arXiv:2602.07911v1 [stat.AP] 8 Feb 2026

高维回归系数的自适应检验程序

Ping Zhao，天津工业大学数学科学学院；
Fengyi Song 和 Huifang Ma，南开大学统计与数据科学学院

2026年2月10日

摘要

我们针对高维回归系数开发了一个统一的 L统计量检验框架，可适应未知稀疏度。所提出的统计量对逐坐标证据度量进行排序，并汇总前 k 个信号，从而衔接经典的最大值型检验与求和型检验。在温和条件下，我们建立了极值分量与标准化 L统计量的联合弱收敛，并得到渐近独立性，这为组合多个k提供了依据。我们通过在 k的二进网格上进行柯西组合，构造了自适应综合检验，并提供具有理论保证的野生自助法校准。模拟结果表明，该方法在稀疏和稠密备择下（包括非高斯设计）均具有准确的检验水平和较强的检验效能。

Keywords：柯西组合检验、高维数据、L统计量、野生自助法

1. 引言

现代应用通常会产生这样的回归问题：协变量的数量与样本量相当，甚至远远超过样本量。在此类情形下，一个核心推断任务是检验大量回归系数是否联合为零，必要时还需对少量干扰协变量进行调整。这类全局或分块检验问题广泛存在于基因组学、神经影像学和金融领域，因为其中的科学问题通常可以自然地表述为：在基准模型之外，一组高维预测变量是否具有额外的解释力。

一个关键难点在于，功效最优的检验策略在很大程度上取决于备择假设的稀疏模式。当信号稀疏时（少数坐标承载了大部分效应），针对最大边际证据的最大值型方法通常较为有效；例如，Feng 等人(2024)提出的最大值型方法及相关渐近独立现象。相比之下，当信号稠密时（许多效应较弱），通常更倾向于采用求和／能量型方法，包括Goeman 等人 (2006)提出的全局检验框架、其扩展，以及针对广义线性模型的第一类错误控制方法 Goeman 等人 (2011)；此外，还有针对高维回归系数的相关求和型检验，例如Zhong 和 Chen (2011)以及Lan 等人 (2014)。然而在实践中，稀疏度通常无法事先获知，而针对单一情形调整的方法

当实际备择假设偏离该情形时，可能会遭受显著的功效损失。Cui 等人 (2018)采用重拟合交叉验证方差估计，以避免高估方差并提升经验检验功效。 Guo 等人 (2022)提出了一种基于随机数据拆分策略、结合筛选程序的两阶段条件 U 统计量检验，以提升经验检验功效。

这促使我们采用自适应检验策略，使其在广泛的备择假设范围内都能保持敏感性。一种自然的方法是构造一族由调节参数索引的统计量，在面向稀疏情形和面向稠密情形的行为之间进行插值，然后综合该索引集上的信息。本文通过 L 统计量框架来实现这一思路：我们对逐坐标证据度量进行排序，并汇总最大的 k 个分量，从而构造一系列连续变化的最高阶 k L 统计量，在最大值型检验与求和型检验之间平滑衔接。这一视角与 Ma 等人(2024) 提出的高维检验自适应 L 统计量最新进展密切相关，也使我们能够以有原则的方式提出一种可适应未知稀疏度的统一方法。

多 k 聚合面临的一项主要理论挑战是，$\{L_k\}$ 在 k 的相邻取值处存在显著依赖，而且最高阶次统计量的极值行为与 k 发散时出现的高斯型波动之间也存在相互作用。如果不在零假设下进行细致的联合刻画，直接组合多个 k 可能导致检验过于保守或检验规模膨胀。基于 Feng 等人 (2024) 关于渐近独立性的见解，以及 Ma 等人(2024) 提出的基于预期损失的正态近似理论，我们建立了联合弱收敛结果，阐明极值分量与标准化的L统计量在零假设下如何因子化。这种渐近独立性为综合不同 k 取值下的证据，同时保持有效的第一类错误控制，提供了严格的理论基础。

在方法上，我们提出一种全能自适应检验，将对应于截断水平二进网格 k 的 p-值聚合起来，并辅以少量极值型分量。由此得到的流程计算简单——所有分量 p-值都可通过统一的自助法流程获得——同时旨在适用于不同的稀疏度情形：较小的 k 着重于稀有/强信号，而较大的 k 则累积广泛/弱信号。为改善有限样本校准并适用于非高斯设计，我们采用野生自助法方案来近似整个 L-统计量族的零假设分布。尽管我们的部分极限理论为技术上的便利而借助了高斯结构，但模拟结果表明，所提出的方法在多种非高斯分布下仍表现良好。

本文其余部分安排如下。第 2 节介绍 top-$k$$L$-统计量族及自助法校准。随后，我们分别在固定-$k$和发散-$k$ 情形下建立渐近理论，并给出启发自适应组合方法的联合极限结果。第 3 节报告模拟研究，将所提出的流程与具有代表性的最大值型和和型竞争方法进行比较。第 4 节以讨论和可能的扩展作结。

2. 检验程序

设 $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$ 为独立同分布的 p 维协变量向量，$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ 为相应的独立响应。为方便记号，假设 $E(\boldsymbol{X}_i) = \boldsymbol{0}$。我们将每个协变量向量划分为

$$\boldsymbol{X}_i = \left(\boldsymbol{X}_{ia}^\top, \boldsymbol{X}_{ib}^\top\right)^\top, \qquad \boldsymbol{X}_{ia} \in \mathbb{R}^q,\ \boldsymbol{X}_{ib} \in \mathbb{R}^{p-q},$$

其中， $q < n$ 为干扰协变量的维数，且 p 可能远大于 n。考虑线性回归模型

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i = \boldsymbol{X}_{ia}^\top \boldsymbol{\beta}_a + \boldsymbol{X}_{ib}^\top \boldsymbol{\beta}_b + \varepsilon_i, \tag{1}$$

其中 $\boldsymbol{\beta} = (\boldsymbol{\beta}_a^\top, \boldsymbol{\beta}_b^\top)^\top \in \mathbb{R}^p$，且 $\boldsymbol{\beta}_a \in \mathbb{R}^q$ 和 $\boldsymbol{\beta}_b \in \mathbb{R}^{p-q}$。噪声 $\{\varepsilon_i\}_{i=1}^n$ 相互独立，满足 $E(\varepsilon_i) = 0$、$\mathrm{Var}(\varepsilon_i) = \sigma^2$，并且与 $\boldsymbol{X}$独立。

我们的目标是检验高维分块 $\boldsymbol{\beta}_b$ 是否为零:

$$H_0 : \boldsymbol{\beta}_b = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_b \neq \mathbf{0}, \tag{2}$$

当 $\boldsymbol{X}_{ia} \equiv \mathbf{0}$时，这可化为全局检验 H_0： $\boldsymbol{\beta} = \mathbf{0}$。

为消除干扰分量 $\boldsymbol{\beta}_a$，定义设计矩阵

$$\mathbf{X}_a = (\boldsymbol{X}_{1a}, \ldots, \boldsymbol{X}_{na})^\top, \qquad \mathbf{X}_b = (\boldsymbol{X}_{1b}, \ldots, \boldsymbol{X}_{nb})^\top, \qquad m = p - q,$$

以及投影到 $\mathbf{X}_a$的列空间上的投影矩阵，$\mathbf{H}_a = \mathbf{X}_a(\mathbf{X}_a^\top \mathbf{X}_a)^{-1}\mathbf{X}_a^\top$.。接下来，我们通过 $\tilde{\mathbf{X}}_b = (\mathbf{I}_n - \mathbf{H}_a)\mathbf{X}_b = \big(\tilde{\mathbf{X}}_{\cdot 1}, \ldots, \tilde{\mathbf{X}}_{\cdot m}\big)$对 $\mathbf{X}_b$ 相对于 $\mathbf{X}_a$ 进行残差化，并定义相应的残差和方差估计量

$$\hat{\boldsymbol{\varepsilon}} = (\mathbf{I}_n - \mathbf{H}_a)\mathbf{Y}, \qquad \hat{\sigma}^2 = (n-q)^{-1}\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}.$$

基于 残差化设计矩阵，我们考虑标准化得分型统计 cs

$$Z_j = \frac{\tilde{\mathbf{X}}_{\cdot j}^\top \hat{\boldsymbol{\varepsilon}}}{\hat{\sigma}\|\tilde{\mathbf{X}}_{\cdot j}\|_2}, \qquad W_j := Z_j^2, \qquad j = 1, \ldots, m,$$

并令 $W_{(1)} \geq \cdots \geq W_{(m)}$ 为次序统计量。对于 $1 \leq k \leq m$，定义最高k L阶统计量

$$L_k = \sum_{j=1}^{k} W_{(j)}.$$

该族在两种经典情形之间架起了统一的桥梁：当 $k = 1$时， L_k 退化为 ℓ_∞型（最大值）统计量，与Feng 等人（2024）中的最大值型检验密切相关；当 $k = m$时， L_k 成为 ℓ_2型（和/能量）统计量，从而与诸如Goeman 等人（2006，2011）；Zhong 和 Chen （2011）；Lan 等人（2014）等和型方法相联系。更一般地， k 的中间取值在稀疏备择与稠密备择之间插值，从而促使同时研究固定k 与发散k 渐近。

下文中，我们在两种互补情形下推导 L_k 的零假设极限分布： (i) k 固定，从而最高阶次统计量具有极值/泊松点过程极限； (ii) k 随 m发散，此时 L_k 具有渐近正态近似。这些结果为构造跨多个k综合证据的自适应方法提供了理论基础。

2.1 固定参数 k

定义部分协方差矩阵 $\boldsymbol{\Sigma}_{b|a} := E[\mathrm{Cov}(\boldsymbol{X}_{ib} \mid \boldsymbol{X}_{ia})] = \big(\sigma_{jk}^*\big) \in \mathbb{R}^{m \times m}$. 不失一般性，我们将 $\boldsymbol{\Sigma}_{b|a}$ 规范化，使得 $\sigma_{jj}^* = 1$ 对所有 $j \in \{1, \ldots, m\}$ 成立。令 B 为总体

$\boldsymbol{X}_{ib}$ 对 $\boldsymbol{X}_{ia}$ 的回归系数，$\mathrm{B} := \mathrm{Cov}(\boldsymbol{X}_{ib}, \boldsymbol{X}_{ia}) \left[\mathrm{Cov}(\boldsymbol{X}_{ia})\right]^{-1} \in \mathbb{R}^{m\times q}$，并定义残差（创新）向量 $\boldsymbol{X}_{ib}^{*} := \boldsymbol{X}_{ib} - \mathrm{B}\boldsymbol{X}_{ia} = (X_{i1}^{*}, \cdots, X_{im}^{*})^{\top} \in \mathbb{R}^{m}$. 由于 $E(\boldsymbol{X}_{i}) = \mathbf{0}$，由此可得 $E(\boldsymbol{X}_{ib}^{*}) = \mathbf{0}$ 以及 $\mathrm{Cov}(\boldsymbol{X}_{ib}^{*}) = \boldsymbol{\Sigma}_{b|a}$. 最后，将残差收集到矩阵 $\mathrm{X}_{b}^{*} := \left(\boldsymbol{X}_{1b}^{*}, \ldots, \boldsymbol{X}_{nb}^{*}\right)^{\top} \in \mathbb{R}^{n\times m}$.

对于一组多元随机向量 $\boldsymbol{z} = \{z_j : j \geq 1\}$ 和整数 $a < b$，令 $\mathcal{Z}_{a}^{b}$ 为由 $\{z_j : j \in [a, b]\}$ 生成的 σ-代数。对于每个 $s \geq 1$，定义 α-混合系数 $\alpha_{Z}(s) = \sup_{t\geq 1}\left\{|P(A\cap B) - P(A)P(B)| : A \in \mathcal{Z}_{1}^{t}, B \in \mathcal{Z}_{t+s}^{\infty}\right\}$。

下面陈述后文将用到的条件。

(A1) $\boldsymbol{X}_{ib}^{*} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_{b|a}\right)$ 以及 $\boldsymbol{\Sigma}_{b|a}$ 的对角元素均等于 1；

(A2) 存在一个常数 $\tau > 1$，使得 $\tau^{-1} < \lambda_{\min}\left(\boldsymbol{\Sigma}_{b|a}\right) \leqslant \lambda_{\max}\left(\boldsymbol{\Sigma}_{b|a}\right) < \tau$。存在 $r_1 > 0$，使得最大值 $_{1\leq i,\ j\leq m}\, \sigma_{ij}^{*} \leq r_1 \leq 1$。

(A3) 假设 $p = o\left(n^3\right), q = o(p)$ $q \leqslant n^{\delta}$，其中某个 $\delta \in \in (0, 1)$，且 $E\left(\left|\varepsilon_1\right|^{\ell}\right) < \infty$ 满足 $\ell = 14(1-\delta)^{-1}$。

(A4) $\left\{\left(X_{ij}^{*}, i = 1, \cdots, n\right) : 1 \leq j \leq m\right\}$ 是 α-混合的，且 $\alpha_{X^{*}}(s) \leq C\delta^{s}$，其中 $\delta \in \in (0, 1)$，$C > 0$ 为某个常数。

假设 (A1)–(A3) 与Feng 等人 (2024) 的定理 S2 中的条件一致，其本质要求是坐标间依赖不能过强。特别是，(A1) 中的高斯要求主要是为了在建立联合极限理论时，利用高斯向量所特有的条件独立性。未来研究的一个有趣且具有挑战性的方向，是放宽这一高斯性假设，以适用于一般的非高斯设计。尽管如此，我们的模拟结果表明，在多种非高斯分布下，所提出的方法仍然准确且具有较高检验力。

假设（A4）与Ma 等人（2024）中的条件（A3）相对应，引入该假设是为了便于运用预期损失（ES）极限理论，推导相关L统计量的渐近正态性。

现在我们陈述第一个主要结果，该结果涉及每个 $W_{(j)}$ 的极限分布以及所有 $W_{(j)}$ 的联合极限分布。

定理 1. 假设假设 (A1)-(A4) 成立。当 $\min(n, p) \to \infty$ 时，我们有 (i) 对所有整数 $1 \leq s \leq m$ 和 $x \in \mathbb{R}$,

$$\mathrm{P}_{H_0}\left(W_{(s)} - b_m \leq x\right) \to \Lambda(x)\sum_{i=0}^{s-1}\frac{\{\log\Lambda^{-1}(x)\}^{i}}{i!},$$

(ii) 对所有整数 $2 \leq k \leq m$ 和 $x_1 \geq \ldots \geq x_k \in \mathbb{R}$

$$\begin{aligned}&\mathrm{P}_{H_0}\left(\bigcap_{j=1}^{k}\left(W_{(j)} - b_m \leq x_j\right)\right)\\ \to&\Lambda\left(x_k\right)\sum_{\sum_{i=2}^{j}k_i \leq j-1, j=2,\ldots,k}\prod_{i=2}^{k}\frac{\{\log\Lambda^{-1}(x_i) - \log\Lambda^{-1}(x_{i-1})\}^{k_i}}{k_i!},\end{aligned}$$

其中 $b_m = 2\log m - \log(\log\ m)$，且 $\Lambda(x) = \exp\{-\pi^{-1/2}\exp(-x/2)\}$。

根据定理 1，通过卷积，对于固定的 k ，L_k 的渐近零分布存在。特别地，$\mathrm{P}_{H_0}(L_1 - b_m \le x) \to \Lambda(x)$。当 $k > 1$ 时，T_k 的渐近零分布较为复杂。因此，我们采用自助法来获得 L_k的零假设分布。

我们通过野生自助法方案生成自助响应变量：

$$Y_i^* = \mathbf{X}_{ia}^\top \hat{\boldsymbol{\beta}}_a + \epsilon_i \hat{\varepsilon}_i, \qquad i = 1, \dots, n, \tag{3}$$

其中，$\hat{\boldsymbol{\beta}}_a = (\mathbf{X}_a^\top \mathbf{X}_a)^{-1} \mathbf{X}_a^\top \mathbf{Y}$ 是 H_0 下的最小二乘估计量，$\hat{\varepsilon}_i = Y_i - \mathbf{X}_{ia}^\top \hat{\boldsymbol{\beta}}_a$ 是相应的残差，而 $\{\epsilon_i\}_{i=1}^n$ 是独立同分布的 Rademacher 随机变量，以概率 $1/2$ 取值 ± 1 ，且与数据相互独立。给定自助样本 $\{(\mathbf{X}_{ia}, Y_i^*)\}_{i=1}^n$，我们计算检验统计量的自助法版本，记为 $L_{k,l}^*$ ，对应第 l 次自助法重复抽样。将式 (3) 独立重复 B 次，得到 $\{L_{k,l}^*\}_{l=1}^B$，并按如下方式计算 L_k 的 Monte Carlo p 值：

$$p_k = \frac{1}{B} \sum_{l=1}^{B} \mathbf{1}\big(L_{k,l}^* \ge L_k\big) . \tag{4}$$

2.2 发散参数 $k = \lceil \gamma m \rceil$

为检测稠密备择，我们考虑一种基于 L_k 的检验，其中 $k = \lceil \gamma m \rceil$ 发散，而 $\gamma \in$ 固定于 $(0, 1)$。定义 $\boldsymbol{U} = (U_1, \cdots, U_m)^\top \sim N(0, \boldsymbol{\Sigma}_{b|a})$。令 $v_\gamma = U^2_{(\lceil (1-\gamma)m \rceil + 1)}$ 为 $\{U_i^2\}_{1 \le i \le m}$ 的第 $(1-\gamma)$ 个样本分位数，v_γ 为 χ_1^2 的第 $(1-\gamma)$ 个分位数。定义 $\beta_j^* = \sum_{l=1}^m \sigma_{jl}^* \beta_{l,b}, j = 1, \cdots, m$ ，相应的前 k 个次序统计量为 $\beta^*_{(1)}, \cdots, \beta^*_{(m)}$。

定理 2. 在假设 (A1)-(A4)下，对于所有固定的 $\gamma_l \in (0, 1)$ $l = 1, \dots, s$ （ s 固定），若 $p = o(n^2)$ 且 $\gamma_l \log(p) n \sum_{i=1}^m \beta_{(i)}^{*2} = o(1)$，则当 $\min(n, p) \to \infty$ 时，

$$\left(\frac{L_{\lceil \gamma_1 m \rceil} - \mu_{\gamma_1} - n \sum_{i=1}^{\lceil \gamma_1 m \rceil} \beta_{(i)}^{*2}}{\sqrt{m \sigma_{\gamma_1 \gamma_1}}}, \dots, \frac{L_{\lceil \gamma_s m \rceil} - \mu_{\gamma_s} - n \sum_{i=1}^{\lceil \gamma_s m \rceil} \beta_{(i)}^{*2}}{\sqrt{m \sigma_{\gamma_s \gamma_s}}} \right) \xrightarrow{d} \mathcal{N}_s(\mathbf{0}, \boldsymbol{\Xi}),$$

其中

$$\begin{aligned}
\mu_\gamma :=& \sum_{i=1}^m \mathrm{E}_{H_0} \left\{ (U_i^2 - v_\gamma) \mathbb{I}(U_i^2 \ge v_\gamma) + \gamma v_\gamma \right\}, \\
\sigma_{\gamma_t \gamma_l} :=& \mathrm{cov}_{H_0} \left\{ \frac{1}{\sqrt{m}} \sum_{i=1}^m (U_i^2 - v_{\gamma_t}) \mathbb{I}(U_i^2 \ge v_{\gamma_t}), \frac{1}{\sqrt{m}} \sum_{i=1}^m (U_i^2 - v_{\gamma_l}) \mathbb{I}(U_i^2 \ge v_{\gamma_l}) \right\} \\
=& \frac{1}{m} \sum_{i=1}^m \mathrm{cov}_{H_0} \{ (U_i^2 - v_{\gamma_t}) \mathbb{I}(U_i^2 \ge v_{\gamma_t}), (U_i^2 - v_{\gamma_l}) \mathbb{I}(U_i^2 \ge v_{\gamma_l}) \} \\
&+ \frac{2}{m} \sum_{1 \le i < j \le m} \mathrm{cov}_{H_0} \{ (U_i^2 - v_{\gamma_t}) \mathbb{I}(U_i^2 \ge v_{\gamma_t}), (U_j^2 - v_{\gamma_l}) \mathbb{I}(U_j^2 \ge v_{\gamma_l}) \} = O(1),
\end{aligned}$$

且 $\boldsymbol{\Xi} = (\Xi_{tq})_{s \times s}$ 对 $1 \le t, q \le s$, $\Xi_{tq} = \sigma_{\gamma_t \gamma_l} / \sqrt{\sigma_{\gamma_t \gamma_t} \sigma_{\gamma_l \gamma_l}}$满足条件。

尽管发散的-k 统计量 $L_{\lceil \gamma m \rceil}$ 在H_0下满足渐近正态近似，但其极限均值和方差涉及干扰量（例如阈值相关矩

和长期协方差），这使得在有限样本中直接实施较为繁琐。因此，我们也采用（3）中的同一野生自助法程序来近似 $L_{\lceil \gamma m \rceil}$ 的零假设分布并计算其 p值。具体而言，对于每次自助法重复抽样 $l = 1, \ldots, B$，我们基于自助样本 $\{(\mathbf{X}_{ia}, Y_i^*)\}_{i=1}^n$ 计算相应的自助法统计量 $L^*_{\lceil \gamma m \rceil, l}$，然后定义

$$p_{\lceil \gamma m \rceil} = \frac{1}{B} \sum_{l=1}^{B} \mathbf{1}\Big(L^*_{\lceil \gamma m \rceil, l} \geq L_{\lceil \gamma m \rceil} \Big). \tag{5}$$

由于固定k 统计量已通过同一自助法流程加以计算，针对多个 γ 取值纳入$L_{\lceil \gamma m \rceil}$ 只需很少的额外编码，并能在不同 k选择下得到统一且稳定的校准。

2.3 自适应程序

在实践中，备择假设的稀疏度未知，因此 L统计量中截断参数 k 的选择本质上依赖于数据。较小的 k 通常对稀疏备择（少数较强坐标）更为敏感，而较大的 k 在稠密备择（许多较弱坐标）下往往更为有效。由于没有任何单一的 k 在这些情形下一致最优，自然的做法是通过综合一组 L统计量 $\{L_k\}$ 在一系列 k 取值上的证据来构造自适应程序。

然而，这种多k 聚合需要理解零假设下$\{L_k\}$ 之间的依赖结构。若缺乏对这些联合关系的精确刻画，简单组合可能因邻近 k之间的强相关而导致过于保守或膨胀的第一类错误。下一定理通过建立相关次序统计量过程与由固定 γ索引的标准化 L统计量的联合弱收敛来解决这一问题，进而蕴含渐近独立性，从而为有原则地组合多个 k提供依据。

定理3. 假设假设 (A1)–(A4) 成立。如果 $p = o(n^2)$，那么当 $\min(n, p) \to \infty$ 时，在H_0条件下，我们有
(i) 对所有整数 $1 \leq s \leq p$ 和固定的 $\gamma \in (0, 1)$,

$$\mathrm{P}_{H_0}\left(W_{(s)} - b_m \leq x, \frac{L_{\lceil \gamma m \rceil} - \mu_\gamma}{\sqrt{m\sigma_{\gamma\gamma}}} \leq y \right) \to \Lambda(x) \sum_{i=0}^{s-1} \frac{\{\log \Lambda^{-1}(x)\}^i}{i!} \Phi(y),$$

(ii) 对所有整数 $2 \leq k \leq p$ 和固定的 $\gamma \in (0, 1)$,

$$\mathrm{P}_{H_0}\left(\bigcap_{j=1}^{k} \left(W_{(j)} - b_m \leq x_j \right), \frac{L_{\lceil \gamma m \rceil} - \mu_\gamma}{\sqrt{m\sigma_{\gamma\gamma}}} \leq y \right)$$
$$\to \Lambda\left(x_k\right) \sum_{\sum_{i=2}^{j} k_i \leq j-1, j=2,\ldots,k} \prod_{i=2}^{k} \frac{\{\log \Lambda^{-1}(x_i) - \log \Lambda^{-1}(x_{i-1})\}^{k_i}}{k_i!} \Phi(y).$$

受定理 3的启发，我们汇总一组由截断水平的二进序列索引的 L-统计量的信息。我们选择

$$k_i = \left\lceil 2^{-i} m \right\rceil, \qquad i = 1, \ldots, K, \quad \text{where} \quad K = \left\lfloor \frac{\log(m/20)}{\log 2} \right\rfloor,$$

使网格覆盖从中度稀疏到相对稠密的范围，同时将最小截断水平保持在至少约 20 的量级。对于每个选定的 k_i，令 p_{k_i} 为相应的 p-值，其对应于 L_{k_i} （按（5）中所述的方法计算）。此外，令 p_1 和 p_{10} 分别表示与两个固定极值型分量相关联的 p-值（例如， $W_{(1)}$ 和 $W_{(10)}$；精确定义见（4））。随后，我们定义柯西组合统计量

$$T_C = \frac{1}{K+2}\tan\left\{\left(\frac{1}{2}-p_1\right)\pi\right\} + \frac{1}{K+2}\tan\left\{\left(\frac{1}{2}-p_{10}\right)\pi\right\} + \frac{1}{K+2}\sum_{i=1}^{K}\tan\left\{\left(\frac{1}{2}-p_{k_i}\right)\pi\right\}. \quad (6)$$

定理3 表明，在 H_0下，标准化的 L统计量与由 $\{W_j\}$的上阶统计量所产生的极值分量渐近独立；对于由固定 γ索引的有限集合，同样的解耦性质也成立。因此，若 $T_C \leq \alpha$，我们便拒绝零假设。因此，通过柯西变换将$\{p_{T_{k_i}}\}_{i=1}^{K}$ 与两个极值型 p值结合起来，如（6）所示，可在广泛的备择假设下提供稳健且有力的综合检验程序。此外，众所周知，当只有少数分量检验显著时，柯西组合仍能保持良好的检验效能；而当信号分布更分散时，它也能保持稳定，因此非常适合针对不同稀疏程度进行自适应检验。

3. 模拟

本节将我们的方法（简称 CC）与以下四种方法进行比较：（1）Feng 等人（2024）提出的 MAX检验；（2）Goeman 等人（2006）提出的 EB检验；（3）Lan 等人（2014）提出的 PF检验；（4）Feng 等人（2024）提出的 COM检验。

我们根据（1）生成数据，其中， β_j $j \in \{1, 2, \cdots, q\}$ 的回归系数由标准正态分布模拟得到，然后设定 $\beta_j = 0$ $j > q$。此外，对于 $i = 1 \cdots$，预测变量向量为 $\boldsymbol{X}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$ n，且 $\boldsymbol{z}_i$的每个分量均独立地从以下三种分布中生成：(i) 正态分布 $N(0, 1)$；(ii) 指数分布 exp(1) -1；(iii) 混合正态分布$V/\sqrt{1.8}$，其中 V服从 $0.9N(0, 1) + 0.1N(0, 9)$。我们基于 1000 次蒙特卡洛重复试验，报告名义水平为 0.05 时的经验尺寸。我们考虑（n， p） $\in \{100, 200\} \times \{200, 400, 600\}$，并固定 $q = 5$ 个基线协变量。协变量具有 AR(1)相关结构 $\boldsymbol{\Sigma} = (0.7^{|i-j|})_{1 \leq i,\ j \leq p}$。

表1 报告了上述五种检验的经验尺寸。总体而言，MAX 在所有设定下都较为保守，在大多数情况下经验尺寸远低于 5%。在高斯设计下，EB 和 PF 的尺寸控制效果均令人满意；在非高斯设计下，两者总体上也表现良好，但在某些重尾或偏斜配置中观察到轻微膨胀。COM 方法在三种设计下的第一类错误控制最为稳定，但略显保守，这与其更严格的组合规则一致。相比之下，CC 在所有情况下都能很好地控制经验尺寸。

我们使用 $n = 100, p = 200$比较这些检验的功效。定义 $\boldsymbol{\beta}_b = \kappa \cdot (\beta_{q+1}, \cdots, \beta_p)^T$。令 s表示非零系数数量。对于 $s = 1, \cdots, m$，我们考虑 $\beta_j \sim N(0, 1)$ $q < j \leqslant q + s$和 $\beta_j = 0, j > q + s$。参数 κ 的选取使得 $\|\boldsymbol{\beta}_b\|^2 = 0.8$。

图1 报告了经验检验功效随非零系数数量变化的情况 s。可以观察到明显的稀疏性–密度权衡。当备择假设极其稀疏（ s非常小时）时，最大型程序（最大值）具有最高的检验功效，体现了其对少数强信号的敏感性。然而，随着 s 增大，其检验功效持续下降，在中度稀疏至稠密备择假设下显著逊色。相比之下，EB 和 PF 在

表 1：五种检验方法的经验尺寸。

	(i)			(ii)			(iii)		
p	200	400	600	200	400	600	200	400	600
				$n = 100$					
MAX	2.7	3.0	1.8	3.0	2.0	2.9	2.9	2.8	2.3
EB	4.8	5.3	3.4	4.4	4.7	5.3	5.8	6.2	5.4
PF	5.7	5.4	3.8	5.5	4.4	6.1	5.5	6.7	5.5
COM	3.7	3.9	2.2	4.1	3.7	4.3	4.4	5.0	4.4
CC	5.9	6.0	4.4	5.2	6.3	6.2	5.9	5.5	5.0
				$n = 200$					
MAX	3.4	3.5	2.9	2.5	2.5	1.8	3.5	2.9	2.1
EB	4.2	4.8	4.9	6.0	4.9	4.8	5.1	5.4	3.7
PF	4.6	4.9	4.9	6.6	5.3	4.9	5.8	5.5	3.9
COM	3.6	4.7	4.1	4.3	4.4	3.0	4.6	4.6	3.4
CC	5.2	5.5	5.1	6.1	5.4	5.3	6.4	6.9	5.3

非常小时 s 检验功效较低，但随着 s 增大而迅速提升，并在信号达到中度稠密后稳定在较高水平，这表明它们在聚合许多弱至中等效应方面具有优势。组合方法 COM 和 CC 在不同稀疏度情形下更为稳健：其功效曲线在较宽的 s范围内相对稳定。CC 始终具有竞争力，并且在中等和较大的 s下通常表现最佳或接近最佳；COM 则略显保守，但仍保持较强的检验功效。总体而言，结果支持自适应性的动机：最大型检验在稀疏信号下表现出色，求和/聚合型检验在稠密信号下表现出色，而组合策略（尤其是 CC）在稀疏度未知时能持续保持出色表现。

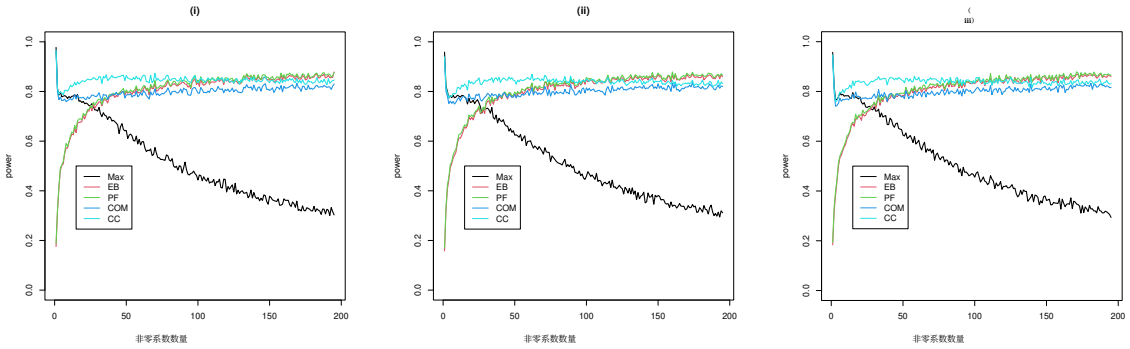

图1：在不同分布下，各检验的功效曲线， $n = 100, p = 20$ 0.

4. 结论

本文提出了一类基于 L 统计量的检验方法，用于检验高维线性模型中的回归系数，并基于柯西组合提出了一种自适应综合检验。理论分析和数值研究均表明，在备择假设下，该检验在广泛的稀疏程度范围内都具有强大且稳定的性能。

有几个方向值得进一步研究。首先，我们的部分极限理论依赖于协变量满足高斯性假设 X，这一假设主要用于简化依赖性论证。放宽这一要求，并在更一般的非高斯设计下（例如椭圆设计或其他弱依赖框架下）建立类似结果，将具有重要意义（Guo 等人，2016）。其次，我们当前的研究对回归误差施加了矩条件，以控制逐坐标证据度量的尾部行为。一个自然的拓展方向是通过以稳健方法取代经典的得分型构造，来适应尾部更重的误差。具体而言，可以采用基于秩的得分检验，例如Feng 等人 (2013)；Xu (2017)；Xu 和 Zhou (2021)提出的检验，以提高稳健性，同时保留高维情形下的检验效能。最后，将我们的框架推广到线性模型之外也值得研究。一个自然的方向是发展适用于广义线性模型的对应方法，沿着Goeman 等人 (2006)的全局检验研究路线，以及Yang 等人(2024)的近期进展。另一个有前景的方向是将所提出的 L统计量及组合策略推广到非线性回归情形，包括针对非参数模型或基于核的模型的现代检验方法；例如参见Li 等人 (2023)和Yin 等人 (2026)。

5. 附录

5.1 定理 1的证明

定义 $\boldsymbol{u}_j = (X_{1j}^*, \cdots, X_{nj}^*)^\top$ 和 $\boldsymbol{e}_2 = (\mathbf{I}_n - \mathbf{H}_a)\boldsymbol{\varepsilon} / \|(\mathbf{I}_n - \mathbf{H}_a)\boldsymbol{\varepsilon}\|$。根据假设 (A1)，$\{\mathbf{u}_j;\ 1 \le j \le m\}$与 $\boldsymbol{X}_a$ 和 $\boldsymbol{\varepsilon}$相互独立，因此也与单位随机向量 $\mathbf{e}_2$相互独立。于是，根据Feng 等人（2024）中的引理 S16，我们有 $\left(\mathbf{u}_1^T\mathbf{e}_2, \cdots, \mathbf{u}_m^T\mathbf{e}_2\right)^T$服从分布 $N\left(\mathbf{0}, \boldsymbol{\Sigma}_{b|a}\right)$，且也与 $\|(\mathbf{I}_n - \mathbf{H}_a)\boldsymbol{\varepsilon}\|$相互独立。定义 $\tilde{W}_{(k)}$ 为 $\{\tilde{Z}_i^2\}_{i=1}^m$ 和 $\tilde{Z}_i = \mathbf{u}_i^\top \boldsymbol{e}_2$中的前$k$ 个次序统计量。

按照Feng 等人（2024）证明定理 S2 时采用的相同分解方法，可以将 $\tilde{\mathbf{X}}_{\cdot j} = (\mathbf{I}_n - \mathbf{H}_a)\mathbf{u}_j$ 表示为 $\mathbf{u}_j \sim N(0, \mathbf{I}_n)$。于是关键恒等式

$$(\tilde{\mathbf{X}}_{\cdot j}^\top \boldsymbol{\varepsilon})^2 = \|(\mathbf{I}_n - \mathbf{H}_a)\boldsymbol{\varepsilon}\|^2 \cdot (\mathbf{u}_j^\top \boldsymbol{e}_2)^2 \tag{7}$$

成立。利用 $\hat{\boldsymbol{\varepsilon}} = (\mathbf{I}_n - \mathbf{H}_a)\boldsymbol{\varepsilon}$ 和 $\hat{\sigma}^2 = r^{-1}\|(\mathbf{I}_n - \mathbf{H}_a)\boldsymbol{\varepsilon}\|^2$, $r = n - q$，我们得到

$$Z_j^2 = \frac{(\tilde{\mathbf{X}}_{\cdot j}^\top \hat{\boldsymbol{\varepsilon}})^2}{\hat{\sigma}^2 \|\tilde{\mathbf{X}}_{\cdot j}\|^2} = \frac{\|(\mathbf{I}_n - \mathbf{H}_a)\boldsymbol{\varepsilon}\|^2 (\mathbf{u}_j^\top \boldsymbol{e}_2)^2}{\hat{\sigma}^2 \|\tilde{\mathbf{X}}_{\cdot j}\|^2} = \frac{r\,(\mathbf{u}_j^\top \boldsymbol{e}_2)^2}{\|\tilde{\mathbf{X}}_{\cdot j}\|^2}.$$

定义 $A_j := r^{-1}\|\tilde{\mathbf{X}}_{\cdot j}\|^2$. 则我们得到便捷表示

$$W_j = Z_j^2 = \frac{\tilde{W}_j}{A_j}. \tag{8}$$

给定 X_a（等价地，给定 $\mathrm{P} = \mathrm{I}_n - \mathrm{H}_a$），$\mathrm{P}u_j$ 服从协方差为 P 的高斯分布。由于 P 是秩为 r 的正交投影，我们有标准恒等式

$$rA_j = \|Pu_j\|^2 \overset{d}{=} \chi^2_r, \qquad \text{conditionally on } \mathbf{P}. \tag{9}$$

接下来证明最大值$\max_{1\le j\le m} |A_j - 1|$ 很小。由 χ^2_r的 Laurent–Massart 尾界，对任意$t > 0$,

$$\mathrm{P}\Big(\chi^2_r - r \ge 2\sqrt{rt} + 2t\Big) \le e^{-t}, \qquad \mathrm{P}\Big(r - \chi^2_r \ge 2\sqrt{rt}\Big) \le e^{-t}.$$

将其应用于（9），对每个固定的 j,

$$\mathrm{P}\left(|A_j - 1| \ge 2\sqrt{\frac{t}{r}} + \frac{2t}{r}\right) \le 2e^{-t}.$$

令 $t = c\log\, m$，其中 $c > 2$为常数。对 $j = 1, \ldots, m$应用并集界,

$$\mathrm{P}\left(\max_{1\le j\le m} |A_j - 1| \ge \delta_{m,r}\right) \le 2me^{-c\log m} = 2m^{\,1-c} \to 0, \tag{10}$$

其中

$$\delta_{m,r} := 2\sqrt{\frac{c\log m}{r}} + \frac{2c\log m}{r}. \tag{11}$$

因此,

$$\Delta_n := \max_{1\le j\le m} |A_j - 1| = O_p\left(\sqrt{\frac{\log m}{r}}\right). \tag{12}$$

在事件 $\{\Delta_n \le 1/2\}$发生时，对每个 j，都有 $1 - \Delta_n \le A_j \le 1 + \Delta_n$，因此

$$\frac{1}{1+\Delta_n}\tilde{W}_j \le \frac{\tilde{W}_j}{A_j} \le \frac{1}{1-\Delta_n}\tilde{W}_j.$$

利用（8）以及次序统计量的单调性，该逐分量不等式蕴含了对于 k阶次序统计量的相同不等式:

$$\frac{1}{1+\Delta_n}\tilde{W}_{(k)} \le W_{(k)} \le \frac{1}{1-\Delta_n}\tilde{W}_{(k)}. \tag{13}$$

因此,

$$|W_{(k)} - \tilde{W}_{(k)}| \le \max\left\{\frac{1}{1-\Delta_n} - 1,\ 1 - \frac{1}{1+\Delta_n}\right\}\tilde{W}_{(k)} \le \frac{\Delta_n}{1-\Delta_n}\tilde{W}_{(k)} \le 2\Delta_n\,\tilde{W}_{(k)}, \tag{14}$$

˜其中最后一个不等式在 $\{\Delta_n \le 1/2\}$ 上成立。我们证明 $W_{(k)} = O_p(\log m)$，这对于固定的 k 已经足够。事实上，由并集界（无需独立性），对任意 $t > 0$,

$$\mathrm{P}\Big(\tilde{W}_{(1)} > t\Big) = \mathrm{P}\left(\max_{1\le j\le m} \tilde{Z}^2_j > t\right) \le \sum_{j=1}^{m} \mathrm{P}(\tilde{Z}^2_j > t) \le m\mathrm{P}(\chi^2_1 > t).$$

利用当 t 较大时 $\mathrm{P}(\chi_1^2 > t) \le 2e^{-t/2}$，只要 $C > 2$，我们便得到 $\mathrm{P}(\tilde{W}_{(1)} > C\log m) \le 2m^{1-C/2} \to 0$。因此 $W_{(1)} = O_p(\log m)$，从而

$$\tilde{W}_{(k)} \le \tilde{W}_{(1)} = O_p(\log m). \tag{15}$$

结合（14）、（12）和（15）可得

$$|W_{(k)} - \tilde{W}_{(k)}| = O_p\left(\sqrt{\frac{\log m}{r}}\right) \cdot O_p(\log m) = O_p\left(\frac{(\log m)^{3/2}}{\sqrt{r}}\right).$$

由于 $r = n - q \asymp n$ 和 $m = o(n^2)$，我们有 $\log m = O(\log n)$，因此 $(\log m)^{3/2}/\sqrt{r} \to 0$，这意味着

$$|W_{(k)} - \tilde{W}_{(k)}| = o_p(1). \tag{16}$$

因此，在假设 (A1)–(A4) 下，Ma 等人（2024）的定理 1 所要求的条件均得到满足，直接应用该定理即可立即得到所需的极限结果。 □

5.2 定理2的证明

在给定 $\mathbf{H}_a$ 的条件下，$\mathbf{X}_{\cdot j} = (\mathbf{I}_n - \mathbf{H}_a)\mathbf{u}_j$ 服从高斯分布，其协方差为 $(\mathbf{I}_n - \mathbf{H}_a)$，这是一个秩为 r 的正交投影。因此

$$rA_j = \|\tilde{\mathbf{X}}_{\cdot j}\|^2 \overset{d}{=} \chi_r^2, \qquad \text{conditionally on } \mathbf{H}_a.$$

因此，

$$E(A_j \mid \mathbf{H}_a) = 1, \qquad \mathrm{var}(A_j \mid \mathbf{H}_a) = \frac{2}{r}, \tag{17}$$

且当 r 足够大时，

$$E\left[(A_j^{-1} - 1)^2 \mid \mathbf{H}_a\right] = O(r^{-1}), \qquad E\left[|A_j^{-1} - 1| \mid \mathbf{H}_a\right] = O(r^{-1}). \tag{18}$$

此外，Feng 等人(2024）中定理 S2 的证明对这些范数项进行了统一控制，即最大值 $_{q+1\le j\le p}\left|1 - (r^{-1}\|\tilde{\mathbf{X}}_{\cdot j}\|^2)^{-1}\right|$ 以对数速率渐近可忽略，这意味着乘以或除以 A_j 不会实质性地改变次序统计量结构。

对于固定的 $\gamma \in (0, 1)$ 内，前 $\lceil \gamma m \rceil$ 项和是一个预期损失函数。施加假设（A4）正是为了对 $\{W_j\}$ 和 $\{W_j\}$调用 ES 极限理论。特别是，可以采用用于推导 $L_{\lceil \gamma m \rceil}$ (Ma 等人，2024）渐近正态性的 ES 近似步骤，写作

$$L_{\lceil \gamma m \rceil} = \sum_{j=1}^{m} W_j \, \mathbf{1}(W_j \ge \xi_\gamma) + o_p(m^{1/2}), \qquad \tilde{L}_{\lceil \gamma m \rceil} = \sum_{j=1}^{m} \tilde{W}_j \, \mathbf{1}\left(\tilde{W}_j \ge \tilde{\xi}_\gamma\right) + o_p(m^{1/2}), \tag{19}$$

其中 ξ_γ、ξ_γ 是相应的总体 $(1-\gamma)$ 分位数。

由（8），

$$W_j - \tilde{W}_j = \tilde{W}_j(A_j^{-1} - 1).$$

代入 t　　将其代入（19），并利用 $A_j - 1$的一致小量性，我们将问题归结为证明　　ng

$$\sum_{j=1}^{m} \tilde{W}_j (A_j^{-1} - 1)\mathbf{1}(\tilde{W}_j \geq \tilde{\xi}_\gamma) = o_p(m^{1/2}). \tag{20}$$

Let

$$\Delta_j := \tilde{W}_j (A_j^{-1} - 1)\mathbf{1}(\tilde{W}_j \geq \tilde{\xi}_\gamma).$$

根据柯西–施瓦茨不等式和（18），

$$E(\Delta_j^2 \mid H_a, e_2) \leq E(\tilde{W}_j^2 \mid e_2) \cdot E((A_j^{-1} - 1)^2 \mid H_a) = O(1) \cdot O(r^{-1}) = O(r^{-1}),$$

由于 $\tilde{W}_j = \tilde{Z}_j^2$ 的二阶矩一致有界。此外，由（$\Delta_j \mid H_a$， e_2） $= O$（r^{-1}），根据（18）。在（A1）–（A3）条件下， j 之间的依赖足够弱，因此

$$\operatorname{var}\left(\sum_{j=1}^{m} \Delta_j \,\middle|\, H_a, e_2\right) = O\left(\frac{m}{r}\right), \qquad E\left(\sum_{j=1}^{m} \Delta_j \,\middle|\, H_a, e_2\right) = O\left(\frac{m}{r}\right).$$

因此，根据切比雪夫不等式，

$$\sum_{j=1}^{m} \Delta_j = O_p\left(\frac{m}{r}\right) + O_p\left(\sqrt{\frac{m}{r}}\right).$$

由于 $r = n - q \asymp n$ 和 $m = o(n^2)$，我们有

$$\frac{m/r}{m^{1/2}} = \frac{m^{1/2}}{r} = o(1), \qquad \frac{\sqrt{m/r}}{m^{1/2}} = \frac{1}{\sqrt{r}} = o(1),$$

这证明了（20）。结合（19）可得

$$L_{\lceil \gamma m \rceil} = \tilde{L}_{\lceil \gamma m \rceil} + o_p(m^{1/2}), \tag{21}$$

其中 $\tilde{L}_{\lceil \gamma m \rceil} = \sum_{j=1}^{\lceil \gamma m \rceil} \tilde{W}_{(j)}$。因此，由于（$\tilde{Z}_1$， $\cdots$， $\tilde{Z}_m$）$^\top \sim N$（$\mathbf{0}$， $\mathbf{\Sigma}_{b|a}$），根据 Ma 等人（2024）中的定理3，我们即可得到该结果。 □

5.3 定理 3的证明

由（16）和（21），根据 Ma 等人（2024）中的定理4，我们可以容易地得到该结果。 □

参考文献

Cui, H.、W. Guo 和 W. Zhong（2018）。使用重拟合交叉验证方差估计检验高维回归系数。统计年刊 46 (3)，958–988。

Feng, L.、T. Jiang、X. Li 和 B. Liu（2024）。相依随机变量的和与最大值的渐近独立性及其在高维检验中的应用。《中华统计学志》34，1745–1763。

Feng, L.、C. Zou、Z. Wang 和 B. Chen（2013）。高维回归系数的基于秩的得分检验。《电子统计学期刊》7，2131–2149。Goeman, J. J.、S. A. van de Geer 和 H. C. van Houwelingen（2006）。针对高维备择假设的检验。《英国皇家统计学会杂志：B辑（统计方法）》68（3），477–493。Goeman, J. J.、H. C. van Houwelingen 和 L. Finos（2011）。广义线性模型中针对高维备择假设的检验：渐近 I 类错误控制。《生物统计学》98（2），381–390。Guo, W.、Y. Chen 和 H. Cui（2016）。使用重拟合交叉验证方差估计的高维回归系数稳健 u 型检验。《中国科学：数学》59（12），2319–2334。Guo, W.、W. Zhong、S. Duan 和 H. Cui（2022）。超高维线性回归系数的条件检验。《中华统计学志》32（3），1381–1409。Lan, W.、H. Wang 和 C.-L. Tsai（2014）。高维回归中的协变量检验。《统计数学研究所年刊》66（2），279–301。Li, R.、K. Xu、Y. Zhou 和 L. Zhu（2023）。通过累积协方差聚合检验高维协变量的效应。《美国统计学会杂志》118（543），2184–2194。Ma, H.、L. Feng 和 Z. Wang（2024）。高维检验问题的自适应 L 统计量。arXiv 预印本，arXiv:2410.14308。Xu, K.（2017）。一种新的高维回归系数非参数检验。《统计计算与模拟杂志》87（5），855–869。Xu, K. 和 Y. Zhou（2021）。使用 Wilcoxon 得分的高维回归系数最大型检验。《统计规划与推断杂志》211，221–240。Yang, W.、X. Guo 和 L. Zhu（2024）。一般协方差结构下广义线性模型的高维检验。《计算统计与数据分析》199，108026。Yin, H.、Y. Wang 和 A. Xu（2026）。高维情形下基于核的协变量效应边际检验。《斯堪的纳维亚统计学杂志》。Zhong, P.-S. 和 S. X. Chen（2011）。析因设计下的高维回归系数检验。《美国统计学会杂志》106（493），260–274。