高维线性回归中的迁移学习：预测、估计和极小极大最优性

李赛, 蔡天东, 李洪哲

摘要

本文研究了在迁移学习设定下高维线性回归的估计和预测问题。在这种设定中，除了目标模型的观测数据外，还可以获得来自不同但可能相关的回归模型的辅助样本。当已知信息性辅助研究集合时，我们提出了一种估计量和预测器，并证明了它们的最优性。结果表明，当信息性辅助样本集非空且对比向量足够稀疏时，预测和估计的最优收敛率比不使用辅助样本的相应速率更快。这意味着来自信息性辅助样本的知识可以被迁移以提高目标问题的学习性能。当信息性辅助样本集未知时，我们提出了一种数据驱动的迁移学习程序，称为Trans-Lasso，并证明了它对非信息性辅助样本的鲁棒性以及在知识迁移方面的效率。所提出的方法在数值研究中得到了验证，并应用于关于基因表达关联的数据集。结果表明，通过整合来自多个不同组织的数据作为辅助样本，Trans-Lasso在目标组织的基因表达预测中取得了改进的性能。

关键词

辅助研究、数据聚合、领域适应、GTEx数据、多任务学习、Q-聚合

1 引言

现代科学研究的特点是数据集庞大且多样化。整合不同的数据集以进行更准确的预测和统计推断具有重要意义。给定一个需要解决的目标问题，迁移学习（Torrey & Shavlik, 2010）旨在从不同但相关的样本中迁移知识，以提高目标问题的学习性能。迁移学习的一个典型例子是，通过使用不仅有汽车的标记数据，还有一些卡车的标记数据，可

以提高识别汽车的准确性（Weiss等，2016）。除了分类之外，另一个重要的迁移学习问题是带有辅助样本的线性回归。在生物医学研究中，由于伦理或成本问题，一些临床或生物学结果难以获取，在这种情况下，可以利用迁移学习通过有效利用相关研究中的信息来提高预测和估计性能。

迁移学习已应用于医学和生物学研究中的问题，包括蛋白质定位预测（Mei等，2011）、生物成像诊断（Shin等，2016）、药物敏感性预测（Turki等，2017）以及"多组学"数据的整合分析，例如Sun和Hu（2016）、Hu等（2019）和Wang等（2019）。它也被应用于机器学习中的自然语言处理（Daumé III，2007）和推荐系统（Pan & Yang，2013）。促使本文的应用是使用基因型-组织表达（GTEx）数据（https://gtexportal.org/）整合不同组织中的基因表达测量，以理解基因调控。这些数据集通常是高维的，样本量相对较小。当研究特定组织或细胞类型的基因调控关系时，可以整合来自其他组织的信息以提高学习准确性。这促使我们考虑高维线性回归中的迁移学习。

1.1 高维线性回归中的迁移学习

回归分析是最广泛使用的统计方法之一，用于理解结果与一组协变量之间的关联。在许多现代应用中，协变量的维度通常与样本量相比非常高。典型的例子包括全基因组关联和基因表达研究。在本文中，我们考虑高维线性模型中的迁移学习。形式上，目标模型可以写为：

$$y_i^{(0)} = (x_i^{(0)})^T\beta + \varepsilon_i^{(0)}, i = 1, ..., n_0,$$

其中$(x_i^{(0)})_{i=1,...,n_0}$和$(y_i^{(0)})_{i=1,...,n_0}$是独立样本，$\beta \in \mathbb{R}^p$是感兴趣的系数向量，$\varepsilon_i^{(0)}, i = 1, ..., n_0$是独立同分布的随机噪声，满足$E[\varepsilon_i^{(0)}|x_i^{(0)}] = 0$。在高维情况下，p可能大于甚至远大于$n_0$，通常假设β是稀疏的，即β的非零元素数量s远小于p。

在迁移学习的背景下，我们观察来自K个辅助研究的额外样本。也就是说，我们观察到$(x_i^{(k)}, y_i^{(k)})$，这些样本来自辅助模型：

$$y_i^{(k)} = (x_i^{(k)})^T w^{(k)} + \varepsilon_i^{(k)}, i = 1, ..., n_k, k = 1, ..., K,$$

其中$w^{(k)} \in \mathbb{R}^p$是第k个研究的回归向量，$\varepsilon_i^{(k)}$是满足$E[\varepsilon_i^{(k)}|x_i^{(k)}] = 0$的随机噪声。回归系数$w^{(k)}$未知且与我们的目标β不同。辅助研究的数量K可以增长，但实际上K可能不会太大。我们将研究利用主要数据$(x_i^{(0)}, y_i^{(0)})$, $i = 1, ..., n_0$以及来自K个辅助研究的数据$(x_i^{(k)}, y_i^{(k)})$, $i = 1, ..., n_k$, $k = 1, ..., K$来估计和预测目标模型(1)。

如果一个辅助模型与目标模型"相似"，我们称该辅助样本/研究是信息性的。在本文中，我们使用对比向量$w^{(k)}$和β之间差异的稀疏性来表征第k个辅助研究的信息性水平。令$\delta^{(k)} = \beta - w^{(k)}$表示$w^{(k)}$和β之间的对比。信息性辅助样本集是那些对比足够稀疏的样本：

$$A_q = \{1 \leqslant k \leqslant K : ||\delta^{(k)}||_0 \leqslant h\},$$

其中$q \in [0, 1]$。集合A_q包含对比向量在ℓ_0范数上最多有h个非零元素的辅助研究，被称为信息性集合。稍后将看到，只要h相对于β的稀疏度s较小，A_q中的研究就可以用于改进β的预测和估计。在$q = 0$的情况下，集合A_q对应于对比向量最多有h个非零元素的辅助样本。我们还考虑近似稀疏约束($q \in (0, 1)$)，这允许所有系数都非零，但它们的幅度以相对快的速度衰减。对于任何$q \in [0, 1]$，较小的h意味着辅助样本在A_q中更具信息性；较大的$|A_q|$意味着更多的信息性辅助样本。因此，较小的h和较大的$|A_q|$应该是有利的。我们允许A_q为空，在这种情况下，没有辅助样本是信息性的。对于A_q外的辅助样本，我们不假设$\delta^{(k)}$是稀疏的，因此$w^{(k)}$可能与β非常不同，对于$k \notin A_q$。

2 已知信息性辅助样本的估计

在本节中，我们考虑已知信息性集合A_q时高维线性回归的迁移学习。重点是对比向量的ℓ_q稀疏特征。在后续内容中，符号A_q将简写为A，除非特别强调。补充材料C节将稀疏对比从ℓ_0约束推广到ℓ_q约束，其中$q \in [0, 1]$，并在此设定下提出了一个速率最优估计器。

2.1 Oracle Trans-Lasso算法

我们提出一种迁移学习算法，称为Oracle Trans-Lasso，用于已知A的情况下的估计和预测。作为概述，我们首先使用所有信息性辅助样本计算初始估计器。然而，其概率极限偏离了β，因为$w^{(k)} \neq \beta$，即使对于$k \in A$。我们然后使用主要数据在第二步中校正其偏差。算法1正式展示了我们提出的Oracle Trans-Lasso算法。

在步骤1中，$\hat{w}^A$是基于Lasso（Tibshirani，1996）使用所有信息性辅助样本实现的。其概率极限是w^A，可以通过以下矩条件定义：

$$E[\Sigma_{k\in A}(X^{(k)})^T(y^{(k)} - X^{(k)}w^A)] = 0.$$

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

$$w^A = \beta + \delta^A$$

其中$\delta^A = \Sigma_{k\in A}\alpha_k\delta^{(k)}$且$\alpha_k = n_k/n_A$，给定$\Sigma^{(k)} = \Sigma^{(0)}$对所有$k \in A$。也就是说，概率极限$w^A$有偏差$\delta^A$，这是$\delta^{(k)}$的加权平均。步骤1与高维错误指定模型（Bühlmann & van de Geer，2015）和矩估计器的方法相关。估计器$\hat{w}^A$收敛相对较快，因为样本量在步骤1中相对较大。步骤2校正偏差δ^A，使用主要样本。实际上，δ^A是一个稀疏的高维向量，其ℓ_q范数不大于h。因此，步骤2的误差在对相对较小的h的控制下。调整参数λ_w和λ_δ将在定理1中进一步指定。

我们将提出的Oracle Trans-Lasso方法与多任务回归方法进行比较，例如Agarwal等（2012）和Danaher等（2014）的3.4.3节。Oracle Trans-Lasso不对辅助研究中的回归系数之间的差异进行惩罚。这是因为迁移学习的重点只是目标研究。理论上，额外的惩罚项和多个估计器的联合分析可能无助于提高感兴趣参数的估计准确性。

2.2 理论保证

我们现在建立Oracle Trans-Lasso估计器的理论保证。我们首先介绍一些符号和条件。对于任何向量$v \in \mathbb{R}^p$和指标集$S \subset \{1, ..., p\}$，令v_S表示v在S上的限制，即$(v_S)_j = v_j$如果$j \in S$，否则为0。对于$q \in [0,$

1]，向量v的ℓ_q范数定义为$\|v\|_q = (\Sigma_{j=1}^{p}|v_j|^q)^{1/q}$，特别地，$\|v\|_0$表示v中非零元素的数量。

我们假设设计矩阵满足以下条件。

条件1（次高斯设计）。存在常数$M > 0$，使得对于所有$k = 0, 1, ..., K$和所有单位向量$v \in \mathbb{R}^p$，有$P(|v^T x_i^{(k)}| > t) \leqslant 2\exp(-t^2/M^2)$对所有$t > 0$成立。

条件2（受限特征值条件）。存在常数$\kappa > 0$，使得对于所有$k = 0, 1, ..., K$和所有满足$\|v_{S^c}\|_1 \leqslant 3\|v_S\|_1$的向量$v \in \mathbb{R}^p$，有$v^T\Sigma^{(k)}v \geqslant \kappa\|v\|_2^2$，其中$S = \mathrm{supp}(\beta)$是β的支撑集。

条件1要求设计矩阵的行是次高斯的，这是高维统计中的标准假设。条件2是受限特征值条件，确保了设计矩阵在与β支撑相关的方向上具有足够的曲率。

我们还假设噪声项满足以下条件。

条件3（次高斯噪声）。存在常数$\sigma > 0$，使得对于所有$k = 0, 1, ..., K$和所有$i = 1, ..., n_k$，有$P(|\varepsilon_i^{(k)}| > t|x_i^{(k)}) \leqslant 2\exp(-t^2/\sigma^2)$对所有$t > 0$成立。

在这些条件下，我们可以建立Oracle Trans-Lasso估计器的理论保证。

定理1。假设条件1-3成立，且β是s稀疏的。令$A = A_0$表示信息性辅助样本集，其中对比向量$\delta^{(k)} = \beta - w^{(k)}$最多有h个非零元素。设$\lambda_w = c_1\sigma\sqrt{(\log p/n_A)}$和$\lambda_\delta = c_2\sigma\sqrt{(\log p/n_0)}$，其中$c_1, c_2 > 0$是足够大的常数。那么，以概率至少$1 - p^{-c}$（对于某个常数$c > 0$），Oracle Trans-Lasso估计器$\beta = \hat{w}^A + \delta^A$满足：

$$\|\beta - \beta\|_2^2 \leqslant C(s + h)\log p/n_0$$

$$\|\beta - \beta\|_1 \leqslant C\sqrt{(s + h)\log p/n_0}$$

$$\|X^{(0)}(\beta - \beta)\|_2^2/n_0 \leqslant C\sigma^2(s + h)\log p/n_0$$

其中C > 0是仅依赖于κ, M, σ的常数。

定理1表明，Oracle Trans-Lasso估计器在ℓ_2和ℓ_1范数以及预测误差方面都达到了(s + h)log p/n_0的收敛率。当h < s且A非空时，这个收敛率快于仅使用目标样本的标准Lasso估计器的s log p/n_0收敛率。这表明，当对比向量足够稀疏时，利用信息性辅助样本可以显著提高估计和预测性能。

3 未知信息性辅助样本的估计

在实际应用中，信息性辅助样本集A通常是未知的。在本节中，我们考虑q = 1的情况，并提出一种数据驱动的程序，称为Trans-Lasso，用于在A未知时进行迁移学习。

3.1 Trans-Lasso算法

我们的目标是开发一种方法，即使在A未知的情况下，也能有效地从信息性辅助样本中迁移知识，同时对非信息性辅助样本保持鲁棒性。我们提出的Trans-Lasso算法基于Q-聚合方法（Dai等，2012；Rigollet & Tsybakov，2012），这是一种将多个估计器聚合成单一估计器的方法。

具体来说，我们首先为每个可能的辅助样本子集构建候选估计器，然后使用Q-聚合方法将这些候选估计器聚合成最终估计器。算法2展示了Trans-Lasso算法的详细步骤。

在算法2中，我们首先为每个可能的辅助样本子集B构建候选估计器β^B。对于每个子集B，我们使用类似于Oracle Trans-Lasso的两步法：首先使用子集B中的辅助样本计算初始估计，然后使用目标样本校正偏差。然后，我们使用Q-聚合方法将这些候选估计器聚合成最终估计器β^Q。

Q-聚合方法的关键是权重的选择。我们使用指数权重：

$$w(B) \propto \exp(-||y^{(0)} - X^{(0)}\beta^B||_2^2/(4\sigma^2))$$

这些权重基于候选估计器在目标样本上的预测误差，预测误差小的估计器获得更高的权重。这确保了算法能够自动识别和利用信息性辅助样本，同时减轻非信息性辅助样本的影响。

3.2 理论保证

我们现在建立Trans-Lasso算法的理论保证。我们的主要结果表明，Trans-Lasso估计器在预测风险方面是鲁棒的，并且在信息性辅助样本足够稀疏时是有效的。

定理2（鲁棒性）。假设条件1-3成立，且β是s稀疏的。那么，以概率至少$1 - p^{-c}$（对于某个常数$c > 0$），Trans-Lasso估计器β^{Q}满足：

$$E[||X^{(0)}(\beta^{Q} - \beta)||_2^2/n_0] \leqslant C\sigma^2 s \log p/n_0 \cdot (1 + \log(2^K))$$

其中$C > 0$是仅依赖于κ, M, σ的常数，期望是相对于用于构建β^{Q}的数据集D_1。

定理2表明，即使在最坏的情况下（没有信息性辅助样本），Trans-Lasso估计器的预测风险也只比标准Lasso估计器高出一个对数因子$\log(2^K)$。这保证了算法对非信息性辅助样本的鲁棒性。

定理3（效率）。在定理2的条件下，如果信息性辅助样本集A非空且对于所有$k \in A$，对比向量$\delta^{(k)}$最多有h个非零元素，其中$h < s$，那么以概率至少$1 - p^{-c}$，Trans-Lasso估计器β^{Q}满足：

$$E[||X^{(0)}(\beta^{Q} - \beta)||_2^2/n_0] \leqslant C\sigma^2 (s + h)\log p/n_0 \cdot (1 + \log(2^K))$$

定理3表明，当存在信息性辅助样本且对比向量足够稀疏时，Trans-Lasso估计器可以达到与Oracle Trans-Lasso估计器相似的收敛率，只是多了一个对数因子。这证明了算法在知识迁移方面的效率。

4 异质设计下的理论表现

在前面的章节中，我们假设所有样本的设计矩阵具有相同的分布。在本节中，我们研究设计矩阵在不同样本中分布不同的情况下迁移学习的表现。

当设计矩阵的分布在不同样本中不同时，我们需要修改信息性辅助样本的定义。具体来说，我们定义信息性辅助样本集为：

$$A = \{1 \leqslant k \leqslant K : ||\Sigma^{(k)1/2}\Sigma^{(0)-1/2} - I||_{op} \leqslant \varepsilon, ||\delta^{(k)}||_0 \leqslant h\}$$

其中$||\cdot||_{op}$表示算子范数，$\varepsilon > 0$是一个小常数。这个定义要求信息性辅助样本不仅在回归系数上与目标模型相似，而且在设计矩阵的分布上也要相似。

在这个设定下，我们可以修改Oracle Trans-Lasso和Trans-Lasso算法，以适应异质设计矩阵。修改后的算法在理论上仍然具有鲁棒性和效率，但收敛率可能会受到设计矩阵差异的影响。

具体来说，当设计矩阵足够相似（ε足够小）时，修改后的算法可以达到与同质设计情况下相似的收敛率。当设计矩阵差异较大时，算法的性能可能会下降，但仍然优于仅使用目标样本的方法，只要对比向量足够稀疏。

5 数值研究

在本节中，我们通过模拟研究和真实数据分析来评估所提出方法的性能。

5.1 模拟研究

我们生成具有以下设置的合成数据：

- 维度：$p = 500$
- 目标样本量：$n_0 = 100$
- 辅助样本量：$n_k = 200, k = 1, ..., 5$
- 真实目标系数：β的前10个元素非零，值从均匀分布$U(0.5, 1.5)$中抽取
- 辅助系数：$w^{(k)} = \beta + \delta^{(k)}$，其中$\delta^{(k)}$的稀疏度根据k变化

我们比较以下方法：

- Lasso：仅使用目标样本的标准Lasso

- Oracle：假设已知真实β的Oracle方法
- Oracle-TL：Oracle Trans-Lasso，假设已知信息性辅助样本集
- Trans-Lasso：我们提出的Trans-Lasso算法
- Pooling：将所有样本池化在一起的Lasso

结果显示，当存在信息性辅助样本时，Oracle-TL和Trans-Lasso显著优于仅使用目标样本的Lasso。特别是，当对比向量足够稀疏时，这些方法可以接近Oracle方法的性能。此外，Trans-Lasso对非信息性辅助样本表现出很强的鲁棒性，即使在存在多个非信息性辅助样本的情况下，其性能也不会显著下降。

5.2 异质设计下的表现

我们还研究了设计矩阵在不同样本中分布不同的情况。结果表明，当设计矩阵足够相似时，Trans-Lasso仍然可以有效地从辅助样本中迁移知识。当设计矩阵差异较大时，算法的性能会下降，但在对比向量足够稀疏的情况下，仍然优于仅使用目标样本的方法。

6 GTEx数据应用

我们将所提出的方法应用于基因型-组织表达（GTEx）数据，这是一个包含来自多个人体组织的基因表达测量的数据集。我们的目标是预测目标组织中的基因表达，利用来自其他组织的数据作为辅助样本。

具体来说，我们选择皮肤组织作为目标，其他9个组织作为辅助样本。对于每个目标基因，我们使用与其相关的100个基因作为预测变量。我们比较了Trans-Lasso与仅使用目标样本的Lasso以及将所有样本池化在一起的方法。

结果显示，Trans-Lasso在大多数情况下都优于其他方法，特别是当目标样本量较小时。这表明，通过整合来自多个组织的数据，Trans-Lasso可以有效地提高基因表达预测的准确性。此外，我们观察到，Trans-Lasso自动给予与目标组织更相似的辅助组织更高的权重，这证实了算法能够自动识别和利用信息性辅助样本。

7 结论与讨论

本文研究了高维线性回归中的迁移学习问题，特别关注如何利用辅助样本来改进目标模型的估计和预测。我们提出了两种算法：Oracle Trans-Lasso（用于已知信息性辅助样本集的情况）和Trans-Lasso（用于未知信息性辅助样本集的情况）。

我们的理论结果表明，当信息性辅助样本集非空且对比向量足够稀疏时，这些算法可以达到比仅使用目标样本的方法更快的收敛率。此外，Trans-Lasso对非信息性辅助样本表现出很强的鲁棒性，并且在信息性辅助样本足够稀疏时是有效的。

我们的数值研究和实际数据应用证实了这些理论结果，并展示了所提出方法在实际问题中的有效性。特别是，在GTEx数据应用中，Trans-Lasso通过整合来自多个组织的数据，显著提高了基因表达预测的准确性。

本文的工作为高维统计中的迁移学习提供了理论基础和实用算法，可以应用于各种需要整合多个数据源的问题。未来的研究方向包括扩展到非线性模型、处理更复杂的依赖结构，以及开发更高效的计算算法以处理大规模数据集。

参考文献

1. Agarwal, A., Daumé III, H., & Gerber, S. (2012). Learning multiple tasks using manifold regularization. Advances in Neural Information Processing Systems, 25.
2. Bühlmann, P., & van de Geer, S. (2015). High-dimensional inference in misspecified linear models. Electronic Journal of Statistics, 9(1), 1449-1473.
3. Dai, D., Rigollet, P., & Zhang, T. (2012). Deviation optimal learning using greedy Q-aggregation. The Annals of Statistics, 40(3), 1878-1905.
4. Danaher, P., Wang, P., & Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. Journal of the Royal Statistical Society: Series B, 76(2), 373-397.

5. Daumé III, H. (2007). Frustratingly easy domain adaptation. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 256-263.
6. Hu, Y., Li, Y., Yang, M., & Shen, X. (2019). Integrative analysis of multi-omics data incorporating pathway information. Biometrics, 75(4), 1082-1092.
7. Mei, S., Wang, F., & Zhou, H. (2011). Gene ontology based transfer learning for protein subcellular localization. BMC Bioinformatics, 12(1), 44.
8. Pan, S. J., & Yang, Q. (2013). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359.
9. Rigollet, P., & Tsybakov, A. B. (2012). Sparse estimation by exponential weighting. Statistical Science, 27(4), 558-575.
10. Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... & Summers, R. M. (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.
11. Sun, Y., & Hu, Y. (2016). Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. Advances in Genetics, 93, 147-190.
12. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267-288.
13. Torrey, L., & Shavlik, J. (2010). Transfer learning. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques (pp. 242-264). IGI Global.
14. Turki, T., Wei, Z., & Wang, J. T. (2017). Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. IEEE Access, 5, 7381-7393.
15. Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., ... & Goldenberg, A. (2019). Similarity network fusion for

aggregating data types on a genomic scale. Nature Methods, 11(3), 333-337.

16. Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. Journal of Big Data, 3(1), 9.