源域和目标域的样本维度均为 $p=1000$，可迁移的源域数 $K=6$，且样本量为 $n_1,n_2,\cdots,n_K=150$，目标域 $n_0=100$。设置目标域的参数 $\beta=\left(0.51_s,0_{p-s}\right)^{\mathrm{T}}$，设置 $\beta^{(k)}=\beta+\frac{h}{p}R_p^{(k)}, h=5, k=1,2,\cdots,K$，其中 $R_p^{(k)}$ 表示以等概率取−1 或 1 的 p 维随机向量，s 的设置为 25。我们以 l_2 估计误差($\left\|\hat{\beta}-\beta\right\|_2$)为评价指标，所有的模拟实验都被重复 100 次，然后取其平均值。具体实验结果见下图 2。

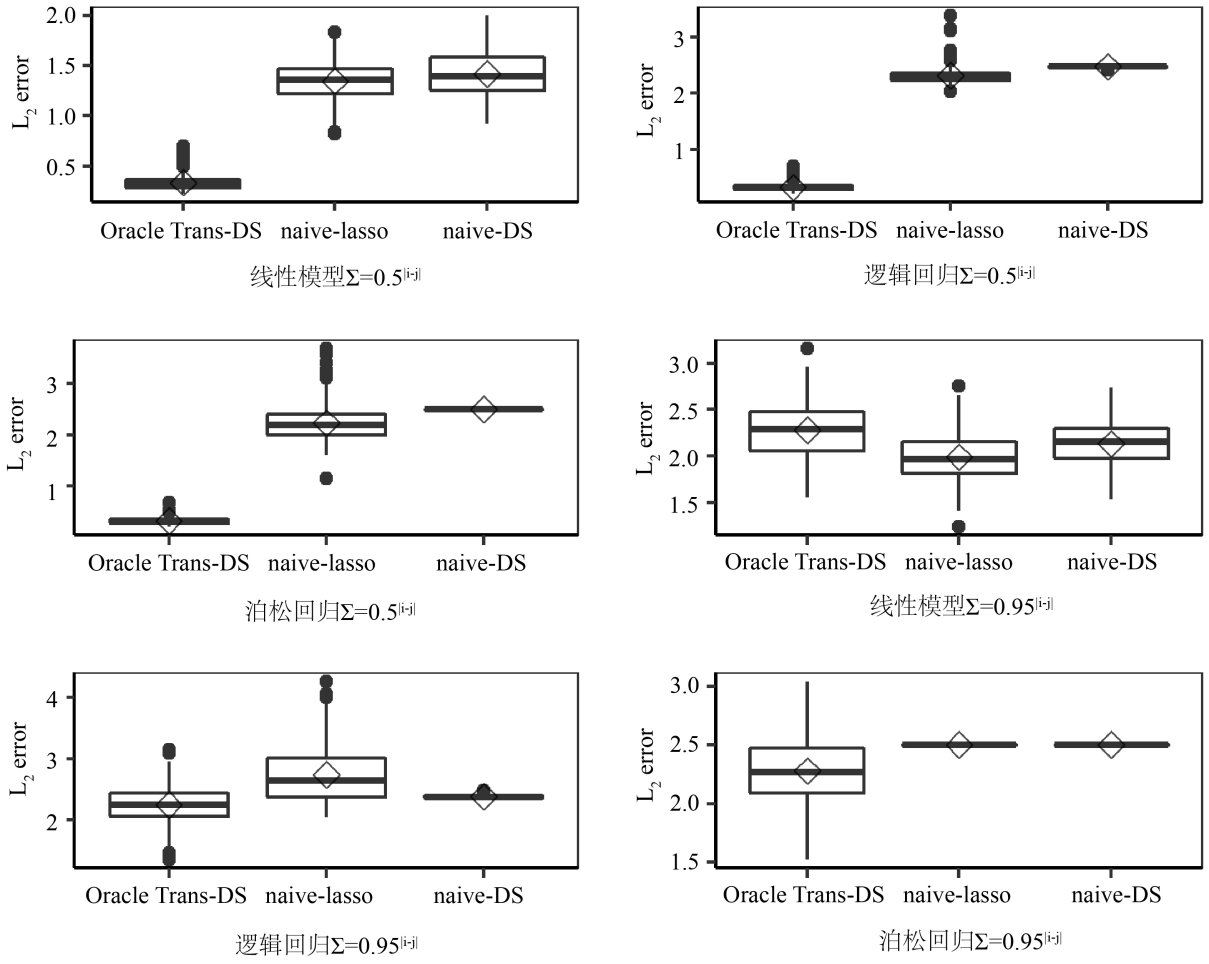

Figure 2. l_2 errors for different models and sparsity
图 2. 不同模型及稀疏度的 l_2 误差

通过图 2 可以发现，当数据之间存在较弱的相关性时，我们提出的 Oracle Trans-DS 方法是远远优于传统的变量选择方法 naïve-Lasso、naïve-dantzig selector。当数据之间存在较强的相关性时，普遍来说，我们提出的算法的也是较优的，这说明我们提出的算法在解决高维数据的参数估计问题时具有稳健性和有效性。

5. 结论

在本篇论文中，我们提出了一种基于 Dantzig selector 的两步迁移学习算法，由数值模拟的分析可知，我们提出的方法的效率是较优的，并且将其运用到了广义线性模型的场合，大大提高了实际的应用价值。

容易看出，我们提出迁移学习算法也是利用传统的统计学习方法。因此，将新的学习框架与经典的统计学习模型相结合，可以很好的解决统计学习中的 n-p 问题，这一思路同样启示我们，对于统计学习方法的应用应该与时俱进，从而使得统计模型能够适应新的时代要求。

6. 讨论

如上所述，利用两步迁移学习算法能够解决不同分布、训练数据较少的问题，但是在这篇论文中，我们局限于源域数据和目标域数据服从均分布参数不同但类型相同的分布，其原因是，若源域数据和目标域数据服从的分布和参数均不同，在此种情况下，使用 l_1 距离度量相似性不再合适，故进一步的研究中，我们可以考虑利用其他的相似性度量准则，从而结合来自不同分布的源域数据，比如，在广义线性模型的框架的下，可以考虑源域部分源域数据来自线性模型，部分源域数据来自逻辑回归。这篇论文我们是在稀疏假定条件下，进行模型的迁移，但是现在生活中，稀疏假定这一情形往往是难以满足的。如何构建将新的统计学习模型与非稀疏数据的预测相结合，也是一个非常值得研究的问题。

迁移学习的效果依赖于相似性的度量准则，在这篇论文中，我们是用参数之间的 l_1 距离度量相似性，这种相似性的度量方式在理论上分析较为简单，且实际的运用中也较为容易实现，但是针对与某些场合，需要采用不同的相似性度量准则。文献[21]考虑了学习模型对于分布的稳健优化，可以将这一思想与迁移学习相结合，从而构建出对相似性度量不太敏感的迁移学习模型，这亦是一个研究重点。

基金项目

本篇论文由国家社会科学基金项目(No. 21BTJ045)资助。

参考文献

[1] 王晋东. 迁移学习简明手册[M/OL]. https://www.doc88.com/p-79899021072337.html, 2020-08-13.

[2] Zhao, L., Pan, S.J., Xiang, E.W., Zhong, E., Lu, Z. and Yang, Q. (2013) Active Transfer Learning for Cross-System Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **27**, 1205-1211. https://doi.org/10.1609/aaai.v27i1.8458

[3] Erbe, R., Kessler, M.D., Favorov, A.V., Easwaran, H., Gaykalova, D.A. and Fertig, E.J. (2020) Matrix Factorization and Transfer Learning Uncover Regulatory Biology across Multiple Single-Cell ATAC-Seq Data Sets. *Nucleic Acids Research*, **48**, e68. https://doi.org/10.1093/nar/gkaa349

[4] Mesnil, G., Dauphin, Y., Glorot, X., *et al*. (2012) Unsupervised and Transfer Learning Challenge: A Deep Learning Approach. *Proceedings of the* 2011 *International Conference on Unsupervised and Transfer Learning Workshop*, Vol. 27, Washington DC, 2 July 2011, 97-111.

[5] Hu, W., Qian, Y., Soong, F.K., *et al*. (2015) Improved Mispronunciation Detection with Deep Neural Network Trained Acoustic Models and Transfer Learning Based Logistic Regression Classifiers. *Speech Communication*, **67**, 154-166. https://doi.org/10.1016/j.specom.2014.12.008

[6] Qi, G.-J., Liu, W., Aggarwal, C. and Huang, T. (2016) Joint Intermodal and Intramodal Label Transfers for Extremely Rare or Unseen Classes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1360-1373. https://doi.org/10.1109/TPAMI.2016.2587643

[7] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[8] Candes, E. and Tao, T. (2007) The Dantzig Selector: Statistical Estimation When *p* Is Much Larger than *n*. *The Annals of Statistics*, **35**, 2313-2351. https://doi.org/10.1214/009053606000001523

[9] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. https://doi.org/10.1198/016214506000000735

[10] Yuan, M. and Lin, Y. (2006) Model Selection and Estimation in Regression with Grouped Variables, *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **68**, 49-67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

[11] Zhang, C.-H. (2010) Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of statistics*, **38**,

894-942. https://doi.org/10.1214/09-AOS729

[12] Tucker, D.C., Wu, Y. and Muller, H.-G. (2021) Variable Selection for Global Frëchet Regression. *Journal of the American Statistical Association*, 1-15. https://doi.org/10.1080/01621459.2021.1969240

[13] Ben-David, S., Blitzer, J., Crammer, K. and Pereira, F. (2006) Analysis of Representations for Domain Adaptation. *Advances in Neural Information Processing Systems*, **19**, 137-144.

[14] Blitzer, J., McDonald, R. and Pereira, F. (2006) Domain Adaptation with Structural Correspondence Learning. *Proceedings of the* 2006 *Conference on Empirical Methods in Natural Language Processing*, Sydney, 22-23 July 2006, 120-128. https://doi.org/10.3115/1610075.1610094

[15] Gong, B., Shi, Y., Sha, F. and Grauman, K. (2012) Geodesic Flow Kernel for Unsupervised Domain Adaptation. 2012 *IEEE Conference on Computer Vision and Pattern Recognition*, 16-21 June 2012, Providence, 2066-2073. https://doi.org/10.1109/CVPR.2012.6247911

[16] Li, S., Cai, T.T. and Li, H. (2022) Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation, and Minimax Optimality. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **84**, 149-173. https://doi.org/10.1111/rssb.12479

[17] Bastani, H. (2021) Predicting with Proxies: Transfer Learning in High Dimension. *Management Science*, **67**, 2964-2984. https://doi.org/10.1287/mnsc.2020.3729

[18] Cai, T.T. and Wei, H. (2021) Transfer Learning for Nonparametric Classification: Minimax Rate and Adaptive Classifier, *The Annals of Statistics*, **49**, 100-128. https://doi.org/10.1214/20-AOS1949

[19] Tian, Y. and Feng, Y. (2022) Transfer Learning under High-Dimensional Generalized Linear Models. *Journal of the American Statistical Association*, 1-30. https://doi.org/10.1080/01621459.2022.2071278

[20] Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society*: *Series A* (*General*), **135**, 370-384. https://doi.org/10.2307/2344614

[21] Duchi, J. and Namkoong, H. (2018) Learning Models with Uniform Performance via Distributionally Robust Optimization. arXiv preprint arXiv:1810.08750.