稀疏网络数据联合潜在空间建模的协变量选择

作者：E.G. Crenshaw

哈佛大学陈曾熙公共卫生学院生物统计学系，马萨诸塞州波士顿市亨廷顿大道*677*号，邮编*02115*。emma crenshaw@g.harvard.edu

–

张宇华哈佛大学陈曾熙公共卫生学院生物统计学系，马萨诸塞州波士顿市亨廷顿大道*677*号，邮编*02115*。yuhuazhang@hsph.harvard.edu

Jukka-Pekka Onnela

哈佛大学陈曾熙公共卫生学院生物统计学系，马萨诸塞州波士顿市亨廷顿大道*677*号，邮编*02115*。onnela@hsph.harvard.edu

arXiv:2602.04682v1 [stat.ME] 4 Feb 2026

摘要

网络数据在社会科学和传染病流行病学中日益普遍。分析通常将网络结构与节点级协变量联系起来，但现有方法在处理稀疏网络和高维节点特征时表现不佳。我们提出一种适用于含高维二元节点协变量的稀疏网络的联合潜在空间建模框架，该框架在进行协变量选择的同时，考虑估计潜在位置的不确定性。基于通过共享潜在位置耦合边与节点变量的联合潜在空间模型，我们引入组套索筛选步骤，并加入测量误差感知稳定项，以减轻将估计的潜在位置用作预测变量所导致的偏差。我们针对协变量部分建立了预测误差率，既考虑将潜在位置视为已观测值的情况，也考虑潜在位置存在有界误差估计的情况；在对 q 个协变量和 n 个节点进行一致控制的条件下，该误差率的阶为 $O(\log\, q/n)$，另加一项由潜在位置估计误差导致的误差。我们的方法应对了三个挑战：（1）纳入孤立节点的信息；这类节点在稀疏网络中很常见，却常被忽略；（2）从高维空间中筛选相关协变量；以及（3）考虑估计的潜在位置所存在的不确定性。模拟结果表明，随着协变量稀疏性增加，该方法的预测性能仍然稳定，而朴素方法的表现则会下降。我们以 75 个印度村庄的家庭社交网络为例，说明该方法如何支持高效的研究设计：模拟试点研究筛选了大量协变量组合，在不牺牲网络预测准确性的前提下，大幅减少后续所需的数据收集量。

关键词：潜在空间模型；网络建模；稀疏网络；组套索；测量误差；

1. 引言

网络数据广泛出现于社会科学和传染病流行病学中，其科学目标通常是将互动模式与节点级属性联系起来。除图数据固有的复杂依赖关系外，还有两个实际特征使分析更加复杂：边稀疏性和高维节点协变量。首先，网络可能是稀疏的，存在许多孤立节点或低度节点（例如接触网络和性关系网络），因此单条连边携带不成比例的信息，而相当一部分节点几乎不提供边信息（Weiss et al., 2020）。其次，研究者经常收集大量节点测量指标，其中只有一部分可能与塑造该网络的社会或流行病学机制相关。在此类情形下，不加区分地纳入协变量的方法可能会稀释信号、增加计算负担，并妨碍可解释性。

潜在空间模型提供了一种灵活且易于解释的框架，用于表示网络结构。该框架假设节点位于一个低维潜在空间中，且关系形成的倾向由这些潜在位置之间的相似性决定（Hoff 等人，2002；Hoff，2005）。在给定潜在位置的条件下，关系被建模为相互独立，从而能够进行基于似然的估计和推断，同时通过几何结构捕捉同质性、传递性和社区结构等显著的网络特征（Hoff 等人，2002；Smith 等人，2019；Bogu˜n´a 等人，2021）。大量文献研究了潜在位置模型及相关潜在结构随机图模型的估计和推断（Athreya 等人，2018、2021；Hoff 等人，2002；Hoff，2005）。这些模型日益被用作下游任务的第一阶段表示，因此，可靠地估计潜在位置对于后续推断至关重要（Athreya 等人，2021）。

在许多实证研究中，除了网络数据外，还会观测到丰富的节点协变量。一种常见策略是通过二元相似性或边级协变量纳入协变量（Hoff 等人，2002；Hoff，2005；Ma 等，2020；Krivitsky 等人，2009；Ma 和 Kundu，2022），但这通常需要选择一种相似性度量，而且对于孤立节点而言，信息可能较少，因为基于边的似然贡献会消失。利用边协变量拟合潜在空间模型的方法可以提高可扩展性和预测性能，但其设计目的并非从高维节点属性中选择节点协变量，也无法直接解决用作预测变量的潜在位置本身也需要估计这一问题（Ma 等，2020）。近期，Zhang 等人（2022）提出了一种联合潜在空间框架，其中网络和高维节点变量均以共享的潜在位置为条件生成。这种联合建模方式对稀疏网络尤其具有吸引力，因为即使节点处于孤立状态，它也能利用节点变量，从而使所有节点都能参与估计。

尽管已有这些进展，实践中仍存在两个缺口。首先，仅靠联合建模并不能解决这样一种常见情形：虽然收集了 q 节点变量，但只有一小部分与网络生成潜在结构存在有意义的关联。在高维情形下，保留大量无关协变量会降低可解释性，并可能损害节点协变量部分的预测性能。其次，任何将估计得到的潜在˜位置 Z 视为固定预测变量的选择程序都有夸大证据的风险：潜在位置是由同一数据估计得到的，在稀疏网络中可能存在噪声，从而给基于 $\hat{Z}$构建的协变量模型带来变量误差问题。

我们通过引入一种面向联合潜在空间模型的协变量选择程序来应对这些挑战，该程序将组稀疏正则化与对潜在位置估计误差的校正相结合。具体而言，我们将每个节点协变量视为一个组，对应于将该协变量与 k维潜在空间相联系的 k 系数，并在高维逻辑回归中应用组套索惩罚进行选择（Meier 等人，2008；Negahban 等人，2012）。为减轻使用估计潜在位置所带来的影响，我们在组套索目标函数中加入一个起稳定作用的岭型项，其动机来自面向高维广义线性模型的测量误差感知正则化（Sørensen 等人，2018，2015）。我们的整体程序是一个两阶段工作流程：（i）按照 Zhang 等人（2022）的方法，使用全部协变量进行联合潜在空间估计；（ii）进行组套索筛选（并作测量误差校正），随后在所选协变量子集上重新拟合联合模型。

本文有三项贡献。首先，我们提出了一种具有实际应用动机、针对稀疏网络和孤立节点量身设计的联合潜在空间建模协变量选择策略。其次，我们将测量误差感知惩罚方法调整到这样一种情形：潜在位置经过估计后，被用作高维节点属性的预测变量（Sørensen 等人，2018，2015）。第三，我们为协变量部分提供了预测误差保证，既包括将潜在位置视为已观测值的预言机情形，也包括潜在位置估计存在有界误差的情形；在对 q 协变量和 n 节点进行一致控制的条件下，我们在标准高维条件下获得了阶为 $O(\log q/n)$ 的收敛速率（Negahban 等人，2012；Meier 等人，2008）。最后，我们展示了一个受研究设计启发的应用案例：通过模拟试点研究，在全面开展网络数据收集之前确定一组精简的协变量组合，并利用来自 75 个印度村庄的社会网络加以演示（Banerjee 等人，2013）。

2. 模型

2.1. 模型设定与符号说明

设 $G = (V, E)$ 为包含 n 个节点的无向网络，其对称邻接矩阵为 $A \in \mathbb{R}^{n\times n}$，满足：若边存在，则 $A_{ii'}, i\neq i' = 1$；否则为 0。对于每个节点 i，我们观测到 q 个二元协变量 $Y \in \{0, 1\}^{n\times q}$。我们对网络数据和节点协变量数据采用联合潜在空间模型，其中每个节点 i 都与一个未观测到的潜在位置 $Z_i \in \mathbb{R}^k$ 相关联，该位置位于一个 k 维潜在空间中。我们假设网络边和节点协变量均

是在给定这些潜在位置的条件下生成的，从而在网络拓扑与节点属性之间形成自然的依赖结构。

我们假设，在给定节点的潜在位置后，节点之间的边相互独立（Ma 等，2020；Zhang 等人，2022；Hoff 等人，2002），因此

$$A_{ii'} = A_{i'i} \sim \text{Bernoulli}(P^A_{ii'}), \quad \text{logit}P^A_{ii'} = \theta^A_{ii'} = \alpha_i + \alpha_{i'} + Z_i^\top Z_{i'},$$

其中，$P^A_{ii'}$ 是节点 i 和 i' 之间存在一条边的概率，$\alpha = (\alpha_1, \ldots\ldots, \alpha_n)$是节点特有的“社交性”因子，用于刻画节点在不受其潜在位置影响的情况下形成边的倾向。该表述允许存在同质性（即潜在位置相似的节点更有可能相连）以及度异质性。

对于 q 个协变量，我们假设，在给定 Z的条件下，每个协变量 Y对于 $j \in \{1, \ldots\ldots$ 服从相互独立的逻辑回归

$$Y_{ij} \sim \text{Bernoulli}(P^Y_{ij}), \quad \text{logit}P^Y_{ij} = \theta^Y_{ij} = \gamma_j + Z_i\beta_{\cdot j},$$

其中，P^Y_{ij} 表示节点 i 的协变量 j 等于 1 的概率，$\gamma \in \mathbb{R}^q$ 是截距项向量，$\beta \in \mathbb{R}^{k\times q}$ 是将潜在位置与节点协变量联系起来的回归系数。

关键在于，我们假设 β 具有组稀疏性：β的许多列为零，表明相应的协变量与网络生成潜在空间无关。我们将 β的活跃集定义为$S = \{ j \in \{1、\ldots\ldots、 q\}: \|\beta_{\cdot j}\|_2 > 0\}$，其基数为 $s = |S|$。这一假设反映出，尽管可能收集了许多协变量，但只有其中一部分协变量与网络结构真正相关。

2.2. *Estimation*

我们的估计策略分为两个阶段：（1）使用完整数据对所有参数进行初始联合估计；（2）进行协变量选择，随后仅使用选定的协变量进行改进估计。

初始估计

为保证可识别性，我们通过减去行均值对潜变量进行中心化，并假设 $\frac{1}{n}Z^\top Z$ 是对角矩阵而非单位矩阵。因此，Z 在正交变换下可识别。

联合估计目标结合了网络数据和协变量数据的负对数似然，使得$L(Z, \alpha, \gamma, \beta) = L_A + \Lambda L_Y$，其中：

$$L_A = -\log P(A|Z,\alpha) = \sum_{1\le i<i'\le n} \{A_{ii'}\theta^A_{ii'} - \log(1+\exp(\theta^A_{ii'}))\},$$

$$L_Y = -\log P(Y|\gamma,\beta) = \sum_{1\le i\le n, 1\le j\le q} \{Y_{ij}\theta^Y_{ij} - \log(1+\exp(\theta^Y_{ij}))\}.$$

权重参数 $\Lambda \ge 0$ 用于平衡网络结构和协变量所提供的信息，是一个超参数。实际应用中，我们采用较小的值 0.1，以便在估计过程中赋予网络信息更高的重要性。投影梯度下降可使用 ADAM（Kingma 和 Ba，2015）或 Adagrad（Duchi 等人，2011）执行。

算法 1 概述了该算法。我们采用无信息初始化：$Z^0 \sim N(0, 1)$、$\alpha^0 \sim \text{Uniform}(-1, 1)$，以及将 Y和 Z^0 用于逻辑回归所得的输出，作为 β^0和 γ^0。此外，我们还需要设置若干超参数。每次投影梯度下降迭代中，Z 和 α各自有对应的学习率。按照 Zhang 等人（2022）的建议，我们令 $\eta_Z = \eta_t/\|Z^{t-1}\|_F^2$和 $\eta_\alpha = \eta_t/(2n)$，具体取值取决于步长对应的学习率 η_t；该学习率可通过余弦退火进行指数衰减。考虑到指数衰减，我们建议将初始 η_t 设得相对较大。如果 A和 Y的平均对数损失在连续 500 次迭代中的变化不超过 10^{-6}，则允许算法提前停止。

协变量选择

给定初始估计值 $\hat{Z}$、$\hat{\alpha}$、$\hat{\beta}$ 和 $\hat{\gamma}$，每个列 $\beta_{\cdot j}$ 都通过组套索作为一个大小为 k 的单独组进行估计，从而将 $\hat{Z}$ 中的估计误差纳入考量。假设给定 Z 时，每个 Y_j 相互独立，那么这就转化为 q 个相互独立的逻辑回归。将每一列视为一个单独的组，可确保对每个回归系数的估计值进行惩罚，从而在协变量层面诱导稀疏性。

当潜在位置固定且已知时，对于每个 q 协变量，其组套索估计量为

$$\widehat{\beta}_{\cdot j} = \arg\min_{\beta_{\cdot j}} \left\{ L_{Y_j}(\beta) + \lambda\sqrt{k}\|\beta_{\cdot j}\|_2 \right\},$$

然而，由于必须估计 Z，我们对测量误差进行校正：

$$\widehat{\beta}_{\cdot j} = \arg\min_{\beta_{\cdot j}} \left\{ L_{Y_j}(\beta) + \lambda\sqrt{k}\|\beta_{\cdot j}\|_2 + \delta\sqrt{k}\|\beta_{\cdot j}\|_2^2 \right\},$$

正如 Sørensen 等人（2018）所示，这个额外的岭型惩罚能在存在测量误差的情况下稳定选择过程。对于适当的阈值 τ，该选择过程可识别活跃集 $\hat{S} = \{ j \in \{1, \ldots\ldots, q\} : \|\widehat{\beta}_{\cdot j}\| > \tau\}$。最后，我们仅使用 $\hat{S}$ 中的协变量重新估计所有参数，从而得到更精细的估计，并利用协变量选择带来的信噪比提升。

算法的摘要版本见算法 2。除了算法 1 中使用的超参数外，我们还指定 τ 为一个非常接近 0 的数值、一组候选 Lasso 惩罚参数 λ，以及一组候选 δ 惩罚项。我们将候选 λ 值设为一组对数线性点，包括 $\sqrt{k/n}$；将候选 δ 值设为一组介于 0 和 0.5 之间的点，具体遵循 (Sørensen 等人，2018) 的方法。最终的 λ 值选为使赤池信息准则最小、同时仍保留部分变量的值，以兼顾简约性与避免变量选择过于激进。鉴于这是一个非凸优化问题，拟合算法的结果取自估计步骤达到 500 次且估计值稳定后（变化不超过 10^{-6}），或取逐参数损失最佳的结果；逐参数损失计算为 $\frac{1}{n(n-1)}$ 对数损失(A) + $\Lambda \frac{1}{nq}$ 对数损失(Y)。对于所选的 λ，δ 的取值选为使 Y 的对数损失最小的值。

算法 1。联合潜在空间估计器

输入： 网络 $A \in \mathbb{R}^{n\times n}$；节点协变量 $Y \in \mathbb{R}^{n\times q}$；潜在维度 k；初始值 $(Z^0, \alpha^0, \beta^0, \gamma^0)$；加权参数 Λ；全局步长 η；最大迭代次数 T。

对于 $t = 1, \ldots, T$，执行：计算余弦退火步长 η_t 并设置 $\eta_z = \eta_t/\|Z^{t-1}\|_F^2$，$\eta_\alpha = \eta_t/(2n)$；$(Z, \alpha$ 步)：使用步长 (η_z, η_α) 对 (Z, α) 执行一步优化；$Z^t \leftarrow JZ^t$；$\alpha^t \leftarrow \alpha^t - \bar{\alpha}^t \mathbf{1}_n$；对于 $j = 1, \ldots, q$，执行：对 Z^t 进行 $Y_{\cdot j}$ 的逻辑回归拟合，得到 $(\gamma_j^t, \beta_{\cdot j}^t)$；$\widehat{\ }$ $\widehat{\ }$输出：估计值 $(Z, \widehat{\alpha}, \beta, \widehat{\gamma})$。

算法 2。联合潜在空间估计器的 Lasso 与测量误差感知的细化

输入： 网络邻接矩阵 $A \in \mathbb{R}^{n\times n}$；节点协变量 $Y \in \mathbb{R}^{n\times q}$；潜在维度 k；由算法 1 得到的估计量 $Z^{(1)}$；加权参数 Λ；全局步长 η；容差水平 τ；执行套索回归的迭代次数 T；候选 Lasso 惩罚参数网格 λ；候选 δ 惩罚项网格 Δ。

对每个 $\lambda_\ell \in \lambda$ 执行

通过按列进行组套索逻辑回归计算 $(\gamma^{\text{lasso}}(\lambda_\ell), \beta^{\text{lasso}}(\lambda_\ell))$；使用 $\gamma^{\text{lasso}}(\lambda_\ell)$、$\beta^{\text{lasso}}(\lambda_\ell)$，计算 $\bar{Y}$，$\mathrm{l}_Y(\lambda_\ell)$ 的平均对数损失；$\lambda^* = \arg\min_{\lambda_\ell} \mathrm{l}_Y(\lambda_\ell)$（假设在 λ^* 处仍有变量被选中）；在 λ^* 处计算 $(\gamma^{\text{lasso}}, \beta^{\text{lasso}})$；移除含有 $\|\beta_{\cdot j}^{\text{lasso}}\|_2 < \tau$ 的列 j，得到 $Y_{\text{keep}}, \beta_{\text{keep}}, \gamma_{\text{keep}}$；使用 $Z^{(1)}$ 和 $(\gamma_{\text{keep}}, \beta_{\text{keep}})$，对 $\delta \in \Delta$ 执行迭代重加权过程（参见 Sørensen 等人（2018））；选择使 Y_{keep} 的平均对数损失最小的 $(\gamma^{\text{GMUL}}, \beta^{\text{GMUL}})$；从 $(Z(1), a^{(1)}, \beta^{\text{GMUL}}, \gamma^{\text{GMUL}})$ 开始，仅使用 Y_{keep}，重新运行算法 1 共 T 次迭代，$\widehat{\ }$ $\widehat{\ }$得到 $(Z, \widehat{a}, \beta, \widehat{\gamma})$；$\widehat{\ }$ $\widehat{\ }$输出：估计值 $(Z, \widehat{a}, \beta, \widehat{\gamma})$。

3. 理论结果

我们在两种情形下阐明估计量的理论性质：将潜在位置视为无误差观测值（第 3.1 节），以及考虑估计误差（第 3.2 节）。我们陈述必要的假设，并给出每种情形下的预测误差率。

3.1. *Perfectly observed Z*

首先，我们考虑理想化的情形：Z被完美观测到（已知），这有助于深入了解我们选择程序的行为。假设各协变量相互独立，该问题便简化为逻辑回归的组套索。组套索估计量为 $\widehat{\beta}_{\cdot j} \in \arg\min_{\beta_j \in \mathbb{R}^k} \left\{ L_{Y_j}(\beta_{\cdot j}) + \tilde{\lambda}\sqrt{k}\,\|\beta_{\cdot j}\|_2 \right\}$。

~由于 k 对所有 β均为常数，自由度项可以并入 λ，从而得到熟悉的逻辑回归组套索：

$$\widehat{\beta}_{\cdot j} \in \arg\min_{\beta_j \in \mathbb{R}^k} \left\{ L_{Y_j}(\beta_{\cdot j}) + \lambda \|\beta_{\cdot j}\|_2 \right\}.$$

我们采用 Meier 等人（2008）的框架，并针对我们的可辨识性约束作了修改。

假设 *1*（有界线性预测子）。存在 $M_1 > 0$，使得 $-M_1 < \Theta^A_{ii'} < M_1$ 对 $1 \le i,\ i' \le n$，并且存在 $M_2 > 0$，使得 $-M_2 < \Theta^Y_j < M_2$ 对 $1 \le j \le q$。

假设 *2*（稀疏分组结构）。每一列 $\beta_{\cdot j}$均视为一个大小为 k的组，真正活跃的协变量数量 $s = |\{ j\colon \|\beta_{\cdot j}\|_2 > 0\}|$满足 $s = o(n/\log q)$。

假设 *3*（设计归一化）。经验协方差 $D = \frac{1}{n} Z^\top Z$ 是对角的（但不是单位矩阵），其特征值 $0 < \nu_{\min}(D) \le \nu_{\max}(D) < \infty$。

假设 *4*（受限强凸性）。设估计的回归系数与真实值之间的误差为 $\widehat{\beta}_{\cdot j} - \beta_{\cdot j} = \Delta_j$。符号 $\cdot,\cdot$ 表示内积。对于每个 j，存在 $\kappa > 0$，使得对于所有 $\Delta_j \in \mathbb{R}^k$，

$$\frac{1}{n}\left(L_{Y_j}(\beta_{\cdot j} + \Delta_j; Z) - L_{Y_j}(\beta_{\cdot j}; Z) - \langle \nabla L_{Y_j}(\beta_{\cdot j}; Z), \Delta_j \rangle \right) \ge \frac{\kappa}{2}\|\Delta_j\|_2^2.$$

假设 *5*（得分集中性）。对于每个 j，当μ_{ij}（Z）＝逻辑斯蒂函数$^{-1}$（$\gamma_j + Z_i^\top \beta_{\cdot j}$）时，有

$$\left\| \frac{1}{n} Z^\top \{ Y_{\cdot j} - \mu_j(Z) \} \right\|_2 \le C_0 \sqrt{\frac{\log q}{n}}$$

以高概率成立。

假设 A1 与 Zhang 等人（2022）所使用的假设相同；假设 A3 提高了 Z 的可识别性。假设 A4 是高维 M 估计量的标准条件，可确保损失函数具有强凸性。逻辑损失二阶可微；因此，该条件等价于海森矩阵 $\nabla^2 L_{Y_j}(\beta_{\cdot j},\ Z)$ 的特征值具有下界。这一点已由 Negahban 等人（2012）和 Meier 等人（2008）充分确立。假设 A5 源自标准的次指数集中（Meier 等人，2008）。

Define the prediction error $d^2(\widehat{\eta}_j, \eta_j) = \frac{1}{n}\|Z^\top(\widehat{\beta}_{\cdot j} - \beta_{\cdot j})\|_F^2, \eta_j = Z\beta_{\cdot j}.$

定理 1（预测误差率，完全观测的 Z）。在假设 *A1–A5* 下：

(a) 对任意固定的 $j \in \{1, \ldots, q\}$，且 $\lambda_j = C\sqrt{k/n}$，其中某个常数 C：

$$d^2(\widehat{\eta}_{\cdot j}, \eta_{\cdot j}) = O_p\left(\frac{k}{n}\right).$$

(b) 对所有 $j \in \{1, \ldots, q\}$ 进行一致控制，且 $\lambda = C\sqrt{\log q/n}$，其中某个常数 C：

$$\max_{j=1,\ldots,q} d^2(\widehat{\eta}_{\cdot j}, \eta_{\cdot j}) = O_p\left(\frac{\log q}{n}\right).$$

Meier 等人（2008）指出，$d^2(\hat{\eta},\ \eta) = O_p(\underline{\log}_n{}^G)$，其中 G是组的数量。对于每次单独的优化，组大小固定为 1，因此我们得到参数收敛速率 $O_p(1/n)$。然而，要对 q列进行一致控制，收敛速率中还需要额外的 $\log q$因子。证明见附录。

3.2. 考虑含测量误差的 Z

在实践中，必须估计潜在位置，这会引入加性测量误差：$\hat{Z} = Z + U$，其中 Z 是预言机潜在位置，U 是某个未知的估计误差（Sørensen 等人，2018，2015）。针对单个组的估计感知组套索估计量为：

$$\widehat{\beta}_{\cdot j} \in \arg\min_{\beta_{\cdot j} \in \mathbb{R}^k} \left\{ L_{Y_j}(\beta_{\cdot j}) + \lambda \|\beta_{\cdot j}\|_2 + \frac{\delta}{2} \|\beta_{\cdot j}\|_2^2 \right\}. \tag{1}$$

令 $\delta = 0$ 即可从简化情形中恢复忽略测量误差的组套索。

假设 *6*（有界测量误差）。假设联合估计阶段的估计误差满足 $\|U\|_\infty \le \delta_\infty$。

注意，Zhang 等人（2022）指出 $\frac{1}{n}\|\widehat{Z} - Z\|_F = O_p$（$n^{1/2}$），这使我们可以假设 $\delta_\infty = O$（$n^{-1/2}$）。因此，假设测量误差有界并非不合理。对于 U 的分布，没有作出其他假设。

定理 2（预测误差率，估计的 $\hat{Z}$）。在假设 *A1–A6*：

(a) 对于任意固定的 $j \in \{1, \ldots, q\}$，在调优参数 $\lambda_j = C_1\sqrt{k/n}$ 和 $\overline{\delta} = C_2\delta_\infty$ 下，其中 C_1、C_2 为正数常量：

$$\frac{1}{n}\|\hat{Z}(\hat{\beta}_{\cdot j} - \beta_{\cdot j})\|_2^2 = O_p\left(\frac{1}{n} + \delta_\infty^2\right).$$

(b) 为了对所有 $j \in \{1, \ldots, q\}$进行一致控制，在调优参数 $\lambda = C_1\sqrt{\log q/n}$ 和 $\overline{\delta} = C_2\delta_\infty$ 下：

$$\max_{j=1,\ldots,q} \frac{1}{n}\|\hat{Z}(\hat{\beta}_{\cdot j} - \beta_{\cdot j})\|_2^2 = O_p\left(\frac{\log q}{n} + \delta_\infty^2\right).$$

当来自第一阶段估计的 $\delta_\infty = O(n^{-1/2})$ 时，两种速率均降至各自的统计误差率。

推论 1： 当 $\delta = O(n^{-1/2})$ 时，测量误差项被统计误差所主导，我们达到与 Z 被完全观测时相同的速率。

4. 模拟

我们通过大量模拟评估我们的方法，比较三种方法：(1) 忽略协变量的仅基于网络的估计，(2) 不进行变量选择的联合估计（Zhang 等人，2022），以及 (3) 我们提出的带有协变量选择的方法。

4.1. *Simulation Design*

我们为 $n = 200$ 个节点、$q = 25$ 个节点协变量和 $k = 2$ 个潜在维度生成了模拟数据。为生成网络和节点协变量值，首先将节点划分为 k 个子集，其中心为 $\mu_k \sim$Uniform(−1, 1)。随后生成潜在位置 $Z \in \mathbb{R}^{n\times k}$，使得对于子集 j 中的节点 i，$Z_i \sim N(\mu_j, 1)$。这在网络中模拟了一定程度的同质性。对 Z 进行变换，使其以 $Z = JZ$ 为中心并通过 $\|ZZ^\top\|_F = n$ 进行标准化，同时具有对角、非单位的经验协方差。对于活跃分量，生成 β 值为 $\beta \sim N$(1，0。1)，其他情况下则设为 0。最后，根据上述方程生成 A 和 Y。通过将矩阵的下三角设置为等于上三角，强制 A 为对称矩阵。通过调整节点社会性参数 $\alpha = (\alpha_1、\ldots\ldots、\alpha_n)$，生成了两个密度水平的网络。$\alpha_i \sim$Uniform(−1，−0。5) 的分布生成平均密度为 0.37 的网络（下文称为较不稀疏网络情景），而 $\alpha_i \sim$Uniform(−2，−1) 生成平均密度为 0.013 的网络（下文称为稀疏网络情景）。我们在模拟过程中改变协变量稀疏度，每种配置重复 30 次。

4.2. 模拟结果

图1展示了在不同协变量稀疏性水平下，采用四种不同方法对30个相互独立的模拟网络进行建模的结果：仅使用网络数据、纳入网络数据但不调整稀疏性（Zhang 等人，2022）、调整稀疏性（套索回归），以及同时调整稀疏性和测量误差（meLasso）。我们使用网络信息 A and Y 的曲线下面积（AUC）指标，并对网络中的所有节点取平均，以评估模型性能。值得注意的是，结果以模型中纳入的所有协变量的平均 AUC 呈现；如果考虑稀疏性的方法剔除了某个协变量，则该协变量不计入最终的平均 AUC。结果见表3。

模拟结果呈现在图1和表3中。对于较不稀疏网络和稀疏网络，所有方法在预测网络方面的模型性能对节点协变量的稀疏性均具有稳健性，可以看出

在图1的A和B面板中。鉴于我们使用的网络与协变量信息平衡参数较小，为 $\Lambda = 0.1$，这一点在预期之中；估计过程对网络信息的加权要强得多，因此来自协变量的信息对刻画网络边的潜在空间位置估计影响有限。此外，在比较稀疏网络与较不稀疏网络的结果时，当底层网络更稠密时，模型能更好地捕捉网络信息，从而使网络信息的曲线下面积更接近0.65，且不受建模方法影响，而稀疏网络则更接近0.57。

在对节点协变量建模时，模型性能同时受到网络稀疏性和节点协变量中噪声量的影响。图1的D面板表明，对于较不稀疏网络，潜在位置得到了足够好的建模，因此平均曲线下面积的主要决定因素是建模方法中所纳入的、实际与潜在空间相关的协变量数量。当噪声协变量数量为0时，所有模型平均都能很好地预测节点协变量。然而，随着噪声协变量数量增加，不考虑节点协变量稀疏性的方法性能会下降，因为曲线下面积仍在对与潜在空间无关的协变量进行平均。不过，套索回归和meLasso方法能够去除约86%的噪声协变量（表2），从而获得高得多的平均曲线下面积。

如果网络是稀疏的，在联合估计过程中纳入节点协变量信息的益处会更加明显（图1中的面板C）。即使所有协变量都与潜在空间相关，仅使用网络数据的建模方法的表现也明显差于纳入协变量信息的其他方法。随着噪声协变量数量的增加，除非我们考虑潜在的稀疏性，否则在估计过程中纳入协变量信息的益处会下降。一旦大多数节点协变量为噪声，我们可以看到套索回归和meLasso方法比不针对稀疏性进行调整的方法更为稳健，尽管结果变异很大。这在很大程度上是因为当网络信息稀疏时，套索回归过程难以识别噪声协变量；当25个节点协变量中有20个为噪声时，套索回归平均只能剔除12%的噪声协变量，而meLasso平均只能剔除56%（表2）。

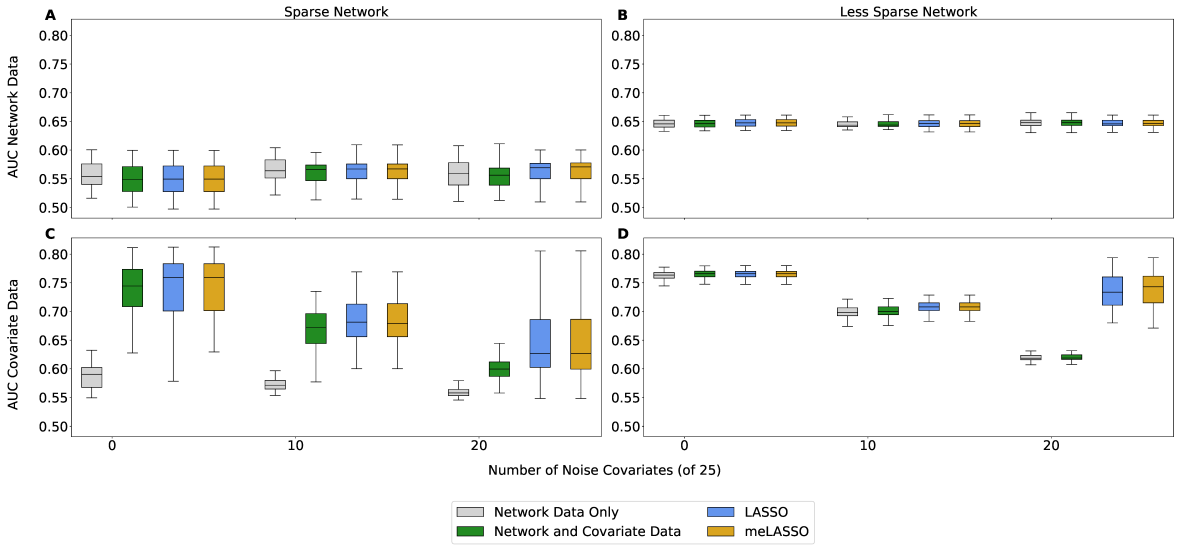

图1。**模拟中的模型性能**。箱线图表示30个独立模拟的、包含 $n = 200$ 个节点的网络所得结果。稀疏网络的平均密度为0.013（平均度为 = 2.6）；较不稀疏网络的平均密度为0.37（平均度为74）。
A和B面板分别显示稀疏网络和较不稀疏网络的网络数据曲线下面积。C和D面板分别显示稀疏网络和较不稀疏网络的节点协变量数据曲线下面积。

表1。 在不同数量的噪声协变量下，各方法对 $q = 25$ 节点协变量的预测性能（ $n = 200$个节点）。

方法	0个噪声协变量 均值（标准差）	10个噪声协变量 均值（标准差）	20个噪声协变量 均值（标准差）
较不稀疏网络（密度 = 0.37）			
仅网络数据	0.76（0.01）	0.70 (0.01)	0.62 (0.01)
网络和协变量数据	0.77 (0.01)	0.70 (0.01)	0.62 (0.01)
套索回归	0.77 (0.01)	0.71 (0.02)	0.73 (0.04)
meLasso	0.77 (0.01)	0.71 (0.01)	0.74 (0.04)
稀疏网络（密度 = 0.013）			
仅网络数据	0.59 (0.02)	0.57 (0.01)	0.56 (0.01)
网络和协变量数据	0.73 (0.06)	0.67 (0.04)	0.60 (0.02)
套索回归	0.74 (0.06)	0.68 (0.05)	0.69 (0.14)
meLasso	0.74 (0.06)	0.68 (0.05)	0.69 (0.14)

表2. 使用套索回归和测量感知*Lasso*进行变量选择。每个单元格报告噪声协变量被剔除的平均比例（真阴性，*TN*）以及真实协变量保留的比例（真阳性，*TP*），基于*30*次独立实现。

方法	0个噪声协变量 TN, TP	10个噪声协变量 TN, TP	20个噪声协变量 TN, TP
较不稀疏网络（密度 = 0.37）			
套索回归	NA, 1	0.12, 1	0.86, 1
meLasso	NA, 1	0.12, 1	0.87, 1
稀疏网络（密度 = 0.013）			
套索回归	NA, 1	0.03, 1	0.12, 0.97
meLasso	NA, 1	0.12, 0.97	0.56, 0.64

5. 真实数据应用

我们展示了该方法如何指导计划开展大型网络研究的研究者。鉴于确定网络既耗时又费力，减少数据收集可减轻参与者负担，并使研究者能够专注于更好地获取更有意义的信息。

我们将模型应用于Banerjee等人（2013）在印度南部农村地区75个村庄开展的2006年基线调查中的社会网络数据; [数据集] 等人（2013）。每个家庭都接受了关于生活设施的调查，例如电力接入。约一半家庭接受了更详细的调查，其中个人被询问其社会网络信息，社会网络定义为到访其家或他们到访其家的人、家庭关系、他们给予或接受建议的人、给予或接受物质支持（金钱、物品）的人，以及与他们一起祈祷的人。这些信息被用于为每个村庄生成无向的家庭层面社会网络。由于各村庄之间距离较远，村庄网络实际上相互独立。平均网络密度为0.05。

该数据集包含数值数据和分类数据。我们将节点（家庭）协变量二值化，使二值变量在可能的情况下相对均衡，并且与经济状况相关的变量（如接入类型、屋顶材料）通常代表较高或较低的社会经济地位。这一点见表2。我们通过创建24个服从伯努利分布（0.5）的随机变量，对研究中的8个真实变量进行了扩充。随后，我们通过随机选取10个村庄来模拟一项试点研究。测量误差感知 Lasso识别出与网络潜在空间相关的协变量。除非协变量较为罕见（在任一试点村庄中的患病率为 < 10%），否则被至少70%的试点村庄剔除的协变量将被排除在完整分析之外。由此剔除了1个原始协变量以及24个噪声协变量中的21个，使最终所需的数据收集减少了69%。然后，我们仅使用试点研究中选出的协变量，分析了其余65个村庄的数据，并将结果与使用全部32个协变量的结果进行了比较。在此过程中被剔除的、来自原研究的变量是家庭宗教信仰，这很可能是因为超过一半的村庄全部为印度教徒。因此，鉴于缺乏变异性，该变量不太可能提供有用信息。

网络预测AUC未发生变化（均值差异为 = 0.00，标准差（SD）为 = 0.002）。比较预测全部32个协变量的均值AUC与预测所选6个协变量的均值AUC可见，受限变量集的AUC平均高出0.09（SD：0.02）。当仅在受限变量集中计算AUC时，这一差异消失了，如图2所示。对于大多数村庄，所估计的

表3. 家庭协变量的二值化。

变量	值	定义	平均患病率
宗教	1	家庭宗教信仰为印度教	96.0%
	0	其他任何宗教	4%
种姓	1	种姓或亚种姓为 ’附表种姓’或’表列部落	20.5%
	0	所有其他种姓	79.5%
房间数量	1	房间数量 > 2	34.8%
房间	0	两间或更少房间	65.2%
数量	1	床位数量 > 0	48.8%
beds	0	没有床铺	51.2%
房屋有电	1	有，私人供电	61.3%
	0	没有，或由政府提供	38.7%
厕所	1	有厕所	26.0%
	0	None	74.0%
屋顶类型	1	茅草或瓦片	33.9%
	0	石材，金属板 ，或钢筋混凝土混凝土	66.1%
社区领导者	1	Yes	12.6%
	0	No	87.4%

在受限变量集上使用meLasso与不使用meLasso所得的结果相当。在估计值差异超过0.01的三个案例中，结果均有利于meLasso。这表明，利用测量误差感知 Lasso选择变量子集不会对估计造成不利影响，而且确实能够筛除与潜在空间不相关、因而无法利用现有信息进行预测的协变量。

这些数据可通过 Harvard Dataverse 公开获取，网址为 https://doi.org/10.7910/DVN/U3BIHX。本项目使用的所有代码均可在 https://github.com/onnela-lab/lsm-lasso 获取。

6. 讨论

我们提出了一种协变量选择框架，用于对具有高维二元节点属性的网络数据进行联合潜在空间建模。我们利用 Zhang 等人（2022）的联合模型结构，该结构使节点属性能够为潜在位置估计提供信息，即使网络稀疏且许多节点彼此孤立；同时，我们引入了组稀疏选择步骤，以识别与网络生成潜在空间关联最密切的节点协变量。我们采用逻辑回归的组套索方法（Meier 等人，2008），并结合受测量误差感知正则化启发的岭型稳定化项（Sørensen 等人，2018，2015），以处理估计潜在位置的不确定性。

我们的理论结果为协变量部分建立了预测误差界，既适用于将潜在位置视为固定或已观测的情形，也适用于潜在位置存在有界估计误差的情形。在对 q 个协变量进行一致控制的条件下，所得收敛速率为 $O(\log q/n)$ 量级，另加一项反映潜在位置估计误差的项；这与采用可分解正则化项的广义线性模型的高维 M 估计理论相一致（Negahban 等人，2012；Meier 等人，2008）。这些结果为采用所提出的筛选与重拟合策略、在高维情形下保持预测准确性提供了理论依据。

在各项模拟中，考虑协变量稀疏性的选择程序比纳入所有节点协变量的方法稳健得多，尤其是在噪声协变量所占比例增加时。真实数据分析展示了一个互补且具有实际动机的应用场景：在全面开展网络调查之前，通过模拟试点研究筛选大量协变量组合。在调查问卷负担较重、网络测量成本较高的情境下，这种做法有助于提高数据收集效率：将测量重点放在与网络生成潜在结构关联最强的变量上，同时保持对网络分量的预测性能（Banerjee 等人，2013）。尽管试点村庄完全是随机选取的，这些结果依然表现良好；而在真实的试点研究中，很可能会更加谨慎地确保所选村庄能够代表目标总体。也许最重要的是，模拟和真实数据示例表明，在潜在空间建模中纳入节点协变量信息时，采用考虑稀疏性的估计技术可能带来显著收益，而且这样做几乎不会产生不利影响。在我们的模拟中，真实协变量很少会被剔除，除非网络和节点协变量信息都非常稀疏；

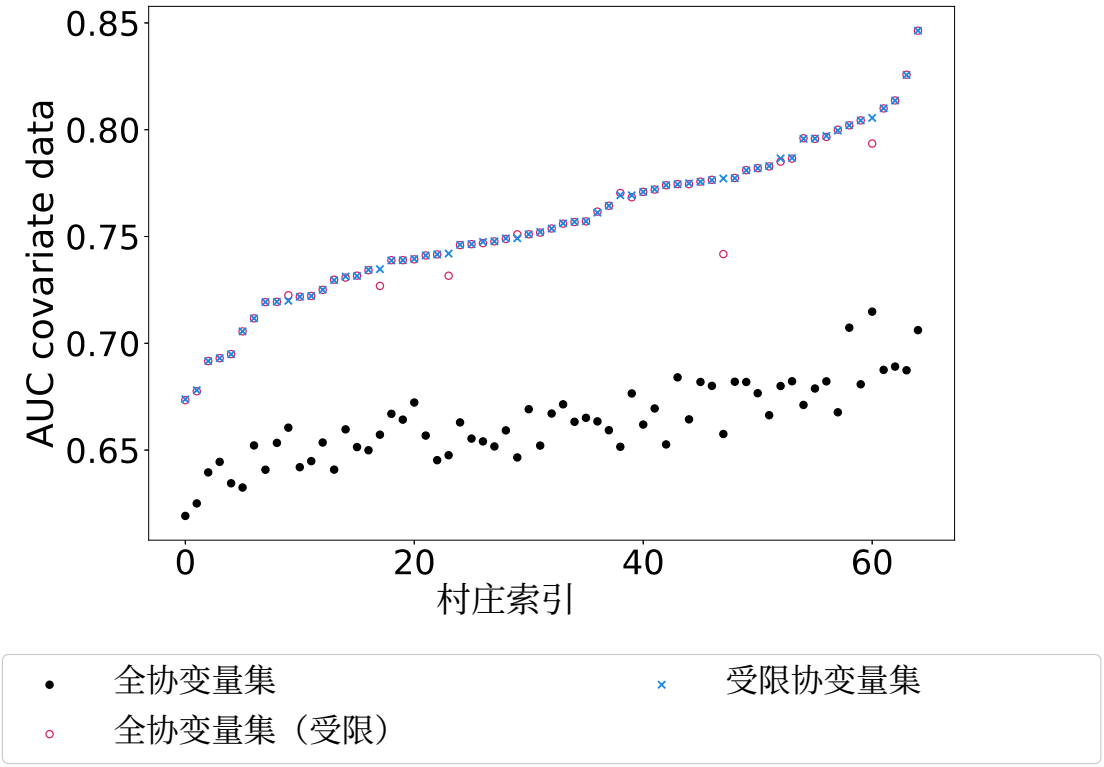

图 2. **使用和不使用模拟试点研究来预测节点协变量的模型表现**。该图显示了在未被选入试点研究的 65 个村庄中，节点协变量的平均曲线下面积。黑色标记表示在全协变量集上估计时，全部 32 个变量的平均曲线下面积。蓝色“X”标记表示仅对模拟试点研究中确定为感兴趣的协变量建模时的平均曲线下面积。红色标记（标注为“全协变量集（受限）”）表示试点研究保留的变量在使用全部协变量进行估计时的平均曲线下面积。村庄按受限协变量集的曲线下面积从低到高排序。

只有在这种情况下，可用于估计模型参数的信息才最少。此外，真实数据结果还表明，唯一被剔除的真实变量很可能因缺乏变异而几乎不包含网络信息；并且，收集完整数据并未改善测量感知Lasso所选协变量的估计或预测。

模型性能取决于少数几个调参与建模选择。首先，潜在维度 k 可通过设定一个适度规模的候选集，并选择使网络分量的样本外预测性能达到最优的取值来确定，这符合潜在位置模型中的常见做法（Hoff 等人，2002；Hoff，2005）。其次，平衡参数 Λ 控制节点属性与网络在联合目标函数中的相对影响（Zhang 等人，2022）。在稀疏网络中，适中的 Λ 值有助于利用节点协变量稳定潜在位置估计，尤其是对于低度节点和孤立节点；但 Λ 过大，则可能使潜在空间偏向解释协变量似然，从而牺牲网络拟合。第三，选择惩罚项 λ 决定稀疏性：λ 越大，选出的协变量越少；我们发现，在试点筛选情境中，将搜索范围限制为至少保留一个协变量的解很有用，这样可避免退化的“全零”拟合。最后，测量误差稳定参数 δ 需要在偏差与方差之间权衡：较大的 δ 会更强地收缩系数，并可降低噪声 $\widehat{Z}$导致的不稳定性，这与测量误差感知惩罚的既有研究结果一致（Sørensen 等人，2018，2015）。

有若干局限值得关注。首先，我们聚焦于采用逻辑回归模型处理二元节点协变量。在广义线性建模框架内，扩展到连续型或混合类型协变量在概念上较为直接，但实践中需要仔细进行尺度调整和正则化校准，以确保不同类型协变量的惩罚项具有可比性（Zhang 等人，2022）。其次，我们假设网络为无向网络。对于有向网络，则需要对非对称关系形成进行建模，并可能需要为发送和接收分别设定不同的潜在表示，同时调整度异质性项（Hoff 等人，2002；Krivitsky 等人，2009）。第三，我们的估计目标函数是非凸的；尽管 ADAM 和 Adagrad 等一阶方法在实践中效果良好（Kingma 和 Ba，2015），但局部最优解的存在意味着初始化和优化诊断仍然十分重要。建议在应用分析中报告对多次随机重启的敏感性，并评估所选协变量集的稳定性。

有几个值得探索的扩展方向。其一是加强模型选择理论，刻画在潜在位置作为回归变量的设定下，实现精确支持集恢复所需的条件。另一个方向是提升计算扩展性。对于更大的网络，开发近似方法或利用稀疏性感知优化，或许能大幅缩短运行时间。最后，研究模型设定错误下的稳健性也很有价值，包括考察偏离假定的潜在空间几何结构，以及当前设定未能捕捉的其他结构特征，例如聚类和度异质性（Athreya 等人，2018、2021）。

网络数据与丰富的节点级测量日益普及，这促使人们为稀疏网络开发兼具统计严谨性与实用性的方法。我们的框架将联合潜在空间建模、组稀疏协变量筛选与考虑测量误差的稳定化相结合，提供了一种易于处理的方法，可识别并利用与网络生成潜在结构相契合的节点属性，从而支持基于网络的社会学和流行病学研究中的可解释性与高效研究设计。

声明

数据和材料的可获得性

本数据可通过 Harvard Dataverse 公开获取，网址为 https://doi.org/10.7910/DVN/U3BIHX。本项目所用全部代码可在 https://github.com/onnela-lab/lsm-lasso 获取。

资助

本项目由美国国立卫生研究院资助项目 R01AI138901 和 T32AI007358 支持。资助方未参与研究设计、数据收集、分析或解释、文稿撰写，以及决定是否投稿发表。

作者贡献

E.G.C. 与 J.P.O. 设计了本研究。E.G.C. 撰写并编辑了理论证明。E.G.C.、Y.Z. 与 J.P.O. 撰写并编辑了论文。J.P.O. 指导了本研究。

致谢

我们感谢 Onnela 实验室成员对本项目的见解与反馈。

关于使用生成式人工智能和人工智能辅助技术的声明

在本工作的准备过程中，作者使用 ChatGPT 5.0 协助进行 LaTeX 排版以及微调图的外观。使用该工具后，作者根据需要对内容进行了审阅和编辑，并对出版物的内容承担全部责任。

补充材料

补充信息包含理论结果（定理 1 和 2）的证明。

参考文献

参考文献

Athreya, A.、Fishkind, D. E.、Tang, M.、Priebe, C. E.、Park, Y.、Vogelstein, J. T.、Levin, K.、Lyzinski, V.、Qin, Y. 和 Sussman, D. L.（2018）。随机点积图上的统计推断：综述。机器学习研究杂志，18(226):1–92。Athreya, A.、Tang, M.、Park, Y. 和 Priebe, C. E.（2021）。潜在结构随机图中的估计与推断。统计科学，36(1):68–88。Banerjee, A.、Chandrasekhar, A. G.、Duflo, E. 和 Jackson, M. O.（2013）。小额信贷扩散。《科学》，341(6144)。Bogu˜n´a, M.、Bonamassa, I.、De Domenico, M.、Havlin, S.、Krioukov, D. 和 Serrano, M. ˜A.（2021）。网络几何。《自然·物理评论》，3(2):114–135。[数据集], Banerjee, A.、Chandrasekhar, A. G.、Duflo, E. 和 Jackson, M. O.（2013）。小额信贷扩散数据。Duchi, J.、Hazan, E. 和 Singer, Y.（2011）。用于在线学习和随机优化的自适应次梯度方法。机器学习研究杂志，12(61):2121–2159。Hoff, P. D.（2005）。用于二元组数据的双线性混合效应模型。美国统计学会杂志，100(469):286–295。Hoff, P. D.、Raftery, A. E. 和 Handcock, M. S.（2002）。用于社会网络分析的潜在空间方法。美国统计学会杂志，97(460):1090–1098。Kingma, D. P. 和 Ba, J. L.（2015）。Adam：一种随机优化方法。第 *3* 届国际学习表征会议，第 1–13 页。Krivitsky, P. N.、Handcock, M. S.、Raftery, A. E. 和 Hoff, P. D.（2009）。利用潜在聚类随机效应模型表示社会网络中的度分布、聚类和同质性。社会网络，31(3):204–213。Ma, X. 和 Kundu, S.（2022）。利普希茨损失的高维测量误差模型。*arXiv*，第 1–35 页。Ma, Z.、Ma, Z. 和 Yuan, H.（2020）。具有边协变量的大型网络的通用潜在空间模型拟合。机器学习研究杂志，21(4):1–67。Meier, L.、Van De Geer, S. 和 B¨uhlmann, P.（2008）。逻辑回归中的组套索。《英国皇家统计学会杂志》。*B* 辑：统计方法学，70(1):53–71。Negahban, S. N.、Ravikumar, P.、Wainwright, M. J. 和 Yu, B.（2012）。具有可分解正则化项的 M 估计量高维分析的统一框架。统计科学，27(4):538–557。Smith, A. L.、Asta, D. M. 和 Calder, C. A.（2019）。网络数据连续潜在空间模型的几何。《统计科学》，34(3):428–453。Sørensen, ˜A.、Frigessi, A. 和 Thoresen, M.（2015）。套索回归中的测量误差：影响与似然偏差校正。《中华统计学志》，25(2):809–829。Sørensen, ˜A.、Hellton, K. H.、Frigessi, A. 和 Thoresen, M.（2018）。具有测量误差的高维广义线性模型中的协变量选择。《计算与图形统计学杂志》，27(4):739–749。Weiss, K. M.、Goodreau, S. M.、Morris, M.、Prasad, P.、Ramaraju, R.、Sanchez, T. 和 Jenness, S. M.（2020）。美国男男性行为者的自我中心性性网络：ARTnet研究结果。《流行病》，30:100386。Zhang, X.、Xu, G. 和 Zhu, J.（2022）。具有高维节点变量的网络数据联合潜在空间模型。《生物统计学》，109(3):707–720。

附录 1

A.1. 模型符号说明与假设回顾

设 G = (V,E) 为一个包含 n 个节点的无向网络，其对称邻接矩阵 $A \in \mathbb{R}^{n\times n}$满足：若边存在，则 $A_{ii',i\neq i'} = 1$，否则为 0。对于每个节点 i，我们观测到 q 个二元协变量$Y \in \{0,1\}^{n\times q}$，并且每个节点 i 与某个固定 k下的未观测低维潜在位置 $Z_i \in \mathbb{R}^k$相关联。我们通过减去行均值对潜变量进行中心化，并要求 Z, $\frac{1}{n}Z^\top Z$ 的经验协方差为对角阵但不是单位阵。

We 假设在给定潜在位置的条件下，节点之间的边相互独立，使得 t

$$A_{ii'} = A_{i'i} \sim \text{Bernoulli}(P^A_{ii'}), \quad \text{logit}P^A_{ii'} = \theta^A_{ii'} = \alpha_i + \alpha_{i'} + Z_i^\top Z_{i'},$$

其中 $P^A_{ii'}$ 是节点 i 与 i' 之间存在边的概率，而 $\alpha = (\alpha_1, \ldots, \alpha_n)$ 是节点特异的“社交性”因子，用于刻画节点独立于其潜在位置形成边的倾向。

对于二元协变量，我们假设在以 Z为条件时，每个协变量Y_j 对于 $j \in \{1, \ldots\ldots, q\}$服从独立的逻辑回归

$$Y_{ij} \sim \text{Bernoulli}(P^Y_{ij}), \quad \text{logit}P^Y_{ij} = \theta^Y_{ij} = \gamma_j + Z_i\beta_{\cdot j},$$

其中 P^Y_{ij} 是节点 i 的协变量 j 等于1的概率，$\gamma \in \mathbb{R}^q$ 是截距项向量，$\beta \in \mathbb{R}^{k\times q}$ 是将潜在位置与节点协变量联系起来的回归系数。

关键在于，我们假设 β 呈现组稀疏性：β 的许多列为零，表明相应的协变量与网络生成潜在空间无关。我们将 β的活跃集定义为 $S = \{ j \in \{1, \ldots\ldots, q\} : \|\beta_{\cdot j}\|_2 > 0\}$，其基数为 $s = |S|$。

我们的假设如下：

假设A I $M_1 > 0$ $-M_1 < \Theta^A_{ii'} < M_1$ $1 \leq$ （有界线性预测子）。存在使 $i\ i' \leq n$ $M_2 > 0$ $-M_2 < \Theta^Y_j < M_2$ $1 \leq j \leq q$ 得对于，并且使得对于。

假设 *A2*（稀疏分组结构）。每一列 $\beta_{\cdot j}$ 均视为一个大小为 k 的组，真正活跃协变量的数量 $s = |\{ j : \|\beta_{\cdot j}\|_2 > 0\}|$ 满足 $s = o(n/\log q)$。

假设 *A3*（设计归一化）。经验协方差 $D = \frac{1}{n}Z^\top Z$ 为对角矩阵（但不是单位矩阵），其特征值为 $0 < \nu_{\min}(D) \leq \nu_{\max}(D) < \infty$。

假设 *A4*（受限强凸性）。设估计的回归系数与真实值之间的误差为 $\beta_{\cdot j} - \beta_{\cdot j} = \Delta_j$。符号 $\cdot, \cdot$ 表示内积。对于每个 j，存在 $\kappa > 0$，使得对于所有 $\Delta_j \in \mathbb{R}^k$,

$$\frac{1}{n}\left(L_{Y_j}(\beta_{\cdot j} + \Delta_j; Z) - L_{Y_j}(\beta_{\cdot j}; Z) - \langle \nabla L_{Y_j}(\beta_{\cdot j}; Z), \Delta_j \rangle\right) \geq \frac{\kappa}{2}\|\Delta_j\|_2^2.$$

假设 *A5*（得分集中性）。对于每个 j，当 $\mu_{ij}(Z)$ =逻辑斯蒂函数$^{-1}(\gamma_j + Z_i^\top\beta_{\cdot j})$时，我们有

$$\left\|\frac{1}{n}Z^\top\{Y_{\cdot j} - \mu_j(Z)\}\right\|_2 \leq C_0\sqrt{\frac{\log q}{n}}$$

以高概率。

假设 *A6*（有界测量误差）。假设联合估计阶段的估计误差满足 $\|U\|_\infty \leq \delta_\infty$。

假设 A1 与 Zhang 等人（2022）所用的假设相同；假设 A3 使 Z 具有更强的可识别性。假设 A4 是高维 M 估计量的一项标准条件，确保损失函数具有强凸性。逻辑损失可二次微分；因此，这等价于对海森矩阵 $\nabla^2 L_{Y_j}(\beta_{\cdot j}, Z)$ 的特征值设定下界。Negahban 等人（2012）和 Meier 等人（2008）已对此进行了充分论证。假设 A5 可由标准的次指数集中性质推出（Meier 等人，2008）。最后，Zhang 等人（2022）指出 $\frac{1}{n}\|\widehat{Z} - Z\|_F = O_p(n^{1/2})$，这使我们能够

假设 $\delta_\infty = O(n^{-1/2})$。因此，假设 A6（测量误差有界）并非不合理。对于 U 的分布，没有作出其他假设。

联合估计目标结合了网络数据和协变量数据的负对数似然，使得 $L(Z,\ \alpha,\ \gamma,\ \beta) = L_A + \Lambda L_Y$，其中：

$$L_A = -\log P(A|Z,\alpha) = \sum_{1\le i<i'\le n} \{A_{ii'}\Theta^A_{ii'} - \log(1+\exp(\Theta^A_{ii'}))\}\ ,$$

$$L_Y = -\log P(Y|\gamma,\beta) = \sum_{1\le i\le n, 1\le j\le q} \{Y_{ij}\Theta^Y_{ij} - \log(1+\exp(\Theta^Y_{ij}))\}\ .$$

权重参数 $\Lambda \ge 0$平衡网络结构和协变量中的信息。

A.2. *Proof of Theorem A1*

Define the prediction error $d^2(\widehat{\eta}_j, \eta_j) = \frac{1}{n}\|Z^\top(\widehat{\beta}_{\cdot j} - \beta_{\cdot j})\|_F^2$, $\eta_j = Z\beta_{\cdot j}$.

定理 A1（预测误差率，完全观测的 Z）。在假设 ***A1–A5*** 下：（a）对于任意固定的 $j \in \{1,\ \ldots,\ q\}$，当 $\lambda_j = C\sqrt{k/n}$ 对某个常数 C 成立时：

$$d^2(\widehat{\eta}_{\cdot j}, \eta_{\cdot j}) = O_p\Big(\frac{k}{n}\Big).$$

（b）对于所有 $j \in \{1,\ \ldots,\ q\}$ 的一致控制，当 $\lambda = C\sqrt{\log q/n}$ 对某个常数 C 成立时：

$$\max_{j=1,\ldots,q} d^2(\widehat{\eta}_{\cdot j}, \eta_{\cdot j}) = O_p\Big(\frac{\log q}{n}\Big).$$

证明。**（a）部分：单个回归速率。**

对于固定的 j，通过调整 Meier 等人（2008）针对一个大小为 k 的组进行单次回归的方法，并使用假设 A5 中速率为 $\sqrt{k/n}$ 的得分集中性，惩罚项为 $\lambda_j = C\sqrt{k/n}$ 的组套索估计量满足：

$$d^2(\widehat{\eta}_j, \eta_j) = O_p\Big(\frac{k}{n}\Big).$$

由于 $D = n^{-1}Z^\top Z$ 是对角矩阵且特征值有界，因此范数等价

$$\nu_{\min}(D)\,\frac{1}{n}\|\tilde{Z}^\top(\widehat{\beta}_{\cdot j} - \beta_{\cdot j})\|_F^2 \le d^2(\widehat{\eta}_j, \eta_j) \le \nu_{\max}(D)\,\frac{1}{n}\|\tilde{Z}^\top(\widehat{\beta}_{\cdot j} - \beta_{\cdot j})\|_F^2\ ,$$

其中 $\tilde{Z} = ZD^{-1/2}$。因此，对于 Z，该速率在一个依赖于 D 的常数范围内成立。

（b）部分：通过并集界实现一致控制。

为确保在 β 的所有 q 列上都能控制优化误差，惩罚项必须大于所有 q 个回归中的最大得分波动。选择 $\lambda = C\sqrt{\log q/n}$，且令 C 足够大（具体而言，$C > C_0\sqrt{k/\log q}$），即可得到

$$\max_{j=1,\ldots,q} d^2(\widehat{\eta}_j, \eta_j) = O_p\Big(\frac{\log q}{n}\Big)$$

通过标准的并集界论证即可得到。 □

A.3. *Proof of Theorem 2*

在实践中，潜在位置未知，必须加以估计，从而引入加性测量误差。该定理的主要困难在于使用 $\hat{}Z$ 而非 Z所引入的额外偏差。

设$\hat{Z} = Z + U$，其中 Z 为预言机潜在位置， U 为某种未知的估计误差（Sørensen 等人，2018, 2015）。单个组的估计感知组套索估计量为：

$$\widehat{\beta}_{\cdot j} \in \arg\min_{\beta_{\cdot j} \in \mathbb{R}^k} \left\{ L_{Y_j}(\beta_{\cdot j}) + \lambda \|\beta_{\cdot j}\|_2 + \frac{\delta}{2} \|\beta_{\cdot j}\|_2^2 \right\}. \tag{A1}$$

设定 $\delta = 0$ 可从简化情形恢复忽略测量误差的组套索。

定理 A2. 在假设 *A1–A6* 下：

（*a*）对于任意固定的 $j \in \{1, \ldots, q\}$，在调优参数 $\lambda_j = C_1\sqrt{k/n}$ 与 $\delta = C_2\delta_\infty$ 下，其中 C_1、C_2 为正数常量：

$$\frac{1}{n}\|\hat{Z}(\hat{\beta}_{\cdot j} - \beta_{\cdot j})\|_2^2 = O_p\left(\frac{1}{n} + \delta_\infty^2\right).$$

（*b*）为对所有 $j \in \{1, \ldots, q\}$ 进行一致控制，在调优参数 $\lambda = C_1\sqrt{\log q/n}$ 与 $\delta = C_2\delta_\infty$ 下：

$$\max_{j=1,\ldots,q} \frac{1}{n}\|\hat{Z}(\hat{\beta}_{\cdot j} - \beta_{\cdot j})\|_2^2 = O_p\left(\frac{\log q}{n} + \delta_\infty^2\right).$$

当 $\delta_\infty = O(n^{-1/2})$ 来自第一阶段估计时，两种速率均退化为各自的统计误差率。

证明. 我们首先证明（a）部分，然后通过若干附加步骤证明（b）部分。

令$\Delta_j = \widehat{\beta}_{\cdot j} - \beta_{\cdot j}$表示协变量 j 的估计误差。由于 k 是固定的，因此我们在全文中将所有涉及 k 的因子视为常数。我们将 $\|\cdot\|_{op}$ 记为矩阵的算子范数。

步骤 1：最优性基本不等式。根据 $\beta_{\cdot j}$ 的最优性，我们有

$$L_{Y_j}(\widehat{\beta}_{\cdot j}; \hat{Z}) + \lambda\sqrt{k}\|\widehat{\beta}_{\cdot j}\|_2 + \frac{\delta}{2}\|\widehat{\beta}_{\cdot j}\|_2^2 \le L_{Y_j}(\beta_{\cdot j}; \hat{Z}) + \lambda\sqrt{k}\|\beta_{\cdot j}\|_2 + \frac{\delta}{2}\|\beta_{\cdot j}\|_2^2.$$

重新排列后可得：

$$\begin{aligned} L_{Y_j}(\widehat{\beta}_{\cdot j}; \hat{Z}) - L_{Y_j}(\beta_{\cdot j}; \hat{Z}) &\le \lambda\sqrt{k}(\|\beta_{\cdot j}\|_2 - \|\widehat{\beta}_{\cdot j}\|_2) + \frac{\delta}{2}(\|\beta_{\cdot j}\|_2^2 - \|\widehat{\beta}_{\cdot j}\|_2^2) \\ &\le \lambda\sqrt{k}\|\Delta_j\|_2 + \frac{\delta}{2}(\|\beta_{\cdot j}\|_2^2 - \|\widehat{\beta}_{\cdot j}\|_2^2). \end{aligned} \tag{A2}$$

步骤 2：通过真实设计 Z 进行分解。加上并减去 $L_{Y_j}(\beta_{\cdot j};\ Z)$ 和 $L_{Y_j}(\beta_{\cdot j};\ Z)$ ：

$$L_{Y_j}(\widehat{\beta}_{\cdot j}; \hat{Z}) - L_{Y_j}(\beta_{\cdot j}; \hat{Z}) = L_{Y_j}(\widehat{\beta}_{\cdot j}; \hat{Z}) - L_{Y_j}(\widehat{\beta}_{\cdot j}; Z) + L_{Y_j}(\widehat{\beta}_{\cdot j}; Z) \tag{A3}$$

$$- L_{Y_j}(\beta_{\cdot j}; Z) + L_{Y_j}(\beta_{\cdot j}; Z) - L_{Y_j}(\beta_{\cdot j}; \hat{Z})\ . \tag{A4}$$

这些项可以重新排列：

$$L_{Y_j}(\widehat{\beta}_{\cdot j}; \hat{Z}) - L_{Y_j}(\beta_{\cdot j}; \hat{Z}) = \underbrace{L_{Y_j}(\widehat{\beta}_{\cdot j}; Z) - L_{Y_j}(\beta_{\cdot j}; Z)}_{\text{Loss on true design Z}} \tag{A5}$$

$$+ \underbrace{L_{Y_j}(\widehat{\beta}_{\cdot j}; \hat{Z}) - L_{Y_j}(\widehat{\beta}_{\cdot j}; Z) + L_{Y_j}(\beta_{\cdot j}; Z) - L_{Y_j}(\beta_{\cdot j}; \hat{Z})}_{\text{Effect of measurement error}}\ . \tag{A6}$$

真实设计上的损失 Z 可分解为一个涉及得分函数 $\nabla L_{Y_j}(\beta_{\cdot j};\ Z)$ 和海森矩阵 $\nabla^2 L_{Y_j}(\beta_{\cdot j};\ Z)$ 的表达式。

首先，我们可以对 $L_{Y_j}(\beta_{\cdot j};\ Z)$ 进行泰勒展开：

$$L_{Y_j}(\widehat{\beta}_{\cdot j}; Z) = L_{Y_j}(\beta_{\cdot j}; Z) + \langle \nabla L_{Y_j}(\beta_{\cdot j}; Z), \Delta_j \rangle + \frac{1}{2}\nabla^2 L_{Y_j}(\tilde{\beta}_{\cdot j}; Z) \tag{A7}$$

~ ^对于 $\beta_{\cdot j}$ 和 $\beta_{\cdot j}$ 之间的某个 $\beta_{\cdot j}$。我们可以将其代入真实设计上的损失表达式：

$$L_{Y_j}(\widehat{\beta}_{\cdot j}; Z) - L_{Y_j}(\beta_{\cdot j}; Z) = \langle \nabla L_{Y_j}(\beta_{\cdot j}; Z), \Delta_j \rangle + L_{Y_j}(\widehat{\beta}_{\cdot j}; Z) - L_{Y_j}(\beta_{\cdot j}; Z) - \langle \nabla L_{Y_j}(\beta_{\cdot j}; Z), \Delta_j \rangle \ . \quad \text{(A8)}$$

最后，我们有

$$\begin{aligned} L_{Y_j}(\widehat{\beta}_{\cdot j}; \hat{Z}) - L_{Y_j}(\beta_{\cdot j}; \hat{Z}) = & \underbrace{L_{Y_j}(\widehat{\beta}_{\cdot j}; Z) - L_{Y_j}(\beta_{\cdot j}; Z) - \langle \nabla L_{Y_j}(\beta_{\cdot j}; Z), \Delta_j \rangle}_{\text{(I): Curvature term}} \\ & + \underbrace{\langle \nabla L_{Y_j}(\beta_{\cdot j}; Z), \Delta_j \rangle}_{\text{(II): Score term}} \\ & + \underbrace{L_{Y_j}(\widehat{\beta}_{\cdot j}; \hat{Z}) - L_{Y_j}(\widehat{\beta}_{\cdot j}; Z) + L_{Y_j}(\beta_{\cdot j}; Z) - L_{Y_j}(\beta_{\cdot j}; \hat{Z})}_{\text{(III): Measurement error bias}} . \end{aligned} \quad \text{(A9)}$$

Step 3：曲率项的下界（I）。根据假设 A4（受限强凸性）：

$$\text{(I)} \geq \frac{n\kappa}{2} \|\Delta_j\|_2^2 .$$

步骤 4：得分项的界（II）。根据假设 A5（得分集中性）：

$$|\text{(II)}| = |\langle Z^\top (Y_{\cdot j} - \mu_j(Z)), \Delta_j \rangle| \leq nC_0 \sqrt{\frac{k}{n}} \|\Delta_j\|_2 = C_0 \sqrt{nk} \|\Delta_j\|_2 \ .$$

步骤 5：测量误差偏差项的界（III）。为了理解如何界定项（III），我们首先替换 $\hat{Z} = Z + U$：

$$L_{Y_j}(\widehat{\beta}_{\cdot j}; \hat{Z}) - L_{Y_j}(\widehat{\beta}_{\cdot j}; Z) = L_{Y_j}(\widehat{\beta}_{\cdot j}; Z + U) - L_{Y_j}(\widehat{\beta}_{\cdot j}; Z) = g(\widehat{\beta}_{\cdot j}) \ , \quad \text{(A10)}$$

$$L_{Y_j}(\beta_{\cdot j}; Z) - L_{Y_j}(\beta_{\cdot j}; \hat{Z}) = L_{Y_j}(\beta_{\cdot j}; Z) - L_{Y_j}(\beta_{\cdot j}; Z + U) = g(\beta_{\cdot j}) \ . \quad \text{(A11)}$$

~ ^根据中值定理，在线段 $\beta_{\cdot j}$ 与 $\beta_{\cdot j}$ 之间存在 β，使得

$$g(\widehat{\beta}_{\cdot j}) - g(\beta_{\cdot j}) = \langle \nabla_\beta g(\tilde{\beta}_{\cdot j}), \Delta \rangle \ . \quad \text{(A12)}$$

关于 β 的梯度为

$$\nabla_\beta L(\beta, Z) = Z^\top (\mu(\beta, Z) - Y) \ ,$$

其中 $\mu(\cdot)$ 是 logit 链接函数。因此，

$$\begin{aligned} \nabla_\beta g(\tilde{\beta}_{\cdot j}) &= \nabla_\beta L(\tilde{\beta}, \hat{Z}) - \nabla_\beta L(\tilde{\beta}, Z) && \text{(A13)} \\ &= \hat{Z}^\top (\mu(\tilde{\beta}, \hat{Z}) - Y) - Z^\top (\mu(\tilde{\beta}, Z) - Y) && \text{(A14)} \\ &= (Z + U)^\top (\mu(\tilde{\beta}, \hat{Z}) - Y) - Z^\top (\mu(\tilde{\beta}, Z) - Y) && \text{(A15)} \\ &= U^\top (\mu(\tilde{\beta}, \hat{Z}) - Y) - Z^\top (\mu(\tilde{\beta}, \hat{Z}) - \mu(\tilde{\beta}, Z)) \ , && \text{(A16)} \end{aligned}$$

并且我们有

$$\langle \nabla_\beta g(\tilde{\beta}_{\cdot j}), \Delta \rangle = \underbrace{-\langle U^\top (Y - \mu(\tilde{\beta}, \hat{Z})), \Delta \rangle}_{\text{(A)}} - \underbrace{\langle Z^\top (\mu(\tilde{\beta}, \hat{Z}) - \mu(\tilde{\beta}, Z)), \Delta_j \rangle}_{\text{(B)}} . \tag{A17}$$

步骤 5.1：界定项 (A) 为简便起见，我们可以将项 (A) 和 (B) 拆开，分别对它们进行界定。

$$(A) = |-\langle U^\top (Y - \mu(\tilde{\beta}, \hat{Z})), \Delta \rangle| \le \|U^\top (Y - \mu(\tilde{\beta}, \hat{Z}))\|_2 \, \|\Delta\|_2 \tag{A18}$$

每个 $Y - \mu(\tilde{\beta}, \hat{Z}) \in [-1, 1]$，因此 $\|Y - \mu(\tilde{\beta}, \hat{Z})\|_2 \le \sqrt{n}$。根据假设 A6，我们有 $\|U\|_\infty \le \delta_\infty$ 因此，U 的每一列的 L2 范数均为 $\le \sqrt{n}\delta_\infty$ 且 $\|U\|_{op} \le \sqrt{nk}\delta_\infty$
将其代入上面的表达式，可得

$$|-\langle U^\top (Y - \mu(\tilde{\beta}, \hat{Z})), \Delta \rangle| \le \sqrt{nk}\delta_\infty \cdot \sqrt{n} \cdot \|\Delta\|_2 \tag{A19}$$

$$= n\sqrt{k}\delta_\infty \|\Delta\|_2 . \tag{A20}$$

步骤 5.2：界定项 (B)： 现在聚焦于 (B)：$(B) = Z^\top(\mu(\tilde{\beta}, \hat{Z}) - \mu(\tilde{\beta}, Z)), \Delta_j$。
我们可以将均值向量 μ_j 关于线性预测器的变化线性化：

$$\eta(Z) = \gamma_j \mathbf{1}_n + Z\tilde{\beta} , \tag{A21}$$

$$\eta(\hat{Z}) = \gamma_j \mathbf{1}_n + \hat{Z}\tilde{\beta} = \eta(Z) + U\tilde{\beta} . \tag{A22}$$

logit 链接函数可微，其导数为 $\mu'(\xi_i) \in (0, 1/4]$。根据中值定理，

$$\mu_j(\tilde{\beta}; \hat{Z}) - \mu_j(\tilde{\beta}; Z) = W(U\tilde{B}) ,$$

其中 $W = \text{diag}(\mu'_1, \dots \mu'_n)$ 且 $\|W\|_{op} \le 1/4$。B
y Cauchy-Sch$_{\text{warz,}}$

$$|(B)| \le \|Z^\top WU\|_{op} \|\tilde{\beta}\|_2 \|\Delta_j\|_2 . \tag{A23}$$

给定 $\|W\|_{op} \le 1/4$：

$$\|Z^\top WU\|_{op} \le \|Z\|_{op} \|W\|_{op} \|U\|_{op} \le \frac{1}{4} \|Z\|_{op} \|U\|_{op} . \tag{A24}$$

我们可以用 $D = \frac{1}{n} Z^\top Z$ 来界定 $\|Z\|_{op}$：对于任意 v，$\frac{1}{n}\|Zv\|_2^2 = v^\top Dv \le \nu_{\max}(D)\|v\|_2^2$：

$$\|Z\|_{op} \le \sqrt{n\nu_{\max}(D)} .$$

令 $\|U\|_{op} \le \sqrt{nk\delta_\infty}$ 如上所述，可将其合并：

$$\|Z^\top WU\|_{op} \le \frac{1}{4}\sqrt{n\nu_{\max}(D)}\sqrt{nk}\delta_\infty = \frac{n}{4}\sqrt{k\nu_{\max}(D)}\delta_\infty .$$

由于 k 是固定的，$\|\tilde{\beta}\|$ 有界，且 $\|\Delta_j\|_2$ 在目标速率下会很小，

$$|(B)| \le C_B n\delta_\infty \|\Delta_j\|_2$$

其中 C_B为某个常数。
结合项 (A) 和 (B)，

$$|(\text{III})| \le C_3 n\delta_\infty \|\Delta_j\|_2 , \tag{A25}$$

其中 C_3 是一个取决于 k 和 $\nu_{\max}(D)$的常数。

步骤 6：合并界。 将步骤 3–5 代入步骤 1 中的不等式，并除以 n：

$$\frac{\kappa}{2}\|\Delta_j\|_2^2 \le C_0\sqrt{\frac{k}{n}}\|\Delta_j\|_2 + C_3\delta_\infty\|\Delta_j\|_2 + \frac{\lambda\sqrt{k}}{n}\|\Delta_j\|_2 + \frac{\delta}{2n}(\|\beta_{\cdot j}\|_2^2 - \|\widehat{\beta}_{\cdot j}\|_2^2). \tag{A26}$$

利用 $\|\beta_{\cdot j}\|_2^2 - \|\widehat{\beta}_{\cdot j}\|_2^2 \le 2\|\beta_{\cdot j}\|_2\|\Delta_j\|_2$、我们对 $\lambda_j = C_1\sqrt{k/n}$ 的选择以及 $\delta = C_2\delta_\infty$（常数吸收了 k 的因子）：

$$\frac{\kappa}{2}\|\Delta_j\|_2^2 \le C_4\sqrt{\frac{1}{n}}\|\Delta_j\|_2 + C_5\delta_\infty\|\Delta_j\|_2,$$

其中 C_4 和 C_5 是常数（取决于 k、C_0、C_1、C_2、C_3 和 $\|\beta_{\cdot j}\|_2$）。

若 $\Delta_j = 0$，则问题得解。否则，两边除以 $\|\Delta_j\|_2$ 并求解：

$$\|\Delta_j\|_2 = O\left(\sqrt{\frac{1}{n}} + \delta_\infty\right).$$

因此，对于任意固定的 j，适当选择 λ_j 即可：

$$\|\Delta_j\|_2^2 = O\left(\frac{1}{n} + \delta_\infty^2\right).$$

步骤 7：关于 $\hat{Z}$ 的预测误差。最后，我们对估计设计矩阵上的预测误差进行界定：

$$\begin{aligned}\frac{1}{n}\|\hat{Z}\Delta_j\|_2^2 &= \frac{1}{n}\|(Z+U)\Delta_j\|_2^2 \\ &\le \frac{2}{n}\|Z\Delta_j\|_2^2 + \frac{2}{n}\|U\Delta_j\|_2^2 \\ &\le 2\nu_{\max}(D)\|\Delta_j\|_2^2 + 2C_6\delta_\infty^2\|\Delta_j\|_2^2,\end{aligned} \tag{A27}$$

其中 C_6 是一个取决于 k的常数，利用以下界

$$\frac{1}{n}\|U\Delta_j\|_2^2 \le \frac{1}{n}\cdot nk\delta_\infty^2\|\Delta_j\|_2^2 = k\delta_\infty^2\|\Delta_j\|_2^2 = O(\delta_\infty^2\|\Delta_j\|_2^2).$$

结合步骤 6 中关于 $\|\Delta_j\|_2^2$ 的界：

$$\frac{1}{n}\|\hat{Z}\Delta_j\|_2^2 = O_p\left(\frac{1}{n} + \delta_\infty^2\right).$$

这证明了（a）部分中的速率。为证明（b）部分，我们还需要以下步骤：

步骤 7：一致控制 上述分析对任意固定的 j均以高概率成立。为获得一致控制，注意到得分项表现为相互独立的次高斯随机变量。因此，我们需要 $\lambda = C_1\sqrt{\log_n q}$ 来同时控制所有 q 回归。

采用这一较大的惩罚项后，一致界变为：

$$\max_{j=1,\dots,q}\|\Delta\|_2^2 = O_p\left(\frac{\log q}{n} + \delta_\infty^2\right).$$

当第一阶段估计中的 $\delta_\infty = O(n^{-1/2})$时：

$$\frac{1}{n}\|\hat{Z}\Delta_j\|_2^2 = O_p\left(\frac{\log q}{n} + \frac{1}{n}\right) = O_p\left(\frac{\log q}{n}\right),$$

因为 $\log q \gg 1$ 适用于高维情形。