迁移学习充分性检验

ZiqianLin[1],YuanGao[2],FeifeiWang[3,4*]和 Hansheng Wang[1]中国北京，北京大学光华管理学院； [2]中国上海，华东师范大学统计学院及 *KLATASDS-MOE*； [3]中国北京，中国人民大学应用统计研究中心； [4]中国北京，中国人民大学统计学院。

arXiv:2304.05636v1 [stat.ME] 12 Apr 2023

摘要

现代统计分析经常面临高维模型但样本量有限的情况。这使得基于目标数据的统计估计非常困难。那么，如何从规模更大的源数据中借用信息，以更准确地估计目标模型，便成为一个有趣的问题。这引出了迁移学习这一有用理念。近期，已有多种相关估计方法得到发展。在本文中，我们从不同角度研究迁移学习。具体而言，我们考虑迁移学习充分性检验问题。迁移学习充分性（记为零假设）是指，在源数据的帮助下，能够充分提取目标数据特征向量中包含的有用信息，以预测所关注的目标响应变量。因此，拒绝零假设意味着目标数据的特征向量中仍存在有助于预测的信息，因而需要进一步探索。为此，我们提出了一种新颖的检验程序和一个集中化、标准化的检验统计量，并对其渐近零分布进行了解析推导。我们通过模拟研究展示了所提方法的有限样本性能。最后，我们通过一个与深度学习相关的真实数据示例加以说明。

关键词：深度学习；高维数据；逻辑回归；统计假设检验；迁移学习；迁移学习充分性

通讯作者。电子邮箱：feifei.wang@ruc.edu.cn

1. 引言

现代统计分析经常面临棘手的情况：目标模型维度很高，但可获取的样本量却非常有限。医学、遗传学和图像相关研究尤其如此，因为伦理或成本方面的问题，往往难以获取支持性样本（Li 等，2022a）。以皮肤黑色素瘤（SKCM）疾病的分析为例。广泛使用的皮肤黑色素瘤数据集[†] 仅包含 347 名患者的样本，但每个样本测量的基因表达水平总数达到 18,335 个。现代深度学习研究是另一个例子。在这一研究领域，小样本学习（即用少量样本训练模型）这一课题受到研究人员越来越多的关注，因为数据收集和标注通常成本高昂（Wang 等，2021）。考虑到深度学习模型通常需要估计数千甚至数十万个参数，“小样本量与高维度”的问题便不可避免。于是，如何在样本量有限的情况下准确训练高维目标模型，就成了一个具有重要实际意义的问题。

解决这一问题的一种看似很有前景的方法是迁移学习（Pan and Yang，2009）。通过迁移学习，我们假设存在另一个规模庞大的辅助数据集，它与目标数据集有所不同，但又存在一定关联。为方便起见，我们将这个辅助数据集称为源数据集。生成源数据集和目标数据集的概率分布分别称为源总体和目标总体。值得注意的是，源总体通常不同于目标总体。为明确这一概念，设 $(X,\ Y)$ 和 $(X^*,\ Y^*)$ 分别为从目标总体和源总体中生成的观测值。目标总体中的感兴趣响应变量 Y 可能与源总体中的 Y^* 完全不同。例如， Y^* 通常被假设为取值属于 $\{0, 1, ..., K\}$ 的分类响应，但 Y 则允许为连续型。尽管如此，

[†]可通过 https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga 从癌症基因组图谱（TCGA）获取。

协变量 X 和 X^* 必须具有相同的维度 p，否则就无法进行迁移学习。尽管目标数据的样本量通常有限，但源数据的规模预计会非常庞大。因此，可以利用源数据以合理的精度估计各种高维模型。那么，如何将从源数据中学到的宝贵信息迁移到目标模型中，便成为关键问题。为此，人们已经开发出各种迁移学习方法。

在计算机科学文献中，存在多种迁移学习策略；有关出色的综述，请参见 Weiss 等（2016）和 Farahani 等（2020）。其中，最典型的迁移学习策略或许是基于特征的方法（Oquab 等，2014；Shin 等，2016；Guo 等，2018）。其核心思想是找到一个同时适用于源数据、并且对目标模型具有预测能力的共同特征空间。以分类响应 Y^* 为例，深度学习模型通常可以写为 $P(Y^*=k) = \exp\{Z_{\mathbb{B}}^\top(X^*)\gamma_k\}/1+\sum_{k=1}^K \exp\{Z_{\mathbb{B}}^\top(X^*)\gamma_k\}$，其中 $Z_{\mathbb{B}}(\cdot)$ 是一个特征生成函数（Donahue 等，2014；He 等人，2016；Ji 等，2021）。它将高维向量 X^* 映射为特征空间中维度较低的向量 $Z_{\mathbb{B}}(X^*)$。由于 X^* 的维度很高，特征生成函数的未知参数 $\mathbb{B}$ 的维度也很高。幸运的是，源数据的样本量通常很大。因此，未知参数 $\mathbb{B}$ （或其线性变换）可以得到较为准确的估计。$\widehat{}$一旦估计出 $\mathbb{B}$ （记为 $\mathbb{B}$），估计得到的特征生成函数 $Z_{\widehat{\mathbb{B}}}(\cdot)$ 就可以直接应用于具有高维协变量 X 的目标数据。由此得到新估计的特征向量 $Z_{\widehat{\mathbb{B}}}(X)$。通过对 $Z_{\widehat{\mathbb{B}}}(X)$ 与 Y之间的关系建模，就可以将从源数据中学习到的宝贵信息迁移到目标模型中。

近年来，迁移学习在统计学文献中引起了广泛关注，但研究视角略有不同。许多研究者针对不同模型开发了新的算法，例如线性回归（Li 等，

2022a）、广义线性回归（Tian 和 Feng，2022）、非参数回归（Cai 和 Wei，2021）、多重检验（Liang 等，2022）和高斯图模型（Li 等，2022b）。这里的总体思路是同时估计目标模型和源模型，然后利用源模型与目标模型之间的相似性，提高目标模型的估计精度。例如，Li 等（2022a）提出了一种用于高维线性回归模型迁移学习的 Trans-Lasso 方法。其基本思路是将源数据和目标数据合并，以获得一个初始估计量。随后，利用目标数据校正初始估计量的偏差。Tian 和 Feng（2022）也为广义线性回归模型开发了类似的方法。

受计算机科学文献和统计学文献中这些先驱性研究的启发，我们在此采取一种非常独特的视角来研究迁移学习。一方面，我们的方法与计算机科学文献中的方法具有相似的理念。具体而言，我们假设存在一个特征生成函数 $Z_{\mathbb{B}}(X)$，该函数可以通过源数据进行学习，然后迁移到目标模型中，用于生成低维特征向量。我们主要遵循这一方向，是因为它可能是实践中应用最广泛的迁移学习策略。事实上，许多最常用的深度学习框架（例如 TensorFlow 和 PyTorch）都已实现了这一迁移学习策略。另一方面，我们提出的方法也受益于统计学文献中的先驱性工作。我们受到其中发展起来的统计理论的启发。借助这些富有启发性的统计理论，人们可以更好地理解源数据与目标数据之间的信息传递动态。因此，可以在统计保证下发展出新的估计与推断程序。

作为我们的首次尝试，我们遵循以往文献，从最简单但可能也是最重要的特征生成函数入手，即线性变换（Lu and Zhang, 2010;Fan and Fine, 2013）。具体而言，我们假设对于源数据

一个标准的多项逻辑回归模型为 $P(Y^* = k) = \exp\{X^{*\top}\beta_k\}/[1 + \sum_{k=1}^{K}\exp\{X^{*\top}\beta_k\}]$，其中 $1 \le k \le K$。随后，最简单的特征生成函数或许可以定义为 $Z^* = Z_{\mathbb{B}}(X^*) = \mathbb{B}^\top X^*$，其中 $\mathbb{B} = (\beta_1, \beta_2, ..., \beta_K) \in \mathbb{R}^{p\times K}$，且 $K \ll p$。接下来，我们假设目标数据服从如下二元逻辑回归模型：$P(Y = 1) = \exp(X^\top\theta)/\{1 + \exp(X^\top\theta)\}$，其中 $\theta \in \mathbb{R}^p$ 是我们感兴趣的目标参数。然后，我们考虑如何迁移从源数据中学到的关于 $\mathbb{B}$ 的知识，以帮助估计 θ。这里的关键问题是，$\theta \in \mathcal{S}(\mathbb{B})$是否成立，其中$\mathcal{S}(\mathbb{B})$表示由 $\mathbb{B}$的列向量生成的线性空间。如果 $\theta \in \mathcal{S}(\mathbb{B})$确实成立，那么对于某个 $\gamma \in \mathbb{R}^K$，我们应有 $\theta = \mathbb{B}\gamma$。因此，目标模型可以改写为 $P(Y = 1) = \exp(Z^\top\gamma)/\{1 + \exp(Z^\top\gamma)\}$，其中$Z = \mathbb{B}^\top X \in \mathbb{R}^K$。此后，我们无需直接处理高维向量$X$，而可以处理降维向量 Z，且不会损失预测准确率。在这种情况下，我们称迁移学习在如下意义上是充分的：$P(Y = 1|\mathbb{B}^\top X) = P(Y = 1|X)$。否则，如果 $\theta \notin \mathcal{S}(\mathbb{B})$，那么对于相当多的 X，我们应有$P(Y = 1|\mathbb{B}^\top X) \neq P(Y = 1|X)$。因此，预测准确率会出现不可忽略的损失。在这种情况下，我们称迁移学习是不充分的。换言之，关于 Y、有助于实现更准确预测的宝贵信息仍保留在 X 中，尚未被 $\mathbb{B}^\top X$完全捕获。

在实际应用中，$\mathbb{B}$ 是一个未知参数。因此，如何估计 $\mathbb{B}$ 就成了一个重要问题。借助源数据，我们可以轻松得到一个估计量$\widehat{\mathbb{B}} = (\widehat{\beta}_1, \widehat{\beta}_2, ..., \widehat{\beta}_K)$，其中 $\widehat{\beta}_k$是（例如）基于源数据计算得到的β_k 最大似然估计量。在适当的正则条件下，我们应当有 $\mathbb{B}$ 在 $N \to \infty$时渐近地满足一致性和渐近正态性。由于源数据通常具有很大的样本量，我们可以预期 $\mathbb{B}$ 具有足够的精度。利用估计得到的 $\widehat{\mathbb{B}}$，可以用 $\widehat{Z} = \widehat{\mathbb{B}}^\top X$替代低维向量 Z。然后，相应地可以使用 Z 作为输入来估计目标模型。由此得到系数 γ的最大似然估计量 $\widehat{\gamma}$。结合关系式 $\theta = \mathbb{B}\gamma$，

$\widehat{\quad}$ $\widehat{\quad}$可以将 θ 的估计量定义为 $\theta = \mathbb{B}\widehat{\gamma}$，我们称之为迁移学习（TL）估计量。为从理论上支持我们的方法，我们建立了严格的渐近理论。我们发现，在适当的正则条件下，所得的$\widehat{\ }$迁移学习估计量 θ 在渐近意义下可与预言机估计量同样有效。这里，预言机估计量定义为利用完美恢复的低维特征向量 $Z_{\mathbb{B}}(X)$得到的理想估计量。

由于目标数据通常极具价值，我们希望借助迁移学习，充分利用目标特征中包含的有用信息。否则，目标特征中仍会有大量有价值的信息有待进一步挖掘。这引出了一个有趣的理论问题：如何检验迁移学习充分性。为解决这一问题，我们在此提出一种新颖的检验程序。其核心思想如下。如前所述，在迁移学习充分性原假设成立时，借助源数据可以为目标模型得到回归系数 θ 的一致估计量。需要注意的是，仅使用目标数据很难完成这一任务。随后，可以对伪残差进行求差。在适当的正则条件下，渐近地，伪残差应与每个特征都不相关。因此，我们预期伪残差与每个预测变量之间的样本协方差都接近于 0。由此，我们想到计算每个预测变量对应的样本协方差的平方，并将它们相加。这便得到一个有趣的检验统计量，其思路与 Lan 等人 (2014) 的方法相似。然而，正如 Lan 等人 (2014) 所展示的，要使这种形式的检验统计量具有非退化的渐近分布，必须进行适当的中心化操作。由此得到的中心化检验统计量显著降低了偏差和变异性。我们严格证明，在零假设下，经过适当的位置—尺度变换后，该中心化检验统计量渐近服从标准正态分布。我们通过大量模拟研究展示该

检验程序的有限样本性能。我们还分析了一个与深度学习相关的真实数据集，以作说明。

本文其余部分安排如下。第 2 节介绍我们针对迁移学习的理论框架。基于所提出的理论框架，我们为目标模型构建了一种基于迁移学习的最大似然估计量，并对其渐近性质进行了严格研究。随后，我们研究迁移学习充分性检验问题。为此，我们基于伪残差与预测变量之间样本协方差的平方和，构建了一种高维检验程序。经过适当的中心化和标准化后，我们严格推导了其渐近零分布。第 3 节通过对模拟数据集和真实数据集进行大量数值研究，评估所提方法的有限样本性能。最后，第 4 节对全文作简要讨论并总结。

2. 迁移学习模型

2.1. 迁移学习估计量

考虑一个总样本量为 n 的目标数据集。设$(X_i,\ Y_i)$为从目标数据集中第 i个受试者（$1 \leq i \leq n$）收集到的观测值。这里，$Y_i \in \{0,1\}$是一个二元响应变量，$X_i = (X_{i1}、\ldots\ldots、X_{ip})^\top \in \mathbb{R}^p$是相应的 p维特征。不失一般性，我们假设 X_i 中的所有特征都已中心化，使得 $E(X_i) = 0$。进一步假设，不同受试者的$(X_i,\ Y_i)$独立同分布。为刻画 X_i 与 Y_i之间的回归关系，我们考虑如下标准逻辑回归模型

$$P\Big(Y_i = 1\big|X_i\Big) = \frac{\exp(X_i^\top\theta)}{1+\exp(X_i^\top\theta)} = g\Big(X_i^\top\theta\Big), \tag{2.1}$$

其中 $\theta = (\theta_1,\ldots,\theta_p)^\top \in \mathbb{R}^p$是 p维回归系数，且 $g(x) =$

对于任意标量 x，$\exp(x)/\{1+\exp(x)\}$ 是 Sigmoid函数。我们将 (2.1) 称为目标模型。要估计目标模型，可以使用最大似然估计法。具体而言，可以构造如下对数似然函数

$$\mathcal{L}_n^{\text{Target}}(\theta) = \sum_{i=1}^{n} \left[Y_i \left(X_i^\top \theta \right) - \log \left\{ 1 + \exp \left(X_i^\top \theta \right) \right\} \right]. \tag{2.2}$$

于是，目标模型的最大似然估计量（最大似然估计）可表示为$\widehat{\theta}^{\text{Target}}{}_{\text{最大似然估计}} = \text{argmax}_\theta \mathcal{L}_n^{\text{Target}}(\theta)$。如果特征维度 p 固定且目标样本量为 $n \to \infty$，那么我们应有 $\widehat{\theta}^{\text{Target}}{}_{\text{最大似然估计}}$ 具有 $\sqrt{n}$-一致性且渐近正态（McCullagh and Nelder, 1989; Shao, 2003）。遗憾的是，我们经常会遇到这种情况：目标样本量 n 非常有限，而特征维度 p却极高。这会导致 $n \approx p$ 或 $n < p$所描述的棘手情形。在这种情况下，最大似然估计量可能存在偏差，甚至在计算上不可行（Sur and Candes, 2019; Candes and Sur, 2020）。

为了解决这个问题，我们可以借助迁移学习。其基本思路是从源数据中借用信息。具体而言，定义 (X_i^*, Y_i^*) 为源数据中第 i 个样本的观测值，其中 $1 \le i \le N$。假设协变量 $X_i^* = (X_{i1}^*, \ldots, X_{ip}^*)^\top \in \mathbb{R}^p$ 与 X_i具有相同的维度。同样，我们也假设 $E(X_i^*) = 0$。不过，与目标数据不同，我们假设源数据中的响应变量 Y_i^*是一个 $(K+1)$ 级分类变量，即 $Y_i^* \in \{0, 1, \ldots, K\}$。我们通常希望 K 相对较大，从而能从 Y_i^*中获得充足的信息。同时， K 也不应过大；相反，它应远小于目标样本量 n。否则，在样本量有限的情况下，就无法将 Y_i^*提供的丰富信息便捷地迁移到目标模型中。

为研究 X_i^* 和 Y_i^* 之间的回归关系，我们作出如下假设

多分类逻辑回归模型为

$$P(Y_i^* = k|X_i^*) = \frac{\exp(X_i^{*\top}\beta_k)}{1 + \sum_{j=1}^{K} \exp(X_i^{*\top}\beta_j)} \text{ for } 1 \le k \le K, \tag{2.3}$$

其中 $\beta_k = (\beta_{k1}, \ldots, \beta_{kp})^\top \in \mathbb{R}^p$ 是相对于基类0 的类别 k 的 p维回归系数。我们将 (2.3) 称为源模型，并定义$\mathbb{B} = (\beta_1, \ldots, \beta_K) \in \mathbb{R}^{p\times K}$ 为相应的系数矩阵。于是，源模型的对数似然函数可写为

$$\mathcal{L}_N^{\text{Source}}(\mathbb{B}) = \sum_{i=1}^{N} \left[\sum_{k=1}^{K} I(Y_i^* = k) X_i^{*\top}\beta_k - \log\left\{ 1 + \sum_{k=1}^{K} \exp\left(X_i^{*\top}\beta_k \right) \right\} \right].$$

此后，源模型的最大似然估计量可计算为$\widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_K) = \text{argmax}_{\mathbb{B}} \mathcal{L}_N^{\text{Source}}(\mathbb{B})$。在适当的正则条件下，我们应有 $\widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}}$是一致且渐近正态的 $N \to \infty$（Fan 和 Peng，2004）。值得注意的是，在本研究中，我们假设源样本量 N 远大于目标样本量 n 和特征维度p。因此，基于大量源数据，参数 $\mathbb{B}$ 可以得到足够精确的估计。接下来，我们将探讨如何通过从 $\widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}}$迁移有用信息来估计目标模型的参数 θ 。

为了通过 $\widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}}$ 从源数据中借用信息，我们需要定义一个特征生成函数 $Z_{\mathbb{B}}(X)$，它可以通过源数据学习，然后迁移到目标数据。给定源模型 (2.3)，我们可以为源数据定义一个可以说是最简单的特征生成函数 $Z_i^* = Z_{\mathbb{B}}(X_i^*) = \mathbb{B}^\top X_i^* \in \mathbb{R}^K$ 。我们通常假设 $K \ll p$，因此特征生成函数生成的新特征 Z_i^* gen 的维度远小于原始特征 X_i^*。接下来的问题是，这个特征生成函数 $Z_{\mathbb{B}}(X)$ 是否适用于目标数据。换句话说，目标数据中原始高维特征 X_i 所包含的丰富信息，能否被充分表

示为降维特征 $Z_i = Z_{\mathbb{B}}(X_i) = \mathbb{B}^\top X_i \in \mathbb{R}^K$。如果答案是肯定的，我们应该能够写出 $X_i^\top \theta = Z_{\mathbb{B}}^\top(X_i)\gamma = (X_i^\top \mathbb{B})\gamma = Z_i^\top \gamma$，其中 $\gamma = (\gamma_1, \ldots\ldots, \gamma_K)^\top \in \mathbb{R}^K$。相应地，我们应该有 $\theta = \mathbb{B}\gamma$。如果出现这种情况，我们应该有 $P(Y_i = 1|X_i) = P(Y_i = 1|Z_i)$。因此，我们称 $Z_{\mathbb{B}}(X)$ 所提供的迁移学习是充分的。否则，我们称迁移学习是不充分的。

回顾一下，Z_i 的维度远小于 X_i的维度。因此，我们无需直接处理高维向量 X_i，而可以处理带有回归系数 γ的降维向量 Z_i。于是，目标模型 (2.1) 可重写为

$$P\Big(Y_i = 1\big|Z_i\Big) = \frac{\exp(Z_i^\top \gamma)}{1 + \exp(Z_i^\top \gamma)}.$$

由于 $\mathbb{B}$ 是一个未知参数，我们必须用某个恰当的估计量替代它。如前所述，一个自然的选择是最大似然估计量 $\widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}}$。相应地，我们可以用 $\widehat{Z}_i = \widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}\top} X_i$替代 Z_i。由此得到如下工作模型：

$$P\Big(Y_i = 1\big|Z_i\Big) \approx \frac{\exp(\widehat{Z}_i^\top \gamma)}{1 + \exp(\widehat{Z}_i^\top \gamma)}.$$

相应地，工作对数似然函数可写为

$$\mathcal{L}_n(\gamma; \widehat{Z}) = \sum_{i=1}^{n} \left[Y_i\Big(\widehat{Z}_i^\top \gamma\Big) - \log\left\{1 + \exp\Big(\widehat{Z}_i^\top \gamma\Big)\right\}\right]. \tag{2.4}$$

接着，通过最大化 (2.4)$\widehat{\ }$，即可得到关于 γ 的极大似然型估计量：$\widehat{\gamma} = \text{argmax}_\gamma \mathcal{L}_n(\gamma; Z)$。回顾一下，$\theta = \mathbb{B}\gamma$。于是，可将 θ 的一个估计量定义为 $\widehat{\theta}^{\text{Transfer}} = \widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}}\widehat{\gamma}$；该估计量不仅利用了从目标数据中学习到的信息，也利用了从源数据迁移而来的信息。为方便起见，我们称其为迁移学习（TL）估计量。

2.2. 迁移最大似然估计量

接下来，我们研究迁移学习估计量的渐近性质。令 $\lambda_{\min}(A)$ 和$\lambda_{\max}(A)$ 分别表示任意方阵 A 的最小和最大特征值。定义 $\pi_{ik} = P(Y_i^* = k|X_i^*)$ 和 $\pi_i = (\pi_{i1}, \ldots, \pi_{iK})^\top \in \mathbb{R}^K$。记作$W_i = \text{diag}(\pi_i) - \pi_i\pi_i^\top$。定义 $\dot{\mathcal{L}}_n(\gamma; Z) \in \mathbb{R}^K$ 为预言机对数似然函数 $\mathcal{L}_n(\gamma; Z)$ 关于 γ的一阶导数。这里，预言机对数似然函数 $\mathcal{L}_n(\gamma; Z)$ 是基于真实特征向量 Z_i 而非其估计值 $\widehat{Z}_i$计算的。进一步定义 $I(\gamma) = n^{-1}E\big[\dot{\mathcal{L}}_n(\gamma; Z)\,\dot{\mathcal{L}}_n(\gamma; Z)^\top\big] = E\big[g(Z_i^\top\gamma)\{1 - g(Z_i^\top\gamma)\}Z_iZ_i^\top\big]$，作为预言机对数似然函数的费 Fisher 信息矩阵。接下来，对于随机变量U，将其次高斯范数定义为 $\|U\|_{\psi_2} = \inf\{t > 0: E\exp(U^2/t^2) \le 2\}$；对于随机向量 V，则定义为$\|V\|_{\psi_2} = \sup_{\|u\|=1}\|u^\top V\|_{\psi_2}$。于是，若随机向量 V满足$\|V\|_{\psi_2} < \infty$，我们就称其为次高斯随机向量；关于次高斯范数的更详细讨论，参见 Vershynin (2018)。对于任意矩阵 $A = (a_{ij}) \in \mathbb{R}^{m\times n}$，将其 Frobenius 范数定义为$\|A\|_F = \sqrt{\sum_{i=1}^m\sum_{j=1}^n a_{ij}^2}$。为研究迁移学习估计量的理论性质，必须满足以下技术条件：

(C1) 当 $N \to \infty$ 时，应有 $n^2 \log N/N \to 0$ 以及 $p \to \infty$。此外，存在一个正数常量 $0 < C < \infty$，使得 $p \le Cn$。

(C2) 随机向量 X_i^*和 X_i 均为次高斯。

（C3）记 $\otimes$ 为克罗内克积。作为 $p \to \infty$，存在两个正数常量 $0 < C_{\min} \le C_{\max} < \infty$，使得：（1） $C_{\min} \le \lambda_{\min}\{E(X_i^*X_i^{*\top})\} \le \lambda_{\max}\{E(X_i^*X_i^{*\top})\} \le C_{\max}$，（2） $C_{\min} \le \lambda_{\min}[E\{W_i \otimes (X_i^*X_i^{*\top})\}] \le \lambda_{\max}[E\{W_i \otimes (X_i^*X_i^{*\top})\}] \le C_{\max}$，（3） $C_{\min} \le \lambda_{\min}\{E(X_iX_i^\top)\} \le \lambda_{\max}\{E(X_iX_i^\top)\} \le C_{\max}$，以及（4） $C_{\min} \le \lambda_{\min}(\Sigma_\gamma) \le \lambda_{\max}(\Sigma_\gamma) \le C_{\max}$。这里 $\Sigma_\gamma = E(\varepsilon_i^2X_iX_i^\top)$，其中 $\varepsilon_i = Y_i - E(Y_i|X_i)$ 是伪残差。

（C4）作为 $p \to \infty$，存在一个正数常量 $\tau_{\max} > 0$，使得 $E\|Z_i\|^4 \leq \tau_{\max}$ 且对任意 $i \neq j$，$E(X_i^\top X_j)^4 \leq \tau_{\max} p^2$。

(C5) 作为 $p \to \infty$，我们有 $\lambda_{\max}(\mathbb{B}^\top \mathbb{B}) \leq \tau_{\max}$，其中 $\tau_{\max}$ 与 (C4) 中相同。

条件（C1）意味着，源数据的样本量应足够大，即 $n^2 \log N/N \to 0$。条件（C1）还限制特征维度 p 的发散速度，使其阶数不高于目标样本量 n。不过，我们允许 $p > n$。在这种情况下，应有 $C > 1$。条件（C2）略强于标准矩条件。若 X_i 服从多元正态分布，则该条件显然成立。以往文献中曾采用过类似但更强的条件 $\sup_i \|X_i\| = O_p(\sqrt{p})$（Portnoy，1985；Welsh，1989；He 和 Shao，2000；Wang，2011）。条件（C3）对维度不断增大的各种正定矩阵 p 的最大和最小特征值施加了一致界。条件（C4）对 Z_i 和 X_i 提出了两个矩条件，这些条件相当合理。举例来说，考虑 X_i 服从多元标准正态分布这一特殊情形。此时可以验证，$E(X_i^\top X_j)^4 = 3\text{tr}^2\{E(X_i X_i^\top)\} + 6\text{tr}\{E^4(X_i X_i^\top)\} = 3p^2 + 6p \leq \tau_{\max} p^2$，其中 $\tau_{\max} = 9$。条件（C5）同样是一个相当合理的假设。为了快速理解该条件，假设条件（C5）不成立。那么必然存在一个向量 γ，其长度为单位长度 $||\gamma|| = 1$。同时，我们有$\gamma^\top(\mathbb{B}^\top \mathbb{B})\gamma = \lambda_{\max}(\mathbb{B}^\top \mathbb{B}) \to \infty$，当 $p \to \infty$ 时。于是有 $\text{var}(Z_i^\top \gamma) = \text{var}(X_i^\top \mathbb{B}\gamma) = \gamma^\top \mathbb{B}^\top E(X_i X_i^\top)\mathbb{B}\gamma \geq C_{\min}\{\gamma^\top(\mathbb{B}^\top \mathbb{B})\gamma\} = C_{\min}\lambda_{\max}(\mathbb{B}^\top \mathbb{B}) \to \infty$。在这种情况下，以 $\text{var}(Z_i^\top \gamma)$ 衡量的 Z_i 中的信息会发散至无穷大。因此，我们可以通过 Z_i对 Y_i 进行近乎完美的预测。这显然不合理。因此，条件（C5）必须成立。在上述条件下，我们有如下定理。

定理1. 假设条件 *(C1)-(C5)* 成立，则 $\|\widehat{\mathbb{B}}_{mle}^{Source} - \mathbb{B}\|_F = O_p(\sqrt{p/N})$。

定理1的技术证明与 Fan 和 Peng（2004）的证明类似。不过，为了保证理论上的完整性，我们在附录A中给出了详细证明。根据定理1，我们知道，源模型的最大似然估计量 $\widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}}$是 $\sqrt{N/p}$-相合的。这意味着，源样本量很大时，参数$\mathbb{B}$ 可以得到较为精确的估计。计算出 $\widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}}$后，可以得到一个低维特征向量 $\widehat{Z}_i = \widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}\top} X_i$。随后，可以建立一个以 Z_i 为输入的修正目标模型。相应估计量 $\widehat{\gamma}$ 的理论性质总结如下定理所示。

定理 2. 假设条件 *(C1)-(C5)* 成立，则有：*(1)* $\sqrt{n}(\widehat{\gamma} - \gamma) \xrightarrow{d} N(0, I(\gamma)^{-1})$; *(2)* $\sqrt{n} v^\top \left(\widehat{\theta}^{Transfer} - \theta \right) / \sigma(v) \xrightarrow{d} N(0, 1)$，其中 v 是 $\mathcal{S}(\mathbb{B})$ 中任意一个具有单位长度的 p 维向量，且 $\sigma^2(v) = v^\top \mathbb{B} I(\gamma)^{-1} \mathbb{B}^\top v$。

定理 2 的证明见附录 B。定理 2 表明，借助源模型估计量 $\widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}}$，目标模型的估计量 $\widehat{\gamma}$ 应当是 $\sqrt{n}$-相合且渐近正态的。特别是，其渐近分布与预言机估计量的渐近分布相同，该预言机估计量定义为 $\widehat{\gamma}_{\text{oracle}} = \text{argmax}_\gamma \mathcal{L}_n(\gamma; Z)$。这是使用真实的低维特征向量 Z_i，而非其估计量 Z_i 得到的最大似然估计量。此外，结合关系 $\theta = \mathbb{B}\gamma$，可得到 θ 的迁移学习估计量 $\widehat{\theta}^{\text{Transfer}} = \widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}} \widehat{\gamma}$。因此，定理 2 表明，对于任意 $v \in \mathcal{S}(\mathbb{B})$，估计量 $v^\top \widehat{\theta}^{\text{Transfer}}$ 对 $v^\top \theta$ 是 $\sqrt{n}$-相合的，且渐近正态。

2.3. Testing Sufficiency for Transfer Learning

上一节中获得的良好理论性质依赖于一个关键假设。也就是说，X_i 关于 Y_i 在目标数据中的所有信息都能由 $Z_i = \mathbb{B}^\top X_i$充分表示。换言之，迁移学习是充分的。然而，对于实际数据集而言，迁移学习是否充分并不立即清楚。因此，亟需发展一种有原则的统计

测试迁移学习充分性的程序。随后，本节将着手开发一种针对此问题的新颖解决方案。

我们的方法受到以下有趣观察的启发。如果迁移学习在这样的意义下是充分的：X_i 中关于 Y_i 的所有信息都已被 Z_i充分捕获，那么我们应有 $E(Y_i|Z_i, X_i) = E(Y_i|Z_i) = g(Z_i^\top\gamma)$。于是，伪残差 $\varepsilon_i = Y_i - E(Y_i|Z_i, X_i) = Y_i - g(Z_i^\top\gamma)$ 应当与特征向量 X_i 不相关，即 $E(\varepsilon_i X_i) = 0$。实际上，真实的特征向量Z_i 和系数参数 γ 都是未知的。然而，借助源数据，它们都可以分别估计为 $\widehat{Z}_i = \widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}\top} X_i$ 和 $\widehat{\gamma} = \text{argmax}_\gamma \mathcal{L}_n(\gamma;\ \widehat{Z})$。于是，可以得到估计的伪残差 $\widehat{\varepsilon}_i = Y_i - g(\widehat{Z}_i^\top\widehat{\gamma})$。因此，我们可以合理地预期 $n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_i X_i \approx 0$。这引出了初步的检验统计量 $T_1 = \|n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_i X_i\|^2 = n^{-2}\widehat{\varepsilon}^\top\mathbb{X}\mathbb{X}^\top\widehat{\varepsilon}$，其中 $\widehat{\varepsilon} = (\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n)^\top \in \mathbb{R}^n$ 是估计的伪残差向量，而 $\mathbb{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n\times p}$ 是设计矩阵。如果迁移学习充分性原假设不成立，我们应预期 $\widehat{\varepsilon}_i$ 与 X_{ij} 在某个 $1 \le j \le p$下存在实质性相关。这应使 T_1 的值异常地大。Lan 等人（2014）也构造了类似的检验统计量，但用于测试高维回归模型的统计显著性。为了理解 T_1的渐近行为，计算迁移学习充分性原假设下的均值和方差十分重要。这引出了以下定理。

定理 *3.* 假设条件 *(C1)-(C4)* 和迁移学习充分性原假设成立。定义 $\Sigma_\gamma = E(\varepsilon_i^2 X_i X_i^\top) = E[g(Z_i^\top\gamma)\{1 - g(Z_i^\top\gamma)\}X_i X_i^\top]$。我们进一步假设 $n^2 p/N \to 0$。则有 $T_1 = T_1^* + o_p(\sqrt{p/n})$，其中 $T_1^* = n^{-2}\varepsilon^\top\mathbb{X}\mathbb{X}^\top\varepsilon$ 且 $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$ 是真实的伪残差向量。对于 T_1^*，有 $E(T_1^*) = n^{-1}tr(\Sigma_\gamma)$，且 $var(T_1^*) = n^{-3}var(\varepsilon_i^2 X_i^\top X_i) + 2n^{-2}(1 - n^{-1})tr(\Sigma_\gamma^2)$。

定理 3的证明见附录 C。根据定理 3和条件(C3)，我们有 $E(T_1^*) \ge (2n)^{-1}\text{tr}(\Sigma_\gamma) \ge 0.5C_{\min}n^{-1}p$。进一步定义 $\text{SE}(T_1^*) =$

{var(T_1^*)}$^{1/2}$。然后根据定理3和条件（C3），我们可以得到 SE(T_1^*) $\leq\{4n^{-3}E($ $X_iX_i^\top)^4+4n^{-2}(1-n^{-1})$tr($\Sigma_\gamma^2$)$\}^{1/2}\leq 2c_{\max}(n^{-3/2}p+n^{-1}p^{1/2})$，其中 $c_{\max}$ =max $\{C_{\max},\tau_{\max}\}$。因此，当 $n\to\infty$ 时，应有 $E(T_1^*)$/SE(T_1^*) $\to\infty$。因此，随机变量 T_1/SE(T_1)不可能依分布收敛到任何非退化的概率分布。Lan等人（2014）也得到了类似的有趣现象。为了解决这个问题，需要对 T_1 进行适当的中心化。

2.4. 集中式检验统计量

如上文所述， $E(T_1)$ 的阶显著大于 SE(T_1) 的阶。因此，必须对 T_1 进行适当的中心化处理。以往文献中也采用过类似操作（Li 和 Racine，2007；Lan 等人，2014）。注意到， $E(T_1)$ 的首项与 $tr(\Sigma_\gamma)$ 线性相关，其中 $\Sigma_\gamma=E(\varepsilon_i^2X_iX_i^\top)$。因此，只要能够获得 Σ_γ 的准确估计量，就可以通过减去该项来显著降低 T_1 的偏差。为此，可以将 Σ_γ 的简单矩估计量构造为 $\widehat{\Sigma}_\gamma=n^{-1}\sum_{i=1}^n\widehat{\varepsilon}_i^2X_iX_i^\top$。由此得到集中式检验统计量 $T_2=T_1-n^{-1}$tr($\widehat{\Sigma}_\gamma$)。随后， T_2 的均值和方差由以下定理给出。

定理 4。假设条件 *(C1)-(C4)*以及迁移学习充分性原假设成立。进一步假设 $n^2p/N\to 0$。则有 $T_2=T_2^*+o_p(\sqrt{p}/n)$，其中 $T_2^*=n^{-2}\varepsilon^\top(\mathbb{X}\mathbb{X}^\top-\mathbb{D}_X)\varepsilon$ 且 $\mathbb{D}_X$= ***diag***$\{X_1^\top X_1,\ldots,X_n^\top X_n\}\in\mathbb{R}^{n\times n}$。对于 T_2^*，有 $E(T_2^*)=0$ 且 ***var***$(T_2^*)=2n^{-2}$***tr***(Σ_γ^2) $\{1+o(1)\}$。

定理 4 的详细证明见附录 D。根据定理 4，我们得到以下有趣的发现。首先，我们发现，集中式统计量 T_2^* 的渐近方差显著小于 T_1^* 的渐近方差，相差 n^{-3}var($\varepsilon_i^2X_i^\top X_i)=O_p(p^2/n^3)$。其次，注意到 SE($T_2^*$) $=\{$var(T_2^*)$\}^{1/2}=\{2n^{-2}$tr(Σ_γ^2) $\}^{1/2}\geq C_{\min}n^{-1}p^{1/2}$。与 SE($T_2^*$) 的阶相比，我们发现该阶

$E(T_2^*)$ 是一个小阶项。因此，可以构造形式上的标准化检验统计量 $T_3 = T_2/\sqrt{2n^{-2}\text{tr}(\Sigma_\gamma^2)}$，其渐近行为总结如下定理。

定理 5. 假设条件（*C1*）-（*C4*）和迁移学习充分性原假设成立。进一步假设 $n^2p/N \to 0$，则当 $N \to \infty$ 时，$T_3 \xrightarrow{d} N(0,\ 1)$。

定理 5 的详细证明见附录 E。定理 5 表明，中心化且标准化的检验统计量 T_3 渐近地服从标准正态分布。注意，项 $\text{tr}(\Sigma_\gamma^2)$ 出现在 T_3 中。因此，要在实际中应用这一检验程序，需要一个 $\text{tr}(\Sigma_\gamma^2)$ 的比率一致估计量。为解决这一问题，一个自然的想法是使用插件估计量 $\text{tr}(\widehat{\Sigma}_\gamma^2)$。该插件估计量的期望可计算为 $E\{\text{tr}(\widehat{\Sigma}_\gamma^2)\} = n^{-1}E\{\varepsilon_i^4(X_i^\top X_i)^2\} + (1-n^{-1})\text{tr}(\Sigma_\gamma^2)\{1+o(1)\}$。根据条件（C3），我们可以推导出 $n^{-1}E\{\varepsilon_i^4(X_i^\top X_i)^2\} \geq (2n)^{-1}\{E(\varepsilon_i^2 X_i^\top X_i)\}^2 = (2n)^{-1}\text{tr}^2(\Sigma_\gamma) \geq C_{\min}^2(2n)^{-1}p^2$。我们还可得 $\text{tr}(\Sigma_\gamma^2) \leq C_{\max}^2 p$。这表明 $E\{\text{tr}(\widehat{\Sigma}_\gamma^2)\}/\text{tr}(\Sigma_\gamma^2) \geq p/n$。因此，当 p 大于 n时，$\text{tr}(\widehat{\Sigma}_\gamma^2)$ 不能作为 $\text{tr}(\Sigma_\gamma^2)$ 的比率一致估计量，即当 $N\to\infty$ 时，$\text{tr}(\widehat{\Sigma}_\gamma^2)/\text{tr}(\Sigma_\gamma^2) \not\xrightarrow{p} 1$。$\text{tr}(\widehat{\Sigma}_\gamma^2)$ 的不一致性主要源于额外的偏差项 $n^{-1}E\{\varepsilon_i^4(X_i^\top X_i)^2\}$。因此，我们旨在构造 $\text{tr}(\Sigma_\gamma^2)$ 的偏差校正估计量，形式如下：

$$\widehat{\text{tr}(\Sigma_\gamma^2)} = \text{tr}(\widehat{\Sigma}_\gamma^2) - n^{-2}\sum_{i=1}^{n}\widehat{\varepsilon}_i^4\left(X_i^\top X_i\right)^2 = n^{-2}\sum_{i\neq j}\widehat{\varepsilon}_i^2\widehat{\varepsilon}_j^2\left(X_i^\top X_j\right)^2.$$

$\text{tr}(\Sigma_\gamma^2)$ 的比率一致性性质由以下定理概述。

定理6。假设条件 *(C1)-(C4)* 和迁移学习充分性原假设成立。进一步假设 $n^2p/N \to 0$，则 $tr(\Sigma_\gamma^2)/tr(\Sigma_\gamma^2) \xrightarrow{p} 1$ 当 $N\to\infty$。定理6的详细证明见附录F。定理6意味着当 $N\to\infty$ 时，$\text{tr}(\Sigma_\gamma^2)$ 是 $\text{tr}(\Sigma_\gamma^2)$ 的比值相合估计量。利用这一比值相合估计量，我们可以构造一个实际可计算的检验统计量： $T_4 =$

$T_2/\{2n^{-2}\widehat{\text{tr}}(\Sigma_\gamma^2)\}^{1/2}$。在迁移学习充分性原假设下，我们应有 T_4 渐近地服从标准正态分布。因此，如果 T_4 的值足够大，就应拒绝迁移学习充分性原假设。具体而言，在给定显著性水平 α 下，如果 $T_4 > z_{1-\alpha}$，则应拒绝原假设，其中 z_α 是标准正态分布的 α 分位数。否则，应接受迁移学习充分性原假设。

3. 数值研究

3.1. 模拟设置

为展示迁移学习估计量和迁移学习充分性检验统计量的有限样本性能，我们在此进行若干数值研究。整个研究分为两部分。第一部分重点考察迁移学习估计量 $\widehat{\theta}^{\text{Transfer}}$ 的估计性能，旨在数值研究特征维度 p、目标样本量 n 和源样本量N 对迁移学习充分性原假设下迁移学习估计量统计效率的影响。第二部分研究所提出的迁移学习充分性检验统计量 T_4 的有限样本性能。在不同的实验设置下，评估检验尺寸（即第一类错误）和检验功效（p，n，N）。在整个研究中，我们固定源数据中的类别数 $K = 8$。随后，给定特征维度 p、目标样本量 n 和源样本量 N，详细的数据生成过程如下。

Source Data. 我们首先生成协变量$X_i^* \in \mathbb{R}^p$，其 $1 \le i \le N$ 独立同分布地来自均值为0、协方差矩阵为 $\Sigma = (\sigma_{j_1 j_2})_{p\times p}$ 的多元正态分布 $\sigma_{j_1 j_2} = 0.5^{|j_1 - j_2|}$。然后我们生成一个随机向量 $\beta_k \in \mathbb{R}^p$，其针对 $0 \le k \le K$。β_k 的所有元素都独立同分布地来自标准正态分布。接下来我们定义 $\beta_k = \beta_k/\|\beta_k\| - \beta_0/\|\beta_0\|$。此后，源响应 Y_i^* 按照

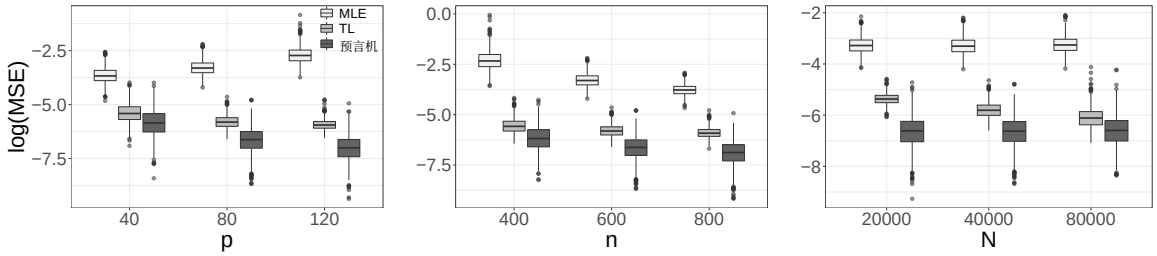
$P(Y_i^* = k|X_i^*) = \exp(X_i^{*\top}\beta_k)/\{1 + \sum_{j=1}^{K} \exp(X_i^{*\top}\beta_j)\}$ 对于 $1 \le k \le K$。因此，我们有 $P(Y_i^* = 0|X_i^*) = 1/\{1 + \sum_{j=1}^{K} \exp(X_i^{*\top}\beta_j)\}$。令 $\mathbb{B} = (\beta_1, \ldots, \beta_k) \in \mathbb{R}^{p\times K}$。于是，我们得到特征生成函数 $Z_{\mathbb{B}}(X_i^*) = \mathbb{B}^\top X_i^*$。

Target Data. 首先，我们从与源数据相同的多元正态分布中，独立同分布地生成目标协变量 $X_i \in \mathbb{R}^p$。接着，计算目标数据的低维特征向量 $Z_i = Z_{\mathbb{B}}(X_i) = \mathbb{B}^\top X_i \in \mathbb{R}^K$。然后，考虑如何生成响应变量 Y_i。具体而言，在迁移学习充分性原假设下，我们按照 $P(Y_i = 1|X_i) = P(Y_i = 1|Z_i) = \exp(Z_i^\top\gamma)/\{1 + \exp(Z_i^\top\gamma)\}$ 生成 Y_i，其中 $\gamma = (0.5, 0.5, 0.5, 0.5, 0.5, -1.25, 0, 0)^\top \in \mathbb{R}^K$。因此，目标模型中与 X_i 相关的真实系数 θ 为 $\theta = \mathbb{B}\gamma$。相反，如果迁移学习不充分，则生成 Y_i，使其满足 $P(Y_i = 1|X_i) = \exp(Z_i^\top\gamma + X_{i1}\delta)/\{1 + \exp(Z_i^\top\gamma + X_{i1}\delta)\}$，其中 $\delta > 0$ 为固定常数。

3.2. 估计性能的结果

我们研究了所提出的迁移学习估计量在迁移学习充分性零假设下的有限样本性能。我们从 $p < n$ 的情形开始。在此情形下，我们比较 θ 的三种不同估计量，分别为：(1) 我们提出的迁移学习估计量 $\widehat{\theta}^{\text{Transfer}}$；(2) 预言机估计量 $\widehat{\theta}^{\text{Oracle}}$，它也是一种迁移学习估计量，但使用真实系数矩阵 $\mathbb{B}$，而非其估计值 $\widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}}$；(3) 根据 (2.2) 得到的基于目标数据的最大似然估计量 $\widehat{\theta}^{\text{Target}}$ 最大似然估计。令特征维度 $p \in$（40，80，120）、目标样本量 $n \in$（400，600，800）以及源样本量 $N \in$（20000，40000，80000）。接着，研究 (p, n, N) 的不同组合。对于每一种 (p, n, N) 组合，我们按照第 3.1 节所述的生成过程生成目标数据和源数据。实验随机重复共 $B = 1,000$ 次。对于第 b 次重复，将 θ 的一个特定估计量记为 $\widehat{\theta}^{(b)}$。然后我们评估其估计.....

通过均方误差（均方误差）表示估计性能，即均方误差$_{(b)} = p^{-1}\sum_{j=1}^{p}(\widehat{\theta}_j^{(b)} - \theta_j)^2$。这总共产生 $B = 1,000$ 个均方误差值。随后对这些值进行对数变换，并在图1中绘制箱线图。



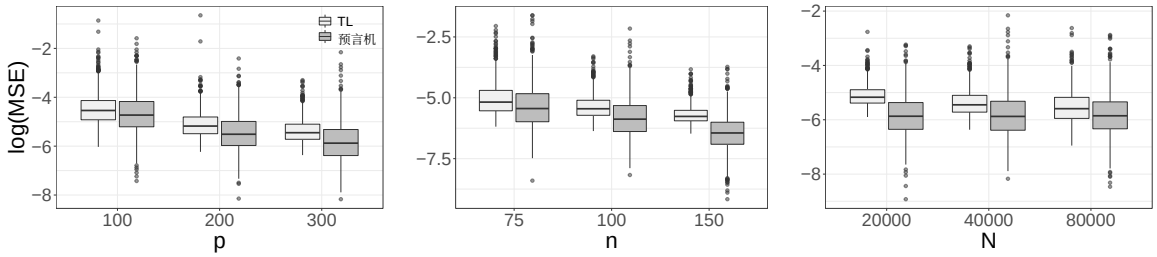
(a) $n = 600$ 和 $N = 40000$　　(b) $p = 80$ 和 $N = 40000$　　(c) $p = 80$ 和 $n = 600$

图1： $p < n$ 情形下最大似然估计量（MLE）、我们提出的迁移学习（TL）估计量和预言机迁移学习估计量的对数变换后的均方误差箱线图。考虑（p, n, N）的不同组合，以探究（p, n, N）对三个估计量的估计性能的影响。

根据图1，我们可以得出以下结论。首先，在所有情况下，最大似然估计的log(均方误差)值始终大于我们提出的迁移学习估计量和预言机估计量的值。这是预期之中的，因为最大似然估计是在没有任何源数据帮助的情况下，通过直接估计高维模型获得的。因此，其性能比这两种迁移学习估计量更差。其次，在特征维度 p 和源样量固定时，随着目标样本量 n 增加，迁移学习估计量以及其他估计量的log(均方误差) 值都会降低。这些结果证明了迁移学习估计量的一致性，这与定理2一致。最后，在特征维度 p 和目标样本量 n 固定时，我们发现，随着 $N \to \infty$，迁移学习估计量的 log(均方误差)值会变得越来越接近预言机估计量的值。这一结果也进一步印证了我们在定理2中的理论发现。也就是说，当源样本量足够大时，迁移学习估计量在渐近意义下可以达到与预言机估计量同样的性能。

接下来考虑 $n \le p < N$ 的情况。在这种情况下，最大似然估计 $\widehat{\theta}^{\text{Target}}{}_{\text{mle}}$ 无法计算。因此，我们仅比较所提出的迁移学习

估计量 $\widehat{\theta}^{\text{Transfer}}$ 与预言机估计量 $\widehat{\theta}^{\text{Oracle}}$的估计性能。我们考虑特征维度$p \in$(100，200，300)、目标样本量 $n \in$(75、 100、 150)以及源样本量 $N \in$(20000、40000、 80000)。随后研究 (p， n， N) 的不同组合；详情见图 2。对于每种组合，我们随机重复实验，共进行 $B = 1$，000 次。对于每次实验，计算两个迁移学习估计量，并计算均方误差（MSE）值，以评估估计性能。这样，每个估计量都会得到 $B = 1$，000 个 MSE 值，随后对这些值进行对数变换，并在图 2 中绘制箱线图。与 $p < n$ 的情况类似，我们发现，随着目标样本量 n 增加，两个估计量的 log(MSE) 值均下降，这表明两个估计量具有一致性。此外，随着 $N \to \infty$，两个估计量变得更加接近。这一发现表明，只要我们提供具有足够大样本量的源数据，θ 的高维效应对估计性能的影响就非常有限。



(a) $n = 100$ 和 $N = 40000$　(b) $p = 300$ 和 $N = 40000$　(c) $p = 300$ 和 $n = 100$

图 2：在 $n \leq p < N$ 情形下，我们提出的迁移学习（TL）估计量和预言迁移学习估计量的对数变换估计误差箱线图。考虑 (p， n， N) 的不同组合，以探究 (p， n， N) 对两个估计量估计性能的影响。

3.3. 测试性能结果

本节中，我们研究所提出的迁移学习充分性检验程序的有限样本性能。我们首先评估该检验程序的检验尺寸。为此，我们应在迁移学习充分性原假设下生成目标数据。具体而言，我们按照以下方式生成 Y_i

$P(Y_i = 1|X_i) = P(Y_i = 1|Z_i) = \exp(Z_i^\top \gamma)/\{1 + \exp(Z_i^\top \gamma)\}$。在这种情况下，我们将所提出的检验统计量与预言机检验统计量进行比较，后者使用真实的 Z_i 而非 Z_i 计算。我们考虑特征维度 $p \in$(1000，1500，2000，3000)、目标样本量 $n \in$(200，300，500)以及源样本量 $N \in$(400000，800000)。对于每种(p，n，N)组合，我们随机重复实验 $B = 1$，000次。令 $T_4^{(b)}$ 为第 b次随机重复实验中计算得到的所提出的检验统计量。随后，我们将经验拒绝概率计算为EJP $= B^{-1} \sum_{b=1}^{B} I(T_4^{(b)} > z_{1-\alpha})$。由于目标数据是在原假设下生成的，因此EJP对应于经验检验尺寸（即第一类错误）。预言机检验统计量的EJP也可用类似方法计算。表1给出了显著性水平为$\alpha = 0.05$时，所提出的检验统计量和预言机检验统计量的经验检验尺寸。我们发现，只要源样本量足够大，这两种检验统计量的经验检验尺寸就与名义零假设下的5%相当接近。这些结果证实了我们在第2.4节中的理论结论。

表 1：我们提出的迁移学习充分性检验以及预言机检验的经验检验尺寸，显著性水平为 $\alpha = 0.05$

		预言机测试				实证检验			
n	N \ p	1000	1500	2000	3000	1000	1500	2000	3000
200	4×10^5	0.025	0.037	0.053	0.036	0.033	0.047	0.053	0.039
	6×10^5	0.030	0.032	0.038	0.031	0.037	0.048	0.045	0.032
	8×10^5	0.034	0.039	0.047	0.038	0.033	0.038	0.041	0.039
300	4×10^5	0.037	0.043	0.040	0.040	0.053	0.044	0.048	0.058
	6×10^5	0.032	0.045	0.042	0.040	0.036	0.049	0.044	0.053
	8×10^5	0.022	0.039	0.041	0.039	0.028	0.040	0.047	0.045
500	4×10^5	0.033	0.038	0.033	0.032	0.053	0.048	0.057	0.052
	6×10^5	0.039	0.037	0.045	0.047	0.049	0.049	0.055	0.067
	8×10^5	0.026	0.041	0.033	0.030	0.033	0.048	0.042	0.041

接下来，我们评估所提出的迁移学习充分性检验程序的检验功效。因此，我们应在备择假设下生成目标数据，即迁移学习不充分的情形。在这种情况下，我们生成 Y_i according

至 $P(Y_i = 1|X_i) = \exp(Z_i^\top \gamma + X_{i1}\delta)/\{1 + \exp(Z_i^\top \gamma + X_{i1}\delta)\}$，其中 $\delta > 0$。我们将 δ称为信号强度。较大的 δ 值表示 X_{i1} 对 Y_i的影响更大。因此， X_i中存在更多无法被 Z_i 捕获的信息，从而产生更强的证据支持备择假设。我们固定 $p = 2{,}000$，并将$N = 400{,}000$。考虑三个目标样本量，分别为 $n \in (200{,}300{,}500)$。我们还让 δ 以 0.2 为步长从 1 变化到 5。对于(n, δ) 的每种组合，我们总共重复实验 $B = 1{,}000$次。对于每次实验，我们计算所提出的检验统计量 T_4 ，然后在显著性水平 $\alpha = 0.05$下，计算 B 次重复实验中的经验拒绝概率EJP。由于备择假设成立，EJP对应于假设检验程序的经验检验功效。图3 展示了所提出检验统计量在不同实验设置下的经验检验功效。由图 3可知，较大的信号强度 δ 会带来更高的经验检验功效。此外，较大的目标样本量 n 也会带来更高的经验检验功效。

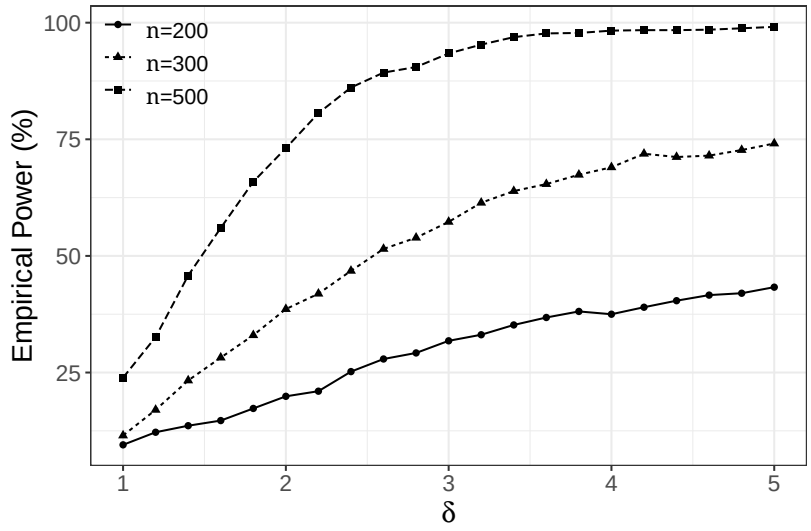


图 3：在不同目标样本量 n 和信号强度 δ下，我们提出的迁移学习充分性检验的经验检验功效。此处设定 $p = 2000$、 $N = 400000$，以及显著性水平为 $\alpha = 0.05$。

3.4. 一个真实数据示例

To 进一步展示所提出的测试方法，我们在此展示一个真实数据示例。 -

ample。此处使用的目标数据是$Dogs\ vs. Cats$数据集，可在 Kaggle 竞赛（`https://www.kaggle.com/competitions/dogs-vs-cats/`）中下载。该数据集包含 $n = 25{,}000$ 张属于两个类别（即狗或猫）的图像。我们的目标是将每张图像分为这两个类别之一，这可用二元响应变量 Y_i表示。这是一个标准的图像分类问题，可通过各种深度学习模型来解决（Krizhevsky 等人，2012；He 等人，2016；Howard 等人，2017）。为展示我们的方法，我们首先将非结构化图像数据转换为结构化向量。为此，我们使用各种预训练深度学习模型。具体而言，我们将预训练模型倒数第二层生成的特征向量作为协变量 X_i，其维度通常较高。鉴于目标样本量相对于特征维度有限，我们转而借助迁移学习。

我们将$ImageNet$数据集（Russakovsky 等，2014）作为源数据。这是一个包含总计 $N = 1{,}281{,}167$ 张图像、涵盖 $K + 1 = 1000$ 个类别的数据集。为了将源数据中的图像转换为结构化向量，我们使用了与目标数据相同的预训练深度学习模型。由此得到高维源协变量 X_i^*。随后，可以针对 X_i^* 和 Y_i^* 拟合标准多分类逻辑回归模型，其中 $Y_i^* \in \{0, 1, ..., K\}$ 表示源数据中第 i 张图像的类别标签。由此得到基于源数据的最大似然估计量 $\widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}}$。为了从源数据迁移信息，可以为目标数据构建低维特征向量 $\widehat{Z}_i = \widehat{\mathbb{B}}_{\text{mle}}^{\text{Source}\top} X_i \in \mathbb{R}^K$。随后，针对 $\backsim Z_i$ 和 Y_i 拟合标准二元逻辑回归模型。因此，可以应用所提出的迁移学习充分性检验程序，并评估预测准确率。

为进行全面评估，我们使用若干经典的预训练深度学习模型提取高维特征向量 X_i 和 X_i^*。具体而言，它们分别是具有两种

不同尺寸的 VGG（Simonyan 和 Zisserman，2014）模型、具有三种不同尺寸的 ResNet（He 等人，2016）模型，以及 InceptionV3（Szegedy 等人，2016）。得到的特征维度分别为 VGG 的 $p = 4$，096，以及 ResNet 和 InceptionV3 的 $p = 2$，048。原始目标数据集被随机分为两部分，其中一部分用于训练，样本量为 15,000；另一部分用于测试，样本量为 10,000。随后，我们在训练数据集上应用所提出的迁移学习充分性检验程序。每个预训练深度学习模型对应的 p 值列于表 2。我们还评估了预测准确率，但评估使用的是测试数据集。详细结果见表 2。由表 2 可知，所有检验的 p 值均大于显著性水平 $\alpha = 0.05$。这些结果支持迁移学习充分性的零假设。这样的检验结果似乎很合理，因为在测试数据集上评估的预测准确率都很高。

表 2：报告了不同预训练深度学习模型下所提出的检验程序的 p-值以及测试数据集上的预测准确率。

模型	p	p 值	预测准确率
VGG16	4096	0.846	97.97%
VGG19	4096	0.809	97.93%
ResNet50	2048	0.827	98.06%
ResNet101	2048	0.777	98.40%
ResNet152	2048	0.755	98.50%
InceptionV3	2048	0.691	98.97%

4. 总结性意见

本文提出了一种有理论依据的迁移学习充分性检验方法。为此，我们开发了一种迁移学习估计程序，并构造了一个检验统计量，其渐近零分布也经过了理论推导。我们通过大量数值研究展示了所提方法的有限样本性能。最后，本文还探讨了一些有趣的

未来研究方向。首先，本文研究的检验程序要求源数据的样本量远大于目标数据的样本量。然而，在实际应用中，源数据的样本量可能与目标数据相当。因此，在这种情况下如何检验迁移学习充分性，仍有待进一步研究。其次，如果迁移学习充分性原假设被拒绝，如何进一步提高目标模型的估计精度便成为下一个重要问题。最后，我们在特征生成函数中使用了线性变换，这可能限制性过强。因此，如何针对非线性特征生成函数检验迁移学习充分性，也是值得进一步研究的重要问题。

参考文献

Cai, T. T. 和 Wei, H.（2021），“非参数分类的迁移学习：极小极大速率与 AdaptiveClassifier”，《统计学年鉴》，49，100–128。

Candes, E. J. 和 Sur, P.（2020），“高维逻辑回归中极大似然估计存在性的相变”，《统计学年鉴》，48，27–42。

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. 和 Darrell, T.（2014），“Decaf：用于通用视觉识别的深度卷积激活特征”，载于国际机器学习会议，PMLR，第647–655页。

Fan, C. 和 Fine, J. P.（2013），“具有参数协变量变换的线性变换模型”，《美国统计学会会刊》，108，701–712。

Fan, J. 和 Li, R.（2001），“通过非凹惩罚似然进行变量选择及其预言机性质”，《美国统计学会会刊》，96，1348–1360。

Fan, J. 和 Peng, H. (2004)，“具有发散参数数量的非凹惩罚似然”，《统计学年鉴》，32，928–961。

Farahani, A.、Pourshojae, B.、Rasheed, K. 和 Arabnia, H. R. (2020)，“迁移学习简明综述”，载于*2020*年国际计算科学与计算智能会议（*CSCI*），第344–351页。

Guo, L.、Lei, Y.、Xing, S.、Yan, T. 和 Li, N. (2018)，“深度卷积迁移学习网络：一种利用无标签数据进行机器智能故障诊断的新方法”，*IEEE*工业电子学汇刊，66，7316–7325。

Hall, P. 和 Heyde, C. C. (1980)，鞅极限定理及其应用，学术出版社。

He, K.、Zhang, X.、Ren, S. 和 Sun, J. (2016)，“面向图像识别的深度残差学习”，载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页。

He, X. 和 Shao, Q.-M.(2000)，《递增维度参数》，多元分析期刊，73，120–135。

Howard, A. G.、Zhu, M.、Chen, B.、Kalenichenko, D.、Wang, W.、Weyand, T.、Andreetto, M. 和 Adam, H. (2017)，《MobileNets：用于移动视觉应用的高效卷积神经网络》，$ArXiv$，abs/1704.04861。

Ji, J., Guo, Y., Yang, Z., Zhang, T., 和 Lu, X.（2021），“利用注意力模型进行细粒度图像分类的多级字典学习”，神经计算，453，403–412。

Krizhevsky, A., Sutskever, I., 和 Hinton, G. E.（2012），“使用深度卷积神经网络进行 ImageNet 分类”，载于神经信息处理系统进展，编者：Pereira, F., Burges, C., Bottou, L., 和 Weinberger, K.，Curran Associates, Inc.，第25卷。

Lan, W.、Wang, H. 和 Tsai, C. L.（2014），“高维回归中的协变量测试”，《统计数学研究所年刊》，66，279–301。

Li, Q. 和 Racine, J. S. (2007)，非参数计量经济学：理论与实践，普林斯顿大学出版社。

Li, S.、Cai, T. T. 和 Li, H.（2022a），“高维线性回归的迁移学习：预测、估计与极小极大最优性”，《皇家统计学会杂志》*B*辑，84，149–173。

—（2022b），《采用错误发现率控制的大规模高斯图模型迁移学习》，《美国统计学会会刊》，即将刊出。

Liang, Z., Cai, T. T., Sun, W., and Xia, Y.（2022），《用于大规模多重检验的局部自适应迁移学习算法》，*ArXiv* 预印本：*2203.11461*。

Lu, W. 和 Zhang, H. H.（2010），“关于部分线性变换模型的估计”，《美国统计学会会刊》，105，683–691。

McCullagh, P. 和 Nelder, J.（1989），《广义线性模型》第*2*版，Chapman & Hall/CRC。

Oquab, M.、Bottou, L.、Laptev, I.和Sivic,J.（2014），“使用卷积神经网络学习和迁移中级图像表示”，载于*2014*年*IEEE*计算机视觉与模式识别会议论文集，第1717–1724页。

Pan, S. J. 和 Yang, Q.（2009），“迁移学习综述”，*IEEE* 知识与数据工程汇刊，22，1345–1359。

Portnoy, S. (1985)，“当 p^2/n 较大时 p 回归参数的 M 估计量的渐近行为；II. 正态近似”，《统计学年鉴》，13，1403–1417。

Russakovsky, O., Deng, J., Su, H., Krause,J., Satheesh, S., Ma, S., Huang, Z., and et al. (2014),“ImageNet 大规模视觉识别挑战赛,”in $CoRR, abs/1409.0575$.

Seber, G.A. (2008), 统计学家的矩阵手册, John Wiley & Sons.Shao, J. (2003),数理统计学, Springer Science & Business Media.

Shin,H.-C., Roth, H. R.,Gao, M.,Lu, L., Xu, Z.,Nogues, I., Yao,J., Mollura,D. 和 Summers,R.M.（2016），“用于计算机辅助检测的深度卷积神经网络：卷积神经网络架构、数据集特征和迁移学习”，*IEEE* 医学成像学报，35，1285–1298。

Simonyan, K. 和 Zisserman, A. (2014)，“用于大规模图像识别的超深卷积网络”，*arXiv* 预印本 *arXiv:1409.1556*。

Sur,P. 和 Candes, E. J.（2019），“高维逻辑回归的现代最大似然理论”，《美国国家科学院院刊》，119，14516–14525。

Szegedy, C., Vanhoucke, V., Ioffe,S., Shlens,J.和 Wojna, Z. (2016)，“重新思考用于计算机视觉的 Inception 架构”，载于 *IEEE*计算机视觉与模式识别会议论文集，第 2818–2826 页。

Tian, Y.和Feng, Y.（2022），《高维广义线性模型下的迁移学习》，《美国统计学会会刊》，1，1–14。

Van der Vaart, A. W.（1998），渐近统计学，《剑桥统计与概率数学丛书》，剑桥大学出版社。

Vershynin，R. (2018)，高维概率：数据科学应用导论，第47卷，剑桥大学出版社。

Wainwright, M.J.(2019)，《高维统计学：非渐近视角》，第48卷，剑桥大学出版社。

Wang, L. (2011)，“协变量数量不断增长时聚类二元数据的 GEE 分析”，《统计学年鉴》，39，389–417。

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M.（2021），“从少量示例中泛化：小样本学习综述”，*ACM Computing Surveys*，53，1–34。

Weiss, K., Khoshgoftaar, T. M. 和 Wang（2016），“迁移学习综述”，大数据期刊，3，1–34。

Welsh, A. (1989), “论 M 过程与 M 估计”，《统计学年鉴》，337– 361。

Zhong, P.-S. 和 Chen, S.X.（2011），“因子设计下高维回归系数的检验”，《美国统计学会会刊》，106，260–274。

附录

附录 A. 定理1的证明

设 $\mathbb{U} = (u_1, \ldots, u_K) \in \mathbb{R}^{p\times K}$ 是一个单位长度矩阵，即 $\|\mathbb{U}\|_F = 1$，其中 $u_k = (u_{k1}, \ldots, u_{kp})^\top \in \mathbb{R}^p$ 是 $\mathbb{U}$ 的第 k 列向量。接下来，根据 Fan 和 Li（2001），只需证明对于任意 $\varepsilon > 0$，都有

$$\liminf_{n\to\infty} P\Big\{ \sup_{\|\mathbb{U}\|_F=1} \mathcal{L}_N^{\text{Source}}\big(\mathbb{B} + C\sqrt{p/N}\mathbb{U}\big) < \mathcal{L}_N^{\text{Source}}\big(\mathbb{B}\big)\Big\} \geq 1 - \varepsilon \tag{A.1}$$

f或一个足够大但固定的常数 $C > 0$。通过泰勒展开，我们有

$$\begin{aligned} p^{-1}\Big\{\mathcal{L}_N^{\text{Source}}\Big(\mathbb{B} + C\sqrt{p/N}\mathbb{U}\Big) - \mathcal{L}_N^{\text{Source}}\big(\mathbb{B}\big)\Big\} &= C(Np)^{-1/2}\sum_{k=1}^{K}\left(\frac{\partial\mathcal{L}_N^{\text{Source}}(\mathbb{B})}{\partial\beta_k}\right)^\top u_k \\ &+2^{-1}C^2N^{-1}\sum_{k=1}^{K}\sum_{l=1}^{K} u_k^\top\Big(\frac{\partial^2\mathcal{L}_N^{\text{Source}}(\mathbb{B})}{\partial\beta_k\partial\beta_l^\top}\Big)u_l \\ &+6^{-1}C^3p^{1/2}N^{-3/2}\sum_{j=1}^{K}\frac{\partial}{\partial\beta_j}\Big\{\sum_{k=1}^{K}\sum_{l=1}^{K} u_k^\top\Big(\frac{\partial^2\mathcal{L}_N^{\text{Source}}(\bar{\mathbb{B}})}{\partial\beta_k\partial\beta_l^\top}\Big)u_l\Big\}^\top u_j, \end{aligned} \tag{A.2}$$

其中 $\bar{\mathbb{B}} = \mathbb{B} + \nu C\sqrt{p/N}\mathbb{U}$ 对于某个 $\nu \in (0, 1)$。接下来我们分别研究 (A.2) 中的三个项。根据柯西-施瓦茨不等式，我们有 $|\sum_{k=1}^{K}(\partial\mathcal{L}_N^{\text{Source}}(\mathbb{B})/\partial\beta_k)^\top u_k| \leq \{\sum_{k=1}^{K}\|u_k\|^2\}^{1/2}\{\sum_{k=1}^{K}\|\partial\mathcal{L}_N^{\text{Source}}(\mathbb{B})/\partial\beta_k\|^2\}^{1/2} = \{\sum_{k=1}^{K}\|\partial\mathcal{L}_N^{\text{Source}}(\mathbb{B})/\partial\beta_k\|^2\}^{1/2}$。直接计算得到 $E\|\partial\mathcal{L}_N^{\text{Source}}(\mathbb{B})/\partial\beta_k\|^2 = E(\sum_{i=1}^{N}\pi_{ik}(1-\pi_{ik})X_i^{*\top}X_i^*) \leq NE(X_i^{*\top}X_i^*) = NE\{\text{tr}(X_i^{*\top}X_i^*)\} = N\text{tr}\{E(X_i^*X_i^{*\top})\} \leq Np\lambda_{\max}\{E(X_i^*X_i^{*\top})\} = O(Np)$，根据条件（C3）。这导致 $\|\partial\mathcal{L}_N^{\text{Source}}(\mathbb{B})/\partial\beta_k\| = O_p(\sqrt{Np})$。因此，我们有 $\sqrt{1/Np}\sum_{k=1}^{K}(\partial\mathcal{L}_N^{\text{Source}}(\mathbb{B})/\partial\beta_k)^\top u_k = O_p(1)$。然后我们考虑 (A.2) 中的第二项。对于任意 $\eta > 0$ 和 $k \in \{1, \ldots, K\}$，根据切比雪夫不等式，我们有

$$P\left[\Big\|N^{-1}\sum_{i=1}^{N}\pi_{ik}(1-\pi_{ik})X_i^*X_i^{*\top} - E\Big\{\pi_{ik}(1-\pi_{ik})X_i^*X_i^{*\top}\Big\}\Big\|_F > \eta\right]$$

$$\leq \eta^{-2}\sum_{j=1}^{p}\sum_{l=1}^{p} E\Big[N^{-1}\sum_{i=1}^{N}\pi_{ik}(1-\pi_{ik})X_{ij}^{*}X_{il}^{*} - E\Big\{\pi_{ik}(1-\pi_{ik})X_{ij}^{*}X_{il}^{*}\Big\}\Big]^{2} = O(p^2/N).$$

根据条件（C1），我们知道 $p^2/N \leq C^2n^2/N \to 0$。因此，我们有$\|N^{-1}\sum_{i=1}^{N}\pi_{ik}(1-\pi_{ik})X_i^*X_i^{*\top} - E\{\pi_{ik}(1-\pi_{ik})X_i^*X_i^{*\top}\}\|_F = o_p(1)$。类似地，对于 k，$l \in \{1,\ldots,K\}$ 和 $k \neq l$，我们可以证明 $\|N^{-1}\sum_{i=1}^{N}\pi_{ik}\pi_{il}X_i^*X_i^{*\top} - E(\pi_{ik}\pi_{il}X_i^*X_i^{*\top})\|_F = o_p(1)$。回想一下，$W_i = \mathrm{diag}(\pi_i) - \pi_i\pi_i^\top \in \mathbb{R}^{K\times K}$，其中 $\pi_i = (\pi_{i1}, \ldots, \pi_{iK})^\top$ 且$\pi_{ik} = P(Y_i^* = k|X_i^*)$。结合以上结果，我们有

$$\begin{aligned}
&\lambda_{\min}^2\Big[N^{-1}\sum_{i=1}^{N}W_i\otimes\big(X_i^*X_i^{*\top}\big) - E\Big\{W_i\otimes\big(X_i^*X_i^{*\top}\big)\Big\}\Big]\\
\leq&\Big\|N^{-1}\sum_{i=1}^{N}W_i\otimes\big(X_i^*X_i^{*\top}\big) - E\Big\{W_i\otimes\big(X_i^*X_i^{*\top}\big)\Big\}\Big\|_F^2\\
=&\sum_{k=1}^{K}\Big\|N^{-1}\sum_{i=1}^{N}\pi_{ik}(1-\pi_{ik})X_i^*X_i^{*\top} - E\Big\{\pi_{ik}(1-\pi_{ik})X_i^*X_i^{*\top}\Big\}\Big\|_F^2\\
+&\sum_{k\neq l}\Big\|N^{-1}\sum_{i=1}^{N}\pi_{ik}\pi_{il}X_i^*X_i^{*\top} - E\Big(\pi_{ik}\pi_{il}X_i^*X_i^{*\top}\Big)\Big\|_F^2 = o_p(1)
\end{aligned}$$

根据韦尔不等式（Seber，2008），我们知道，对于任意两个对称矩阵 A 和 B，$\lambda_{\min}(A+B) \geq \lambda_{\min}(A) + \lambda_{\min}(B)$。因此，对于 (A.2) 的第二项，我们有

$$\begin{aligned}
&-N^{-1}\sum_{k=1}^{K}\sum_{l=1}^{K}u_k^\top\frac{\partial^2\mathcal{L}_N^{\mathrm{Source}}(\mathbb{B})}{\partial\beta_k\partial\beta_l^\top}u_l - \lambda_{\min}\Big[E\Big\{W_i\otimes\big(X_i^*X_i^{*\top}\big)\Big\}\Big]\\
=&\mathrm{vec}(\mathbb{U})^\top\Big\{N^{-1}\sum_{i=1}^{N}W_i\otimes\big(X_i^*X_i^{*\top}\big)\Big\}\mathrm{vec}(\mathbb{U}) - \lambda_{\min}\Big[E\Big\{W_i\otimes\big(X_i^*X_i^{*\top}\big)\Big\}\Big]\\
\geq&\lambda_{\min}\Big\{N^{-1}\sum_{i=1}^{N}W_i\otimes\big(X_i^*X_i^{*\top}\big)\Big\} - \lambda_{\min}\Big[E\Big\{W_i\otimes\big(X_i^*X_i^{*\top}\big)\Big\}\Big]\\
\geq&\lambda_{\min}\Big[N^{-1}\sum_{i=1}^{N}W_i\otimes\big(X_i^*X_i^{*\top}\big) - E\Big\{W_i\otimes\big(X_i^*X_i^{*\top}\big)\Big\}\Big] \xrightarrow{p} 0.
\end{aligned} \tag{A.3}$$

根据条件（C3），我们知道 $E\{W_i\otimes(X_i^*X_i^{*\top})\}$ 是一个正定矩阵，且对于某个正数常量 $C_{\min} > 0$，有 $\lambda_{\min}\{E(W_i\otimes X_i^*X_i^{*\top})\} \geq C_{\min}$。因此

以趋于 1 的概率有 $-N^{-1}\sum_{k=1}^{K}\sum_{l=1}^{K}u_k^\top\partial^2\mathcal{L}_N^{\text{Source}}(\mathbb{B})/\partial\beta_k\partial\beta_l^\top u_l>0$。对于（A.2）的第三项，根据柯西-施瓦茨不等式，

$$\Big|\frac{\partial}{\partial\beta_j}\Big\{\sum_{k=1}^{K}\sum_{l=1}^{K}u_k^\top\Big(\frac{\partial^2\mathcal{L}_N^{\text{Source}}(\bar{\mathbb{B}})}{\partial\beta_k\partial\beta_l^\top}\Big)u_l\Big\}^\top u_j\Big|=\Big|\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{N}\Big(\frac{\partial p_{ikl}}{\partial\beta_j}\Big)^\top u_j\Big(X_i^{*\top}u_k\Big)\Big(X_i^{*\top}u_l\Big)\Big|$$
$$\leq\sum_{k=1}^{K}\sum_{l=1}^{K}\sup_i\Big\|\frac{\partial p_{ikl}}{\partial\beta_j}\Big|_{\beta_j=\bar{\beta}_j}\Big\|\Big\{\sum_{i=1}^{N}(u_k^\top X_i^*X_i^{*\top}u_k)\Big\}^{1/2}\Big\{\sum_{i=1}^{N}(u_l^\top X_i^*X_i^{*\top}u_l)\Big\}^{1/2},$$

其中 $p_{ikl}=\pi_{ik}(1-\pi_{ik})$，若 $k=l$ 且 $p_{ikl}=-\pi_{ik}\pi_{il}$ 若 $k\neq l$。通过简单计算，在 $k=l$的情况下，我们有 $\partial p_{ikl}/\partial\beta_j=\pi_{ik}(1-\pi_{ik})(1-2\pi_{ik})X_i^*$若 $j=k=l$，且 $\partial p_{ikl}/\partial\beta_j=-\pi_{ik}\pi_{ij}(1-2\pi_{ik})X_i^*$若 $j\neq k$。在 $k\neq l$的情况下，我们有 $\partial p_{ikl}/\partial\beta_j=2\pi_{ik}\pi_{il}\pi_{ij}X_i^*$若 $j\neq k$，且 $j\neq l$ 和 $\partial p_{ikl}/\partial\beta_j=-\pi_{ik}\pi_{il}(1-2\pi_{ij})X_i^*$若 $j=k$或 $j=l$。在所有情况下，我们都有 $\sup_i\|\partial p_{ikl}/\partial\beta_j\|\leq 2\sup_i\|X_i^*\|$。注意，根据条件（C2），我们知道 X_i^* 的每个分量都是次高斯的。因此，根据 Vershynin (2018) 中的命题2.5.2，我们知道 $E\exp(X_{ij}^{*2}/\|X_{ij}^*\|_{\psi_2}^2)\leq 2$。令$j_i^*=\max\{j:\ X_{ij}^{*2}/\|X_{ij}^*\|_{\psi_2}^2\}$。那么根据 $\|X_i^*\|_{\psi_2}\geq\|X_{ij}^*\|_{\psi_2}$这一事实，我们有$E\exp\{\|X_i^*\|^2/(p\|X_i^*\|_{\psi_2}^2)\}\leq E\exp(X_{ij^*}^{*2}/\|X_{ij^*}^*\|_{\psi_2}^2)\leq 2$。于是我们知道 $\|X_i^*\|/\sqrt{p}$是次高斯的。根据 Vershynin(2018)中练习 2.5.10 的结果，我们于是有 $\sup_i\|X_i^*\|=O_p(\sqrt{p\log N})$。因此，我们有$\sup_i\|\partial p_{ikl}/\partial\beta_j\|=O_p(\sqrt{p}\log N)$。通过与(A.3)类似的证明，我们知道 $\lambda_{\max}(N^{-1}\sum_{i=1}^{N}X_i^*X_i^{*\top})-\lambda_{\max}\{E(X_i^*X_i^{*\top})\}\xrightarrow{p}0$。根据条件（C3），我们知道 $\lambda_{\max}\{E(X_i^*X_i^{*\top})\}\leq C_{\max}$。注意，$\|u_k\|\leq 1$。于是我们有 $N^{-1}\sum_{i=1}^{N}u_k^\top X_i^*X_i^{*\top}u_k\leq\lambda_{\max}(N^{-1}\sum_{i=1}^{N}X_i^*X_i^{*\top})=O_p(1)$。因此，我们有 $\sum_{i=1}^{N}u_k^\top X_i^*X_i^{*\top}u_k=O_p(N)$。因此，根据条件（C1），(A.2)的第三项为 $p^{1/2}N^{-3/2}O_p(\sqrt{p}\log N)O_p(N)=O_p(\sqrt{p^2}\log N/N)=O_p(\sqrt{n^2}\log N/N)=o_p(1)$阶。于是我们可以选择一个充分大的 C 使得$\mathcal{L}_N^{\text{Source}}(\mathbb{B}+C\sqrt{p/N}\mathbb{U})-\mathcal{L}_N^{\text{Source}}(\mathbb{B})<0$，从而完成定理证明。

附录 B. 定理 2 的证明

Th该定理的结论可分三步得出。第一步，e

证明存在一个局部优化器 $\hat{\gamma}$，且其具有 $\sqrt{n}$-一致性。¯第二步，研究 $\hat{\gamma}$ 的渐近正态性。第三步，分析 $\widehat{\theta}^{\text{Transfer}}$ 的渐近行为。

Step 1. $\sqrt{n}$-consistency

回顾一下，$\mathcal{L}_n(\gamma;\widehat{Z}) = \sum_{i=1}^{n}[Y_i(\widehat{Z}_i^{\top}\gamma) - \log\{1+\exp(\widehat{Z}_i^{\top}\gamma)\}]$ 是工作对数似然。根据 Fan 和 Li（2001），我们希望证明，对于任意但固定的 $\varepsilon > 0$，有

$$\liminf_{n\to\infty} P\Big\{\sup_{\|u\|=1}\mathcal{L}_n\big(\gamma + C_\varepsilon n^{-1/2}u;\widehat{Z}\big) < \mathcal{L}_n\big(\gamma;\widehat{Z}\big)\Big\} \geq 1-\varepsilon, \tag{B.1}$$

其中常数 $C_\varepsilon > 0$ 足够大。为此，进行泰勒展开如下：

$$\begin{aligned}&\mathcal{L}_n\Big(\gamma + C_\varepsilon n^{-1/2}u;\widehat{Z}\Big) - \mathcal{L}_n\Big(\gamma;\widehat{Z}\Big)\\ &= C_\varepsilon n^{-1/2}\dot{\mathcal{L}}_n(\gamma;\widehat{Z})^{\top}u - 2^{-1}C_\varepsilon^2 u^{\top}\Big\{-n^{-1}\ddot{\mathcal{L}}_n(\gamma;\widehat{Z})\Big\}u + o_p(1),\end{aligned} \tag{B.2}$$

. ^ ¨ ^其中，$\mathcal{L}_n(\gamma,Z)$ 和 $\mathcal{L}_n(\gamma,Z)$ 分别是工作对数似然函数 $\mathcal{L}_n(\gamma;Z)$ 关于 γ 的一阶和二阶偏导数。为验证 (B.1)，只需证明以下两个关键结果：

$$(1)\ \sup_{\|u\|=1} n^{-1/2}\dot{\mathcal{L}}_n(\gamma;\widehat{Z})^{\top}u = O_p(1), \tag{B.3}$$

$$(2)\lambda_{\min}\{-n^{-1}\ddot{\mathcal{L}}_n(\gamma;\widehat{Z})\} \geq \tau_{\min} > 0, \tag{B.4}$$

对于某个正数常量 $\tau_{\min}$，上述结论以趋近于 1 的概率成立。接下来，我们通过以下两个子步骤验证这些结论。

Step 1.1. 我们从 (B.3) 开始。直接计算可得 $n^{-1/2}\dot{\mathcal{L}}_n(\gamma,\widehat{Z})^{\top}u = n^{-1/2}\dot{\mathcal{L}}_n(\gamma,Z)^{\top}u + n^{-1/2}\{\dot{\mathcal{L}}_n(\gamma,\widehat{Z}) - \dot{\mathcal{L}}_n(\gamma,Z)\}^{\top}u$。第一项 $n^{-1/2}\dot{\mathcal{L}}_n(\gamma,Z)^{\top}u$ 可以一致地界定为 $n^{-1/2}\|\dot{\mathcal{L}}_n(\gamma,Z)\| = O_p(1)$；参见 Fan 和 Li（2001）。因此我们

重点考察第二项。根据中值定理和柯西-施瓦茨不等式，我们可以验证

$$\begin{aligned}\left|n^{-1/2}\Big\{\dot{\mathcal{L}}_n(\gamma,\widehat{Z})-\dot{\mathcal{L}}_n(\gamma,Z)\Big\}^\top u\right| &\le \sum_{i=1}^n\left|n^{-1/2}g\Big(\xi_i^\top\gamma\Big)\Big\{1-g\Big(\xi_i^\top\gamma\Big)\Big\}(\widehat{Z}_i-Z_i)^\top\xi_i\gamma^\top u\right| \\ &\quad+\sum_{i=1}^n\left|n^{-1/2}\Big\{Y_i-g\Big(\xi_i^\top\gamma\Big)\Big\}(\widehat{Z}_i-Z_i)^\top u\right| \\ &\le \sqrt{n}\Big\{n^{-1}\sum_{i=1}^n\|\widehat{Z}_i-Z_i\|^2\Big\}^{1/2}\left[n^{-1}\sum_{i=1}^n\left\|g\Big(\xi_i^\top\gamma\Big)\Big\{1-g\Big(\xi_i^\top\gamma\Big)\Big\}\xi_i\right\|^2\right]^{1/2}\|\gamma\| \\ &\quad+\left\{\sum_{i=1}^n\left|Y_i-g\Big(\xi_i^\top\gamma\Big)\right|^2\right\}^{1/2}\Big\{n^{-1}\sum_{i=1}^n\|\widehat{Z}_i-Z_i\|^2\Big\}^{1/2}, \qquad \text{(B.5)}\end{aligned}$$

其中 $\xi_i=\eta_i Z_i+(1-\eta_i)\widehat{Z}_i$ 对某个 $\eta_i\in(0,1)$ 成立。我们首先计算 (B.5) 的第一项。简单计算可得 $n^{-1}\sum_{i=1}^n\|\widehat{Z}_i-Z_i\|^2=n^{-1}\sum_{i=1}^n\|(\widehat{\mathbb{B}}-\mathbb{B})^\top X_i\|^2=\sum_{k=1}^K(\widehat{\beta}_k-\beta_k)^\top(n^{-1}\sum_{i=1}^n X_iX_i^\top)(\widehat{\beta}_k-\beta_k)\le\lambda_{\max}(n^{-1}\sum_{i=1}^n X_iX_i^\top)\sum_{k=1}^K\|\widehat{\beta}_k-\beta_k\|^2$。与此同时，根据 Wainwright (2019) 中的定理 6.5 以及条件（C1），我们知道存在某个常数 $C_1>0$，使得 $\lambda_{\max}\{n^{-1}\sum_{i=1}^n X_iX_i^\top-E(X_iX_i^\top)\}\le C_1(\sqrt{p/n}+p/n)\le C_1(\sqrt{C}+C)$ 以趋于 1 的概率成立。接着由条件（C3），有 $\lambda_{\max}(n^{-1}\sum_{i=1}^n X_iX_i^\top)\le C_1(\sqrt{C}+C)+\lambda_{\max}\{E(X_iX_i^\top)\}\le C_1(\sqrt{C}+C)+C_{\max}$ 以趋于 1 的概率成立。也就是说，$\overline{\lambda_{\max}}(n^{-1}\sum_{i=1}^n X_iX_i^\top)=O_p(1)$。由定理1可得，对于 $k=1,\ldots,K$，有 $\|\widehat{\beta}_k-\beta_k\|=O_p(\sqrt{p/N})$。因此，我们有 $n^{-1}\sum_{i=1}^n\|\widehat{Z}_i-Z_i\|^2=O_p(p/N)$。此外，注意到 $\xi_i=\eta_i(\widehat{Z}_i-Z_i)+Z_i$。由条件（C4），我们知道 $E\|Z_i\|^2\le\{E\|Z_i\|^4\}^{1/2}\le\tau_{\max}^{1/2}$。于是有$n^{-1}\sum_{i=1}^n\|g(\xi_i^\top\gamma)\{1-g(\xi_i^\top\gamma)\}\xi_i\|^2\le n^{-1}\sum_{i=1}^n\|\xi_i\|^2\le 2n^{-1}\sum_{i=1}^n(\|Z_i\|^2+\|\eta_i(\widehat{Z}_i-Z_i)\|^2)=O_p(1)$。因此，(B.5) 的第一项的阶为 $O_p(\sqrt{np/N})$。接下来考虑 (B.5) 的第二项。有 $\{\sum_{i=1}^n|Y_i-g(\xi_i^\top\gamma)|^2\}^{1/2}\le\sqrt{n}$，因为 $|Y_i-g(\xi_i^\top\gamma)|\le 1$。注意 $n^{-1}\sum_{i=1}^n\|\widehat{Z}_i-Z_i\|^2=O_p(p/N)$。因此，(B.5) 的第二项的阶为 $O_p(\sqrt{np/N})$。综合以上结果，我们得到 $|n^{-1/2}\{\dot{\mathcal{L}}_n(\gamma,\widehat{Z})-\dot{\mathcal{L}}_n(\gamma,Z)\}^\top u|=O_p(\sqrt{np/N})$。由条件（C1）可知$np/N\le Cn^2/N\to 0$。于是有 $n^{-1/2}\dot{\mathcal{L}}_n(\gamma,\widehat{Z})^\top u=n^{-1/2}\dot{\mathcal{L}}_n(\gamma,Z)^\top u+o_p(1)=$

$O_p(1)$。这就证明了 (B.3)。

STEP 1.2. 接下来我们研究 (B.4)。回顾一下，$I(\gamma) = -E\{n^{-1}\ddot{\mathcal{L}}_n(\gamma, Z)\}$ 是预言机似然的费 Fisher 信息矩阵。因此，只要我们能够证明

$$\sup_{\|u\|=1}\Big|u^\top\Big\{-n^{-1}\ddot{\mathcal{L}}_n(\gamma;\widehat{Z})-I(\gamma)\Big\}u\Big| \le \sup_{\|u\|=1}\Big|n^{-1}u^\top\ddot{\mathcal{L}}_n(\gamma;\widehat{Z})u-n^{-1}u^\top\ddot{\mathcal{L}}_n(\gamma;Z)u\Big| + \sup_{\|u\|=1}\Big|u^\top\Big\{-n^{-1}\ddot{\mathcal{L}}_n(\gamma;Z)-I(\gamma)\Big\}u\Big| = o_p(1). \tag{B.6}$$

根据大数定律，我们有 $-n^{-1}\ddot{\mathcal{L}}_n(\gamma; Z) \xrightarrow{p} I(\gamma)$。因此，只要我们能够证明 $\sup_{\|u\|=1} n^{-1}u^\top\{\ddot{\mathcal{L}}_n(\gamma;\ \widehat{Z})\ -\ddot{\mathcal{L}}_n(\gamma; Z)\}u = o_p(1)$，(B.4) 即成立。根据中值定理，我们可以写成

$$\begin{aligned}
&\Big|u^\top\Big\{n^{-1}\ddot{\mathcal{L}}_n(\gamma;\widehat{Z})-n^{-1}\ddot{\mathcal{L}}_n(\gamma;Z)\Big\}u\Big| \\
=\ & n^{-1}\Big|\sum_{i=1}^n\Big[g\Big(\widehat{Z}_i^\top\gamma\Big)\Big\{1-g\Big(\widehat{Z}_i^\top\gamma\Big)\Big\}\Big(\widehat{Z}_i^\top u\Big)^2-g\Big(Z_i^\top\gamma\Big)\Big\{1-g\Big(Z_i^\top\gamma\Big)\Big\}\Big(Z_i^\top u\Big)^2\Big]\Big| \\
=\ & \Big[\Big\{1-2g\Big(\tilde{\xi}_i^\top\gamma\Big)\Big\}g\Big(\tilde{\xi}_i^\top\gamma\Big)\Big\{1-g\Big(\tilde{\xi}_i^\top\gamma\Big)\Big\}\Big(\tilde{\xi}_i^\top u\Big)^2\gamma+2g\Big(\tilde{\xi}^\top\gamma\Big)\Big\{1-g\Big(\tilde{\xi}_i^\top\gamma\Big)\Big\}\Big(\tilde{\xi}_i^\top u\Big)u\Big] \\
\times\ & n^{-1}\Big|\sum_{i=1}^n\Big(\widehat{Z}_i-Z_i\Big)^\top\Big| \le \Big\{n^{-1}\sum_{i=1}^n\|\widehat{Z}_i-Z_i\|^2\Big\}^{1/2}\times\Big(I_1+I_2\Big) = O_p(\sqrt{p/N})(I_1+I_2),
\end{aligned}$$

其中 $I_1 = \big\{n^{-1}\sum_{i=1}^n\big\|\big\{1-2g\big(\tilde{\xi}_i^\top\gamma\big)\big\}g\big(\tilde{\xi}_i^\top\gamma\big)\big\{1-g\big(\tilde{\xi}_i^\top\gamma\big)\big\}\big(\tilde{\xi}_i^\top u\big)^2\gamma\big\|^2\big\}^{1/2}$, $I_2 = \big\{n^{-1}\sum_{i=1}^n\big\|2g\big(\tilde{\xi}^\top\gamma\big)\big\{1-g\big(\tilde{\xi}_i^\top\gamma\big)\big\}\big(\tilde{\xi}_i^\top u\big)u\big\|^2\big\}^{1/2}$，且对于某个 $\tilde{\eta}_i \in [0,1]$，$\tilde{\xi}_i = \tilde{\eta}_i\widehat{Z}_i + (1-\tilde{\eta}_i)Z_i$。接下来，我们应依次评估 I_1 和 I_2。对于 I_1，我们可以验证 $I_1^2 \le n^{-1}\|\gamma\|^2\sum_{i=1}^n(\tilde{\xi}_i^\top u)^2 \le n^{-1}\|\gamma\|^2\sum_{i=1}^n\|\tilde{\xi}_i\|^2$。第一个不等式由以下结果成立：(a) $|g(\tilde{\xi}_i^\top\tilde{\gamma})\{1-g(\tilde{\xi}_i^\top\tilde{\gamma})\}| \le 1$ 以及 (b) $|1-2g(\tilde{\xi}_i^\top\tilde{\gamma})| \le 1$。然后，通过与 STEP 1.1 中类似的过程，我们有 $n^{-1}\sum_{i=1}^n\|\tilde{\xi}_i\|^2 = O_p(1)$。注意，假设 γ有界，则有 $I_1 = O_p(1)$。对于 I_2，通过上述类似论证，我们可以证明$I_2^2 \le n^{-1}\sum_{i=1}^n\|\tilde{\xi}_i\|^2 = O_p(1)$。因此，(B.6) 的第一项由 $O_p(\sqrt{p/N}) = o_p(1)$ 有界。结合以上结果，我们有 $\sup_{\|u\|=1}|-n^{-1}u^\top\ddot{\mathcal{L}}_n(\gamma;\ \widehat{Z})u - u^\top I(\gamma)u| = o_p(1)$，且 $\inf_{\|u\|=1}\{-n^{-1}u^\top\ddot{\mathcal{L}}_n(\gamma;\ \widehat{Z})u\} \ge \inf_{\|u\|=1}\{-n^{-1}u^\top\ddot{\mathcal{L}}_n(\gamma;\ \widehat{Z})u - u^\top I(\gamma)u\}$

——

$+\inf_{\|u\|=1}\{u^\top I(\gamma)u\} = o_p(1)+\lambda_{\min}\{I(\gamma)\}$.。注意， $I(\gamma)$ 是一个正定矩阵，其特征值与零之间存在正距离。因此，我们有 $\lambda_{\min}\{I(\gamma)\} > 0$。由此可得 $\lambda_{\min}\{n^{-1}u^\top \ddot{\mathcal{L}}_n(\gamma; \widehat{Z})u\} = \inf_{\|u\|=1}\{n^{-1}u^\top \ddot{\mathcal{L}}_n(\gamma; \widehat{Z})u\} > 2^{-1}\lambda_{\min}\{I(\gamma)\}$，且该结果以趋于一的概率成立。因此，对于上述提到的 ε，我们可以取一个充分大的 C_ε，使得 $\sup_{\|u\|=1}\mathcal{L}_n(\gamma + C_\varepsilon n^{-1/2}u; \widehat{Z}) < \mathcal{L}_n(\gamma; \widehat{Z})$。结合前面的步骤，我们完成了 $\sqrt{n}$-一致性的证明。

–

步骤 *2.* 渐近正态性

接下来，我们研究 $\widehat{\gamma}$ 的渐近分布。注意， $\widehat{\gamma}$ 是 $\mathcal{L}_n(\gamma; Z)$ 的局部优化器 $\widehat{\quad\quad\quad

因此，我们有$-n^{-1}\ddot{\mathcal{L}}_n(\gamma;\ \widehat{Z}) \xrightarrow{p} I(\gamma)$。那么，$n^{-1}\ddot{\mathcal{L}}_n(\gamma;\ \widehat{Z})$以趋于 1 的概率可逆。因此，根据斯卢茨基定理，我们有 $\sqrt{n}(\widehat{\gamma}-\gamma) =\{-n^{-1}\ddot{\mathcal{L}}_n(\gamma;\widehat{Z})\}^{-1}\{n^{-1/2}\dot{\mathcal{L}}_n(\gamma;\ \widehat{Z})\} + o_p(1) \xrightarrow{d} N(0,\ I(\gamma)^{-1})$。这就完成了 $\widehat{\gamma}$ 渐近正态性的证明。

步骤 *3*。$\widehat{\theta}^{Transfer}$ 的一致性与渐近正态性

接下来，我们研究 $\widehat{\theta}^{\text{Transfer}}$ 的渐近行为。证明分为两部分。第一部分验证 $\widehat{\theta}^{\text{Transfer}}$ 的 $\sqrt{n}$-一致性，第二部分则证明 $\widehat{\theta}^{\text{Transfer}}$ 的渐近正态性。

Step 3.1. 我们首先讨论$\sqrt{n}$-一致性。根据柯西-施瓦茨不等式和算子范数的性质（Seber，2008），有 $||\widehat{\theta}^{\text{Transfer}}-\theta|| \le ||\widehat{\mathbb{B}}^{\text{Source}}_{\text{mle}}-\mathbb{B}||_F||\widehat{\gamma}|| + \lambda^{1/2}_{\max}(\mathbb{B}^\top\mathbb{B})||\widehat{\gamma}-\gamma||$。回顾一下，我们已在步骤 *1*中证明了 $\widehat{\gamma}-\gamma = O_p(1/\sqrt{n})$。根据定理1，第一项的阶为 $O_p(\sqrt{p/N})\{\gamma + O_p(1/\sqrt{n})\} = O_p(\sqrt{p/N})$。由条件（C5）可知，第二项的阶为 $O(1)O_p(1/\sqrt{n}) = O_p(1/\sqrt{n})$。此外，由条件（C1），给定 $n^2\log N/N \to 0$ 和 $p \le Cn$，我们可以得到 $||\widehat{\theta}^{\text{Transfer}}-\theta|| = O_p(1/\sqrt{n})$，因此 $\widehat{\theta}^{\text{Transfer}}$ 是 $\sqrt{n}$-一致的。

Step 3.2. 然后我们评估 $\widehat{\theta}^{\text{Transfer}}$的渐近正态性。根据定理条件 $v \in \mathcal{S}(\mathbb{B})$，存在一个向量 $\zeta \in \mathbb{R}^K$ 使得 $v = \mathbb{B}\zeta$。因此，我们有 $\sigma^2(v) = v^\top\mathbb{B}I^{-1}(\gamma)\mathbb{B}^\top v = \zeta^\top\mathbb{B}^\top\mathbb{B}I^{-1}(\gamma)\mathbb{B}^\top\mathbb{B}\zeta \ge \lambda^{-1}_{\max}\{I(\gamma)\}\lambda_{\min}\{\mathbb{B}^\top\mathbb{B}\}$。类似地，我们有 $\sigma^2(v) \le \lambda^{-1}\min\{I(\gamma)\}\lambda_{\max}(\mathbb{B}^\top\mathbb{B})$。根据条件(C5)以及 $I(\gamma)$和 $\mathbb{B}^\top\mathbb{B}$的正定性，我们有 $\sigma^2(v) = O(1)$和 $\sigma^{-2}(v) = O(1)$。注意，$\sqrt{n}v^\top(\widehat{\theta}^{\text{Transfer}}-\theta)/\sigma(v) = \sqrt{n}v^\top(\widehat{\mathbb{B}}^{\text{Source}}_{\text{mle}}-\mathbb{B})\widehat{\gamma}/\sigma(v) + \sqrt{n}v^\top\mathbb{B}(\widehat{\gamma}-\gamma)/\sigma(v)$。根据条件(C1)、定理1和柯西-施瓦茨不等式，我们知道 $\sqrt{n}v^\top(\widehat{\mathbb{B}}^{\text{Source}}_{\text{mle}}-\mathbb{B})\widehat{\gamma}/\sigma(v) \le \sqrt{n}\|v\|\|\widehat{\mathbb{B}}^{\text{Source}}_{\text{mle}}-\mathbb{B}\|_F\|\widehat{\gamma}\|/\sigma(v) = \sqrt{n}O_p(\sqrt{p/N})O_p(1) = O_p(\sqrt{np/N}) = O_p(\sqrt{n^2/N}) = o_p(1)$。注意，$\|\mathbb{B}^\top v\| \le \lambda_{\max}(\mathbb{B}^\top\mathbb{B}) \le \tau_{\max}$。根据 Step 中的结果。

1.1、(B.6) 和 (B.7)，我们知道

$$
\begin{aligned}
0 &= n^{-1/2}\sigma^{-1}(v)v^\top\mathbb{B}\Big\{-n^{-1}\ddot{\mathcal{L}}_n(\gamma;\widehat{Z})\Big\}^{-1}\dot{\mathcal{L}}_n(\gamma;\widehat{Z}) - \sqrt{n}\sigma^{-1}(v)v^\top\mathbb{B}(\widehat{\gamma}-\gamma) + o_p(1) \\
&= n^{-1/2}\sigma^{-1}(v)v^\top\mathbb{B}I(\gamma)^{-1}\dot{\mathcal{L}}_n(\gamma;Z) - \sqrt{n}\sigma^{-1}(v)v^\top\mathbb{B}(\widehat{\gamma}-\gamma) + o_p(1) \\
&= \sum_{i=1}^{n} Y_{ni} - \sqrt{n}\sigma^{-1}(v)v^\top\mathbb{B}(\widehat{\gamma}-\gamma) + o_p(1), \qquad \text{(B.8)}
\end{aligned}
$$

其中 $Y_{ni} = n^{-1/2}\sigma^{-1}(v)\{Y_i - g(Z_i^\top\gamma)\}v^\top\mathbb{B}I(\gamma)^{-1}Z_i$ 当 $1 \le i \le n$ 时成立。于是，根据林德伯格-费勒中心极限定理（Van der Vaart，1998），若能证明对任意 $\eta_0 > 0$，则 (B.8) 的第一项渐近正态。

$$
(1)\ \sum_{i=1}^{n} E\|Y_{ni}\|^2 I\{\|Y_{ni}\| \ge \eta_0\} \to 0 \quad \text{and} \quad (2)\ \sum_{i=1}^{n} \mathrm{var}(Y_{ni}) \to 1. \qquad \text{(B.9)}
$$

随后我们验证这两个结论。对于第一个结论，根据切比雪夫不等式和条件（C5），我们有

$$
\begin{aligned}
\sum_{i=1}^{n} E\|Y_{ni}\|^2 I\{\|Y_{ni}\| \ge \eta_0\} &\le \eta_0^{-2}\sum_{i=1}^{n} E\|Y_{ni}\|^4 \le n^{-2}\sigma^{-4}(v)\sum_{i=1}^{n} E\Big\{v^\top\mathbb{B}I(\gamma)^{-1}Z_i\Big\}^4 \\
&\le n^{-2}\sigma^{-4}(v)\sum_{i=1}^{n}\lambda_{\max}^2(\mathbb{B}^\top\mathbb{B})\lambda_{\min}^{-4}\{I(\gamma)\}E\|Z_i\|^4 = O(n^{-1}) = o(1).
\end{aligned}
$$

因此，（B.9）的第一个结论成立。接下来，我们考虑（B.9）的第二个结论。通过直接计算，我们知道 $E(Y_{ni}) = n^{-1/2}\sigma^{-1}(v)v^\top\mathbb{B}\times I(\gamma)^{-1}E[\{Y_i - g(Z_i^\top\gamma)\}Z_i] = 0$。于是我们有

$$
\sum_{i=1}^{n}\mathrm{var}(Y_{ni}) = \sum_{i=1}^{n} n^{-1}\sigma^{-2}(v)v^\top\mathbb{B}I(\gamma)^{-1}E\left[\Big\{Y_i - g(Z_i^\top\gamma)\Big\}^2 Z_iZ_i^\top\right]I(\gamma)^{-1}\mathbb{B}^\top v = 1.
$$

然后我们验证了（B.9）的第二个结论。结合上述结果，我们有 $\sqrt{n}\sigma^{-1}(v)v^\top(\widehat{\theta}^{\text{Transfer}} - \theta) = \sqrt{n}\sigma^{-1}(v)v^\top\mathbb{B}(\widehat{\gamma}-\gamma) + o_p(1) = n^{-1/2}\sigma^{-1}(v)v^\top\mathbb{B}I(\gamma)^{-1}\dot{\mathcal{L}}_n(\gamma;\ Z) + o_p(1) \xrightarrow{d} N(0,\ 1)$。至此，我们完成了$\widehat{\theta}^{\text{Transfer}}$的渐近正态性证明。定理证明完毕。

附录 C. 定理 3 的证明

定义 $\widetilde{\mathbb{W}}=\mathrm{diag}[g(\widehat{Z}_1^\top\gamma)\{1-g(\widehat{Z}_1^\top\gamma)\},\ldots,g(\widehat{Z}_n^\top\gamma)\{1-g(\widehat{Z}_n^\top\gamma)\}]\in\mathbb{R}^{n\times n}$和 $\mathbb{W}=\mathrm{diag}[g(Z_1^\top\gamma)\{1-g(Z_1^\top\gamma)\},\ldots,g(Z_n^\top\gamma)\{1-g(Z_n^\top\gamma)\}]\in\mathbb{R}^{n\times n}$。令 $\mathbb{Z}=(Z_1,\ldots,Z_n)^\top\in\mathbb{R}^{n\times K}$为降维特征的设计矩阵，$\widehat{\mathbb{Z}}=(\widehat{Z}_1,\ldots,\widehat{Z}_n)^\top\in\mathbb{R}^{n\times K}$为 $\mathbb{Z}$的估计矩阵。令 $\tilde{\varepsilon}_i=Y_i-g(\widehat{Z}_i^\top\gamma)$ 和 $\tilde{\varepsilon}=(\tilde{\varepsilon}_1,\ldots,\tilde{\varepsilon}_n)^\top\in\mathbb{R}^n$。回顾一下，当零假设成立时，伪残差为 $\varepsilon_i=Y_i-g(Z_i^\top\gamma)$。令 $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_n)^\top\in\mathbb{R}^n$为伪残差向量。根据定理 2 的证明，我们有 $\widehat{\gamma}-\gamma=(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}})^{-1}\widehat{\mathbb{Z}}^\top\tilde{\varepsilon}+o_p(n^{-1/2})$。回顾一下，$\widehat{\varepsilon}=(\widehat{\varepsilon}_1,\ldots,\widehat{\varepsilon}_n)^\top$，其中 $\widehat{\varepsilon}_i=Y_i-g(\widehat{Z}_i^\top\widehat{\gamma})$。根据泰勒展开，我们有

$$\widehat{\varepsilon}=\tilde{\varepsilon}-\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}(\widehat{\gamma}-\gamma)+o_p(n^{-1/2})=\left\{\mathbb{I}_n-\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\left(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\right)^{-1}\widehat{\mathbb{Z}}^\top\right\}\tilde{\varepsilon}+o_p(n^{-1/2}),$$

其中 $\mathbb{I}_n$ 是单位矩阵。于是，检验统计量可写为

$$\begin{aligned}T_1&=n^{-2}\tilde{\varepsilon}^\top\left\{\mathbb{I}_n-\widehat{\mathbb{Z}}\left(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\right)^{-1}\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\right\}\mathbb{X}\mathbb{X}^\top\left\{\mathbb{I}_n-\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\left(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\right)^{-1}\widehat{\mathbb{Z}}^\top\right\}\tilde{\varepsilon}+o_p(n^{-1})\\&=n^{-2}\Big\{\varepsilon^\top\mathbb{X}\mathbb{X}^\top\varepsilon+(\tilde{\varepsilon}^\top\mathbb{X}\mathbb{X}^\top\tilde{\varepsilon}-\varepsilon^\top\mathbb{X}\mathbb{X}^\top\varepsilon)+\tilde{\varepsilon}^\top\widehat{\mathbb{Z}}\left(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\right)^{-1}\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\mathbb{X}\mathbb{X}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\left(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\right)^{-1}\widehat{\mathbb{Z}}^\top\tilde{\varepsilon}\\&\quad-2\tilde{\varepsilon}^\top\mathbb{X}\mathbb{X}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\left(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\right)^{-1}\widehat{\mathbb{Z}}^\top\tilde{\varepsilon}\Big\}+o_p(n^{-1})=T_1^*+R_1+R_2-2R_3+o_p(n^{-1}),\end{aligned}$$

其中 $T_1^*=n^{-2}\varepsilon^\top\mathbb{X}\mathbb{X}^\top\varepsilon$，$R_1=n^{-2}(\tilde{\varepsilon}^\top\mathbb{X}\mathbb{X}^\top\tilde{\varepsilon}-\varepsilon^\top\mathbb{X}\mathbb{X}^\top\varepsilon)$ $R_2=n^{-2}\tilde{\varepsilon}^\top\widehat{\mathbb{Z}}(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}})^{-1}\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\times\mathbb{X}\mathbb{X}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}})^{-1}\widehat{\mathbb{Z}}^\top\tilde{\varepsilon}$，且 $R_3=n^{-2}\tilde{\varepsilon}^\top\mathbb{X}\mathbb{X}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}})^{-1}\widehat{\mathbb{Z}}^\top\tilde{\varepsilon}$。回想一下，$\Sigma_\gamma=E(\varepsilon_i^2X_iX_i^\top)=E[g(Z_i^\top\gamma)\{1-g(Z_i^\top\gamma)\}X_iX_i^\top]=(\sigma_{j_1j_2}^{(\gamma)})_{p\times p}$。根据条件（C3），我们知道 $pC_{\min}^2\le\mathrm{tr}(\Sigma_\gamma^2)\le pC_{\max}^2$。因此，只要我们能证明以下五个结果，定理结论便可得出

$$(1)E(T_1^*)=n^{-1}\mathrm{tr}(\Sigma_\gamma),\tag{C.1}$$

$$(2)\mathrm{var}(T_1^*)=n^{-3}\mathrm{var}(\varepsilon_i^2X_i^\top X_i)+2n^{-2}\mathrm{tr}(\Sigma_\gamma^2)\{1+o(1)\},\tag{C.2}$$

$$(3)R_1=o_p\{n^{-1}\sqrt{p}\},\tag{C.3}$$

$$(4) R_2 = o_p\{n^{-1}\sqrt{p}\}, \tag{C.4}$$

$$(5) R_3 = o_p\{n^{-1}\sqrt{p}\}. \tag{C.5}$$

注意，如果上述结果成立，我们应有 $R_i/\mathrm{var}^{1/2}(T_1^*) \xrightarrow{p} 0$，其中 $i = 1, 2, 3$。余项应当可以忽略。随后将验证这五个结论。

Step 1. 我们从 $E(T_1^*)$ 开始。在零假设下，我们知道，给定 $\mathbb{Z}$， Y_i 与 X_i 独立。因此，我们可以计算条件期望为$E(\varepsilon^\top \mathbb{X}\mathbb{X}^\top \varepsilon|\mathbb{Z}) = E\{\mathrm{tr}(\varepsilon\varepsilon^\top \mathbb{X}\mathbb{X}^\top)|\mathbb{Z}\}$ =$\mathrm{tr}\{E(\varepsilon\varepsilon^\top|\mathbb{Z})E(\mathbb{X}\mathbb{X}^\top|\mathbb{Z})\}$ =$\mathrm{tr}\{\mathbb{W}E(\mathbb{X}\mathbb{X}^\top|\mathbb{Z})\}$ =$\mathrm{tr}\{E(\mathbb{X}^\top\mathbb{W}\mathbb{X}|\mathbb{Z})\} = \mathrm{tr}\{E[\sum_{i=1}^n g(Z_i^\top\gamma)\{1 - g(Z_i^\top\gamma)\}X_iX_i^\top|\mathbb{Z}]\} = n\mathrm{tr}\{E[g(Z_i^\top\gamma)\{1 - g(Z_i^\top\gamma)\}X_iX_i^\top|\mathbb{Z}]\}$。因此，我们有 $E(T_1^*) = n^{-2}E\{E(\varepsilon^\top \mathbb{X}\mathbb{X}^\top\varepsilon|\mathbb{Z})\} = n^{-1}\mathrm{tr}(\Sigma_\gamma)$。

Step 2. 我们考虑 $\mathrm{var}(T_1^*) = n^{-4}E(\varepsilon^\top \mathbb{X}\mathbb{X}^\top\varepsilon)^2 - n^{-4}\{E(\varepsilon^\top \mathbb{X}\mathbb{X}^\top\varepsilon)\}^2$。首先考虑 $E(\varepsilon^\top \mathbb{X}\mathbb{X}^\top\varepsilon)^2$。于是我们有

$$\begin{aligned}
& E\Big(\varepsilon^\top \mathbb{X}\mathbb{X}^\top\varepsilon\Big)^2 = E\Big\{\sum_{j=1}^p\Big(\sum_{i=1}^n X_{ij}\varepsilon_i\Big)^2\Big\}^2 = E\Big(\sum_{j=1}^p\sum_{i_1,i_2}\varepsilon_{i_1}\varepsilon_{i_2}X_{i_1j}X_{i_2j}\Big)^2 \\
= & \sum_{j_1,j_2}\sum_{i_1i_2i_3i_4} E\Big(\varepsilon_{i_1}\varepsilon_{i_2}\varepsilon_{i_3}\varepsilon_{i_4}X_{i_1j_1}X_{i_2j_2}X_{i_3j_1}X_{i_4j_2}\Big) \\
= & \sum_{j_1,j_2}\sum_{i=1}^n E\Big(\varepsilon_i^4X_{ij_1}^2X_{ij_2}^2\Big) + \sum_{j_1,j_2}\sum_{i_1\neq i_2} E\Big(\varepsilon_{i_1}^2\varepsilon_{i_2}^2X_{i_1j_1}^2X_{i_2j_2}^2\Big) \\
+ & 2\sum_{j_1,j_2}\sum_{i_1\neq i_2} E\Big(\varepsilon_{i_1}^2\varepsilon_{i_2}^2X_{i_1j_1}X_{i_1j_2}X_{i_2j_1}X_{i_2j_2}\Big) \\
= & nE\Big\{\varepsilon_i^4\Big(X_i^\top X_i\Big)^2\Big\} + n(n-1)\sum_{j_1,j_2} E\Big(\varepsilon_i^2X_{ij_1}^2\Big)E\Big(\varepsilon_{i_2}^2X_{ij_2}^2\Big) \\
+ & 2n(n-1)\sum_{j_1,j_2} E^2\Big(\varepsilon_i^2X_{ij_1}X_{ij_2}\Big) \\
= & nE\Big\{\varepsilon_i^4\Big(X_i^\top X_i\Big)^2\Big\} + n(n-1)\Big\{\sum_j \sigma_{jj}^{(\gamma)}\Big\}^2 + 2n(n-1)\sum_{j_1,j_2}\sigma_{j_1j_2}^{2(\gamma)} \\
= & nE\Big\{\varepsilon_i^4\Big(X_i^\top X_i\Big)^2\Big\} + n(n-1)\mathrm{tr}^2(\Sigma_\gamma) + 2n(n-1)\mathrm{tr}(\Sigma_\gamma^2).
\end{aligned}$$

那么我们有 $\mathrm{var}(T_1^*) = n^{-4}E(\varepsilon^\top \mathbb{X}\mathbb{X}^\top\varepsilon)^2 - n^{-4}\{E(\varepsilon^\top \mathbb{X}\mathbb{X}^\top\varepsilon)\}^2 = n^{-3}E\{\varepsilon_i^4(X_i^\top X_i)^2\} -$

$n^{-3}\mathrm{tr}^2(\Sigma_\gamma) + 2n^{-2}(1-n^{-1})\mathrm{tr}(\Sigma_\gamma^2) = n^{-3}\mathrm{var}(\varepsilon_i^2 X_i^\top X_i) + 2n^{-2}(1-n^{-1})\mathrm{tr}(\Sigma_\gamma^2)$.

Step 3. 接下来我们研究 R_1 的阶。根据中值定理，我们有

$$\begin{aligned}
|R_1| &= n^{-2}\left|\tilde{\varepsilon}^\top \mathbb{X}\mathbb{X}^\top \tilde{\varepsilon} - \varepsilon^\top \mathbb{X}\mathbb{X}^\top \varepsilon\right| = n^{-2}\sum_{i=1}^n\sum_{j=1}^n \left|\left(\tilde{\varepsilon}_i\tilde{\varepsilon}_j - \varepsilon_i\varepsilon_j\right)X_i^\top X_j\right| \\
&\le n^{-2}\sum_{i=1}^n\sum_{j=1}^n \left|\left(\tilde{\varepsilon}_i\tilde{\varepsilon}_j - \tilde{\varepsilon}_i\varepsilon_j\right)X_i^\top X_j\right| + n^{-2}\sum_{i=1}^n\sum_{j=1}^n \left|\left(\tilde{\varepsilon}_i\varepsilon_j - \varepsilon_i\varepsilon_j\right)X_i^\top X_j\right| \\
&\le 2n^{-2}\sum_{i=1}^n\sum_{j=1}^n \left|(\tilde{\varepsilon}_i - \varepsilon_i)X_i^\top X_j\right| \\
&= 2n^{-2}\sum_{i=1}^n\sum_{j=1}^n g\left(\bar{\xi}_i^\top\gamma\right)\left\{1-g\left(\bar{\xi}_i^\top\gamma\right)\right\}\left|(\widehat{Z}_i - Z_i)^\top\gamma X_i^\top X_j\right| \le R_{11}+R_{12},
\end{aligned}$$

其中，对某个 $\bar{\eta}_i \in [0,1]$, $R_{11} = 2n^{-2}\sum_{i=1}^n |(\widehat{Z}_i - Z_i)^\top\gamma X_i^\top X_i|$，有 $\bar{\xi}_i = \bar{\eta}_i\widehat{Z}_i + (1-\bar{\eta}_i)Z_i$，且 $R_{12} = 2n^{-2}\sum_{i\neq j}|(\widehat{Z}_i - Z_i)^\top\gamma X_i^\top X_j|$。接下来，我们依次估计上述两项。对于 R_{11}，由柯西-施瓦茨不等式可得 $R_{11} \le 2n^{-1}\|\gamma\|\left(n^{-1}\sum_{i=1}^n\|\widehat{Z}_i - Z_i\|^2\right)^{1/2}\left\{n^{-1}\sum_{i=1}^n(X_i^\top X_i)^2\right\}^{1/2}$。根据定理 2 的证明，我们有 $n^{-1}\sum_{i=1}^n\|\widehat{Z}_i - Z_i\|^2 = O_p(p/N)$。由 X_i 的次高斯性可得 $E\{n^{-1}\sum_{i=1}^n(X_i^\top X_i)^2\} = E(X_i^\top X_i)^2 \le \kappa_{\max}p^2$，其中某个 $0 < \kappa_{max} < \infty$。因此，我们有 $n^{-1}\sum_{i=1}^n(X_i^\top X_i)^2 = O_p(p^2)$。注意，由条件（C1），我们有 $p^2/N \le n^2/N \to 0$。因此，我们有 $R_{11} = 2n^{-1}O_p(p^{1/2}/N^{1/2})O_p(p) = O_p(n^{-1}p^{3/2}N^{-1/2}) = o_p(n^{-1}p^{1/2})$。

接下来，我们关注 R_{12}。根据柯西-施瓦茨不等式，我们有 $R_{12} \le 2\|\gamma\|\left(n^{-1}\sum_{i=1}^n\|\widehat{Z}_i - Z_i\|^2\right)^{1/2}\left\{n^{-2}\sum_{i\neq j}(X_i^\top X_j)^2\right\}^{1/2}$。根据条件（C4）和柯西-施瓦茨不等式，我们有 $E\{n^{-2}\sum_{i\neq j}(X_i^\top X_j)^2\} = (1-n^{-1})E[\{n(n-1)\}^{-1}\sum_{i\neq j}(X_i^\top X_j)^2] = (1-n^{-1})E(X_i^\top X_j)^2 \le \{E(X_i^\top X_j)^4\}^{1/2} \le \tau_{\max}^{1/2}p$。注意，$n^{-1}\sum_{i=1}^n\|\widehat{Z}_i - Z_i\|^2 = O_p(p/N)$。因此，我们有 $R_{12} = O_p(p^{1/2}N^{-1/2})O_p(p^{1/2}) = O_p(pN^{-1/2}) = o_p(n^{-1}p^{1/2})$，其中最后一个等式来自假设 $n^2p/N \to 0$。结合以上结果，(C.3) 的证明完成。

Step 4. 接下来我们评估 R_2。令 $\widehat{\mathbb{H}} = \widehat{\mathbb{Z}}(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}})^{-1}\widehat{\mathbb{Z}}^\top = (\widehat{h}_{ij})_{K\times K}$ 和 $\mathbb{H} = \mathbb{Z}(\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\mathbb{Z}^\top = (h_{ij})_{K\times K}$。那么 $R_2 = n^{-2}\tilde{\varepsilon}^\top\widehat{\mathbb{H}}\widetilde{\mathbb{W}}\mathbb{X}\mathbb{X}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{H}}\tilde{\varepsilon} \le n^{-2}\lambda_{\max}(\mathbb{X}\mathbb{X}^\top)\tilde{\varepsilon}^\top\widehat{\mathbb{H}}^2\tilde{\varepsilon}$。

接下来我们应分别研究 $\lambda_{\max}(\mathbb{X}\mathbb{X}^\top)$ 和 $\tilde{\varepsilon}^\top\widehat{\mathbb{H}}^2\tilde{\varepsilon}$。

STEP 4.1. 我们首先考虑 $\lambda_{\max}(\mathbb{X}\mathbb{X}^\top)$。根据切比雪夫不等式，对于任意 $t>0$，我们有 $P(\lambda_{\max}(\mathbb{X}\mathbb{X}^\top)>t)\le t^{-4}E[\lambda^4_{\max}\{(\mathbb{X}\mathbb{X}^\top)^4\}]\le t^{-4}E[\mathrm{tr}\{(\mathbb{X}\mathbb{X}^\top)^4\}]$。接下来我们考虑 $E[\mathrm{tr}\{(\mathbb{X}\mathbb{X}^\top)^4\}]$。我们将指标集分解为以下六个子集：(i) $\mathcal{A}_1=\{(i_1,i_2,i_3,i_4)$: $i_1\neq i_2\neq i_3\neq i_4\}$；(ii) $\mathcal{A}_2=\{(i_1,i_2,i_3,i_4)$: $i_{j_1}\neq i_{j_2}\neq i_{j_3}=i_{j_4}$对于某些 j_1、j_2、j_3、$j_4\in\{1,2,3,4\}\}$；(iii) $\mathcal{A}_3=\{(i_1,i_2,i_3,i_4)$: $i_1=i_2\neq i_3=i_4$ 或 $i_1=i_4\neq i_2=i_3\}$；(iv) $\mathcal{A}_4=\{(i_1,i_2,i_3,i_4)$: $i_1=i_3\neq i_2=i_4\}$；(v) $\mathcal{A}_5=\{(i_1,i_2,i_3,i_4)$: $i_{j_1}=i_{j_2}=i_{j_3}\neq i_{j_4}$对于某些 j_1、j_2、j_3、$j_4\in\{1,2,3,4\}\}$；以及 (vi) $\mathcal{A}_6=\{(i_1,i_2,i_3,i_4)$: $i_1=i_2=i_3=i_4\}$。注意，$E[\mathrm{tr}\{(\mathbb{X}\mathbb{X}^\top)^4\}]=\sum_{k=1}^6\sum_{\mathcal{A}_k}E(X_{i_1}^\top X_{i_2}X_{i_2}^\top X_{i_3}X_{i_3}^\top X_{i_4}X_{i_4}^\top X_{i_1})$。随后，我们依次计算不同 $\mathcal{A}_k$'s 中期望值之和。注意，对于 $(i_1,i_2,i_3,i_4)\in\mathcal{A}_1$，应有$X_{i_1}$、$X_{i_2}$、$X_{i_3}$和 X_{i_4}相互独立。因此，我们有

$$\sum_{\mathcal{A}_1}E\Big(X_{i_1}^\top X_{i_2}X_{i_2}^\top X_{i_3}X_{i_3}^\top X_{i_4}X_{i_4}^\top X_{i_1}\Big)=\sum_{i_1\neq i_2\neq i_3\neq i_4}E\Big\{\mathrm{tr}\Big(X_{i_1}X_{i_1}^\top X_{i_2}X_{i_2}^\top X_{i_3}X_{i_3}^\top X_{i_4}X_{i_4}^\top\Big)\Big\}$$
$$=n(n-1)(n-2)(n-3)\mathrm{tr}\Big\{E^4\Big(X_iX_i^\top\Big)\Big\}=O(n^4p).$$

对于 $\mathcal{A}_2$，当$n\ge 2$时，我们有 $|\mathcal{A}_2|\le|\mathcal{A}^c|=n^4-|\mathcal{A}|=n^4-n(n-1)(n-2)(n-3)\le 6n^3$。注意，对于任意正定矩阵 A 和 B，$\mathrm{tr}(AB)\le\mathrm{tr}(A)\mathrm{tr}(B)$（Seber，2008）。注意，由于 X_i具有次高斯性，我们有 $E(X_i^\top X_i)^4\le\kappa'_{\max}p^4$，其中某个 $0<\kappa'_{max}<\infty$。然后，根据柯西-施瓦茨不等式、霍尔德不等式和条件（C3），我们有

$$\sum_{\mathcal{A}_2}E\Big(X_{i_1}^\top X_{i_2}X_{i_2}^\top X_{i_3}X_{i_3}^\top X_{i_4}X_{i_4}^\top X_{i_1}\Big)\le 6n^3E\Big[\mathrm{tr}\Big\{X_{i_1}X_{i_1}^\top X_{i_2}X_{i_2}^\top\big(X_{i_3}X_{i_3}^\top\big)^2\Big\}\Big]$$
$$=6n^3\mathrm{tr}\Big\{E^2\big(X_iX_i^\top\big)E\big(X_iX_i^\top\big)^2\Big\}\le 6n^3\mathrm{tr}^{1/2}\Big\{E^4\big(X_iX_i^\top\big)\Big\}\mathrm{tr}^{1/2}\Big\{E^2\big(X_iX_i^\top\big)^2\Big\}$$
$$\le 6C_{\max}^2n^3p^{1/2}E^{1/2}\Big(X_{i_2}^\top X_{i_1}X_{i_1}^\top X_{i_1}X_{i_1}^\top X_{i_2}X_{i_2}^\top X_{i_2}\Big)$$
$$\le 6C_{\max}^2n^3p^{1/2}\Big\{E\big(X_{i_1}^\top X_{i_2}\big)^4\Big\}^{1/4}\Big\{E\big(X_i^\top X_i\big)^4\Big\}^{1/4}=O(n^3p^2).$$

对于 $\mathcal{A}_3$，注意到 $|\mathcal{A}_3| \le 2n^2$。然后，根据前面的结果，我们有

$$\sum_{\mathcal{A}_3} E\Big(X_{i_1}^\top X_{i_2} X_{i_2}^\top X_{i_3} X_{i_3}^\top X_{i_4} X_{i_4}^\top X_{i_1}\Big) \le 2n^2 \mathrm{tr}\Big[E\Big\{\big(X_{i_1}X_{i_1}^\top\big)^2\big(X_{i_2}X_{i_2}^\top\big)^2\Big\}\Big]$$
$$= 2n^2 \mathrm{tr}\Big\{E^2\big(X_i X_i^\top\big)^2\Big\} \le 2C_{\max}^2 n^2 p^3 = O(n^2p^3).$$

对于 $\mathcal{A}_4$，注意到 $|\mathcal{A}_4| \le n^2$。于是我们有 $\sum_{\mathcal{A}_4} E(X_{i_1}^\top X_{i_2} X_{i_2}^\top X_{i_3} X_{i_3}^\top X_{i_4} X_{i_4}^\top X_{i_1})$ $\le n^2 E(X_{i_1}^\top X_{i_2})^4 \le \tau_{max} n^2 p^2 = O(n^2p^2)$，其中最后一个不等式由条件（C3）得出。对于 $\mathcal{A}_5$，根据柯西-施瓦茨不等式和条件（C4），我们有

$$\sum_{\mathcal{A}_5} E\Big(X_{i_1}^\top X_{i_2} X_{i_2}^\top X_{i_3} X_{i_3}^\top X_{i_4} X_{i_4}^\top X_{i_1}\Big) \le n^2 E\Big\{\big(X_{i_1}^\top X_{i_1}\big)^2\big(X_{i_1}^\top X_{i_2}\big)^2\Big\}$$
$$\le n^2\Big\{E\big(X_i^\top X_i\big)^4\Big\}^{1/2}\Big\{E\big(X_{i_1}^\top X_{i_2}\big)^4\Big\}^{1/2} \le \kappa_{\max}'^{1/2}\tau_{\max}^{1/2} n^2 p^3 = O(n^2p^3).$$

对于 $\mathcal{A}_6$，我们知道 $\sum_{\mathcal{A}_6} E(X_{i_1}^\top X_{i_2} X_{i_2}^\top X_{i_3} X_{i_3}^\top X_{i_4} X_{i_4}^\top X_{i_1}) = nE(X_i^\top X_i)^4 = O(np^4)$。结合前面的结果和条件（C1），我们有 $E[\mathrm{tr}\{(\mathbb{X}\mathbb{X}^\top)^4\}] = O(n^4p) + O(n^3p^2) + O(n^2p^3) + O(np^4) = O(n^4p)$。因此，我们有 $\lambda_{\max}(\mathbb{X}\mathbb{X}^\top) = O_p(np^{1/4})$。

Step 4.2. 我们研究 $\tilde{\varepsilon}^\top \widehat{\mathbb{H}}^2 \tilde{\varepsilon} = R_{21} + R_{22} + R_{23}$，其中 $R_{21} = \tilde{\varepsilon}^\top(\widehat{\mathbb{H}}^2 - \mathbb{H}^2)\tilde{\varepsilon}$、$R_{22} = \tilde{\varepsilon}^\top \mathbb{H}^2 \tilde{\varepsilon} - \varepsilon^\top \mathbb{H}^2 \varepsilon$ 和 $R_{23} = \varepsilon^\top \mathbb{H}^2 \varepsilon$。然后，我们依次评估上述三个术语。

Step 4.2.1. 对于 R_{21}，我们首先考虑 $\|\mathbb{H} - \mathbb{H}\|_F$。注意到 $\|AB\|_F \le \|A\|_F \|B\|_F$ 对于任意矩阵 A 和 B 均成立（Seber，2008）。因此，我们有 $\|\mathbb{H} - \mathbb{H}\|_F \le \mathbb{H}_1 + \mathbb{H}_2$，其中 $\mathbb{H}_1 = \mathrm{tr}(\widehat{\mathbb{Z}}^\top\widehat{\mathbb{Z}})\|(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}})^{-1} - (\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\|_F$，且 $\mathbb{H}_2 = \|\widehat{\mathbb{Z}}(\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\widehat{\mathbb{Z}} - \mathbb{Z}(\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\mathbb{Z}^\top\|_F$。接着我们考虑 $\mathbb{H}_1$。根据定理 2 中的证明，我们有 $\|\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}} - \mathbb{Z}^\top\mathbb{W}\mathbb{Z}\|_F^2 = O_p(np/N)$。因此，由于根据定理 2 中的证明，$n^{-1}\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}} = O_p(1)$ 和 $n^{-1}\mathbb{Z}^\top\mathbb{W}\mathbb{Z} = O_p(1)$，我们有 $\|(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}})^{-1} - (\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\|_F \le \|(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}})^{-1}\|_F \|\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}} - \mathbb{Z}^\top\mathbb{W}\mathbb{Z}\|_F \|(\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\|_F = O_p(\sqrt{p}/n^3N)$。注意，$\mathrm{tr}(\widehat{\mathbb{Z}}^\top\widehat{\mathbb{Z}}) =$

$\sum_{i=1}^{n} \widehat{Z}_i^\top \widehat{Z}_i = O(n)$。因此，我们有 $\mathbb{H}_1 = O_p(\sqrt{p/nN})$。接下来我们研究 $\mathbb{H}_2$ 并得到

$$\begin{aligned}
\mathbb{H}_2 &= \|\widehat{\mathbb{Z}}(\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\widehat{\mathbb{Z}} - \mathbb{Z}(\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\mathbb{Z}^\top\|_F \\
&\le \|\widehat{\mathbb{Z}}(\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\widehat{\mathbb{Z}} - \widehat{\mathbb{Z}}(\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\mathbb{Z}^\top\|_F + \|\widehat{\mathbb{Z}}(\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\mathbb{Z}^\top - \mathbb{Z}(\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\mathbb{Z}^\top\|_F \\
&\le \|\widehat{\mathbb{Z}}\|_F\|(\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\|_F\|\widehat{\mathbb{Z}} - \mathbb{Z}\|_F + \|\widehat{\mathbb{Z}} - \mathbb{Z}\|_F\|(\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\|_F\|\mathbb{Z}\|_F = O_p(\sqrt{p/N}),
\end{aligned}$$

其中，最后一个等号由以下事实得出：(1) $\|\mathbb{Z}\|_F = O_p(\sqrt{n})$，(2) $\|\mathbb{Z}\|_F = O_p(\sqrt{n})$，(3) $\|(\mathbb{Z}^\top\mathbb{W}\mathbb{Z})^{-1}\|_F = O_p(1/n)$，以及 (4) $\|\widehat{\mathbb{Z}} - \mathbb{Z}\|_F = (\sum_{i=1}^n \|\widehat{Z}_i - Z_i\|^2)^{1/2} = O_p(\sqrt{np/N})$。结合上述结果，我们得到 $\|\widehat{\mathbb{H}} - \mathbb{H}\|_F = O_p(\sqrt{p/N})$。根据大数定律，我们有 $n^{-1}\mathbb{Z}^\top\mathbb{W}\mathbb{Z} \xrightarrow{p} I(\gamma)$ 和 $n^{-1}\mathbb{Z}^\top\mathbb{Z} = n^{-1}\sum_{i=1}^n Z_i Z_i^\top \xrightarrow{p} E(Z_i Z_i^\top)$。因此，我们有 $n^{-1}\mathbb{Z}^\top\mathbb{W}\mathbb{Z} = O_p(1)$ 和 $n^{-1}\mathbb{Z}^\top\mathbb{Z} = O_p(1)$。注意，$I(\gamma)$ 是正定矩阵，且 $\lambda_{\min}\{I(\gamma)\} > 0$。因此，$\lambda_{\min}(n^{-1}\mathbb{Z}^\top\mathbb{W}\mathbb{Z}) \ge 2^{-1}\lambda_{\min}\{I(\gamma)\} > 0$ 以趋于 1 的概率成立。于是我们有

$$\begin{aligned}
\|\mathbb{H}\|_F^2 = \mathrm{tr}(\mathbb{H}^2) &= \mathrm{tr}\Big\{\mathbb{Z}\Big(\mathbb{Z}^\top\mathbb{W}\mathbb{Z}\Big)^{-1}\mathbb{Z}^\top\mathbb{Z}\Big(\mathbb{Z}^\top\mathbb{W}\mathbb{Z}\Big)^{-1}\mathbb{Z}^\top\Big\} \\
&= \mathrm{tr}\Big\{\Big(n^{-1}\mathbb{Z}^\top\mathbb{Z}\Big)\Big(n^{-1}\mathbb{Z}^\top\mathbb{W}\mathbb{Z}\Big)^{-1}\Big(n^{-1}\mathbb{Z}^\top\mathbb{Z}\Big)\Big(n^{-1}\mathbb{Z}^\top\mathbb{W}\mathbb{Z}\Big)^{-1}\Big\} = O_p(1)
\end{aligned}$$

同样地，我们有 $\|\widehat{\mathbb{H}}\|_F^2 = O_p(1)$。然后 $\|\widehat{\mathbb{H}}^2 - \mathbb{H}^2\|_F \le \|\widehat{\mathbb{H}} - \mathbb{H}\|_F(\|\widehat{\mathbb{H}}\|_F + \|\mathbb{H}\|_F) = O_p(\sqrt{p/N})$。因此，我们有 $\tilde{\varepsilon}^\top(\widehat{\mathbb{H}}^2 - \mathbb{H}^2)\tilde{\varepsilon} \le \|\widehat{\mathbb{H}}^2 - \mathbb{H}^2\|_F\tilde{\varepsilon}^\top\tilde{\varepsilon} = O_p(\sqrt{n^2p/N}) = o_p(1)$。

Step 4.2.2. 我们研究 R_{22}。根据柯西-施瓦茨不等式，我们有

$$\begin{aligned}
|R_{22}| &= \Big|\sum_{i,j}^n h_{ij}\Big(\tilde{\varepsilon}_i\tilde{\varepsilon}_j - \varepsilon_i\varepsilon_j\Big)\Big| = \Big|\sum_{i,j}^n h_{ij}\tilde{\varepsilon}_i\Big(\tilde{\varepsilon}_j - \varepsilon_j\Big) + \sum_{i,j}^n h_{ij}\varepsilon_j\Big(\tilde{\varepsilon}_i - \varepsilon_i\Big)\Big| \\
&\le \Big|\sum_{i,j}^n h_{ij}\tilde{\varepsilon}_i g\Big(\bar{\xi}_j^\top\gamma\Big)\Big\{1 - g\Big(\bar{\xi}_j^\top\gamma\Big)\Big\}\Big(\widehat{Z}_j - Z_j\Big)^\top\gamma\Big| \\
&+ \Big|\sum_{i,j}^n h_{ij}\varepsilon_j g\Big(\bar{\xi}_i^\top\gamma\Big)\Big\{1 - g\Big(\bar{\xi}_i^\top\gamma\Big)\Big\}\Big(\widehat{Z}_i - Z_i\Big)^\top\gamma\Big| \\
&\le 2\Big(\sum_{i,j}^n h_{ij}^2\Big)^{1/2}\Big(n\sum_{i=1}^n\|\widehat{Z}_i - Z_i\|^2\Big)^{1/2}\|\gamma\|.
\end{aligned}$$

根据前面的证明，我们知道$\sum_{i}^{n}{}_{,j}h_{ij}^{2}=\mathrm{tr}(\mathbb{H}^{2})=O_{p}(1)$。因此，根据条件$n^{2}p/N\to 0$，我们有$|R_{22}|=O_{p}(1)O_{p}(\sqrt{n^{2}p/N})=O_{p}(\sqrt{n^{2}p/N})=o_{p}(1)$。

Step 4.2.3. 我们考虑 R_{23}。通过直接计算，我们得到 $R_{23}=\varepsilon^{\top}\mathbb{H}^{2}\varepsilon=\mathrm{tr}(\varepsilon\varepsilon^{\top}\mathbb{H}^{2})\leq\mathrm{tr}(\mathbb{H}^{2})=O_{p}(1)$。结合 Step 4.2.1 至 Step 4.2.3中的结果，我们知道$\tilde{\varepsilon}^{\top}\widehat{\mathbb{H}}^{2}\tilde{\varepsilon}=O_{p}(1)$。结合前面的结果，我们有 $np^{-1/2}R_{2}\leq n^{-1}p^{-1/2}\lambda_{\max}(\mathbb{X}\mathbb{X}^{\top})\tilde{\varepsilon}^{\top}\widehat{\mathbb{H}}^{2}\tilde{\varepsilon}=n^{-1}p^{-1/2}O_{p}(np^{1/4})O_{p}(1)=O_{p}(p^{-1/4})=o_{p}(1)$。于是$R_{2}=o_{p}(n^{-1}\sqrt{p})$。我们完成了 (C.4) 的证明。

–

Step 5. 现在我们来计算 R_{3}。注意，我们已经证明了 $E[\mathrm{tr}\{(\mathbb{X}\mathbb{X}^{\top})^{4}\}]=O(n^{4}p)$和 $\|\mathbb{H}\|_{F}=O_{p}(1)$。根据柯西-施瓦茨不等式，我们有

$$
\begin{aligned}
|R_{3}| &= n^{-2}\left|\tilde{\varepsilon}^{\top}\mathbb{X}\mathbb{X}^{\top}\widetilde{\mathbb{W}}\widehat{\mathbb{H}}\tilde{\varepsilon}\right| = n^{-2}\left|\mathrm{tr}\left(\tilde{\varepsilon}\tilde{\varepsilon}^{\top}\mathbb{X}\mathbb{X}^{\top}\widetilde{\mathbb{W}}\widehat{\mathbb{H}}\right)\right| \leq n^{-2}\left|\mathrm{tr}\left(\mathbb{X}\mathbb{X}^{\top}\widehat{\mathbb{H}}\right)\right| \\
&\leq n^{-2}\mathrm{tr}^{1/2}\left\{\left(\mathbb{X}\mathbb{X}^{\top}\right)^{2}\right\}\mathrm{tr}^{1/2}\left(\widehat{\mathbb{H}}^{2}\right) \leq n^{-2}\mathrm{tr}^{1/4}\left\{\left(\mathbb{X}\mathbb{X}^{\top}\right)^{4}\right\}\mathrm{tr}^{1/2}\left(\widehat{\mathbb{H}}^{2}\right) \\
&= n^{-2}O_{p}(np^{1/4})O_{p}(1)=O_{p}(n^{-1}p^{1/4})=o_{p}(n^{-1}p^{1/2}).
\end{aligned}
$$

然后我们证明结论（C.5）。由此，我们完成定理的完整证明。

附录 D. 定理 4 的证明

定义 $\mathbb{D}_{X}=\mathrm{diag}\{X_{1}^{\top}X_{1},\ldots,X_{n}^{\top}X_{n}\}\in\mathbb{R}^{n\times n}$。那么我们有 $\mathrm{tr}(\widehat{\Sigma}_{\gamma})=n^{-1}\widehat{\varepsilon}^{\top}\mathbb{D}_{X}\widehat{\varepsilon}$。通过类似于定理 3 中证明的步骤，我们将检验统计量$T_{2}=T_{1}-n^{-1}\mathrm{tr}(\widehat{\Sigma}_{\gamma})$ 分解为

$$
\begin{aligned}
T_{2} = n^{-2}\Big[&\varepsilon^{\top}(\mathbb{X}\mathbb{X}^{\top}-\mathbb{D}_{X})\varepsilon+\{\tilde{\varepsilon}^{\top}(\mathbb{X}\mathbb{X}^{\top}-\mathbb{D}_{X})\tilde{\varepsilon}-\varepsilon^{\top}(\mathbb{X}\mathbb{X}^{\top}-\mathbb{D}_{X})\varepsilon\} \\
&+\tilde{\varepsilon}^{\top}\widehat{\mathbb{Z}}\left(\widehat{\mathbb{Z}}^{\top}\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\right)^{-1}\widehat{\mathbb{Z}}^{\top}\widetilde{\mathbb{W}}(\mathbb{X}\mathbb{X}^{\top}-\mathbb{D}_{X})\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\left(\widehat{\mathbb{Z}}^{\top}\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\right)^{-1}\widehat{\mathbb{Z}}^{\top}\tilde{\varepsilon} \\
&-2\tilde{\varepsilon}^{\top}(\mathbb{X}\mathbb{X}^{\top}-\mathbb{D}_{X})\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\left(\widehat{\mathbb{Z}}^{\top}\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}\right)^{-1}\widehat{\mathbb{Z}}^{\top}\tilde{\varepsilon}\Big]+o_{p}(1)=T_{2}^{*}+\tilde{R}_{1}+\tilde{R}_{2}-2\tilde{R}_{3}+o_{p}(n^{-1}),
\end{aligned}
$$

其中 $T_{2}^{*}=n^{-2}\varepsilon^{\top}(\mathbb{X}\mathbb{X}^{\top}-\mathbb{D}_{X})\varepsilon$，$\tilde{R}_{1}=n^{-2}\{\tilde{\varepsilon}^{\top}(\mathbb{X}\mathbb{X}^{\top}-\mathbb{D}_{X})\tilde{\varepsilon}-\varepsilon^{\top}(\mathbb{X}\mathbb{X}^{\top}-\mathbb{D}_{X})\varepsilon\}$，$\tilde{R}_{2}=$

$n^{-2}\tilde{\varepsilon}^\top\widehat{\mathbb{Z}}(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}})^{-1}\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}(\mathbb{X}\mathbb{X}^\top-\mathbb{D}_X)\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}})^{-1}\widehat{\mathbb{Z}}^\top\tilde{\varepsilon}$，以及 $\tilde{R}_3 = n^{-2}\tilde{\varepsilon}^\top(\mathbb{X}\mathbb{X}^\top-\mathbb{D}_X)\times\widetilde{\mathbb{W}}\widehat{\mathbb{Z}}(\widehat{\mathbb{Z}}^\top\widetilde{\mathbb{W}}\widehat{\mathbb{Z}})^{-1}\widehat{\mathbb{Z}}^\top\tilde{\varepsilon}$。那么，如果我们能够证明以下四个结果，定理结论即可得出

$$(1)E(T_2^*) = 0, \mathrm{var}(T_2^*) = 2n^{-2}\mathrm{tr}(\Sigma_\gamma^2)\{1+o(1)\} \tag{D.1}$$

$$(2)\tilde{R}_1 = o_p\{n^{-1}\sqrt{p}\}, \tag{D.2}$$

$$(3)\tilde{R}_2 = o_p\{n^{-1}\sqrt{p}\}, \tag{D.3}$$

$$(4)\tilde{R}_3 = o_p\{n^{-1}\sqrt{p}\}. \tag{D.4}$$

这四个结论分别对应结论 (C.1)-(C.5)。接下来我们依次证明这四个结果。

Step 1. 通过直接计算，我们有 $E(T_2^*) = E(T_1^*) - n^{-2}E(\varepsilon^\top\mathbb{D}_X\varepsilon) = n^{-1}\mathrm{tr}(\Sigma_\gamma) - n^{-2}E(\sum_{i=1}^n\varepsilon_i^2X_i^\top X_i) = 0$。接着我们考虑方差 $\mathrm{var}(T_2^*) = n^{-4}E(T_2^{*2}) = E(\varepsilon^\top\mathbb{X}\mathbb{X}^\top\varepsilon)^2 + n^{-4}E(\varepsilon^\top\mathbb{D}_X\varepsilon)^2 - 2n^{-4}E(\varepsilon^\top\mathbb{X}\mathbb{X}^\top\mathbb{D}_X\varepsilon)$。根据定理 3 中的证明，我们有 $E(\varepsilon^\top\mathbb{X}\mathbb{X}^\top\varepsilon)^2 = nE\{\varepsilon^4(X_i^\top X_i)^2\} + n(n-1)\mathrm{tr}^2(\Sigma_\gamma) + 2n(n-1)\mathrm{tr}(\Sigma_\gamma^2)$。于是我们有

$$E(\varepsilon^\top\mathbb{D}_X\varepsilon)^2 = E\Big\{\Big(\sum_{i=1}^n\varepsilon_i^2X_i^\top X_i\Big)^2\Big\} = \sum_{i=1}^n E\Big\{\varepsilon_i^4\Big(X_i^\top X_i\Big)^2\Big\} + \sum_{i\neq j}E\Big(\varepsilon_i^2\varepsilon_j^2X_i^\top X_iX_j^\top X_j\Big)$$
$$= nE\Big\{\varepsilon_i^4\Big(X_i^\top X_i\Big)^2\Big\} + n(n-1)E^2\Big(\varepsilon_i^2X_i^\top X_i\Big) = nE\Big\{\varepsilon_i^4\Big(X_i^\top X_i\Big)^2\Big\} + n(n-1)\mathrm{tr}^2(\Sigma_\gamma).$$

通过直接计算，我们得到

$$E(\varepsilon^\top\mathbb{X}\mathbb{X}^\top\mathbb{D}_X\varepsilon) = E\Big\{\Big(\sum_{j_1,j_2}^n\varepsilon_{j_1}\varepsilon_{j_2}X_{j_1}^\top X_{j_2}\Big)\Big(\sum_{i=1}^n\varepsilon_i^2X_i^\top X_i\Big)\Big\}$$
$$= \sum_{i,j_1,j_2}E\Big(\varepsilon_i^2\varepsilon_{j_1}\varepsilon_{j_2}X_i^\top X_iX_{j_1}^\top X_{j_2}\Big) = nE\Big\{\varepsilon_i^4\Big(X_i^\top X_i\Big)^2\Big\} + \sum_{i\neq j}E\Big(\varepsilon_i^2\varepsilon_j^2X_i^\top X_iX_j^\top X_j\Big)$$
$$= nE\Big\{\varepsilon_i^4\Big(X_i^\top X_i\Big)^2\Big\} + n(n-1)\mathrm{tr}^2(\Sigma_\gamma).$$

因此，我们有 $\text{var}(T_2^*) = 2n^{-2}(1-n^{-1})\text{tr}(\Sigma_\gamma^2)$。

STEP 2. 通过与定理 3 中 STEP 3 类似的步骤，我们可以证明$\tilde{R}_1 = O_p(p/N^{1/2}) = o_p(n^{-1}p^{1/2})$。

STEP 3. 通过与定理 3 中 STEP 4 类似的步骤，我们可以证明$\tilde{R}_2 = o_p(n^{-1}p^{1/2})$。

STEP 4. 根据柯西-施瓦茨不等式，我们有 $\tilde{R}_3 = o_p(n^{-1}p^{1/2})$。至此，我们完成了定理的证明。

附录 E. 定理 5 的证明

根据定理 4 的证明，我们只需证明T_2^*的渐近正态性。接下来，我们利用鞅中心极限定理（Hall 和 Heyde，1980；Lan 等人，2014）验证这一结论。设 $\Delta_{n,i} = 2n^{-2}\sum_{j=1}^{i-1}\varepsilon_i\varepsilon_j X_i^\top X_j$ 和 $T_{n,r} = \sum_{i=2}^{r}\Delta_{n,i}$。接着，令 $\mathcal{F}_r = \sigma\{(\varepsilon_1, X_1^\top)^\top, \ldots, (\varepsilon_r, X_r^\top)^\top\}$ 为由 $(\varepsilon_j, X_j)^\top$ 生成的 σ-代数，其中 $1 \le j \le r$。于是我们有 $T_{n,r} \in \mathcal{F}_r$。注意，对于 $j \le q < i$，有 $E(\varepsilon_i\varepsilon_j X_i^\top X_j|\mathcal{F}_q) = E(\varepsilon_i X_i^\top)\varepsilon_j X_j = 0$；对于 $q < j \le i-1$，有$E(\varepsilon_i\varepsilon_j X_i^\top X_j|\mathcal{F}_q) = E(\varepsilon_i\varepsilon_j X_i^\top X_j) = E(\varepsilon_i X_i^\top)E(\varepsilon_j X_j) = 0$。因此，对于任意 $q < i$，有 $E(\Delta_{n,i}|\mathcal{F}_q) = 0$。由此，对于任意 $q < r$，我们有 $E(T_{n,r}|\mathcal{F}_q) = T_{n,q}$。因此，序列 $\{T_{n,r}, 2 \le r \le n\}$ 关于 $\mathcal{F}_r$ 构成一个鞅。令$V_{n,i} = E(\Delta_{n,i}^2|\mathcal{F}_{i-1})$ 且 $V_n = \sum_{i=2}^{n} V_{n,i}$。为应用鞅中心极限定理（Hall 和 Heyde，1980；Lan 等人，2014），我们还需证明：(1)$V_n/\{2n^{-2}\text{tr}(\Sigma_\gamma^2)\} \xrightarrow{p} 1$，以及 (2)

$$\frac{n^2}{2\text{tr}(\Sigma_\gamma^2)}\sum_{i=2}^{n} E\Big[\Delta_{n,i}^2 I\Big\{|\Delta_{n,i}| > \eta\sqrt{2}n^{-1}\text{tr}^{1/2}(\Sigma_\gamma^2)\Big\}\Big|\mathcal{F}_{i-1}\Big] \xrightarrow{p} 0, \tag{E.1}$$

对于一个 y $\eta > 0$。接下来，我们分别验证上述两个结论。对于 V_n，有

$$V_n = \sum_{i=2}^{n} E\Big(\Delta_{n,i}^2|\mathcal{F}_{i-1}\Big) = 4n^{-4}\sum_{i=2}^{n} E\Big(\sum_{1\le j_1, j_2\le i-1}\varepsilon_i^2\varepsilon_{j_1}\varepsilon_{j_2}X_{j_1}^\top X_i X_i^\top X_{j_2}\Big|\mathcal{F}_{i-1}\Big)$$

$$
\begin{aligned}
&= 4n^{-4}\sum_{i=2}^{n}\sum_{1\le j_1,j_2\le i-1}\varepsilon_{j_1}\varepsilon_{j_2}X_{j_1}^{\top}\Sigma_\gamma X_{j_2}\\
&= 4n^{-4}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\varepsilon_j^2 X_j^{\top}\Sigma_\gamma X_j + 8n^{-4}\sum_{i=2}^{n}\sum_{1\le j_1<j_2\le i-1}\varepsilon_{j_i}\varepsilon_{j_2}X_{j_1}^{\top}\Sigma_\gamma X_{j_2} = V_{n1}+V_{n2},
\end{aligned}
$$

其中 $V_{n1}=4n^{-4}\sum_{i=1}^{n-1}(n-i)\varepsilon_i^2 X_i^{\top}\Sigma_\gamma X_i$，且 $V_{n2}=8n^{-4}\sum_{i_1<i_2}^{n-1}(n-i_2)\varepsilon_{i_1}\varepsilon_{i_2}(X_{i_1}^{\top}\Sigma_\gamma X_{i_2})$。然后，我们分别研究 V_{n1} 和 V_{n2}。具体而言，有 $E(V_{n1})=4n^{-4}\sum_{i=1}^{n-1}(n-i)\mathrm{tr}(\Sigma_\gamma^2)$ $=2n^{-2}\mathrm{tr}(\Sigma_\gamma^2)\{1+o(1)\}$。按照定理1中的类似步骤，我们可以证明$\|X_i\|/\sqrt{p}$ 是次高斯的。因此，有 $E(X_i^{\top}X_i)^2=O(p^2)$。接下来，根据柯西-施瓦茨不等式和条件（C3），有 –

$$
\begin{aligned}
&\mathrm{var}(V_{n1}) = 16n^{-8}\sum_{i=1}^{n-1}(n-i)^2\mathrm{var}\Big(\varepsilon_i^2 X_i^{\top}\Sigma_\gamma X_i\Big)\\
&\le 16n^{-8}\sum_{i=1}^{n-1}(n-i)^2 C_{\max}^2 E\Big(X_i^{\top}X_i\Big)^2 = O(n^{-5}p^2) = o(n^{-4}p^2).
\end{aligned}
$$

因此，$\mathrm{var}(V_{n1}) = o\{\mathrm{var}^2(T_2^*)\}$。这意味着 $V_{n1}/\mathrm{var}(T_2^*)\xrightarrow{p}1$。接下来考虑 V_{n2}。可以验证 $E(V_{n2})=0$。于是我们有

$$
\begin{aligned}
\mathrm{var}(V_{n2}) &= 64n^{-8}\sum_{1\le i_1<i_2\le n-1}(n-i_2)^2 E\Big\{\varepsilon_{i_1}^2\varepsilon_{i_2}^2\Big(X_{i_1}^{\top}\Sigma_\gamma X_{i_2}\Big)^2\Big\}\\
&= 64n^{-8}\sum_{1\le i_1<i_2\le n-1}(n-i_2)^2\mathrm{tr}\Big\{E\Big(\varepsilon_{i_1}^2 X_{i_1}X_{i_1}^{\top}\Sigma_\gamma\Big)E\Big(\varepsilon_{i_2}^2 X_{i_2}X_{i_2}^{\top}\Sigma_\gamma\Big)\Big\}\\
&= 64n^{-8}\sum_{1\le i_1<i_2\le n-1}(n-i_2)^2\mathrm{tr}(\Sigma_\gamma^4)\le 64n^{-4}\mathrm{tr}(\Sigma_\gamma^4).
\end{aligned}
$$

根据条件（C3），我们知道 $\mathrm{tr}(\Sigma_\gamma^4)\le C_{\max}^4 p$。因此，我们有 $\mathrm{var}(V_{n2}) = O(n^{-4}p)$ $=o\{\mathrm{var}^2(T_2^*)\}$。接下来证明 $V_{n2}/\mathrm{var}(T_2^*)\xrightarrow{p}0$。

接下来，我们考虑 (E.1) 的第二个结论。对于任意 $\eta>0$，根据切比雪夫不等式，我们有

$$
\frac{n^2}{2\mathrm{tr}(\Sigma_\gamma^2)}\sum_{i=2}^{n}E\Big[\Delta_{n,i}^2 I\Big\{|\Delta_{n,i}|>\eta\sqrt{2}n^{-1}\mathrm{tr}^{1/2}(\Sigma_\gamma^2)\Big\}\Big|\mathcal{F}_{i-1}\Big]\le\frac{n^4}{4\eta^2\mathrm{tr}^2(\Sigma_\gamma^2)}\sum_{i=2}^{n}E(\Delta_{n,i}^4).
$$

因此, 如果我们能够证明 $\sum_{i=2}^{n} E(\Delta_{n,\,i}^{4}) = o(n^{-4}p^{2})$，则该结论成立。由此可得

$$\begin{aligned}\sum_{i=2}^{n} E(\Delta_{n,i}^{4}) &= 16n^{-8}\sum_{i=2}^{n}\sum_{1\le j_1,j_2,j_3,j_4\le i-1} E\Big(\varepsilon_i^4\varepsilon_{j_1}\varepsilon_{j_2}\varepsilon_{j_3}\varepsilon_{j_4}X_i^\top X_{j_1}X_i^\top X_{j_2}X_i^\top X_{j_3}X_i^\top X_{j_4}\Big)\\ &= 16n^{-8}\sum_{i=2}^{n}\sum_{j=1}^{i-1} E\Big\{\varepsilon_i^4\varepsilon_j^4\Big(X_i^\top X_j\Big)^4\Big\} + 96n^{-8}\sum_{i=2}^{n}\sum_{1\le j_1\ne j_2\le i-1} E\Big\{\varepsilon_i^4\varepsilon_{j_1}^2\varepsilon_{j_2}^2\Big(X_i^\top X_{j_1}\Big)^2\Big(X_iX_{j_2}\Big)^2\Big\}.\end{aligned}$$

根据条件（C4），我们有 $16n^{-8}\sum_{i=2}^{n}\sum_{j=1}^{i-1} E\{\varepsilon_i^4\varepsilon_j^4(X_i^\top X_j)^4\} \le 16n^{-8}\sum_{i=2}^{n}\sum_{j=1}^{i-1} E\big(X_i^\top X_j\big)^4 = O(n^{-6}p^2)$。类似地，根据柯西-施瓦茨不等式，我们有

$$\begin{aligned}&16n^{-8}\sum_{i=2}^{n}\sum_{1\le j_1\ne j_2\le i-1} E\Big\{\varepsilon_i^4\varepsilon_{j_1}^2\varepsilon_{j_2}^2\Big(X_i^\top X_{j_1}\Big)^2\Big(X_i^\top X_{j_2}\Big)^2\Big\}\\ \le\ & 8n^{-8}\sum_{i=2}^{n}\sum_{1\le j_1\ne j_2\le i-1}\left[E\Big\{\varepsilon_i^4\varepsilon_{j_1}^4\Big(X_i^\top X_{j_1}\Big)^4\Big\} + E\Big\{\varepsilon_i^4\varepsilon_{j_2}^4\Big(X_i^\top X_{j_2}\Big)^4\Big\}\right]\\ &\le 16n^{-5}E\Big\{\varepsilon_i^4\varepsilon_j^4\Big(X_i^\top X_j\Big)^4\Big\} = O\Big(n^{-5}p^2\Big) = o(n^{-4}p^2).\end{aligned}$$

因此，（E.1）的第二个结论成立。由此，我们验证了 T_2^* 的渐近正态性。然后，根据定理 4 和斯卢茨基定理，我们得出 $T_3 \xrightarrow{d} N(0,1)$。

附录F. 定理6的证明

通过简单计算，我们可以将 $\mathrm{tr}(\Sigma_\gamma^2)$ 改写为 $\mathrm{tr}(\Sigma_\gamma^2) = (1-n^{-1})(U_1+U_2)$，其中 $U_1 = \{n(n-1)\}^{-1}\sum_{i\ne j}\varepsilon_i^2\varepsilon_j^2(X_i^\top X_j)^2$，且 $U_2 = \{n(n-1)\}^{-1}\sum_{i\ne j}(\widehat{\varepsilon}_i^2\widehat{\varepsilon}_j^2 - \varepsilon_i^2\varepsilon_j^2)(X_i^\top X_j)^2$。注意，根据条件（C4），$\mathrm{tr}(\Sigma_\gamma^2) \le C_{\max}^2 p$。那么，如果我们能够证明（1） $U_1/\mathrm{tr}(\Sigma_\gamma^2) \xrightarrow{p} 1$ 以及（2） $U_2 = o_p(p)$，结论便成立。

Step 1. 我们首先评估 U_1。我们从其期望开始。根据不同样本之间的独立性，有 $E(U_1) = \{n(n-1)\}^{-1}\sum_{i\ne j}E\{\varepsilon_i^2\varepsilon_j^2(X_i^\top X_j)^2\} = \mathrm{tr}\{E(\varepsilon_i^2\varepsilon_j^2X_iX_i^\top X_jX_j^\top)\} = \mathrm{tr}\{E^2(\varepsilon_i^2X_iX_i^\top)\} = \mathrm{tr}(\Sigma_\gamma^2)$。因此， $E(U_1) = \mathrm{tr}(\Sigma_\gamma^2)$。然后我们评估 U_1的方差。为此，我们应用Zhong和Chen（2011）提出的高维U统计量的Hoeffding分解。令

$V_i = (\varepsilon_i, X_i^\top)^\top$ 对于 $i = 1, \ldots, n$，且 $h(V_i, V_j) = \varepsilon_i^2\varepsilon_j^2(X_i^\top X_j)^2$ 为核函数。于是 $h_1(V_i) = E\{h(V_i, V_j)|V_i\} = E\{\varepsilon_i^2\varepsilon_j^2(X_i^\top X_j)^2|V_i\} = \mathrm{tr}\{E(\varepsilon_i^2\varepsilon_j^2 X_i X_i^\top X_j X_j^\top|V_i)\} = \mathrm{tr}(\varepsilon_i^2 X_i X_i^\top \Sigma_\gamma) = \varepsilon_i^2 X_i^\top \Sigma_\gamma X_i$。根据条件（C3）以及 X_i 的次高斯性，我们有 $\zeta_1 = \mathrm{var}\{h_1(V_i)\} = \mathrm{var}(\varepsilon_i^2 X_i^\top \Sigma_\gamma X_i) \le E\{\varepsilon_i^4 (X_i^\top \Sigma_\gamma X_i)^2\} \le C_{\max}^2 E(X_i^\top X_i)^2 \le C_{\max}^2 p^2$。同时，根据条件（C4），我们知道 $\zeta_2 = \mathrm{var}\{h(V_i, V_j)\} = \mathrm{var}(\varepsilon_i^2\varepsilon_j^2(X_i^\top X_j)^2) \le E(X_i^\top X_j)^4 \le \tau_{\max} p^2$。因此，我们有 $\mathrm{var}(U_1) = 2(n-2)\{n(n-1)\}^{-1}\zeta_1 + \{n(n-1)\}^{-1}\zeta_2 = O(n^{-1}p^2)$。根据切比雪夫不等式，对于任意 $\eta > 0$，有 $P(|U_1/\mathrm{tr}(\Sigma_\gamma^2) - 1| > \eta) \le \mathrm{var}(U_1)/\{\mathrm{tr}^2(\Sigma_\gamma^2)\eta^2\} = O(n^{-1})$。由此我们证明了第一个结论。

Step 2. 我们考虑 U_2。令 $\phi_i = \psi_i \widehat{Z}_i^\top \widehat{\gamma} + (1-\psi_i) Z_i^\top \gamma$ 和 $\tilde{\phi}_i = \tilde{\psi}_i \widehat{Z}_i^\top \widehat{\gamma} + (1-\tilde{\psi}_i) Z_i^\top \gamma$，其中 ψ_i，$\tilde{\psi}_i \in (0, 1)$。根据中值定理和柯西-施瓦茨不等式，我们有

$$
\begin{aligned}
|U_2| &= \Big|\frac{1}{n(n-1)}\sum_{i\neq j}\Big(\widehat{\varepsilon}_i^2\widehat{\varepsilon}_j^2 - \varepsilon_i^2\widehat{\varepsilon}_j^2 + \varepsilon_i^2\widehat{\varepsilon}_j^2 - \varepsilon_i^2\varepsilon_j^2\Big)\Big(X_i^\top X_j\Big)^2\Big| \\
&\le \frac{1}{n(n-1)}\sum_{i\neq j}|\widehat{\varepsilon}_i^2 - \varepsilon_i^2|\widehat{\varepsilon}_j^2\Big(X_i^\top X_j\Big)^2 + \frac{1}{n(n-1)}\sum_{i\neq j}\varepsilon_i^2|\widehat{\varepsilon}_j^2 - \varepsilon_j^2|\Big(X_i^\top X_j\Big)^2 \\
&\le \frac{2}{n(n-1)}\sum_{i\neq j}\Big|\widehat{\varepsilon}_i^2 - \varepsilon_i^2\Big|\Big(X_i^\top X_j\Big)^2 \le \frac{4}{n(n-1)}\sum_{i\neq j}\Big|g\Big(\widehat{Z}_i^\top\widehat{\gamma}\Big) - g\Big(Z_i^\top\gamma\Big)\Big|\Big(X_i^\top X_j\Big)^2 \\
&+ \frac{2}{n(n-1)}\sum_{i\neq j}\Big|g^2\Big(\widehat{Z}_i^\top\widehat{\gamma}\Big) - g^2\Big(Z_i^\top\gamma\Big)\Big|\Big(X_i^\top X_j\Big)^2 \\
&= \frac{4}{n(n-1)}\sum_{i\neq j}\Big|g(\phi_i)\Big\{1-g(\phi_i)\Big\}\Big(\widehat{Z}_i^\top\widehat{\gamma} - Z_i^\top\gamma\Big)\Big|\Big(X_i^\top X_j\Big)^2 \\
&+ \frac{4}{n(n-1)}\sum_{i\neq j}\Big|g^2(\tilde{\phi}_i)\Big\{1-g(\tilde{\phi}_i)\Big\}\Big(\widehat{Z}_i^\top\widehat{\gamma} - Z_i^\top\gamma\Big)\Big|\Big(X_i^\top X_j\Big)^2 \\
&\le \frac{8}{n(n-1)}\sum_{i\neq j}\Big|\Big(\widehat{Z}_i - Z_i\Big)^\top\widehat{\gamma}\Big|\Big(X_i^\top X_j\Big)^2 + \frac{8}{n(n-1)}\sum_{i\neq j}\Big|Z_i^\top\Big(\widehat{\gamma}-\gamma\Big)\Big|\Big(X_i^\top X_j\Big)^2 \\
&\le 8\Big\{\|\widehat{\gamma}\|\Big(n^{-1}\sum_{i=1}^n\|\widehat{Z}_i - Z_i\|^2\Big)^{1/2} + \|\widehat{\gamma}-\gamma\|\Big(n^{-1}\sum_{i=1}^n\|Z_i\|^2\Big)^{1/2}\Big\} \\
&\times \Big\{\frac{1}{n(n-1)}\sum_{i\neq j}\Big(X_i^\top X_j\Big)^4\Big\}^{1/2}
\end{aligned}
$$

根据定理 2 证明中的 Step 1.1，我们知道 $n^{-1}\sum_{i=1}^{n}\|\widehat{Z}_i - Z_i\|^2 = O_p(p/N)$。回顾一下，我们已在定理 2 中证明了 $\|\widehat{\gamma} - \gamma\| = O_p(n^{-1/2})$。于是有 $\|\widehat{\gamma}\| = O_p(1)$。根据大数定律，我们知道 $n^{-1}\sum_{i=1}^{n}\|Z_i\|^2 \xrightarrow{p} E\|Z_i\|^2$。因此，有 $n^{-1}\sum_{i=1}^{n}\|Z_i\|^2 = O_p(1)$。根据条件（C4），有 $E[\{n(n-1)\}^{-1}\sum_{i\neq j}(X_i^\top X_j)^4] = E(X_i^\top X_j)^4 \leq \tau_{\max}p^2$。因此，有$\{n(n-1)\}^{-1}\sum_{i\neq j}(X_i^\top X_j)^4 = O_p(p^2)$。结合以上结果，我们有 $U_2 = \{O_p(\sqrt{p/N}) + O_p(n^{-1/2})\}O_p(p) = o_p(p)$。至此，定理得证。