arXiv:2601.10357v1 [stat.ME] 15 Jan 2026

监督学习中模型无关且不确定性感知的降维

Yue Yu[1],Guanghui Wang[1], Liu Liu[2], 和 Changliang Zou[11]南开大学统计与数据科学学院，天津，中国[2]成都理工大学数学科学学院，成都，中国

摘要

降维是监督学习中分析高维数据的基础工具。传统的内在序估计方法往往优先考虑特定于模型的结构假设，而非预测效用。本文提出预测阶数确定（POD），这是一种模型无关框架，通过直接评估样本外预测性来确定预测所需的最小维度。POD通过误差界量化高估和低估的不确定性，并在温和条件下实现一致性。通过将降维与预测性能相结合，POD可灵活应用于多种降维任务和监督学习器。模拟和真实数据分析表明，POD能够提供准确且不确定性感知的阶数估计，使其成为以预测为中心的流程中的通用组件。

1 引言

降维是现代统计学中的核心思想。它在保留可靠预测和推断所需信息的同时，为高维数据提供低维摘要（Jolliffe，2002年；Li，2018年）。在监督学习问题中，我们观测到预测变量 $X \in \mathbb{R}^p$ ，目标是尽可能少地使用变换变量来预测响应变量 $Y \in \mathbb{R}^q$ 。满足这一条件的最小数量称为内在序，记作 d^*。

1.1 动机示例

例*1*（因子回归）。潜在因子$f \in \mathbb{R}^{d_F^*}$ 驱动预测变量 X，并通过未知链接函数 m驱动响应变量 Y，其中 $q = 1$ （Fan 和 Gu，2024年；Fan

等，2024年）：

$$Y = m(f) + \varepsilon, \qquad X = Bf + u,$$

其中 B 是载荷矩阵，ε 和 u 是噪声。所有预测信息都包含在 f的 d_{F}^* 坐标中。当 m 是线性函数时，该模型就简化为主成分回归。

例 *2*（充分降维，充分降维）。可能存在一个矩阵 $B \in \mathbb{R}^{p\times d_{\mathrm{CS}}^*}$，其列数尽可能少，并满足 Y（其中 $q = 1$）在给定 X 的条件下与 $B^\top X$相互独立。B 的张成空间是中心子空间，其维度为 d_{CS}^*（Li，1991年；Cook 和 Weisberg，1991年）。仅关注 $\mathbb{E}$（$Y \mid X$）可得到中心均值子空间，其维度为 $d_{\mathrm{CMS}}^* \le d_{\mathrm{CS}}^*$（Cook 和 Li，2002年）。

例 *3*（降秩回归）。对于多元响应变量，

$$Y = A^\top X + \varepsilon, \qquad A = B\Omega,$$

其中 A 的秩为 $d_{\mathrm{R}}^* <$ 最小$\{p, q\}$，$B \in \mathbb{R}^{p\times d_{\mathrm{R}}^*}$，$\Omega \in \mathbb{R}^{d_{\mathrm{R}}^*\times q}$，且 ε 是噪声。所有预测信息都位于 d_{R}^* 的坐标中 $B^\top X$（Izenman，1975）。

尽管这些范式在机制上各不相同，但它们有着共同的目标：找到最小维度 d^*（例如，d_{F}^*、d_{CS}^*、d_{CMS}^*、d_{R}^*），使其保留 X 中有关 Y 的所有预测信息。这一共同点促使我们转换视角：我们不再依赖针对特定问题的结构诊断，而是直接衡量维度如何影响预测性能，以确定 d^*。

1.2 经典阶数估计方法的局限性

经典方法通常将阶数估计视为模型选择问题。基于惩罚似然或特征值阈值的信息准则，在因子模型（Bai 和 Ng，2002）、充分降维（Zhu 等，2006）和降秩回归（Bunea 等，2011）中仍广受欢迎；此外，也有人为充分降维提出了自适应惩罚（Luo 和 Li，2016）。其他常见方法包括交叉验证，应用于充分降维（Xia 等，2002）和降秩回归（Yuan 等，2007）；以及序贯检验，在充分降维中比较特定方法的核矩阵的特征值（例如，Li，1991；Bura 和 Yang，2011），或在因子模型中比较样本协方差矩阵的特征值（通常以无监督方式进行）（例如，Onatski，2009；Yu 等，2025）。尽管缺乏充分的理论支持，碎石图或回归图等直观启发式方法在实践中仍然常见（Cattell，1966；Cook，1994）。

这些方法通常优先考虑结构假设（例如特征值衰减或分布条件），而非监督学习的最终目标：预测准确性。虽然它们针对特定的任务、模型和

算法组合提供点估计，但缺乏一个统一框架，无法从预测充分性的角度评估维度。因此，它们对下游预测任务的指导有限，而且往往无法以系统的方式量化不确定性。近期研究凸显了不确定性量化在阶数估计中的实际重要性。例如，在因子回归中，因子数量的轻微高估可能影响不大（Fan 和 Liao，2022年）；然而，低估可能导致遗漏重要的预测信号，而严重高估则会累积噪声，两者都会损害预测性能。此外，随着现代数据分析流程越来越多地采用深度神经网络等复杂的非线性降维方法（Fan 和 Gu，2024年；Chen 等人，2024年b），传统的基于特征值的诊断方法会因由此产生的复杂谱行为而变得不可靠。这些挑战要求建立一个明确关联降维与预测性能、提供模型无关的不确定性量化，并优先进行监督感知的维度选择的框架。

1.3 我们的贡献

为应对这些挑战，我们提出预测阶数确定（*POD*），并取得了两项主要进展。

- **通过监督实现预测最优维度**。对于任意候选维度d，POD 都会评估预测性差距，即仅使用 d 个维度而非完整表示时，样本外预测风险的增幅。在温和条件下，使这一差距消失的最小 d 即为 d^*。因此，预测性量化了额外坐标对监督任务的效用，并将阶数确定与特定算法或分布假设解耦。

- **误差界与一致性保证**。POD 通过交叉拟合预测性比较，依次检验给定维度是否足够。该程序 (i) 将高估 d^* 的概率控制在预先指定的水平 α，(ii) 对低估的概率给出界限，(iii) 在正则条件下实现一致性。

1.4 相关文献

预测驱动的阶数选择已有研究，但大多局限于特定模型框架。在具有多指标结构的充分降维中，Xia 等人 (2002年) 提出了最小平均方差估计，通过反复进行局部线性平滑并更新指标，来估计中心均值子空间；结构维数则通过交叉

验证来选择，从而得到一致点估计量。在多元线性回归中，Yuan 等 (2007) 通过惩罚最小二乘法获得降秩估计量，并采用广义交叉验证作为实用规则来选择惩罚项，进而确定秩。尽管这些方法着眼于预测，但仍依赖特定模型和估计量，且对所选阶数的不确定性量化能力有限。相比之下，我们的方法与模型和算法无关，并且能够在一致性之外提供错误率控制。

Lei （2020年）提出了带置信度的交叉验证方法，该方法以固定的候选模型集为条件，并返回一个由近乎最优模型组成的置信集。然而，当候选风险相等时，其检验统计量可能会退化 （Chen 等人，2024年a）。相比之下，POD 的目标是按维度索引的总体风险，不依赖于具体模型和学习算法，即使在这种退化情况下也能保证错误率控制。

与之并行的、与算法无关的变量重要性 研究对预测性对比进行了形式化：比较使用全部特征时的总体风险与省略某个子集时的风险，并针对这一目标开发非参数、高效的估计量和有效检验（Williamson 等人，2023；Dai 等人，2024年）。我们的目标和研究设定都与此不同。我们不问“哪些特征重要”，而是问“多少个维度就足够”。POD 通过检测预测性随维度增加趋于平稳的位置，推断最小预测阶数。此外，与直接作用于观测预测变量的变量重要性框架不同， XPOD 通过一种未观测到的低维表示来评估预测性，并将其作为代理学习得到。因此，我们的理论量化了表示误差如何通过交叉拟合预测性差距和序贯检验传递，进而实现不确定性控制（对高估和低估的界定），并在温和的正则性条件下保证一致性。

1.5 结构

本文其余部分的结构如下。第2 节将形式化预测性诱导的序概念，开发我们的交叉拟合检验，并介绍用于估计 d^*的序贯程序。第3节建立误差界和一致性结果。第4 节给出模拟研究和真实数据示例。第5 节总结全文。技术证明和其他示例见补充材料。

记号。对于正整数 n，记 $[n] = \{1, \ldots, n\}$。对于向量 $x \in \mathbb{R}^p$ 和指标集 $l \subseteq [p]$，令 x_l 收集 l 中的坐标，令 x_{-l} 收集其余坐标。欧几里得范数为 $\|x\|_2 = (\sum_{i=1}^{p} x_i^2)^{1/2}$。对于事件 A， $1\{A\}$ 表示其示性函数。对于$a \in \mathbb{R}$，令 $\lfloor a \rfloor$ 表示小于或等于 a的最大整数。符号 δ_z 表示在 z处的点质量分布。

2 方法论

2.1 预测性诱导的序

2.1.1 Oracle 与代理表示

设 $(X,\ Y) \sim \mathbb{Q}$ 定义在 $\mathcal{X}\times\mathcal{Y}$上。我们假设存在一个未观测到的预言表示

$$R^* \in \mathbb{R}^{d^*},$$

内在序未知的 d^*，使得 R^* 包含 X 中与预测 Y相关的全部信息。该设定涵盖例如因子回归，其中 $R^* = f$（示例 1），以及采用 $R^* = B^\top X$ 的充分降维或降秩回归（示例 2 和 3）。为了在共同的环境空间中比较候选阶数，我们固定一个由用户选定的上界 $d_{\max} \geq d^*$，并将 R^* 通过零填充嵌入 $\mathbb{R}^{d_{\max}}$：

$$R \in \mathbb{R}^{d_{\max}} \text{ with } R_{[d^*]} = R^* \text{ and } R_{-[d^*]} = 0.$$

令 $Z = (R,\ X,\ Y)$，并令 $\mathbb{P}$ 表示 Z的联合分布。分布 $\mathbb{P}$ 仅用于定义涉及 R的总体量；观测数据是从 $\mathbb{Q}$中抽取的独立同分布（i.i.d.）样本。

在实际应用中，未观测到 R。我们改为构造一个依赖于数据的代理表示

$$\widehat{R} = \widehat{\varphi}(X) \in \mathbb{R}^{d_{\max}},$$

其中 $\widehat{\varphi}:\ \mathcal{X} \to \mathbb{R}^{d_{\max}}$ 由选定的降维方法生成；例如，将 X 投影到样本协方差矩阵或核矩阵的主特征向量上（Xia 等人，2002年；Fan 和 Gu，2024年）。我们假设 R 的坐标按降维方法返回的重要性分数排序，例如按样本特征值递减排序（Bai 和 Ng，2002年），因此排在前面的坐标旨在包含更多信息。

2.1.2 预测性

设 $\mathcal{G} = \{g:\ \mathbb{R}^{d_{\max}} \to \widehat{\mathcal{Y}}\}$ 为一类预测规则（通常为 $\widehat{\mathcal{Y}} = \mathcal{Y}$）。对于 $Y \in \mathcal{Y}$ 和 $\widehat{Y} \in \widehat{\mathcal{Y}}$，令 $\ell(Y,\ \widehat{Y})$ 为损失函数。例如，对于满足 $\mathcal{Y} = \widehat{\mathcal{Y}} = \mathbb{R}^q$ 的连续响应，平方损失 $\ell(Y,\ \widehat{Y}) = \|Y - \widehat{Y}\|_2^2$；对于满足 $\mathcal{Y} = \widehat{\mathcal{Y}} = \{0,1\}$ 的二元分类，0-1 损失 $\ell(Y,\ \widehat{Y}) = 1\{Y \neq \widehat{Y}\}$；以及当 $\mathcal{Y} = [M]$ 且 $\widehat{\mathcal{Y}} = [0,1]^M$时用于多类别结果的交叉熵 $\ell(Y,\ \widehat{Y}) = -e_Y^\top \log \widehat{Y}$，其中 e_m为第 m 个标准基向量。对于 $g \in \mathcal{G}$，定义总体风险

$$\mathcal{L}(g, \mathbb{P}) = \mathbb{E}_{\mathbb{P}}\{\ell(Y, g(R))\},$$

数值越小，预测性能越好。

对于 $0 \le d \le d_{\max}$，定义一类规则，它们仅通过 R 的前 d 个坐标来依赖 R：

$$\mathcal{G}_d = \left\{ g \in \mathcal{G} : g(r) = g(r') \text{ whenever } r_{[d]} = r'_{[d]} \right\}.$$

我们将阶数为 d 时的预测性定义为

$$\mathcal{L}_d = \min_{g \in \mathcal{G}_d} \mathcal{L}(g, \mathbb{P}),$$

并令

$$g_d \in \arg\min_{g \in \mathcal{G}_d} \mathcal{L}(g, \mathbb{P})$$

表示一个极小化解。

命题 1。 $\mathcal{L}_0 \ge \mathcal{L}_1 \ge \cdots \ge \mathcal{L}_{d^*} = \mathcal{L}_{d^*+1} = \cdots = \mathcal{L}_{d_{\max}}$。

命题1 表明， $\{\mathcal{L}_d\}_{d=0}^{d_{\max}}$ 是非递增的，并且在纳入所有信息性坐标后变为常数。因此，我们定义目标阶数

$$d^*(\mathcal{L}) = \min\{0 \le d \le d_{\max} : \mathcal{L}_d = \mathcal{L}_{d_{\max}}\}.$$

当最终目标面向预测时，这一定义本身就具有内在意义，因为它直接量化了实现最优监督学习性能所需的最小复杂度。在温和正则条件和适当的损失函数下， $d^*(\mathcal{L})$ 在动机示例中与内在序重合。对于采用平方损失的连续响应， $d^*(\mathcal{L}) = d^*_{\mathrm{F}}$ 在例1中， $d^*(\mathcal{L}) = d^*_{\mathrm{CMS}}$ 在例2中，以及 $d^*(\mathcal{L}) = d^*_{\mathrm{R}}$ 在例 3中；而采用负对数似然时， $d^*(\mathcal{L}) = d^*_{\mathrm{CS}}$ 在例2中。对于分类响应，交叉熵同样意味着 $d^*(\mathcal{L}) = d^*_{\mathrm{CS}}$ 在例2中。有关证明和其他示例，请参见补充材料的 B.1节。下文中，我们记作$d^* = d^*(\mathcal{L})$，并省略对 $\mathcal{L}$的显式依赖。

备注 1. 在例 2中，平方损失函数针对条件均值，因此当中心均值子空间严格包含于中心空间时，可能产生 $d^*(\mathcal{L}) = d^*_{\mathrm{CMS}} < d^*_{\mathrm{CS}}$ 。这反映了条件均值独立与条件独立之间的区别（Lundborg 等人，2024年）。当目标是在所选损失函数下进行预测时，$d^*(\mathcal{L})$ 仍是恰当的目标。

2.2 检验候选阶数是否充分

对于每个 $d \in \{0, \ldots, d_{\max}\}$，我们检验前 d 个坐标是否已达到最大预测性：

$$H_{0,d} : \mathcal{L}_d = \mathcal{L}_{d_{\max}} \quad \text{versus} \quad H_{1,d} : \mathcal{L}_d > \mathcal{L}_{d_{\max}}.$$

等价地，我们检验预测性差距 $\mathcal{L}_d - \mathcal{L}_{d_{\max}}$ 是否等于零。

2.2.1 交叉拟合程序

设 $\{(X_i,\ Y_i)\colon\ i \in [n]\}$ 是来自 $\mathbb{Q}$ 的独立同分布观测。预言表示 R 和最优预测规则 g_d 和 $g_{d_{\max}}$ 均不可观测。因此，我们学习一个降维映射，计算代理表示，并在代理坐标上拟合规则。如果我们在用于训练的同一数据上评估预测性能，所得估计通常会有偏，且其渐近行为难以分析。为避免这种"双重使用数据"，我们采用 K折 交叉拟合 (Chernozhukov 等人，2018年；Dai 等人，2024年；Williamson 等人，2023）。

将 $[n]$ 划分为 $K \geq 2$ 个互不相交的折 I_1，……，I_K。对于每个 $k \in [K]$，令 $I_{-k} = [n] \setminus I_k$表示训练索引，并令 I_k 表示测试索引。

1. **在 I_{-k} 上拟合**。仅使用 I_{-k}中的观测，拟合降维映射 $\widehat{\varphi}_{-k}$，并为所有 $\widehat{R}_i^{(k)} = \widehat{\varphi}_{-k}(X_i)$ 计算代理表示。学习两个预测规则$\widehat{g}_{-k;d}$ 和 $\widehat{g}_{-k;d_{\max}}$：分别通过回归 Y_i 到前 d 个和全部 $d_{\max}$ 坐标，来拟合 $\widehat{R}_i^{(k)}$ 。降维映射可根据具体的降维任务和方法进行调整；详细示例见《补充材料》第 C.1 节。预测规则可使用灵活的学习器进行拟合，例如核函数、样条、树模型和神经网络。

2. **在 I_k 上，**第 2.2.2 节说明如何仅在留出折 I_k 上评估风险，以构造非退化的检验统计量。

2.2.2 避免退化的第二次划分

对于拟合得到的降维映射 $\widehat{\varphi}_{-k}$，定义条件风险

$$\widehat{\mathcal{L}}_{-k}(g, \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\{\ell(Y, g(\widehat{\varphi}_{-k}(X))) \mid \widehat{\varphi}_{-k}\}.$$

对差值 $\mathcal{L}_d - \mathcal{L}_{d_{\max}}$ 的朴素代入估计量会在同一个检验折 I_k 上评估两种风险：

$$\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{Q}}_k) - \widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d_{\max}}, \widehat{\mathbb{Q}}_k),$$

其中 $\widehat{\mathbb{Q}}_k = |I_k|^{-1} \sum_{i \in I_k} \delta_{(X_i, Y_i)}$ 是定义在 I_k上的经验分布。然而，在原假设 $H_{0,\ d}$下，这一对比是退化的：两项风险评估的主要影响函数项相同，因此相互抵消（参见补充材料中的（A.1））。在变量重要性推断中也会出现类似的退化现象（例如，Dai 等人，2024年；Williamson 等人，2023；Chen 等人，2024年a）。为获得具有良好检验功效的非退化统计量，我们进一步采用受Chen 等人（2024年a）启发的重叠方案，将每个检验折进一步拆分。

将 I_k 随机划分为三个互不相交的部分 $I_{k,a}$、$I_{k,b}$、$I_{k,o}$，其中 $|I_{k,a}| = |I_{k,b}|$。子集 $I_{k,o}$ 由两项风险评估共享，我们用重叠比例 $\tau = |I_{k,o}|/(|I_{k,a}| + |I_{k,o}|) \in [0,1)$ 概括共享程度。对于 $J \in \{a,b,o\}$，令 $\mathbb{Q}_{k,J} = |I_{k,J}|^{-1} \sum_{i \in I_{k,J}} \delta_{(X_i,Y_i)}$ 为定义在 $I_{k,J}$ 上的经验分布。定义折特定风险估计

$$\widehat{\mathcal{L}}_{d,k} = \tau \widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{Q}}_{k,o}) + (1-\tau)\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{Q}}_{k,a}) \text{ and}$$
$$\widehat{\mathcal{L}}_{d_{\max},k} = \tau \widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d_{\max}}, \widehat{\mathbb{Q}}_{k,o}) + (1-\tau)\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d_{\max}}, \widehat{\mathbb{Q}}_{k,b}),$$

以及交叉拟合的折聚合对比

$$\widehat{\psi}_d = \frac{1}{K}\sum_{k=1}^{K}(\widehat{\mathcal{L}}_{d,k} - \widehat{\mathcal{L}}_{d_{\max},k}). \tag{1}$$

在 $H_{0,d}$，定理 1表明

$$\{n/(2-\tau)\}^{1/2}\widehat{\psi}_d \rightsquigarrow \mathcal{N}(0, \nu_d^2).$$

为实施检验，我们通过 $\widehat{\nu}_d^2 = (1-\tau)K^{-1}\sum_{k=1}^{K}(\widehat{\sigma}_{d,k}^2 + \widehat{\sigma}_{d_{\max},k}^2)$ 估计 ν_d^2，其中 $\widehat{\sigma}_{d,k}^2 = |I_k|^{-1}\sum_{i \in I_k}\{\ell(Y_i, \widehat{g}_{-k;d}(\widehat{R}_i^{(k)})) - \widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{Q}}_k)\}^2$，而 $\widehat{\sigma}_{d_{\max},k}^2$ 的定义类似。命题 2中证明了 $\widehat{\nu}_d^2$ 的一致性。当 α 时，我们以 $H_{0,d}$ 为显著性水平拒绝

$$\widehat{T}_d = \{n/(2-\tau)\}^{1/2}\widehat{\psi}_d/\widehat{\nu}_d \geq z_{1-\alpha},$$

其中，$z_{1-\alpha}$ 是 $\mathcal{N}$（0，1）的（$1-\alpha$）分位数。

注2。渐近理论允许任意固定的 $\tau \in [0,1)$，但有限样本表现体现了效率-退化权衡。较大的 τ 会增加重叠，通常降低方差并提高检验功效；然而，当 $\tau \to 1$ 时，对比会接近退化，且检验尺寸可能膨胀。遵循Chen 等人（2024年a）的建议，我们选择τ，以确保有效独立检验尺寸（$1-\tau$）$n/$（$2-\tau$）足够大。在我们的模拟中，设定 $\tau = 0.8$，由此获得了稳定的检验尺寸和检验功效。关于 τ 的数据驱动选择留待未来研究。

2.3 预测性诱导的序的序贯选择

第2.1–2.2 节为我们的预测阶数确定（*POD*）方法提供了所需的组成部分。POD 将交叉拟合预测性差距检验与前向搜索相结合，以估计预测性诱导的序 d^*。

从 $d = 0$ 开始，每次将 d 增加一步，我们依次进行检验：

- 计算 T_d；

- 若 $T_d < z_{1-\alpha}$（即 $H_{0,\,d}$ 未被拒绝），则设定 $d = d$ 并停止；
- 否则，设定 $d \leftarrow d+1$ 并继续，直至 d最大值。

等价地，

$$\widehat{d} = \min\Big\{ \min\big\{d : 0 \le d \le d_{\max},\ \widehat{T}_d < z_{1-\alpha}\big\}, d_{\max}\Big\}. \tag{2}$$

为便于实现，算法 1概述了 POD 程序。

算法 1： POD 程序

输入： 观测数据 $\{(X_i,\ Y_i)\}_{i=1}^n$；折数 $K \ge 2$；搜索上限 $d_{\max}$；重叠比例 $\tau \in [0,1)$；显著性水平 $\alpha \in (0,\ 1)$；损失函数 ℓ；一种降维方法 $\mathcal{R}$；以及一个预测规则类 $\mathcal{F}$。

将 $[n]$ 随机划分为 K 折 $I_1, \ldots, I_K$。

对于 $k = 1, \ldots, K$ **do**

通过将 $\mathcal{R}$ 应用于 $\{(X_i, Y_i)\}_{i \in I_{-k}}$，拟合降维映射 $\widehat{\varphi}_{-k}\colon \mathcal{X} \to \mathbb{R}^{d_{\max}}$，并计算所有 $i \in [n]$的 $\widehat{R}_i^{(k)} = \widehat{\varphi}_{-k}(X_i)$。将 I_k 随机划分为三个不相交集合 $I_{k,a}$、$I_{k,b}$ 和 $I_{k,o}$，使得$|I_{k,a}| = |I_{k,b}|$ 以及 $\tau = |I_{k,o}|/(|I_{k,a}| + |I_{k,o}|) \in [0,1)$。

for $d = 0, \ldots, d_{\max}$ **do**

for $k = 1, \ldots, K$ **do**

Using $\{(\widehat{R}_i^{(k)}, Y_i)\}_{i \in I_{-k}}$, fit prediction rules $\widehat{g}_{-k;d}$ (using the first d coordinates of $\widehat{R}_i^{(k)}$) and $\widehat{g}_{-k;d_{\max}}$ (using all $d_{\max}$ coordinates). Optionally, one may select within $\mathcal{F}$ using an internal procedure based only on I_{-k}.

Using $\{(\widehat{R}_i^{(k)}, Y_i)\}_{i \in I_k}$, compute the loss values $\ell(Y_i, \widehat{g}_{-k;d}(\widehat{R}_i^{(k)}))$ and $\ell(Y_i, \widehat{g}_{-k;d_{\max}}(\widehat{R}_i^{(k)}))$.

Compute $\widehat{\mathcal{L}}_{d,k}$, $\widehat{\mathcal{L}}_{d_{\max},k}$, $\widehat{\sigma}^2_{d,k}$, and $\widehat{\sigma}^2_{d_{\max},k}$ as in Section 2.2.2.

Compute $\widehat{\psi}_d$, $\widehat{\nu}_d^2$, and $\widehat{T}_d$ as in Section 2.2.2.

输出： 若 $T_d > z_{1-\alpha}$，则拒绝 $H_{0,d}$。▷ `Sequential tests`$\widehat{d}$ = 最小值 {最小值$\{d\colon 0 \le d \le d_{\max},\ \widehat{T}_d < z_{1-\alpha}\}, d_{\max}\}$。▷ `Sequential selection`

显著性水平 α 在简约性与潜在预测增益之间进行透明的权衡。较小的 α 会使拒绝更困难，并倾向于更早停止，从而得到更精简的模型；较大的 α 会降低拒绝阈值，并倾向于选择更大的阶数。

由于每一步都在水平 α下进行检验，且该过程在首次不拒绝时停止，因此高估真实阶数的概率由最后一次检验控制：

$$\Pr(\widehat{d} > d^*) \le \Pr(\widehat{T}_{d^*} \ge z_{1-\alpha}) \to \alpha. \tag{3}$$

因此，高估的渐近概率不超过 α。关于低估的上界见定理 2，定理 3 则证明了 d的一致性

当 $\alpha = \alpha_n \to 0$ 以适当速率；见第 3节。

3 理论结果

3.1 渐近零分布

回顾一下，$(X, Y) \sim \mathbb{Q}$ 在 $\mathcal{X} \times \mathcal{Y}$ 上，且 $Z = (R, X, Y) \sim \mathbb{P}$。令 $\mathcal{Y}$ 为预测空间（通常为 $\mathcal{Y} = \mathcal{Y}$），其范数为 $\|\cdot\|_{\mathcal{Y}}$（例如，当 $\mathcal{Y} = \mathbb{R}^q$ 时为欧几里得范数）。对于可测函数 $h : \mathrm{supp}(\mathbb{P}) \to \mathcal{Y}$（其中 $\mathrm{supp}(\mathbb{P})$ 是 $\mathbb{P}$ 的支撑集），我们用 $\|h\|$ 表示一般的范数，例如 $L_2(\mathbb{P})$-范数 $\|h\|_{L_2(\mathbb{P})} = [\int \|h(z)\|_{\mathcal{Y}}^2 \, d\mathbb{P}(z)]^{1/2}$，或 $L_\infty(\mathbb{P})$-范数 $\|h\|_{L_\infty(\mathbb{P})} = \sup_{z \in \mathrm{supp}(\mathbb{P})} \|h(z)\|_{\mathcal{Y}}$。给定降维映射 $\varphi : \mathcal{X} \to \mathbb{R}^{d_{\max}}$ 和预测规则 $g \in \mathcal{G}$，我们用 $g \circ \varphi$ 表示复合函数 $x \mapsto g(\varphi(x))$。

我们研究式（1）中的交叉拟合对比量 ψ_d 以及检验统计量 T_d 的渐近零分布。由 K 折的可交换性，只需分析一个一般项 $\mathcal{L}_{-k}(\widehat{g}_{-k;d}, \mathbb{Q}_k)$，其中 $\widehat{\varphi}_{-k}$ 和 $\widehat{g}_{-k;d}$ 在 I_{-k} 上训练，而 $\mathbb{Q}_k$ 是 I_k 上的经验分布。我们的目标是在 H_0 下，得到 $\mathcal{L}_{-k}(\widehat{g}_{-k;d}, \mathbb{Q}_k) - \mathcal{L}(g_{d^*}, \mathbb{P})$ 的根-n 展开式：$\mathcal{L}(g_d, \mathbb{P}) = \mathcal{L}(g_{d^*}, \mathbb{P})$，其中 $g_d \in \arg\min_{g \in \mathcal{G}_d} \mathcal{L}(g, \mathbb{P})$。为此，将其分解为

$$\begin{aligned} &\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{Q}}_k) - \mathcal{L}(g_{d^*}, \mathbb{P}) \\ &\quad = \underbrace{\{\mathcal{L}(g_{d^*}, \widehat{\mathbb{P}}_k) - \mathcal{L}(g_{d^*}, \mathbb{P})\}}_{\text{(I)}} + \underbrace{\{\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \mathbb{P}) - \mathcal{L}(g_{d^*}, \mathbb{P})\}}_{\text{(II)}} + r_n, \end{aligned} \tag{4}$$

其中，$\mathbb{P}_k$ 是 I_k 上 Z 的经验分布（仅为理论分析而引入），且 $r_n := \{\mathcal{L}_{-k}(\widehat{g}_{-k;d}, \mathbb{P}_k) - \mathcal{L}_{-k}(\widehat{g}_{-k;d}, \mathbb{P})\} - \{\mathcal{L}(g_{d^*}, \mathbb{P}_k) - \mathcal{L}(g_{d^*}, \mathbb{P})\}$。项（I）在检验折上评估固定的预言规则 g_{d^*}，根据中心极限定理，该项服从渐近正态分布。项 r_n 是一个差中之差余项；在温和正则条件下，交叉拟合通常使其为 $o_{\mathbb{P}}(n^{-1/2})$（参见 Williamson 等人，2023）。项（II）是总体风险效应（以训练折为条件），表示将 g_{d^*} 替换为学习得到的规则 $\widehat{g}_{-k;d} \circ \widehat{\varphi}_{-k}$：

$$\text{(II)} = \mathbb{E}_{\mathbb{P}}\big\{\ell\big(Y, \widehat{g}_{-k;d} \circ \widehat{\varphi}_{-k}(X)\big) \mid \widehat{g}_{-k;d} \circ \widehat{\varphi}_{-k}\big\} - \mathbb{E}_{\mathbb{P}}\big\{\ell\big(Y, g_{d^*}(R)\big)\big\}.$$

以下条件确保项（II）为 $o_{\mathbb{P}}(n^{-1/2})$。其机制在于，$g_{d^*} \in \arg\min_{g \in \mathcal{G}_{d^*}} \mathcal{L}(g, \mathbb{P})$，且 R 足以用于预测。举例来说，考虑标量响应变量的平方损失和条件均值充分性，即 $\mathbb{E}_{\mathbb{P}}(Y \mid X, R) = \mathbb{E}_{\mathbb{P}}(Y \mid R)$。由于 $g_{d^*}(R) = \mathbb{E}_{\mathbb{P}}(Y \mid R)$ 使均方误差最小，对于任意 $g \in \mathcal{G}_d$ 以及

任意 $\varphi : \mathcal{X} \to \mathbb{R}^{d_{\max}}$,

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}}\{\ell(Y, g \circ \varphi(X))\} - \mathbb{E}_{\mathbb{P}}\big\{\ell\big(Y, g_{d^*}(R)\big)\big\} \\ &\qquad = 2\mathbb{E}_{\mathbb{P}}\big\{\big(Y - g_{d^*}(R)\big)\big(g_{d^*}(R) - g \circ \varphi(X)\big)\big\} + \|g \circ \varphi - g_{d^*}\|^2_{L_2(\mathbb{P})}.\end{aligned}$$

由条件均值充分性，交叉项为零，只剩下二次余项。因此，风险差关于 $g \circ \varphi - g_{d^*}$ 是二阶的；若$\|\widehat{g}_{-k;d} \circ \widehat{\varphi}_{-k} - g_{d^*}\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}(n^{-1/4})$，则项（II）可忽略。

我们通过三个条件将这些要求形式化。

(C1)（风险曲率）对于 $d \geq d^*$，存在常数 $C > 0$，使得对任意满足 $\varphi^{(j)}: \mathcal{X} \to \mathbb{R}^{d_{\max}}$ 和 $g^{(j)} \in \mathcal{G}_d$ 的序列$\{(\varphi^{(j)}, g^{(j)})\}_{j \geq 1}$，若 $\|g^{(j)} \circ \varphi^{(j)} - g_{d^*}\| \to 0$，则对所有充分大的 j，有 $|\mathbb{E}_{\mathbb{P}}\{\ell(Y, g^{(j)} \circ \varphi^{(j)}(X)) - \ell(Y, g_{d^*}(R))\}| \leq C\|g^{(j)} \circ \varphi^{(j)} - g_{d^*}\|^2$。

（C2）（学习精度）对于 $d \geq d^*$，$\|\widehat{g}_{-k;d} \circ \widehat{\varphi}_{-k} - g_{d^*}\| = o_{\mathbb{P}}(n^{-1/4})$。

（C3）（一致性）对于 $d \geq d^*$、$\mathbb{E}_{\mathbb{P}}[\{\ell(Y, \widehat{g}_{-k;d} \circ \widehat{\varphi}_{-k}(X)) - \ell(Y, g_{d^*}(R))\}^2 \mid \widehat{g}_{-k;d}, \widehat{\varphi}_{-k}] = o_{\mathbb{P}}(1)$。

条件（C1）要求：（i）由损失函数诱导的风险具有局部曲率（例如，Farrell 等人，2021；Williamson 等人，2023），以及（ii）R、X、Y之间存在结构联系，当预言规则基于 R 而学习得到的规则基于 X时，该结构联系可以消除一阶项。对于平方损失，（ii）的一个充分条件是 $\mathbb{E}_{\mathbb{P}}(Y \mid X, R) = \mathbb{E}_{\mathbb{P}}(Y \mid R)$，这在例2中成立。在例1中，只需满足 $\mathbb{E}(\varepsilon \mid f, u) = 0$；而在例 3中，只需满足 $\mathbb{E}(\varepsilon \mid X) = 0$；参见例如 Yuan 等人（2007）和 Fan和Gu（2024）。除平方损失之外的其他示例及证明见补充材料的 B.2 节。

条件（C2）要求学习得到的规则以 $o_{\mathbb{P}}(n^{-1/4})$ 的速率收敛到 g_{d^*}。结合（C1），这意味着式（4）中的项（II）为 $o_{\mathbb{P}}(n^{-1/2})$。在许多情形下，如果降维阶段能够恢复 R^*（允许旋转），且预测规则足够规则，就能达到这一速率。具体而言，在温和条件下，存在一个旋转矩阵 O，使得 $\|O\widehat{R}_{[d]} - R^*\| = o_{\mathbb{P}}(n^{-1/4})$。补充材料B.3节在两个具有代表性的情形下证明了这一界：例1中，$\widehat{\varphi}$ 通过主成分分析获得；例2中，$\widehat{\varphi}$通过正向回归或逆向回归学习得到。如果 g_{d^*} 满足利普希茨条件，这种表示误差就会传递为预测误差。$\widehat{g}_{-k;d}$ 剩余的估计误差和逼近误差可通过标准的复杂度论证加以控制（Gyorfi 等，2002）。例如，在因子回归（例1）中，Fan和Gu（2024）表明，对于采用神经网络预测规则的主成分降维（在 I_{-k}内部还需进行一次额外划分），如果 g_{d^*} 是 β-光滑的，且 p 与 n同阶，那么$\|\widehat{g}_{-k;d} \circ \widehat{\varphi}_{-k} - g_{d^*}\|^2_{L_2}(\mathbb{P}) = O_{\mathbb{P}}(n^{-2\beta/(2\beta+d^*)})$ 可在相差对数因子的意义下成立；因此，只要 $\beta > d^*/2$，就有$\|\widehat{g}_{-k;d} \circ \widehat{\varphi}_{-k} - g_{d^*}\|_{L_2}(\mathbb{P}) = o_{\mathbb{P}}(n^{-1/4})$；a

平均处理效应推断中结果回归非参数估计的一个密切相关的速率条件见Fan 等（2025）。

条件（C3）是对损失函数的一项温和的一致性要求，与Williamson 等人（2023）中的要求类似。对于第二个参数满足利普希茨条件的损失函数，（C2）通常蕴含（C3）（Farrell 等人，2021）。

定理 1（渐近零分布）。在条件（*C1*）–（*C3*）及 $H_{0,\,d}$ 下，对于任意 $\tau \in [0,1)$，

$$\{n/(2-\tau)\}^{1/2}\widehat{\psi}_d \rightsquigarrow \mathcal{N}(0,\nu_d^2),$$

其中 $\nu_d^2 = 2\,(1-\tau)\,\sigma_{d^*}^2$ 且 $\sigma_{d^*}^2 =$ 方差$\{\ell\,(Y, g_{d^*}\,(R)\,\}$。

命题 2。在条件（*C1*）–（*C3*）及 $H_{0,\,d}$ 下，对于任意 $\tau \in [0,1)$，$\widehat{\nu}_d^2 \to \nu_d^2$ 依概率。

定理 1、命题 2和斯卢茨基引理表明，在 H_0，下，

$$\{n/(2-\tau)\}^{1/2}\widehat{\psi}_d/\widehat{\nu}_d \rightsquigarrow \mathcal{N}(0,1).$$

因此，$\Pr(\widehat{d} > d^*) \le \Pr(T_{d^*} \ge z_{1-\alpha}) \to \alpha$，所以该序贯程序将高估真实阶数的渐近概率控制在 α 水平。

3.2 低估的概率界

接下来，我们控制序贯选择器 $\widehat{d}$ 在 (2) 中过早停止的概率，即返回一个严格小于预测阶数 d^* 的维度。为此，我们研究备择假设 $H_{1,d}$ 下，检验统计量 T_d 的行为，其中 $d \in \{0,\dots,d^*-1\}$。

为简单起见，本小节专门讨论平方损失函数。在 $H_{1,d}$ 且 $d < d^*$ 的条件下，令 $\Delta_d := \mathcal{L}_d - \mathcal{L}_{d^*} = \mathcal{L}(g_d,\mathbb{P}) - \mathcal{L}(g_{d^*},\mathbb{P}) > 0$ 表示预测差距。对于每一折 k，在温和条件下，我们有

$$\begin{aligned}&\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d},\widehat{\mathbb{Q}}_k) - \widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d^*},\widehat{\mathbb{Q}}_k)\\ &\quad= \Delta_d + \underbrace{\{\mathcal{L}(g_d,\widehat{\mathbb{P}}_k) - \mathcal{L}(g_{d^*},\widehat{\mathbb{P}}_k)\} - \Delta_d}_{\text{(I')}} + \underbrace{\{\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d},\mathbb{P}) - \mathcal{L}(g_d,\mathbb{P})\}}_{\text{(II')}} + o_{\mathbb{P}}(n^{-1/2}).\end{aligned}$$

项（I'）为 $O_{\mathbb{P}}(n^{-1/2})$，且渐近正态。项（II'）是以 $\widehat{g}_{-k;d}\circ\widehat{\varphi}_{-k}$ 替换 g_d 所产生的总体风险效应，与项（II）类似；但与零假设情形不同，它不一定是 $o_{\mathbb{P}}(n^{-1/2})$。对于平方损失函数，可以用估计误差$\|\widehat{g}_{-k;d}\circ\widehat{\varphi}_{-k} - g_d\|_{L_2(\mathbb{P})}$ 以及确定性预测差距 Δ_d 来界定它。以下条件是情形 $d < d^*$ 下（C2）–（C3）的对应条件。

（C2'）（学习精度）对于 $1 \le d < d^*$, $\|\widehat{g}_{-k;d}\circ\widehat{\varphi}_{-k} - g_d\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}(n^{-1/4})$。

（C3’）（一致性）对于 $1 \le d < d^*$, $\mathbb{E}_{\mathbb{P}}[\{\|Y - \widehat{g}_{-k;d} \circ \widehat{\varphi}_{-k}(X)\|_2^2 - \|Y - g_d(R)\|_2^2\}^2 \mid \widehat{g}_{-k;d}, \widehat{\varphi}_{-k}] = o_{\mathbb{P}}(1)$。

条件（C2’）假设，对于 $d < d^*$, $R_{[d]}$ 构成了针对$\curlywedge R_{[d]}$的适定目标。其可识别性通常源于总体算子或矩阵的特征值分离。在例1中，广泛性意味着 $p^{-1}B^\top B$的特征值互不相同，且与 0 和 ∞保持正距离；参见 Fan 等人（2013）。在例2中，一个类似的条件是，与充分降维程序相关联的核矩阵具有互不相同的主要特征值；参见 Kim 等人（2020）。因此，Davis-Kahan 定理（Davis 和 Kahan，1970；Yu 等人，2015）保证存在一个旋转矩阵 O，使得 $\|O\widehat{R}_{[d]} - R_{[d]}\| = o_{\mathbb{P}}(n^{-1/4})$。根据在（C2）讨论中采用的相同论证，在标准正则条件以及 n 和 p适当增长的情况下，该界进而支持（C2’）。补充材料的 B.4 节在两种情形下推导了上述界：例1中，$\widehat{\varphi}$ 通过主成分分析构造；例2中，$\widehat{\varphi}$通过基于核矩阵的方法学习。

定理 2（低估界）。假设满足条件（*C1*）–（*C3*）和（*C2’*）–（*C3’*）。固定 $\tau \in [0, 1)$。那么，对于任意 $0 < \delta < (\mathcal{L}_d - \mathcal{L}_{d^*})$,

$$\Pr(\widehat{d} < d^*) \le \sum_{d=0}^{d^*-1} \Phi\left(\frac{\nu_d}{\nu_{d,\eta}} z_{1-\alpha} - \frac{\{n/(2-\tau)\}^{1/2}(\mathcal{L}_d - \mathcal{L}_{d^*} - \delta)}{\nu_{d,\eta}} \right) + o(1),$$

其中 $\nu_d^2 = (1-\tau)(\sigma_d^2 + \sigma_{d^*}^2)$, $\nu_{d,\eta}^2 = \nu_d^2 + \tau\eta_d^2$, $\sigma_d^2 =$方差$\{\|Y - g_d(R)\|_2^2\}$, 以及 $\eta_d^2 =$方差$\{\|Y - g_d(R)\|_2^2 - \|Y - g_{d^*}(R)\|_2^2\}$。

参数 $\delta > 0$充当一个间隔，用于考虑将 g_d 替换为 $\widehat{g}_{-k;d}\circ\widehat{\varphi}_{-k}$时产生的估计误差。定理 2 还表明，从渐近意义上看，较大的重叠 τ 会降低低估 d^*的概率。

3.3 一致性

结合高估控制（3）与定理 2 可得一个简单的一致性判据。设名义水平依赖于 n，并记 $\alpha = \alpha_n \to 0$。定理 1 意味着$\widehat{\ }$ $\widehat{\ }\Pr(d > d^*) \le \Pr(T_{d^*} \ge z_{1-\alpha_n}) \to 0$。若 $z_{1-\alpha_n}/\sqrt{n} \to 0$ (等价地，$(\log \alpha_n)/n \to 0$)，则$\widehat{\ }$低估界迫使 $\Pr(d < d^*) \to 0$。

定理 3 (一致性)。在定理 *1–2*的假设下，若 $\alpha_n \to 0$ 且$\widehat{\ }(\log \alpha_n)/n \to 0$，则 $\lim_{n\to\infty} \Pr(d = d^*) = 1$。

因此，任何比 $\exp(-cn)$ 衰减更慢的趋零序列 α_n，都能保证序贯选择器以趋近于 1 的概率恢复预言机阶数。

4 数值研究

4.1 模拟

我们在两种典型情形下考察 POD 的有限样本表现：因子回归（示例 1）和充分降维（示例 2）。我们的目标是：（i）评估各项预测性差距检验的经验显著性水平和检验功效；（ii）评估由此得到的阶数估计量的准确性。

每种配置重复 500 次，结果以蒙特卡洛平均值呈现。除非另有说明，我们使用 $K = 5$ 折、搜索上限$d_{\max} = 8$和重叠比例 $\tau = 0.8$实现 POD。初步实验表明，在各折内重新拟合降维映射与在完整样本上拟合一次，在所报告的指标上几乎没有差异；因此，为降低计算成本，我们使用所有观测值拟合一次 $\widehat{\varphi}$，并在各折中重复使用。在每个训练折内，我们从一个灵活的类别 $\mathcal{F}$ 中拟合预测规则，该类别包含多个候选学习器（以及适用时的调优参数）。我们仅依据训练数据，在 $\mathcal{F}$ 中使用损失函数 ℓ进行二折交叉验证并据此选择。随后，在$\widehat{}$留出折上评估得到的拟合规则，以构造 T_d；参见算法 1。这种多模型类别可以通过从候选项中自适应地选择表现较强的学习器，提升降维后的预测效果。

4.1.1 因子回归

数据生成机制。 在本实验中，我们设定真实阶数为 $d_{\mathrm{F}}^{*} = 5$，表示驱动响应变量和协变量的潜在因子 $f \sim \mathcal{N}(0, I_{d_{\mathrm{F}}^{*}})$。具体而言，$Y = m(f) + \varepsilon$，其中 $m(f) = f_1 + 2f_2 + f_3 + 3f_4 + 2f_5$且 $X = Bf + u$，其中 $\varepsilon \sim \mathcal{N}(0, 0.1)$ 和 $u \sim \mathcal{N}(0, \mathrm{diag}(v_1^2, \ldots, v_p^2))$ 是相互独立的噪声项。我们取 $n = 500$ 和 $p = 1{,}000$。我们考察载荷和方差的两种不同情形：

- 弱因子：载荷遵循Harding (2013年)，方差为 $v_j = 0.55^2$，适用于 $j \in [d^{*}_{\mathrm{F}}]$。具体而言，对每个 j，$b_{lj} = \sqrt{3}p^{-1/2}$ 适用于$\overline{l \le d^{*}_{\mathrm{F}}}$，且 $b_{lj} = a_{lj}\sqrt{3}(p-j)^{-1}$适用于 $l > d^{*}_{\mathrm{F}}$，其中 $a_{lj} = -1$ 在 $l = rj\,(r = 1, 2, \ldots)$ 时成立，否则为 $a_{lj} = 1$。
- 广泛因子：载荷独立同分布地取自 $b_{lj} \sim \mathrm{Unif}(0, j)$，方差为$v_j^2 = 25$。

实现。 降维步骤对 X 的样本协方差矩阵进行主成分分析。取前 $d_{\max}$ 个特征向量构成 $\widehat{\varphi}$。我们考虑 $\mathcal{F}$，包括普通最小二乘法（OLS）、多元自适应回归样条（MARS，通过 `R` 包 `earth`的默认设置实现），以及一个双层神经网-

络（通过 nnet实现，隐藏层有 5 个神经元）。使用平方损失，因此目标阶数为 $d^*(\mathcal{L}) = d_{\mathrm{F}}^*$。

检验尺寸和检验功效。我们将 POD 与 Onatski (2009) 和 Kapetanios (2010) 提出的两种基于特征值的序贯检验进行比较。这些方法依次检验 $H_{0,d}'$：$d = d_{\mathrm{F}}^*$ 与 $H_{1,d}$：$d < d_{\mathrm{F}}^*$，其中 $d = 0$ 为 $1, \ldots, d_{\max}$。Onatski (2009) 的检验统计量为 $\max_{d<i\le d_{\max}}(\lambda_i - \lambda_{i+1})/(\lambda_{i+1} - \lambda_{i+2})$，其中 λ_i 是矩阵 $\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \widetilde{X}_i \widetilde{X}_i^\top / \lfloor \frac{n}{2} \rfloor$ 中第 i 大的特征值，且 $\widetilde{X}_i = X_i + \sqrt{-1} X_{i+\lfloor \frac{n}{2} \rfloor}$。对于 Kapetanios (2010)，检验统计量为 $\lambda_{d+1} - \lambda_{d_{\max}+1}$，其中 λ_i 是样本协方差矩阵中第 i 大的特征值。按照原论文的建议，两种检验的临界值均通过模拟或子抽样获得。

表1：在 $\alpha = 5\%$ 显著性水平下，序贯检验各步骤的拒绝比例（%）。

情景	方法	$d=0$	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$
弱因子	POD	100	100	100	100	100	5.4	5.4
	Onatski (2009)	22.4	24.4	25.8	29.8	27.2	3.4	3.4
	Kapetanios (2010)	100	100	100	100	100	69.4	38
广泛因子	POD	100	100	100	100	100	6	4.8
	Onatski (2009)	100	90.2	93	96.2	50.2	5	4
	Kapetanios (2010)	100	100	100	100	100	19.8	10.6

表 1 报告了在 $\alpha = 5\%$ 下的拒绝比例。POD 和 Onatski (2009年) 在 $d \ge d_{\mathrm{F}}^*$ 时保持名义尺寸；对于 $d < d_{\mathrm{F}}^*$，POD 的检验功效显著更高。相比之下，Kapetanios (2010年) 严重过度拒绝，这与 Onatski (2009年) 报告的尺寸扭曲一致。

阶数估计。 我们将 POD 与三种基于特征值的替代方法进行比较：来自 Bai 和 Ng (2002年) 的信息准则（IC）（参见补充材料中的（C.1））；来自 Ahn 和 Horenstein (2013年) 的特征值比率（ER）估计量，以及来自 Fan 等 (2022年) 的调整相关性阈值法（ACT）。

图 1 展示了随着 n 增加，因子数量正确估计、高估和低估的概率。POD 很快收敛至真实阶数，其过估计率稳定在预设的 α，而低估率趋近于零，这与第 3 节所述理论一致。相比之下，在所考察的样本量下，信息准则和特征比率均表现出低估，这与 Fan 等 (2022年) 指出的情况一致，而 ACT 收敛缓慢。

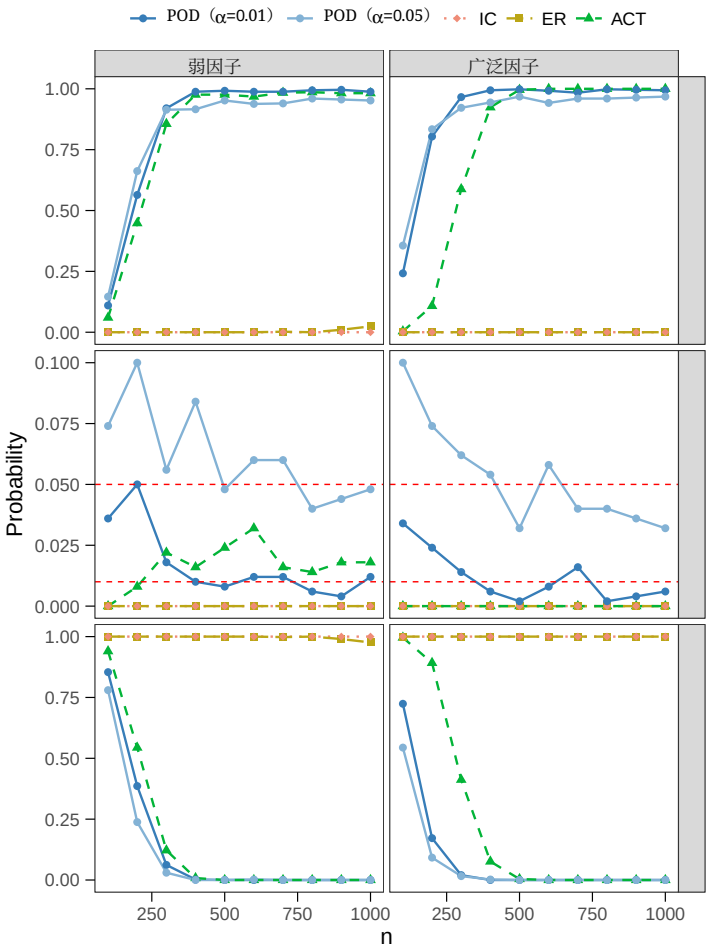


图1：因子回归中 n 下正确估计（$d = d_{\mathrm{F}}^*$）、高估（$d > d_{\mathrm{F}}^*$）和低估（$d < d_{\mathrm{F}}^*$）的经验概率。第二行中的红色水平虚线标示名义显著性水平 $\alpha = 0.01$ 和 $\alpha = 0.05$。

.

4.1.2 充分降维

检验尺寸和检验功效。 令 $X \sim \mathcal{N}(0,\ I_p)$，其中 $p = 10$。考虑

$$\begin{aligned} \text{Model 1}: &\quad Y = X_1 + X_2 + X_3 + X_4 + \sigma\varepsilon, \\ \text{Model 2}: &\quad Y = 0.4(X_1 + X_2 + X_3)^2 + 3\sin[(X_1 + X_9 + 3X_{10})/4] + \sigma\varepsilon, \end{aligned}$$

其中 $\epsilon \sim \mathcal{N}(0,\ 1)$ 与 X 和 $\sigma = 0.5$相互独立。在模型1中，样本量为$n \in \{100, 200\}$；在模型2中，我们使用 $n \in \{200, 300\}$。模型1的结构维数为$d_{\mathrm{CS}}^* = 1$，模型2的结构维数为$d_{\mathrm{CS}}^* = 2$，且两种情况下均有 $d_{\mathrm{CS}}^* = d_{\mathrm{CMS}}^*$。

POD 使用平方损失来量化预测性，因此 $d^*(\mathcal{L}) = d_{\mathrm{CMS}}^*$。对于模型1，降维映射通过切片逆回归（SIR，Li，1991年）进行估计；对于模型2，则通过方向回归（DR，Li 和 Wang，2007）进行估计。SIR 使用 10 个切片，DR 使用 4 个切片。我们设置 $\mathcal{F}$ 以纳入普通最小二乘法、多元自适应回归样条和回归树（通过 R 软件包实现）

`tree` 并使用默认设置）。

为进行基准比较，我们针对每个模型，采用与 POD 相同的降维方法，实现 Bura 和 Yang（2011年）提出的两种序贯检验。这些方法为：(i) 基于该方法对应的核矩阵最小特征值的检验，其渐近极限为加权卡方（χ^2）（BY，加权-χ^2）；(ii) 渐近极限为 χ^2 的 Wald 型检验（BY，χ^2）。每种方法都依次检验 $H'_{0,\,d}$：$d = d^*_{\mathrm{CS}}$ 对比 $H_{1,\,d}$：$d < d^*_{\mathrm{CS}}$，其中 $d = 0,\ 1,\ \ldots\ldots,\ d_{\max}$。

Ta表 2：在 $\alpha = 5\%$ 水平下，序贯检验每一步的拒绝比例（%）

模型	n	方法	$d=0$	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$
1	100	POD	100	5	7.2	7.2	4.6	6.2
		BY，加权-χ^2	100	3.4	0.2	0	0	0
		BY，χ^2	100	100	97.4	42.8	3	0
	200	POD	100	5	5.4	4.4	5.4	4.6
		BY，加权-χ^2	100	5.2	0.2	0	0	0
		BY，χ^2	100	82.6	22.6	1.6	0	0
2	200	POD	99.8	96	6.8	7	7	4.8
		BY，加权-χ^2	100	52	0.8	0.2	0	0
		BY，χ^2	100	100	100	100	100	100
	300	POD	100	99	4.6	4.8	5.2	5.4
		BY，加权-χ^2	100	87	1	0	0	0
		BY，χ^2	99.8	99.8	80	54.4	30.6	13.4

表2 显示了 $\alpha = 0.05$ 序贯检验各步骤的拒绝比例。POD 和采用加权-χ^2 限值的 BY 检验在各自的原假设下均能维持检验水平（POD：$H_{0,\,d}$：$d \geq d^*_{\mathrm{CS}}$，加权-χ^2：$H'_{0,\,d}$：$d = d^*_{\mathrm{CS}}$）。在模型2中，在备择假设χ^2 下，POD 的检验功效高于加权- $d = 1$检验，而后者往往较为保守。相比之下，采用 χ^2 限值的 Wald 型检验在许多配置下会过度拒绝，这与Bura 和 Yang (2011年)的发现一致；他们指出，要使 χ^2 渐近校准准确，可能需要大得多的样本量。

阶数估计。 令 $X \sim \mathcal{N}(0, I_p)$，且满足 $p = 10$。考虑

$$
\begin{aligned}
\text{Model 3}: &\quad Y = \sin(X_1) + \sigma\varepsilon, \\
\text{Model 4}: &\quad Y = X_1^2 + 0.5\sin(X_2) + \sigma\varepsilon, \\
\text{Model 5}: &\quad Y = |X_1| + X_2(X_2 + X_3 + 1) + \sigma\varepsilon,
\end{aligned}
$$

其中 $\epsilon \sim \mathcal{N}(0,1)$ 与 X 和 $\sigma = 0.5$相互独立。结构维数为$d^*_{\text{CS}} = 1$ （模型 3 中，Luo 和 Li，2021），$d^*_{\text{CS}} = 2$ （模型 4 中），以及 $d^*_{\text{CS}} = 3$ （模型 5 中）；在所有情况下，$d^*_{\text{CS}} = d^*_{\text{CMS}}$。

POD 使用平方损失预测能力，因此 $d^*(\mathcal{L}) = d^*_{\text{CMS}}$。模型 3–5 的降维映射通过最小平均方差估计（MAVE；Xia 等人，2002年）进行估计。预测规则包括普通最小二乘法、多元自适应回归样条和回归树。我们将 POD 与三种备择假设进行比较：(i) ladle 估计量（Luo 和 Li，2016），它结合特征值碎石图与核矩阵特征向量的自助法变异性——为指定核，我们对模型 3 使用 10 个切片的切片逆回归，对模型 4–5 使用 4 个切片的方向回归，使目标矩阵的秩与预言值 d^*_{CS}对齐；(ii) 基于 MAVE 的交叉验证（CV）准则，目标为 d^*_{CMS}；以及 (iii) 基于切片回归的交叉验证（CV）准则（SR，Wang 和 Xia，2008），目标为 d^*_{CS}。

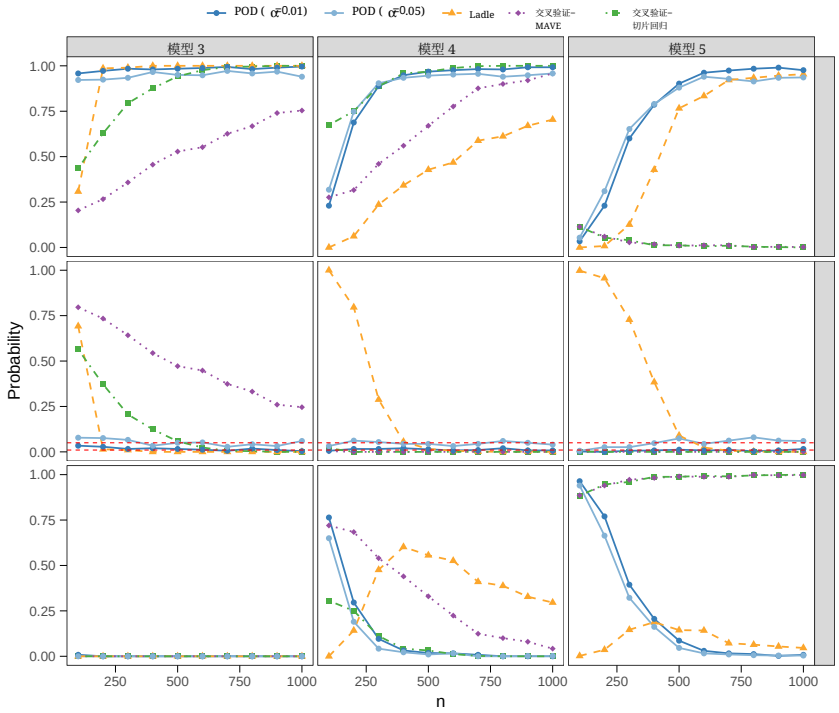


图 2： n 下 SDR 的正确估计（$d = d^*_{\text{CMS}}$）、高估（$d > d^*_{\text{CMS}}$）和低估（$d < d^*_{\text{CMS}}$）的经验概率。第二行中的红色水平虚线标示名义显著性水平 $\alpha = 0.01$ 和 $\alpha = 0.05$。

图 2 展示了随着 d^*_{CMS} 作为 n 不断增大时，正确估计、高估和低估的概率。在所有模型中，POD 都比其他方法更快收敛。其高估概率稳定在名义水平，而低估概率则趋于零。POD 可以受益于采用更丰富的候选类$\mathcal{F}$，因为单个学习器的表现可能取决于数据生成机制。在《补充材料》的 C.3 节中，将 $\mathcal{F}$ 限制为仅使用普通最小二乘法，或

仅使用回归树会导致准确度明显下降，而单独使用多元自适应回归样条仍能取得令人满意的效果；这促使我们在 $\mathcal{F}$ 中纳入多个候选模型。

分类响应的结果将在补充材料的第 C.4 节中介绍。

4.1.3 损失函数的影响

我们通过目标维度 $d^*(\mathcal{L})$ 来研究损失函数的选择如何影响 POD。令 $X \sim \mathcal{N}(0, \Sigma)$，其中 $p = 10$ 和 $\Sigma_{i,j} = (0.5)^{|i-j|}$。考虑伯努利响应变量 Y，其 $\Pr(Y = 1 \mid X_1 > 0) = 1$ 且 $\Pr(Y = 1 \mid X_1 \le 0) = 0.6$ (Cook 和 Yin，2001)，因此 Y 仅取决于 X_1。在这种情况下，$d^*_{\mathrm{CS}} = d^*_{\mathrm{CMS}} = 1$。然而，由于贝叶斯分类器总是预测 $Y = 1$，中心判别子空间（CDS，Cook 和 Yin，2001）的维度为 $d^*_{\mathrm{CDS}} = 0$。我们抽取 $n = 2000$ 个观测值。

对于这个问题，在交叉熵损失下，POD 的目标是 $d^*(\mathcal{L}) = d^*_{\mathrm{CS}} = 1$；在 0-1 损失下，目标是 $d^*(\mathcal{L}) = d^*_{\mathrm{CDS}} = 0$。降维映射通过 DR 估计，预测规则则从支持向量机（通过 R 软件包 `e1071` 实现，采用默认设置）和分类树中选取。我们将 POD 与三种备择方法进行比较：应用于 DR 的 ladle 估计量，以及基于 MAVE 和 SR 的交叉验证准则；所有方法的目标均为 $d^*_{\mathrm{CS}} = d^*_{\mathrm{CMS}} = 1$。

表3：估计结构维数的比例（%）。ave($\hat{d}$) 表示估计值的平均值。

	POD				Ladle	CV-SR	CV-MAVE
	0-1 损失		交叉熵损失				
	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 1\%$	$\alpha = 5\%$			
$\hat{d} = 0$	99.2	96	0	0	0	0	0
$\hat{d} = 1$	0.8	3.4	99	94.2	99.4	61.2	59.4
$\hat{d} \ge 2$	0	0.6	1	5.8	0.6	38.8	40.6
ave($\hat{d}$)	0.01	0.05	1.01	1.07	1.01	1.39	1.41

表 3展示了不同损失函数和方法下估计结构维数所占的比例。结果证实，POD能够适应目标阶数 $d^*(\mathcal{L})$：在0-1 损失下，POD选择 $d = 0$；而在交叉熵损失下，它识别出 $d = 1$。在这两种情况下，POD都将过估计率保持在接近名义水平。相比之下，ladle 估计量始终估计出 $d^*_{\mathrm{CS}} = 1$，而SR和MAVE则表现出明显的过估计。

4.2 真实数据示例：PenDigits 手写数字

我们分析了来自UCI 机器学习知识库的PenDigits数据集，重点确定需要多少个方向回归（DR）方向才能区分最容易混淆的数字 $\{0, 6, 9\}$。训练集包含2,219幅图像，测试集包含1,035幅图像；每幅图像均由一个16维笔迹轨迹向量表示。对根据训练数据估计的DR 核矩阵进行特征值分析显示，前两个特征值之后出现明显下降，表明前两个DR方向捕获了大部分判别信息；见图3A。

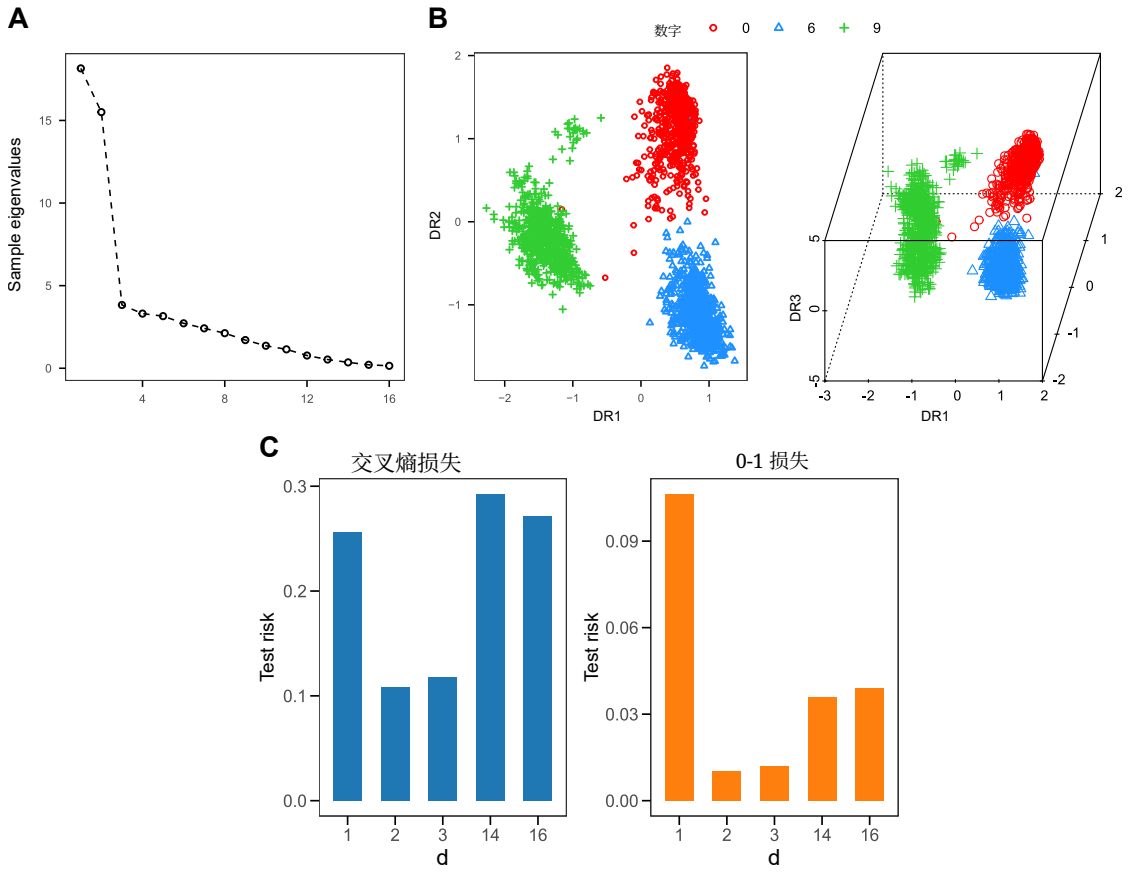


图 3：PenDigits 手写数字分析。面板 A 展示 DR 核矩阵的特征值碎石图。面板 B 展示采用前两个和前三个 DR方向得到的 DR 表示。面板 C 展示不同维度下的测试风险。

我们的 POD 程序检验前几个 DR方向对交叉熵和 0-1损失函数的增量预测价值。预测规则采用双层神经网络。我们考察显著性水平 $\alpha \in \{0.01, 0.05\}$。作为基准，我们实现 ladle 估计量和 BY 加权-χ^2 检验，两者均根据 DR 核矩阵计算。所有方法均将搜索上限设为 $d_{\max} \in \{8, 16\}$。为考虑 POD 交叉拟合、ladle 自助法以及 BY 的

临界值模拟所引入的抽样变异性，每种程序都在固定训练集上重复运行 100 次，$\widehat{}$并报告众数估计值 d （频率详情见补充材料C.5 节）。对于损失函数（交叉熵或 0-1）、显著性$\widehat{}$水平和搜索上限 $d_{\max}$的所有组合，POD 均稳定地选出 $d = 2$。相比之下，ladle$\widehat{}$ $\widehat{}$估计量在 $d = 3$ 时选出 $d_{\max} = 8$，但在 $d = 14$ 时会将估计值膨胀至$d_{\max} = 16$。加权-χ^2 检验在 $\widehat{d} = 14$ 下输出 $\alpha = 0.01$，而在 $\widehat{d} = 16$ 下输出$\alpha = 0.05$。图 3B 展示了 DR 表示的散点图：左图绘制前两个 DR方向，右图则加入第三个方向。即使在二维视图中，各类别也已几乎完全分离，这证实 POD 选出的两个 DR方向几乎捕获了全部判别信息。

为了评估不同维度模型的样本外表现，我们$\widehat{}$按如下步骤进行。对于每个候选维度 $d \in \{1, 2, 3, 14, 16\}$，我们选取相应的降维得分，然后在两种目标损失函数（交叉熵或 0-1 损失）下，使用神经网络在完整训练集上训练分类器，并采用相同的损失函数（重复 500 次以提高神经网络训练的稳定性）。随后在测试集上评估这些模型。图3C 报告了每个选定维度对应的（平均）测试风险。POD 选出的二维表示在两种损失函数下都取得了最低的测试风险。三维模型的表现略差。维度更高的模型会导致测试风险上升，表明出现了明显的过拟合。这些结果证实，POD 在实践中能提供简约而预测能力很强的表示。

5 结论

POD 将阶数确定视为预测充分性问题，而非特定于模型的结构性准则问题。它将维度视为总体风险的指标，并选择这样一个最小阶数：在该阶数下，增加坐标不再改善样本外表现。该方法结合了交叉拟合风险评估、可避免退化的重叠式第二次拆分以及前向序贯检验，并可应用于多种降维方法和学习器之上，从而得到与所选损失函数和预测任务相契合的预测阶数及其对应的不确定性。

有几个方向值得进一步研究。一个重要方向是稳健性，包括将 POD 扩展到相关数据或重尾数据。这可能需要用于时间序列或聚类观测的分块交叉拟合方案，以及在离群值和重尾情况下仍保持稳定的稳健损失函数。第二个方向是在所选阶数下，对预测风险及相关泛函进行后选择推断，开发能够考虑数据驱动选择 d 的置信区间和检验$\widehat{}$。这两项

扩展将拓宽 POD 的适用范围，同时保留其核心原则：根据预测价值确定维度。

参考文献

Ahn, S. C. and Horenstein, A. R. (2013), “因子数量的特征值比率检验”，《计量经济学》, 81, 1203–1227. Bai, J. and Li, K. (2012), “高维因子模型的统计分析”，《统计年鉴》, 40, 436–465. Bai, J. and Ng, S. (2002), “确定近似因子模型中的因子数量”，《计量经济学》, 70, 191–221. Bunea, F., She, Y., and Wegkamp, M. H. (2011), “高维矩阵降秩估计量的最优选择”，《统计年鉴》, 39, 1282–1309. Bura, E. and Yang, J. (2011), “充分降维中的维数估计：一种统一方法”，《多元分析杂志》, 102, 130–142. Cattell, R. B. (1966), “因子数量的碎石检验”，多元行为研究, 1, 245–276. Chen, G., Jia, Y., Wang, G., and Zou, C. (2024a), “Zipper：解决算法无关推断中的退化问题”，《神经信息处理系统进展》, 37, 65884–65911. Chen, Y., Jiao, Y., Qiu, R., and Yu, Z. (2024b), “深度非线性充分降维”，《统计年鉴》, 52, 1201–1226. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), “用于处理效应和结构参数的双重/去偏机器学习”，《计量经济学杂志》, 21, C1–C68. Cook, R. D. (1994), “回归图的解释”，美国统计学会会刊, 89, 177–189. Cook, R. D. and Li, B. (2002), “回归中条件均值的降维”，《统计年鉴》, 30, 455–474. Cook, R. D. and Weisberg, S. (1991), “用于降维的切片逆回归：评论”，美国统计学会会刊, 86, 328–332. Cook, R. D. and Yin, X. (2001), “理论与方法：特邀论文：判别分析中的降维与可视化（附讨论）”，《澳大利亚与新西兰统计学杂志》, 43, 147–199. Dai, B., Shen, X., and Pan, W. (2024), “黑箱学习器的特征相关性显著性检验”，《*IEEE*神经网络与学习系统汇刊》, 35, 1898–1911. Davis, C. and Kahan, W. M. (1970), “扰动引起的特征向量旋转。III”，《工业与应用数学学会数值分析杂志》, 7, 1–46. Doz, C., Giannone, D., and Reichlin, L. (2012), “大型近似动态因子模型的拟极大似然方法”，《经济学与统计学评论》, 94, 1014 1024– . Fan, J. and Gu, Y. (2024), “用于高维回归的因子增强稀疏贯穿深度ReLU神经网络”，美国统计学会会刊, 119, 2680–2694. Fan, J., Guo, J., and Zheng, S. (2022), “通过调整特征值阈值估计因子数量

g,”期刊 f 美国统计学会 , 117, 852–861。

Fan, J., Jana, S., Kulkarni, S，一个d Yin, Q. (2025), "F演员 informed double deep learning f
或平均处理效应估计"，*arXiv*预印本 *arXiv:2508.17136*。Fan, J. 和 Liao, Y.（2022），《从多样化投影中学习潜在因子及其在因子过度估计和弱因子中的应用》，美国统计学会会刊，117，909–924。Fan, J.、Liao, Y. 和 Mincheva, M.（2013），《通过阈值化主正交补空间进行大规模协方差估计》，英国皇家统计学会会刊*B*辑：统计方法学，75，603–680。Fan, J.、Liao, Y. 和 Wang, W.（2016），《因子模型中的投影主成分分析》，《统计年鉴》，44，219–254。Fan, J.、Lou, Z. 和 Yu, M.（2024），《潜在因子回归和稀疏回归是否足够？》，美国统计学会会刊，119，1076–1088。Farrell, M. H.、Liang, T. 和 Misra, S.（2021），《用于估计和推断的深度神经网络》，《计量经济学》，89，181–213。Golub, G. H. 和 Van Loan, C. F.（2013），《矩阵计算》，马里兰州巴尔的摩：约翰斯·霍普金斯大学出版社，第4版。Gy¨orfi, L.、Kohler, M.、Krzy˙zak, A. 和 Walk, H.（2002），《非参数回归的无分布理论》，纽约：施普林格。Harding, M.（2013），《估计高维因子模型中的因子数量》，手稿。Hristache, M.、Juditsky, A.、Polzehl, J. 和 Spokoiny, V.（2001），《用于降维的结构自适应方法》，《统计年鉴》，29，1537–1566。Hsing, T. 和 Carroll, R. J.（1992），《切片逆回归的渐近理论》，《统计年鉴》，20，1040–1061。Izenman, A. J.（1975），《多元线性模型的降秩回归》，《多元分析杂志》，5，248–264。Jolliffe, I. T.（2002），《主成分分析》，纽约：施普林格，第2版。Kapetanios, G.（2010），《一种用于确定大型数据集中近似因子模型因子数量的检验程序》，《商业与经济统计杂志》，28，397 409– 。Kim, K.、Li, B.、Yu, Z. 和 Li, L.（2020），《关于降维后统计推断》，《统计年鉴》，48，1567–1592。Lei, J.（2020），《带置信度的交叉验证》，美国统计学会会刊，115，1978–1997。Li, B.（2018），《充分降维：方法及其在*R*中的应用》，佛罗里达州博卡拉顿：CRC出版社。Li, B. 和 Wang, S.（2007），《关于用于降维的方向回归》，美国统计学会会刊，102，997–1008。Li, K.-C.（1991），《用于降维的切片逆回归》，美国统计学会会刊，86，316–327。—（1992），《关于用于数据可视化和降维的主海森方向：斯坦引理的另一种应用》，美国统计学会会刊，87，1025–1039。

Lin, Q.、Zhao, Z. 和 Liu, J. S.（2019），《通过Lasso实现稀疏切片逆回归》，美国统计学会会刊，114，1726–1739。Lundborg, A. R.、Kim, I.、Shah, R. D. 和 Samworth, R. J.（2024），《用于弱假设变量显著性检验的投影协方差测度》，《统计年鉴》，52，2851–2878。Luo, W. 和 Li, B.（2016），《结合特征值与特征向量变化确定阶数》，*Biometrika*，103，875–887。—（2021），《关于通过预测变量增广确定阶数》，*Biometrika*，108，557–574。Onatski, A.（2009），《检验关于大规模因子模型中因子数量的假设》，《计量经济学》，77，1447–1479。Reinsel, G. C.、Velu, R. P. 和 Chen, K.（2022），《多元降秩回归》，纽约：施普林格，第2版。van der Vaart, A. W.（1998），《渐近统计学》，剑桥：剑桥大学出版社。Wang, H. 和 Xia, Y.（2008），《用于降维的切片回归》，美国统计学会会刊，103，811–821。Williamson, B. D.、Gilbert, P. B.、Simon, N. R. 和 Carone, M.（2023），《算法无关的变量重要性推断通用框架》，美国统计学会会刊，118，1645–1658。Xia, Y.（2007），《估计降维方向的构造性方法》，《统计年鉴》，35，2654–2690。Xia, Y.、Tong, H.、Li, W. K. 和 Zhu, L.-X.（2002），《降维空间的自适应估计》，英国皇家统计学会会刊*B*辑：统计方法学，64，363–410。Yin, X. 和 Li, B.（2011），《基于最小平均方差估计量集成的充分降维》，《统计年鉴》，39，3392–3416。Yu, L.、Zhao, P. 和 Zhou, W.（2025），《在高维因子模型中通过自助法样本协方差矩阵检验公共因子数量》，美国统计学会会刊，120，448–459。Yu, Y.、Wang, T. 和 Samworth, R. J.（2015），《适用于统计学家的Davis–Kahan定理的一个实用变体》，*Biometrika*，102，315–323。Yuan, M.、Ekici, A.、Lu, Z. 和 Monteiro, R.（2007），《多元线性回归中的降维与系数估计》，英国皇家统计学会会刊*B*辑：统计方法学，69，329–346。Zhu, L.、Miao, B. 和 Peng, H.（2006），《关于高维协变量下的切片逆回归》，美国统计学会会刊，101，630–643。Zhu, L.-X. 和 Ng, K. W.（1995），《切片逆回归的渐近性质》，*Statistica Sinica*，5，727–736。Zhu, Y. 和 Zeng, P.（2006），《用于估计回归中的中心子空间和中心均值子空间的傅里叶方法》，美国统计学会会刊，101，1638–1651。Zou, C.、Ke, Y. 和 Zhang, W.（2022），《多响应数据的低秩高维多元线性模型估计》，美国统计学会会刊，117，693 703– 。

《监督学习中模型无关且不确定性感知的降维》补充材料

本补充材料提供正文的补充细节，包括理论结果的证明、对假设和构造选择的进一步讨论，以及扩展的数值证据。A节汇集了文中命题和定理的证明。B节阐明了预测性诱导的序$d^*(\mathcal{L})$，并论证了分析中所用关键条件的合理性，包括动机示例的表示误差界。C节补充了数值研究和真实数据分析。图S.1 概述了本文的组织结构，便于快速查阅。

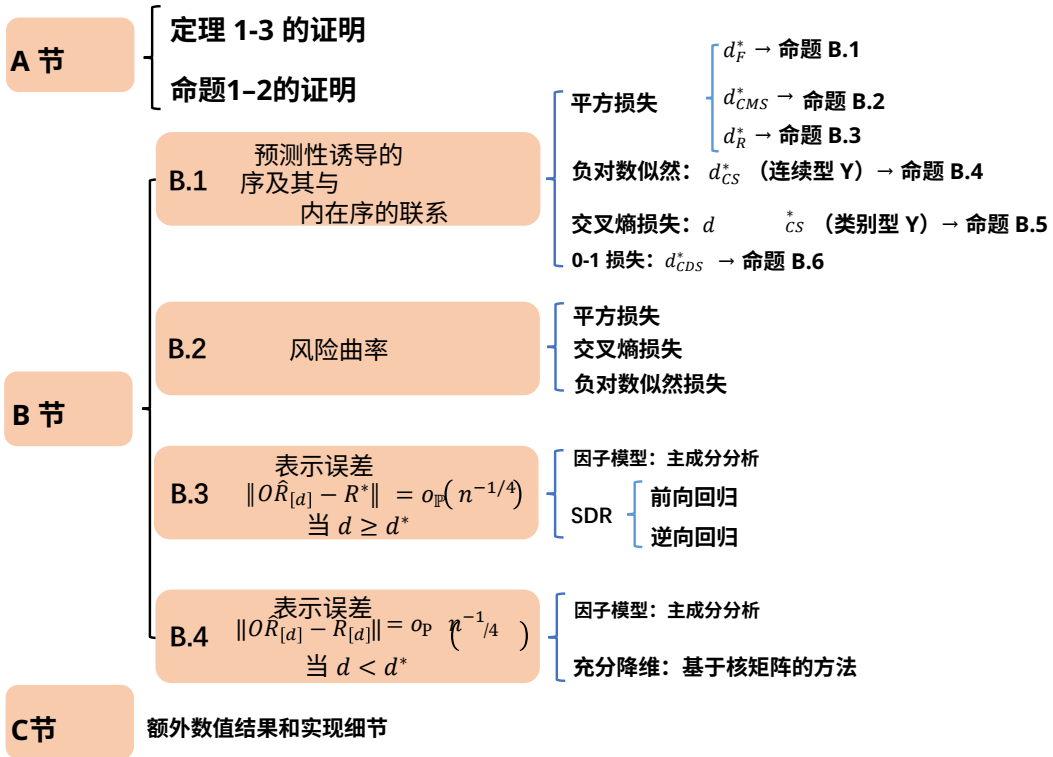


图 S.1：补充材料路线图。

A 证明

A.1 命题1

证明。回顾一下，在第2.1 节中，我们通过零填充将 R^* 嵌入到 $\mathbb{R}^{d_{\max}}$ 中；也就是说，$R_{[d^*]}=R^*$ 并且 $R_{-[d^*]}=0$。一方面，$\mathcal{G}_0 \subseteq \mathcal{G}_1 \subseteq \cdots \subseteq \mathcal{G}_{d_{\max}}$。因此，$\mathcal{L}_0 \geq \mathcal{L}_1 \geq \cdots \geq \mathcal{L}_d$最大值。另一方面，考虑两种情况。情况 1：$d^*=0$。则 $\mathcal{L}_d = \min_{g\in\mathcal{G}_d} \mathbb{E}_{\mathbb{P}}\{\ell(Y,g(0))\}$。对于任意 $g\in\mathcal{G}_d$，定义 $h_g(u)=g(0)$。于是 $h_g\in\mathcal{G}_0$，因此$\mathcal{L}_d = \min_{g\in\mathcal{G}_d} \mathbb{E}_{\mathbb{P}}\{\ell(Y,h_g(R))\} \geq \min_{g\in\mathcal{G}_0} \mathbb{E}_{\mathbb{P}}\{\ell(Y,g(R))\} = \mathcal{L}_0$。情况 2：$d^*\geq 1$。则$\mathcal{L}_d = \min_{g\in\mathcal{G}_d} \mathbb{E}_{\mathbb{P}}\{\ell(Y,g(R^*,0))\}$。对于任意 $g\in\mathcal{G}_d$，定义 $f_g(u)=g(u_{[d^*]},0)$。于是 $f_g\in\mathcal{G}_{d^*}$，因此，$\mathcal{L}_d = \min_{g\in\mathcal{G}_d} \mathbb{E}_{\mathbb{P}}\{\ell(Y,f_g(R))\} \geq \min_{g\in\mathcal{G}_{d^*}} \mathbb{E}_{\mathbb{P}}\{\ell(Y,g(R))\} = \mathcal{L}_{d^*}$。综合这些结果可得 $\mathcal{L}_0 \geq \mathcal{L}_1 \geq \cdots \geq \mathcal{L}_{d^*} = \mathcal{L}_{d^*+1} = \cdots = \mathcal{L}_d$最大值。 □

定理 1的证明

证明。回顾一下，对于所有 $d^* \leq d \leq d_{\max}$，$\mathcal{L}(g_d, \mathbb{P}) = \mathcal{L}(g_{d^*}, \mathbb{P})$。固定 $J\in\{o,a,b\}$ 和 $k\in[K]$。我们首先分解

$$\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_d,\mathbb{P}) = \widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_{d^*}, \widehat{\mathbb{P}}_{k,J}) + \mathcal{L}(g_{d^*}, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_{d^*},\mathbb{P}),$$

其中

$$\mathcal{L}(g_{d^*}, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_{d^*},\mathbb{P}) = \frac{1}{|I_{k,J}|}\sum_{i\in I_{k,J}} \ell(Y_i, g_{d^*}(R_i)) - \mathcal{L}(g_{d^*},\mathbb{P})$$

$\widehat{\ }$ $\widehat{\ }$ 是渐近正态性的主导项。因此，只需证明 $\mathcal{L}_{-k}(\widehat{g}_{-k;d}, \mathbb{P}_{k,\ J}) - \mathcal{L}(g_{d^*}, \widehat{\mathbb{P}}_{k,\ J}) = o_{\mathbb{P}}(n^{-1/2})$。

记

$$\begin{aligned}&\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_{d^*}, \widehat{\mathbb{P}}_{k,J})\\ =&\underbrace{\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \mathbb{P}) - \mathcal{L}(g_{d^*},\mathbb{P})}_{:=A_n} + \underbrace{\{\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{P}}_{k,J}) - \widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d},\mathbb{P})\} - \{\mathcal{L}(g_{d^*},\widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_{d^*},\mathbb{P})\}}_{:=B_n}.\end{aligned}$$

由（C1），$|A_n| \leq C\|\widehat{g}_{-k;d}\circ\widehat{\varphi}_{-k} - g_{d^*}\|^2$。由（C2），$\|\widehat{g}_{-k;d}\circ\widehat{\varphi}_{-k} - g_{d^*}\| = o_{\mathbb{P}}(n^{-1/4})$，因此 $A_n = o_{\mathbb{P}}(n^{-1/2})$。

重写

$$B_n = \frac{1}{|I_{k,J}|}\sum_{i\in I_{k,J}} \underbrace{\left[\left\{\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d},\delta_{Z_i}) - \widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d},\mathbb{P})\right\} - \{\mathcal{L}(g_{d^*},\delta_{Z_i}) - \mathcal{L}(g_{d^*},\mathbb{P})\}\right]}_{:=\widehat{\xi}_{d,k}(Z_i)}.$$

$\widehat{\ }$令 $\mathcal{D}_k=\{Z_i\colon i\in I_k\}$，对于 $k=1,\ldots\ldots,K$。通过交叉拟合，$\mathbb{E}[\xi_{d,k}(Z) \mid \cup_{j\neq k}\mathcal{D}_j] = 0$。

因此，对于任意 $\eta > 0$，由切比雪夫不等式可得

$$\begin{aligned} 0 &\le \Pr(|I_{k,J}|^{-1/2} \sum_{i \in I_{k,J}} \widehat{\xi}_{d,k}(Z_i) > \eta \mid \cup_{j \neq k} \mathcal{D}_j) \\ &\le \frac{\mathrm{var}[\widehat{\xi}_{d,k}(Z) \mid \cup_{j \neq k} \mathcal{D}_j]}{\eta^2} \le \frac{\mathbb{E}[(\widehat{\xi}_{d,k}(Z))^2 \mid \cup_{j \neq k} \mathcal{D}_j]}{\eta^2}. \end{aligned}$$

根据（C1）–（C3），

$$\begin{aligned} &\mathbb{E}[\{\widehat{\xi}_{d,k}(Z)\}^2 \mid \cup_{j \neq k} \mathcal{D}_j] \\ \le& 2\mathbb{E}[\{\ell(Y, \widehat{g}_{-k;d} \circ \widehat{\varphi}_{-k}(X)) - \ell(Y, g_{d^*}(R))\}^2 \mid \cup_{j \neq k} \mathcal{D}_j] \\ &+ 2\mathbb{E}[\{\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \mathbb{P}) - \mathcal{L}(g_{d^*}, \mathbb{P})\}^2 \mid \cup_{j \neq k} \mathcal{D}_j] \\ =& o_{\mathbb{P}}(1). \end{aligned}$$

因此，根据控制收敛，$|I_{k,J}|^{-1/2} \sum_{i \in I_{k^*J}} \widehat{\xi}_{d,k}(Z_i) = o_{\mathbb{P}}(1)$，从而$B_n = o_{\mathbb{P}}(n^{-1/2})$。

综合以上各部分，在（C1）–（C3）条件下，对于 $d^* \le d \le d_{\max}$、 $J \in \{o, a, b\}$ 和$k \in [K]$,

$$\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_d, \mathbb{P}) = \frac{1}{|I_{k,J}|} \sum_{i \in I_{k,J}} \ell(Y_i, g_{d^*}(R_i)) - \mathcal{L}(g_{d^*}, \mathbb{P}) + o_{\mathbb{P}}(|I_{k,J}|^{-1/2}). \quad \text{(A.1)}$$

由于对于 $d \ge d^*$， $\mathcal{L}_d = \mathcal{L}_{d^*}$，因此

$$\begin{aligned} \widehat{\psi}_d = \frac{1}{K} \sum_{k \in [K]} \Big\{ &\tau \cdot \big(\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{P}}_{k,o}) - \widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d_{\max}}, \widehat{\mathbb{P}}_{k,o})\big) \\ &+ (1-\tau) \cdot \big(\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{P}}_{k,a}) - \widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d_{\max}}, \widehat{\mathbb{P}}_{k,b})\big) \Big\} \\ = \frac{1}{n/(2-\tau)} &\left\{ \sum_{k \in [K]} \sum_{i \in I_{k,a}} \ell(Y_i, g_{d^*}(R_i)) - \sum_{k \in [K]} \sum_{i \in I_{k,b}} \ell(Y_i, g_{d^*}(R_i)) \right\} + o_{\mathbb{P}}(n^{-1/2}). \end{aligned}$$

根据中心极限定理,

$$\{n/(2-\tau)\}^{1/2} \widehat{\psi}_d \rightsquigarrow \mathcal{N}(0, 2(1-\tau)\sigma_{d^*}^2),$$

其中 $\sigma_{d^*}^2 = \mathrm{var}\{\ell(Y, g_{d^*}(R))\}$。 □

A.3 命题 2

证明。 在（C2）–（C3）条件下，并根据 Chen 等人 (2024年a）命题 2.1 的证明和注 2.2 中使用的相同论证，可得对于任意 $d^* \le d \le d_{\max}$， $k \in [K]$,

$\widehat{\sigma}^2_{d,\ k} \to \sigma^2_{d^*}$ 依概率成立。由连续映射定理可得 $\widehat{\nu}^2_d$ 的一致性。

□

A.4 定理2的证明

证明。固定 $J \in \{o, a, b\}$ 和 $k \in [K]$。将其分解为

$$\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_d, \mathbb{P}) = \widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_d, \widehat{\mathbb{P}}_{k,J}) + \mathcal{L}(g_d, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_d, \mathbb{P}).$$

由于 $\mathcal{L}(g,\ \mathbb{P}) = \mathbb{E}_{\mathbb{P}}[\ell(Y,\ g(R))] = \mathbb{E}_{\mathbb{P}}[\|Y - g(R)\|^2]$ 关于 $\mathbb{P}$是线性的，因此有

$$\mathcal{L}(g_d, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_d, \mathbb{P}) = \frac{1}{|I_{k,J}|} \sum_{i \in I_{k,J}} \ell(Y_i, g_d(R_i)) - \mathcal{L}(g_d, \mathbb{P}).$$

接下来，写作

$$\begin{aligned}
&\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_d, \widehat{\mathbb{P}}_{k,J}) \\
&= \underbrace{\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \mathbb{P}) - \mathcal{L}(g_d, \mathbb{P})}_{:=A'_n} + \underbrace{\{\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{P}}_{k,J}) - \widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \mathbb{P})\} - \{\mathcal{L}(g_d, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_d, \mathbb{P})\}}_{:=B'_n}.
\end{aligned}$$

我们可以表示为

$$\begin{aligned}
A'_n &= \mathbb{E}_{\mathbb{P}}\{\|Y - \widehat{g}_{-k;d}(\widehat{\varphi}_{-k}(X))\|^2 - \|Y - g_d(R)\|^2 \mid \widehat{g}_{-k;d}, \widehat{\varphi}_{-k}\} \\
&= \underbrace{2\mathbb{E}_{\mathbb{P}}\{\langle Y - g_d(R), g_d(R) - \widehat{g}_{-k;d}(\widehat{\varphi}_{-k}(X))\rangle \mid \widehat{g}_{-k;d}, \widehat{\varphi}_{-k}\}}_{:=A'_{n,k,d}} + \underbrace{\|\widehat{g}_{-k;d} \circ \widehat{\varphi}_{-k} - g_d\|^2_{L_2(\mathbb{P})}}_{=o_{\mathbb{P}}(n^{-1/2})}.
\end{aligned}$$

根据（C2'），$|A'_{n,\ k,\ d}| \leq 2\sqrt{\mathrm{tr}(\mathrm{var}_{\mathbb{P}}(Y))}\|\widehat{g}_{-k;d} \circ \widehat{\varphi}_{-k} - g_d\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}(n^{-1/4})$。在（C2'）–（C3'）条件下，我们同样可以得到 $B'_n = o_{\mathbb{P}}(n^{-1/2})$，论证与定理1的证明中用于推导 $B_n = o_{\mathbb{P}}(n^{-1/2})$ 的相同。

特殊情况：$d = 0$。在平方损失下，令 $g_0 = \mathbb{E}(Y)$，使得 $\mathcal{L}(g_0,\ \mathbb{P}) = \mathrm{tr}(\mathrm{var}_{\mathbb{P}}(Y))$。对于任意 k，$\widehat{g}_{-k;0} = Y_{-k} = \sum_{i \in I_{-k}} Y_i/|I_{-k}|$，因此 Y是在不使用 $(R,\ X)$ 的情况下预测的。在这种情况下，

$$\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;0}, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_0, \mathbb{P}) = \frac{1}{|I_{k,J}|} \sum_{i \in I_{k,J}} \ell(Y_i, g_0(R_i)) - \mathcal{L}(g_0, \mathbb{P}) + O_{\mathbb{P}}(n^{-1}).$$

为方便记号，令 $A'_{n,\ k,\ 0} = 0$，此后对所有 $k \in [K]$ 均如此。

因此，我们得到统一表达式：在（C2'）–（C3'）下，对于 $0 \leq d < d^*$,

$J \in \{o, a, b\}$，以及 $k \in [K]$,

$$\widehat{\mathcal{L}}_{-k}(\widehat{g}_{-k;d}, \widehat{\mathbb{P}}_{k,J}) - \mathcal{L}(g_d, \mathbb{P}) = \frac{1}{|I_{k,J}|} \sum_{i \in I_{k,J}} \ell(Y_i, g_d(R_i)) - \mathcal{L}(g_d, \mathbb{P}) + A'_{n,k,d} + o_{\mathbb{P}}(|I_{k,J}|^{-1/2}). \tag{A.2}$$

将此展开式与定理 1 的结果结合，可得，对于 $0 \le d < d^*$,

$$\begin{aligned}
&\widehat{\psi}_d - (\mathcal{L}_d - \mathcal{L}_{d_{\max}}) - \frac{1}{K}\sum_{k=1}^{K} A'_{n,k,d} \\
=&\frac{1}{n/(2-\tau)} \sum_{k \in [K]} \Bigg[\sum_{i \in I_{k,a}} \{\ell(Y_i, g_d(R_i)) - \mathcal{L}(g_d, \mathbb{P})\} - \sum_{i \in I_{k,b}} \{\ell(Y_i, g_{d^*}(R_i)) - \mathcal{L}(g_{d^*}, \mathbb{P})\} \\
&+ \sum_{i \in I_{k,o}} \{\ell(Y_i, g_d(R_i)) - \ell(Y_i, g_{d^*}(R_i)) - (\mathcal{L}(g_d, \mathbb{P}) - \mathcal{L}(g_{d^*}, \mathbb{P}))\} \Bigg] + o_{\mathbb{P}}\left(n^{-1/2}\right).
\end{aligned}$$

根据标准中心极限定理,

$$\sqrt{n/(2-\tau)} \left(\widehat{\psi}_d - (\mathcal{L}_d - \mathcal{L}_{d^*}) - \frac{1}{K}\sum_{k=1}^{K} A'_{n,k,d} \right) \rightsquigarrow \mathcal{N}\left(0, (1-\tau)\left(\sigma_d^2 + \sigma_{d^*}^2\right) + \tau\eta_d^2\right),$$

其中 $\eta_d^2 = \mathsf{var}_{\mathbb{P}}\{\ell(Y, g_d(R)) - \ell(Y, g_{d^*}(R))\}$。令 $\nu_{d,\eta}^2 = (1-\tau)\left(\sigma_d^2 + \sigma_{d^*}^2\right) + \tau\eta_d^2$和$\nu_d^2 = (1-\tau)\left(\sigma_d^2 + \sigma_{d^*}^2\right)$。通过与命题 2 类似的论证，在（C2'）–（C3'）条件下，对于 $0 \le d < d^*$ 以及任意 $\tau \in [0, 1)$，我们有 $\widehat{\nu}_d^2 \to \nu_d^2$ 依概率成立，当 $n \to \infty$ 时。对于任意$0 \le d < d^*$, $\alpha \in (0, 1)$，以及足够小的 $0 < \delta < (\mathcal{L}_d - \mathcal{L}_{d^*})$,

$$\begin{aligned}
&\Pr\Big(\frac{\sqrt{n/(2-\tau)}\widehat{\psi}_d}{\widehat{\nu}_d} < z_{1-\alpha}\Big) \\
=&\Pr\Big(\underbrace{\frac{\sqrt{n/(2-\tau)}\left(\widehat{\psi}_d - (\mathcal{L}_d - \mathcal{L}_{d^*}) - \sum_{k=1}^{K} A'_{n,k,d}/K\right)}{\nu_{d,\eta}}}_{:=\widehat{\xi}_d} < z_{1-\alpha} \cdot \frac{\widehat{\nu}_d}{\nu_{d,\eta}} \\
&\qquad\qquad - \frac{\sqrt{n/(2-\tau)}\big((\mathcal{L}_d - \mathcal{L}_{d^*}) + \sum_{k=1}^{K} A'_{n,k,d}/K\big)}{\nu_{d,\eta}}\Big) \\
=&\Pr\Big(\widehat{\xi}_d < z_{1-\alpha} \cdot \frac{\widehat{\nu}_d}{\nu_{d,\eta}} - \frac{\sqrt{n/(2-\tau)}\big((\mathcal{L}_d - \mathcal{L}_{d^*}) + \sum_{k=1}^{K} A'_{n,k,d}/K\big)}{\nu_{d,\eta}}, \\
&\qquad\qquad |\sum_{k \in [K]} A'_{n,k,d}/K| \le \delta\Big) + o(1) \\
\le&\Pr\Big(\widehat{\xi}_d < z_{1-\alpha} \cdot \frac{\widehat{\nu}_d}{\nu_{d,\eta}} - \frac{\sqrt{n/(2-\tau)}(\mathcal{L}_d - \mathcal{L}_{d^*} - \delta)}{\nu_{d,\eta}}\Big) + o(1),
\end{aligned}$$

其中，第二个等式成立是因为对所有 $k \in [K]$, $A'_{n,k,d} = o_{\mathbb{P}}(1)$。

最后，由于事件 $\{\widehat{d} < d^*\} = \bigcup_{d=0}^{d^*-1}\{\widehat{T}_d < z_{1-\alpha}\}$，我们得到

$$\begin{aligned}
\Pr(\widehat{d} < d^*) &\leq \sum_{d=0}^{d^*-1} \Pr\left(\frac{\sqrt{n/(2-\tau)}\widehat{\psi}_d}{\widehat{\nu}_d} < z_{1-\alpha}\right) \\
&\leq \sum_{d=0}^{d^*-1} \Pr\left(\widehat{\xi}_d < z_{1-\alpha} \cdot \frac{\widehat{\nu}_d}{\nu_{d,\eta}} - \frac{\sqrt{n/(2-\tau)}(\mathcal{L}_d - \mathcal{L}_{d^*} - \delta)}{\nu_{d,\eta}}\right) + o(1) \\
&\leq \sum_{d=0}^{d^*-1} \Phi\left(\frac{\nu_d}{\nu_{d,\eta}} z_{1-\alpha} - \frac{\sqrt{n/(2-\tau)}(\mathcal{L}_d - \mathcal{L}_{d^*} - \delta)}{\nu_{d,\eta}}\right) + o(1),
\end{aligned}$$

其中最后几步用到了 $\widehat{\xi}_d \rightsquigarrow \mathcal{N}(0,1)$, $\widehat{\nu}_d^2 \to \nu_d^2$依概率以及 Polya定理（例如van derVaart，1998中的引理 2.11）。 □

A.5 定理 3 的证明

证明。 首先，根据检验程序的定义，

$$\begin{aligned}
\Pr(\widehat{d} > d^*) &\leq \Pr\left(\frac{\sqrt{n/(2-\tau)}\widehat{\psi}_{d^*}}{\widehat{\nu}_{d^*}} \geq z_{1-\alpha_n}\right) \\
&\leq \left|\Pr\left(\frac{\sqrt{n/(2-\tau)}\widehat{\psi}_{d^*}}{\widehat{\nu}_{d^*}} \geq z_{1-\alpha_n}\right) - \alpha_n\right| + \alpha_n \\
&\leq \sup_{\beta\in(0,1)} \left|\Pr\left(\frac{\sqrt{n/(2-\tau)}\widehat{\psi}_{d^*}}{\widehat{\nu}_{d^*}} \geq z_{1-\beta}\right) - \beta\right| + \alpha_n \\
&= o(1),
\end{aligned}$$

其中，最后一个等号由 Polya定理和假设 $\alpha_n = o(1)$ 得出。

接下来，根据定理 2 证明中的论证，

$$\begin{aligned}
&\Pr\left(\frac{\sqrt{n/(2-\tau)}\widehat{\psi}_d}{\widehat{\nu}_d} < z_{1-\alpha_n}\right) \\
=&\Pr\left(\frac{\sqrt{n/(2-\tau)}(\widehat{\psi}_d - \psi_d)}{\nu_{d,\eta}} - \frac{(\widehat{\nu}_d - \nu_d)}{\nu_{d,\eta}} z_{1-\alpha_n} < \frac{\nu_d}{\nu_{d,\eta}} z_{1-\alpha_n} - \frac{\{n/(2-\tau)\}^{1/2}(\mathcal{L}_d - \mathcal{L}_{d^*})}{\nu_{d,\eta}}\right).
\end{aligned}$$

由于 $\widehat{\nu}_d - \nu_d = O_{\mathbb{P}}(n^{-1/2})$，如果 $z_{1-\alpha_n} = o(n^{1/2})$，那么

$$\Pr\left(\frac{\sqrt{n/(2-\tau)}\widehat{\psi}_d}{\widehat{\nu}_d} < z_{1-\alpha_n}\right) = \Phi\left(\frac{\nu_d}{\nu_{d,\eta}} z_{1-\alpha_n} - \frac{\{n/(2-\tau)\}^{1/2}(\mathcal{L}_d - \mathcal{L}_{d^*} - \delta)}{\nu_{d,\eta}}\right) + o(1),$$

根据 Slutsky定理和Polya定理。因此，

$$\Pr(\widehat{d} < d^*) \leq \sum_{d=0}^{d^*-1} \Phi\left(\frac{\nu_d}{\nu_{d,\eta}} z_{1-\alpha_n} - \frac{\{n/(2-\tau)\}^{1/2} (\mathcal{L}_d - \mathcal{L}_{d^*} - \delta)}{\nu_{d,\eta}} \right) + o(1).$$

最后，回顾标准正态尾部近似：

$$\Pr\{\mathcal{N}(0,1) > z_{1-\alpha_n}\} \sim (2\pi)^{-1/2} z_{1-\alpha_n}^{-1} \exp\left(-z_{1-\alpha_n}^2/2\right), \text{ as } n \to \infty.$$

设 γ_n是一个序列，满足 $\alpha_n = \exp\left(-z_{1-\gamma_n}^2/2\right)$。则 $z_{1-\alpha_n} < z_{1-\gamma_n} = (-2\log\ \alpha_n)^{1/2}$。在假设 $\alpha_n \to 0$ 和 $(\log \alpha_n)/n \to 0$下，我们有 $z_{1-\alpha_n} = o(n^{1/2})$，因此$\widehat{\Pr}(d < d^*) = o(1)$，因为 $(\mathcal{L}_d - \mathcal{L}_{d^*} - \delta) > 0$ 对 $d < d^*$成立。

结合这两个界可得 $\widehat{\Pr}(d = d^*) = 1$，当 $n \to \infty$时成立。 □

B 预测性诱导的序及其所需的理论条件

B.1 预测性诱导的序及其与内在序的联系

除非另有说明，我们假设 $d^* \geq 1$；当 $d^* = 0$时，$d^*(\mathcal{L})$ 的定义与命题 1 共同意味着 $d^*(\mathcal{L}) = 0$。我们从回归问题开始，其中平方损失函数和负对数似然是标准选择。

命题 B.1。在因子回归中，采用平方损失时，$d^*(\mathcal{L}) = d_{\mathrm{F}}^*$ 前提是噪声 $\varepsilon \perp\!\!\!\perp f$ 均值为零且具有有限二阶矩，并且回归函数满足 $m(f) \neq \mathbb{E}_{\mathbb{P}}(Y \mid f_S)$ 对每个真子集 $S \subsetneq [d_{\mathrm{F}}^*]$ 均成立。

证明。为保持记号一致，采用约定 $\sigma(f_{[0]}) = \sigma(f_\varnothing) = \{\varnothing, \Omega\}$。反设 $d^*(\mathcal{L}) < d_{\mathrm{F}}^*$成立。则 $\mathbb{E}_{\mathbb{P}}(Y - m(f))^2 = \mathbb{E}_{\mathbb{P}}(Y - \mathbb{E}_{\mathbb{P}}(Y \mid f_{[d_{\mathrm{F}}^*-1]}))^2$，，其中 $m(f) = \mathbb{E}_{\mathbb{P}}(Y \mid f)$。这意味着 $m(f) = \mathbb{E}_{\mathbb{P}}(Y \mid f_{[d_{\mathrm{F}}^*-1]})$，与假设矛盾：对于任意真子集 $S \subsetneq [d_{\mathrm{F}}^*]$，$m(f) \neq \mathbb{E}_{\mathbb{P}}(Y \mid f_S)$ 均成立。因此，$d^*(\mathcal{L}) = d_{\mathrm{F}}^*$。 □

命题 B.2。在充分降维中，在平方损失下，$d^*(\mathcal{L}) = d_{\mathrm{CMS}}^*$只要中心均值子空间存在且 $\mathbb{E}_{\mathbb{P}} Y^2 < \infty$。

证明。如果 $d^*(\mathcal{L}) < d_{\mathrm{CMS}}^*$，那么根据定义 $\mathcal{L}_{d_{\mathrm{CMS}}^*-1} = \mathcal{L}_{d_{\mathrm{CMS}}^*}$，这意味着 $\mathbb{E}_{\mathbb{P}}(Y \mid X) = \mathbb{E}_{\mathbb{P}}(Y \mid B_{d_{\mathrm{CMS}}^*-1}^\top X)$，其中 B_d 表示 B的前 d 列；我们采用约定$\sigma(B_0^\top X) = \{\varnothing, \Omega\}$。这与结构维数 d_{CMS}^*的最小性矛盾。因此，结论得证。

□

命题 B.3。 在降秩回归中，采用平方损失函数时， $d^*(\mathcal{L}) = d_{\mathrm{R}}^*$ ，前提是噪声 $\varepsilon \perp\!\!\!\perp X$ 均值为零且具有有限二阶矩。

证明。 反设， $d^*(\mathcal{L}) < d_{\mathrm{R}}^*$。将回归函数写为 $f(B^\top X) \triangleq \Omega^\top B^\top X$。在所述噪声假设下，$\Omega^\top B^\top X = \mathbb{E}_{\mathbb{P}}(Y \mid B_{d_{\mathrm{R}}^*-1}^\top X)$。由于 $\mathrm{span}(B_{d_{\mathrm{R}}^*-1}) \subsetneq \mathrm{span}(B)$，存在 $\alpha \in \mathbb{R}^p$ 使得 $B_{d_{\mathrm{R}}^*-1}^\top \alpha = 0$但 $B^\top \alpha \neq 0$。因此，对于任意标量 $k \in \mathbb{R}$,

$$f(B^\top X + kB^\top \alpha) = \mathbb{E}_{\mathbb{P}}(Y \mid B_{d_{\mathrm{R}}^*-1}^\top X + kB_{d_{\mathrm{R}}^*-1}^\top \alpha) = \mathbb{E}_{\mathbb{P}}(Y \mid B_{d_{\mathrm{R}}^*-1}^\top X) = f(B^\top X).$$

令 $\beta_1 = B^\top \alpha / \|B^\top \alpha\| \in \mathbb{R}^{d_{\mathrm{R}}^*}$，并将其扩充为$\mathbb{R}^{d_{\mathrm{R}}^*}$的一组正交归一基 $\{\beta_1, \beta_2, \ldots, \beta_{d_{\mathrm{R}}^*}\}$。于是

$$f(B^\top X) = f\big(\sum_{j \in [d_{\mathrm{R}}^*]} \langle B^\top X, \beta_j \rangle \beta_j\big) = f\big(\sum_{j=2}^{d_{\mathrm{R}}^*} \langle B^\top X, \beta_j \rangle \beta_j\big),$$

因此 $f(B^\top X) = f(\widetilde{B}^\top X)$，其中 $\widetilde{B} = B \sum_{}^{d_{\mathrm{R}}^*} \beta_j \beta_j^\top$的秩至多为 $d_{\mathrm{R}}^* - 1$。于是，对于所有 X，$\Omega^\top B^\top X = \widetilde{\Omega^\top B^\top X}$，这意味着 $B\Omega = B\Omega$。这与假设 $A = B\Omega$ 的秩为满秩 d_{R}^* 相矛盾。因此， $d_{\mathrm{R}}^*(\mathcal{L}) = d_{\mathrm{R}}^*$。

□

命题 B.4。 令 $\ell(y, g(\cdot \mid r)) = -$对数 $g(y \mid r)$ 为负对数似然，其中$g(\cdot \mid r)$ 表示由 $R = r$ 选定的条件密度，并且属于该类

$$\mathcal{G} = \Big\{ g : \mathbb{R}^{d_{\max}} \to L^1(\mathcal{Y}) \mid g(\cdot \mid r) \geq 0, \int_{\mathcal{Y}} g(y \mid r)\, dy = 1 \Big\}.$$

则 $d^*(\mathcal{L}) = d_{\mathrm{CS}}^*$。

证明。 Y给定 $R = r$ 时的条件密度 $f(\cdot \mid r)$ 是在 $\mathcal{G}$ 上对期望负对数似然进行无约束优化得到的总体最小化解。反设$d^*(\mathcal{L}) < d_{\mathrm{CS}}^*$。则 $\mathcal{L}_{d_{\mathrm{CS}}^*} = \mathcal{L}_{d_{\mathrm{CS}}^*-1}$，这意味着

$$\mathbb{E}_{\mathbb{P}}\{\log f(Y \mid B_{d_{\mathrm{CS}}^*-1}^\top X)\} = \mathbb{E}_{\mathbb{P}}\{\log f(Y \mid B_{d_{\mathrm{CS}}^*}^\top X)\}.$$

因此， $\mathbb{E}_{\mathbb{P}}\Big[\mathbb{E}_{\mathbb{P}}\Big\{ \underline{\text{对数 } f(Y \mid B_{d_{\mathrm{CS}}^*-1}^\top X)\, f(Y \mid B_{d_{\mathrm{CS}}^*}^\top X)} \mid X \Big\}\Big] = 0.$ 根据詹森不等式,

$$\mathbb{E}_{\mathbb{P}}\Big\{ \log \frac{f(Y \mid B_{d_{\mathrm{CS}}^*-1}^\top X)}{f(Y \mid B_{d_{\mathrm{CS}}^*}^\top X)} \mid X \Big\} \leq \log \mathbb{E}_{\mathbb{P}}\Big\{ \frac{f(Y \mid B_{d_{\mathrm{CS}}^*-1}^\top X)}{f(Y \mid B_{d_{\mathrm{CS}}^*}^\top X)} \mid X \Big\} = \log 1 = 0,$$

其中，等式 $\mathbb{E}_{\mathbb{P}}\Big\{ \underline{f(Y \mid B_{d_{\mathrm{CS}}^*-1}^\top X)\, f(Y \mid B_{d_{\mathrm{CS}}^*}^\top X)} \mid X \Big\} = 1$ 是根据中心

子空间的定义得出的。
因此

$$\mathbb{E}_{\mathbb{P}}\Big\{\log\frac{f(Y\mid B_{d^*_{\mathrm{CS}}-1}^{\top}X)}{f(Y\mid B_{d^*_{\mathrm{CS}}}^{\top}X)}\mid X\Big\}=0 \text{ almost surely.}$$

詹森不等式取等号，当且仅当 $f(Y\mid B_{d^*_{\mathrm{CS}}-1}^{\top}X)=f(Y\mid B_{d^*_{\mathrm{CS}}}^{\top}X)$ 几乎处处成立，这与 d^*_{CS} 的定义矛盾。因此，$d^*(\mathcal{L})=d^*_{\mathrm{CS}}$. □

现在我们转向分类问题，此时响应变量 Y 是分类变量。我们考虑两种广泛使用的损失函数：交叉熵和 0-1 损失。

命题 B.5. 在交叉熵损失下，$d^*(\mathcal{L})=d^*_{\mathrm{CS}}$，前提是中心子空间存在。

证明。 反设 $d^*(\mathcal{L})<d^*_{\mathrm{CS}}$。则 $\mathcal{L}_{d^*_{\mathrm{CS}}}=\mathcal{L}_{d^*_{\mathrm{CS}}-1}$，这意味着 $H(Y\mid B_{d^*_{\mathrm{CS}}-1}^{\top}X)=H(Y\mid B_{d^*_{\mathrm{CS}}}^{\top}X)$，其中 $H(Y\mid R)=-\mathbb{E}_{\mathbb{P}}\{\log\Pr(Y\mid R)\}$ 表示条件熵。根据条件熵的数据处理不等式，

$$H(Y\mid B_{d^*_{\mathrm{CS}}-1}^{\top}X)\geq H(Y\mid B_{d^*_{\mathrm{CS}}}^{\top}X),$$

当且仅当.....时等号成立

$$Y\perp\!\!\!\perp B_{d^*_{\mathrm{CS}}}^{\top}X\mid B_{d^*_{\mathrm{CS}}-1}^{\top}X.$$

这意味着

$$\Pr(Y\mid B_{d^*_{\mathrm{CS}}-1}^{\top}X)=\Pr(Y\mid X),$$

这与 d^*_{CS} 作为满足 $\Pr(Y\mid X)=\Pr(Y\mid B_{d^*_{\mathrm{CS}}}^{\top}X)$ 的最小维度的定义相矛盾。因此，结论得证。 □

命题 B.6. 在 *0-1* 损失下，$d^*(\mathcal{L})=d^*_{\mathrm{CDS}}$，前提是中心判别子空间存在，其中 d^*_{CDS} 表示其结构维数。

证明。 定义最优分类器

$$b_{d^*_{\mathrm{CDS}}-1}(X)=\arg\max_y\Pr(Y=y\mid B_{d^*_{\mathrm{CDS}}-1}^{\top}X),$$
$$b_{d^*_{\mathrm{CDS}}}(X)=\arg\max_y\Pr(Y=y\mid B_{d^*_{\mathrm{CDS}}}^{\top}X).$$

根据中心判别子空间的定义（Cook 和 Yin，2001年），$\arg\max_y\Pr(Y=y\mid B_{d^*_{\mathrm{CDS}}}^{\top}X)=\arg\max_y\eta_y(X)$，其中 $\eta_y(X)=\Pr(Y=y\mid X)$。假设 $d^*(\mathcal{L})<d^*_{\mathrm{CDS}}$。定义 $\eta_*(X)=\max_y\eta_y(X)$。则 $\mathcal{L}_{d^*_{\mathrm{CDS}}}=\mathcal{L}_{d^*_{\mathrm{CDS}}-1}$，并且在 0-1 损失下，

$$\mathcal{L}_{d^*_{\mathrm{CDS}}}=\Pr(Y\neq b_{d^*_{\mathrm{CDS}}}(X))=1-\mathbb{E}_{\mathbb{P}}(\eta_*(X))$$
$$\mathcal{L}_{d^*_{\mathrm{CDS}}-1}=\Pr(Y\neq b_{d^*_{\mathrm{CDS}}-1}(X))=1-\mathbb{E}_{\mathbb{P}}(\eta_{b_{d^*_{\mathrm{CDS}}-1}(X)}(X))$$

因此，$\mathbb{E}_{\mathbb{P}}(\eta_*(X)) = \mathbb{E}_{\mathbb{P}}(\eta_{b_{d^*_{\mathrm{CDS}}-1}(X)}(X))$。由于 $\eta_*(X) \geq \eta_{b_{d^*_{\mathrm{CDS}}-1}(X)}(X)$ 逐点成立，期望相等意味着 $\eta_*(X) = \eta_{b_{d^*_{\mathrm{CDS}}-1}(X)}(X)$，几乎处处成立。因此，$b_{d^*_{\mathrm{CDS}}-1}(X) = \arg\max_y \eta_y(X)$ 几乎处处成立，这与 d^*_{CDS} 的定义矛盾。因此，在 0-1 损失下，$d^*(\mathcal{L}) = d^*_{\mathrm{CDS}}$。 □

B.2 常见损失函数的条件（C1）验证

(1) 平方损失。 为验证平方损失下的（C1），我们假设 R 包含了 X 中与预测 Y 的条件均值相关的所有信息，即

$$\mathbb{E}_{\mathbb{P}}(Y \mid X, R) = \mathbb{E}_{\mathbb{P}}(Y \mid R). \tag{B.1}$$

对于任意 $g \in \mathcal{G}$ 和 $\varphi\colon \mathcal{X} \to \mathbb{R}^{d_{\max}}$，我们有

$$\begin{aligned}
&|\mathbb{E}_{\mathbb{P}}\{\|Y - g\circ\varphi(X)\|^2 - \|Y - g_{d^*}(R)\|^2\}| \\
=&\|g\circ\varphi(X) - g_{d^*}(R)\|^2_{L_2(\mathbb{P})} + 2\mathbb{E}_{\mathbb{P}}\{\langle Y - g_{d^*}(R), g_{d^*}(R) - g\circ\varphi(X)\rangle\}.
\end{aligned}$$

由于 $g_{d^*}(r) = \mathbb{E}_{\mathbb{P}}(Y \mid R = r)$ 是均方误差的极小化解，结合（B.1），$\mathbb{E}_{\mathbb{P}}\{\langle Y - g_{d^*}(R)\ g_{d^*}(R) - g\circ\varphi(X)\rangle\} = 0$，因此唯一的贡献来自 $\|g\circ\varphi(X) - g_{d^*}(R)\|^2_{L_2(\mathbb{P})}$，从而证明了平方损失下的（C1）。

(2) 交叉熵损失。 为验证交叉熵损失下的（C1），我们假设连接条件 $\Pr(Y \mid X, R) = \Pr(Y \mid R)$成立。对于任意 $g \in \mathcal{G}$ 和 $\varphi\colon \mathcal{X} \to \mathbb{R}^{d_{\max}}$，我们有

$$\begin{aligned}
&\Big|\mathbb{E}_{\mathbb{P}}\Big[\sum_{m=1}^{M} 1\{Y = m\}\log\Big(\frac{g\circ\varphi(X)_m}{g_{d^*}(R)_m}\Big)\Big]\Big| \\
=&\Big|\mathbb{E}_{\mathbb{P}}\Big[\mathbb{E}_{\mathbb{P}}\Big[\sum_{m=1}^{M} 1\{Y = m\}\log\Big(\frac{g\circ\varphi(X)_m}{g_{d^*}(R)_m}\Big) \mid X, R\Big]\Big]\Big| \\
=&\Big|\mathbb{E}_{\mathbb{P}}\Big[\sum_{m=1}^{M} \Pr(Y = m \mid X, R)\log\Big(\frac{g\circ\varphi(X)_m}{g_{d^*}(R)_m}\Big)\Big]\Big| \\
=&\Big|\mathbb{E}_{\mathbb{P}}\Big[\sum_{m=1}^{M} g_{d^*}(R)_m\log\Big(\frac{g\circ\varphi(X)_m}{g_{d^*}(R)_m}\Big)\Big]\Big|,
\end{aligned}$$

其中最后一个等式用到了 $g_{d^*}(R) = \Pr(Y \mid R)$ 和连接条件。按照 Williamson 等人 (2023) 的方法，对以下内容应用二阶泰勒展开：

在 $t=0$ 处对 $\log(1+t)$ 进行展开，可得对每个类别 $m \in [M]$,

$$\begin{aligned}
&\log\left(\frac{g\circ\varphi(X)_m}{g_{d^*}(R)_m}\right)\\
=&\log\left(1+\frac{g\circ\varphi(X)_m - g_{d^*}(R)_m}{g_{d^*}(R)_m}\right)\\
=&\frac{g\circ\varphi(X)_m - g_{d^*}(R)_m}{g_{d^*}(R)_m} - \frac{(g\circ\varphi(X)_m - g_{d^*}(R)_m)^2}{2g_{d^*}(R)_m^2(1+\xi_m)^2},
\end{aligned}$$

其中，ξ_m 逐点介于 0 与 $\frac{g\circ\varphi(X)_m - g_{d^*}(R)_m}{g_{d^*}(R)_m}$ 之间。因此，$g_{d^*}(R)_m(1+\xi_m)$逐点介于 $g_{d^*}(R)_m$ 与 $g\circ\varphi(X)_m$之间。假设对于某个 $\delta \in (0, 1/2)$,

$$g_{d^*}(R)_m, g\circ\varphi(X)_m \in (\delta, 1-\delta) \text{ almost surely for all } m \in [M]. \tag{B.2}$$

Then

$$\left|\mathbb{E}_{\mathbb{P}}\left[\sum_{m=1}^{M} g_{d^*}(R)_m \log\left(\frac{g\circ\varphi(X)_m}{g_{d^*}(R)_m}\right)\right]\right| \le \frac{1-\delta}{2\delta^2}\|g\circ\varphi - g_{d^*}\|^2_{L_2(\mathbb{P})},$$

由于 $\sum_{m=1}^{M} g\circ\varphi(X)_m = \sum_{m=1}^{M} g_{d^*}(R)_m = 1$。这验证了交叉熵损失在（B.2）条件下满足（C1），并且连接条件 $\Pr(Y \mid X, R) = \Pr(Y \mid R)$ 成立。

在例 2中，中心子空间由 $\Pr(Y \mid X) = \Pr(Y \mid B^\top X)$ 定义。因此，当 $R = B^\top X$ 时，连接条件自动成立。

(3) 负对数似然。 负对数似然的（C1）验证与交叉熵情形类似。只需假设连接条件

$$\Pr(Y \le y \mid X, R) = \Pr(Y \le y \mid R), \text{ for all } y \in \mathbb{R},$$

以及与（B.2）类似的结论：对于某个 $\delta \in (0, 1/2)$，

$$g_{d^*}(y \mid R), g(y \mid \varphi(X)) \in (\delta, 1-\delta), \text{ for all } y \in \mathbb{R}, \text{ almost surely.}$$

在例 2中，中心子空间由以下内容定义：$Y \perp\!\!\!\perp X \mid B^\top X$。因此，当$R = B^\top X$时，连接条件自然再次成立。

B.3 条件（C2）的支持：当 $d \ge d^*$ 时的表示误差界

B.3.1 因子回归

我们的目标是验证，在适当条件下，对于某个依赖于数据的旋转矩阵 O，表示误差满足$\|O\widehat{R}_{[d]} - R^*\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}(n^{-1/4})$。我们重点关注

作为示例 1 的代表性降维方法，主成分分析 (PCA)1。令 $\widehat{v}_1, \ldots, \widehat{v}_{d_{\max}} \in \mathbb{R}^p$ 表示由 $\{X_i$ 计算得到的样本协方差矩阵的前 $d_{\max}$ 个特征向量 $\{X_i : \ i \in I_{-k}\}$。定义

$$\widehat{R} = \widehat{\varphi}_{-k}(X) = p^{-1} W_{d_{\max}}^{\top} X,$$

其中 $W_{d_{\max}} = \sqrt{p}[\widehat{v}_1, \ldots, \widehat{v}_{d_{\max}}] \in \mathbb{R}^{p \times d_{\max}}$。为简洁起见，我们省略 $\widehat{k}$ 上的折索引 R，因为各折的分析相同。回顾例 1和命题B.1 可知 $R^* = f$ 且 $d^* = d_{\mathrm{F}}^*$ 在温和正则条件和平方损失下成立。对于$d \geq d_{\mathrm{F}}^*$，定义 $H_d = p^{-1} W_d^{\top} B \in \mathbb{R}^{d \times d_{\mathrm{F}}^*}$，其中 W_d 是 d 的前几列 $W_{d_{\max}}$，并将 $O = H_d^{\dagger} \in \mathbb{R}^{d_{\mathrm{F}}^* \times d}$ 定义为 H_d的伪逆。下一个命题将在附加假设下建立表示误差界。

(D1)（有界性）在例1中的因子结构下，存在通用常数 C_1 和 C_2，使得：

> 1. 因子载荷矩阵满足 $\max_{i \in [p], j \in [d_0]} |B_{i,j}| \leq C_1$。2. 每个因子 f_i 的均值为零，且对所有 $i \in [d_0]$，在 $[-C_2, C_2]$ 中有界。3. 每个特质成分 u_j 的均值为零，且对所有 $j \in [p]$，在 $[-C_2, C_2]$ 中有界。

(D2)（弱依赖）$\sum_{j, j' \in [p], j \neq j'} |\mathbb{E}[u_j u_{j'}]| = O(p)$。

(D3)（有界投影矩阵）$\max_{i \in [p], j \in [d_{\max}]} |(W_{d_{\max}})_{i,j}| \leq C_3$。

(D4)（广泛性） $p^{-1} B^{\top} \in \mathbb{R}^{\mathbf{d}_{\mathrm{F}}^* \times \mathbf{d}_{\mathrm{F}}^*}$ 的所有特征值均远离 0 和 ∞，当 $p \to \infty$时成立。

这些条件是因子模型文献中的标准条件（Fan 等，2013年；Fan 和 Gu，2024年），并且与（C1）–（C3）相容。

命题B.7。设 $1 \leq d_{\mathrm{F}}^* \leq d \leq d_{\max}$。假设（*D1*）–（*D4*）成立，且 $(\log p)/n = o(1)$当 $n \to \infty$ 时成立。那么，存在 $O = H_d^{\dagger} \in \mathbb{R}^{d_{\mathrm{F}}^* \times d}$ 使得

$$\|O\widehat{R}_{[d]} - f\|_{L_2(\mathbb{P})} = O_{\mathbb{P}}\left(p^{-1/2}\right).$$

证明。这可直接由Fan 和 Gu（2024年）的引理 2 和命题 1 得出。 □

根据命题B.7</style》，只要 (1) 且 (log)(1)，要求 $(\mathbb{P}) = o_{\mathbb{P}}(n^{-1/4})$ 就以 $O = H_d^{\dagger}$ 成立。

B.3.2 充分降维

我们证明，常用的充分降维（SDR）方法能够针对某个依赖于数据的旋转矩阵 O，满足表示要求 $\|O\widehat{R} - R^*\| = o_{\mathbb{P}}(n^{-1/4})$。假设，在用户选择的预测性度量 $\mathcal{L}$ 下，$B^\top X$ 包含了 X 中关于响应变量 $Y \in \mathbb{R}$ 的全部预测信息。根据损失函数 ℓ的不同，B 的张成空间可能表示中心子空间、中心均值子空间或其他任务特定充分子空间，而 d^* 表示其结构维数；参见 B.1节。对于每一折 k，将用户指定的 SDR 方法应用于 $\{(X_i, Y_i) : i \in I_{-k}\}$，以估计目标子空间，从而得到估计的 $d_{\max}$ 维子空间的基矩阵 $\widehat{B}_{d_{\max}} \in \mathbb{R}^{p\times d_{\max}}$。定义 $\widehat{R} = \widehat{B}_{d_{\max}}^\top X \in \mathbb{R}^{d_{\max}}$，并为简洁起见省略折索引。回顾例2，$R^* = B^\top X$。对于 $d \geq d^*$，定义$O = B^\top \widehat{B}_d \in \mathbb{R}^{d^*\times d}$，$\widehat{}$，其中 B_d 是 $B_{d_{\max}}$的前 d 列。下一个命题刻画了表示误差。

命题 B.8。 设 $1 \leq d^* \leq d \leq d_{\max}$。对于 $O = B^\top \widehat{B}_d \in \mathbb{R}^{d^*\times d}$，有

$$\|O\widehat{R}_{[d]} - B^\top X\| \leq \|P_B - P_{\widehat{B}_{d^*}}\|_2 \|X\|, \tag{B.3}$$

其中，P_A 表示投影到 A 的列空间上的正交投影。

证明。S 将 $O = B^\top \widehat{B}_d$ 和 $\widehat{R}_{[d]} = \widehat{B}_d^\top X$ 代入等式（B.3）左侧，得到

$$\begin{aligned}
\|O\widehat{R}_{[d]} - B^\top X\| =& \|B^\top \widehat{B}_d \widehat{B}_d^\top X - B^\top X\| \\
=& \|B^\top (I - P_{\widehat{B}_d}) X\| \\
\leq& \|B^\top (I - P_{\widehat{B}_d})\|_2 \|X\| \\
\leq& \|B^\top (I - P_{\widehat{B}_{d^*}})\|_2 \|X\| \\
=& \|P_B - P_{\widehat{B}_{d^*}}\|_2 \|X\|.
\end{aligned}$$

第二个等号利用了 $P_{\widehat{B}_d} = \widehat{B}_d \widehat{B}_d^\top$，因为 $\widehat{B}_d$ 具有标准正交列。第二个$\widehat{}$ $\widehat{}$不等式成立，因为当 $d \geq d^*$ 时，span$(B_{d^*}) \subseteq$span(B_d)，其中 span(A) 表示 A 的列空间。最后一个等号由恒等式 $\|A^\top Z\|_2 = \|P_A Z\|_2$ 得出，该恒等式适用于任意A （其列为标准正交列）以及任意矩阵 Z，并结合 Golub 和 Van Loan 的定理2.5.1（2013年）。

□

$\widehat{}$命题 B.8 表明，当 $\|X\|$ 有界时，表示误差 $\|OR_{[d]} - B^\top X\|$受 $\|P_B - P_{\widehat{B}_{d^*}}\|_2$或 $\|P_B - P_{\widehat{B}_{d^*}}\|_F$控制，二者拓扑等价（Yin 和 Li，2011年）。下面，我们总结几种常见充分降维方法的已知收敛速率。这些速率表明，条件 $\|O\widehat{R} - R^*\| = o_{\mathbb{P}}(n^{-1/4})$ 要求较宽松，通常大多数充分降维方法都能满足。

前向回归。这类方法包括两种经典方法——梯度外积和最小平均方差估计量，以及许多扩展方法。例如，参见 Xia 等人(2002年）；Hristache 等人(2001年）；Xia(2007）；Wang 和 Xia(2008）；Yin 和 Li(2011年）。根据这些文献中的正则条件，可以证明 B_{d^*} 与

$$\|P_{\widehat{B}_{d^*}} - P_B\|_2 = O_{\mathbb{P}}(h^4 + (nh^{d^*})^{-1}\log n + n^{-1/2}),$$

其中 $h \propto n^{-1/(d^*+4)}$ （Xia 等人，2002年；Wang 和 Xia，2008；Yin 和 Li，2011年）。等价地，$\|P_{\widehat{B}^{d^*}} - P_B\| = O_{\mathbb{P}}$（$r$（$d^*$，$n$））， 其中

$$r(d^*, n) = \begin{cases} n^{-1/2}, & d^* \le 3, \\ (nh^{d^*})^{-1}\log n, & d^* > 3. \end{cases}$$

因此，只要 $r(d^*, n) = o(n^{-1/4})$，恢复要求就成立。在典型应用中，当 $d^* \le 3$ 时，该条件会自动满足；若 $\|X\|$ 有界，前向回归估计量 B_d （满足 $d \ge d^*$）即可达到所需的表示误差。

逆向回归。代表性示例包括切片逆回归（SIR）（Li，1991年）、切片平均方差估计（Cook 和 Weisberg，1991年）以及方向回归（DR）（Li 和 Wang，2007）。这些方法通常会构造一个矩阵值统计量 M，该统计量收敛到一个满足 M 条件的总体矩阵 $P_M = P_{B_{d^*}}$。

当 $p < n$ 固定时，已有研究（Hsing 和 Carroll，1992；Zhu 和 Ng，1995；Li 和Wang，2007）证明，在温和条件下，$n^{1/2}(\widehat{M} - M)$依分布收敛于多元正态分布。设 B_d 为对应于个最大 d 特征值的 M的特征向量。若 M 的第 d^*个特征值与 0 有正距离，则恒等式 $\|P_{\widehat{B}_{d^*}} - P_B\|_F = 2^{1/2}\|\sin\Theta(\widehat{B}_{d^*}, B)\|_F$，结合 Davis-Kahan sin θ定理（Davis 和 Kahan，1970；Yu 等人，2015）以及 $\widehat{M} - M = O_{\mathbb{P}}(n^{-1/2})$，可得

$$\|P_{\widehat{B}_{d^*}} - P_B\|_F = O_{\mathbb{P}}(n^{-1/2}).$$

因此，$\|O\widehat{R} - R^*\| = o_{\mathbb{P}}(n^{-1/4})$立即成立。

当 $p > n$ 且随 n发散时，通常需要稀疏性等额外结构。在此类条件下，Lin 等人(2019) 表明，Lasso-SIR 估计量满足

$$\|P_{\widehat{B}_{d^*}} - P_B\|_F = O_{\mathbb{P}}(\{(s\log p)/(n\lambda)\}^{1/2}),$$

其中，s 是活跃变量的数量，λ 是 $\mathrm{var}_{\mathbb{P}}\{\mathbb{E}_{\mathbb{P}}(X\,|Y)\}$ 的第 d^* 大特征值，且对于某个 $\gamma > 1/2$，$n\lambda = p^{\gamma}$。当 $(s\log p)/(n\lambda) \ll n^{-1/2}$ 时，表示要求成立。

B.4 条件（C2'）的支持： $d < d^*$时的表示误差界

B.4.1 因子回归

我们的目标是证明，在适当条件下，对于某个依赖于数据的旋转矩阵 O，表示误差满足$\|O\widehat{R}_{[d]} - f_{[d]}\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}(n^{-1/4})$。如 B.3.1节所述，我们使用主成分分析估计潜在因子 f。我们沿用 B.3.1节中的记号，并重点讨论 $d < d^*_{\mathrm{F}}$ 的情形。由于 $\widehat{R} = p^{-1}W_{d_{\max}}^{\top}X$，我们有 $\widehat{R}_{[d]} = p^{-1}W_d^{\top}X$。不失一般性，假设 $B = (\beta_1, \ldots, \beta_{d^*_{\mathrm{F}}})$ 的列相互正交，并按 $\|\beta_1\|_2 \geq \cdots \geq \|\beta_{d^*_{\mathrm{F}}}\|_2$ 的顺序排列。令 $B_d = (\beta_1, \ldots, \beta_d)$ 和$B_{-d} = (\beta_{d+1}, \ldots, \beta_{d^*_{\mathrm{F}}})$。对于$1 \leq d < d^*_{\mathrm{F}}$，定义 $H_d = p^{-1}W_d^{\top}B_d \in \mathbb{R}^{d\times d}$，并将 $O = H_d^{-1} \in \mathbb{R}^{d\times d}$ 定义为 H_d 的逆。下面的命题在 (D1)–(D3) 和 (D4') 条件下建立了表示误差界；为解决可识别性问题，我们用一个强化形式替代 (D4)。

（D4'）（广泛性）当 $p \to \infty$ 时， $p^{-1}B^{\top}B \in \mathbb{R}^{d^*_{\mathrm{F}}\times d^*_{\mathrm{F}}}$ 的所有特征值均不相同，且与 0 和 ∞ 均保持有界距离。

命题 B.9。 令 $1 \leq d < d^*_{\mathrm{F}}$。假设（*D1*）–（*D3*）和（*D4'*）成立，且当 $n \to \infty$ 时，$n^{-1}\log p = o$（1）。则存在 $O = H_d^{-1} \in \mathbb{R}^{d\times d}$ 使得

$$\|O\widehat{R}_{[d]} - f_{[d]}\|_{L_2(\mathbb{P})} = O_{\mathbb{P}}(\delta),$$

其中 $\delta = n^{-1/2} +$ $(n/\log p)^{-1/2} + p^{-1/2}$。

我们首先给出一个技术引理，以确保 $\nu_{\min}(H_d)$ 依概率与 0 保持有界距离。

引理 B.1. 假设 *(D1)–(D3)* 和 *(D4')* 成立，且当 $n \to \infty$ 时， $n^{-1}\log p = o(1)$。那么，对于任意 $\varepsilon > 0$，存在一个整数 N 和一个普适常数 $C > 0$，它们与 n、p、d、$d_{\max}$ 无关，使得

$$\Pr(\nu_{\min}(H_d) > C) \geq 1 - \varepsilon, \quad for\ all\ n \geq N.$$

引理 *B.1* 的证明。 在满足 (D1)–(D3) 和 (D4') 的条件下，将 Fan 和 Gu (2024年) 中命题 1 所用的论证直接推广，即可得出该结论。 □

命题的证明 *B.9*。 记

$$\begin{aligned}\widehat{R}_{[d]} &= p^{-1}W_d^\top X \\ &= p^{-1}W_d^\top (Bf+u) \\ &= p^{-1}W_d^\top B_d f_{[d]} + \frac{1}{p}W_d^\top B_{-d} f_{-[d]} + p^{-1}W_d^\top u \\ &= H_d f_{[d]} + p^{-1}W_d^\top B_{-d} f_{-[d]} + p^{-1}W_d^\top u.\end{aligned}$$

因此,

$$H_d^{-1}\widehat{R}_{[d]} - f_{[d]} = p^{-1}H_d^{-1}W_d^\top B_{-d} f_{-[d]} + p^{-1}H_d^{-1}W_d^\top u. \tag{B.4}$$

我们先考察（B.4）右侧的第一项。我们有

$$\|p^{-1}H_d^{-1}W_d^\top B_{-d} f_{-[d]}\|_2 \le p^{-1}\|H_d^{-1}\|_2 \|W_d^\top B_{-d}\|_2 \|f_{-[d]}\|_2.$$

注意，由引理 B.1 可知 $\|H_d^{-1}\|_2 = \nu^{-1}$ $\min(H_d) = O_{\mathbb{P}}(1)$。此外，(D1) 意味着，对于某个普适常数 C，有 $\mathbb{E}(\|f_{-[d]}\|_2) \le C$。剩下需要控制 $\|W_d^\top B_{-d}\|_2$。令 $B^{(u)} = B \cdot \mathrm{diag}(\|\boldsymbol{\beta}_1\|^{-1}, \ldots, \|\boldsymbol{\beta}_{d_0}\|^{-1})$，使得 $B^{(u)}$ 具有标准正交列，并类似地定义 $B_d^{(u)}$ 和 $B_{-d}^{(u)}$。令 $\widehat{W}_d = W_d/\sqrt{p} = [\widehat{v}_1, \ldots, \widehat{v}_d]$。于是 $\|W_d^\top B_{-d}\|_2 \le \sqrt{p}\|\widehat{W}_d^\top B_{-d}^{(u)}\|_2 \|\boldsymbol{\beta}_{d+1}^-\|_2$。根据奇异值分解，存在正交矩阵 O_1、$O_2 \in \mathbb{R}^{d\times d}$，使得

$$O_1^\top \widehat{W}_d^\top B_d^{(u)} O_2 = \cos\Theta\big(\widehat{W}_d, B_d^{(u)}\big).$$

因此，$\|\widehat{W}_d^\top B_{-d}^{(u)}\|_2 = \|(\widehat{W}_d - B_d^{(u)}O_2O_1^\top)^\top B_{-d}^{(u)}\|_2 \le \|\widehat{W}_d - B_d^{(u)}O_2O_1^\top\|_2 \|B_{-d}^{(u)}\|_2$，因为 B具有正交列。此外，$\|B_{-d}^{(u)}\|_2 = 1$，并且

$$\begin{aligned}\|\widehat{W}_d - B_d^{(u)}O_2O_1^\top\|_2^2 &\le \|\widehat{W}_d - B_d^{(u)}O_2O_1^\top\|_F^2 \\ &= \mathrm{tr}\{(\widehat{W}_d - B_d^{(u)}O_2O_1^\top)^\top(\widehat{W}_d - B_d^{(u)}O_2O_1^\top)\} \\ &= 2\cdot\mathrm{tr}\{\mathbf{I}_d - O_1^\top\widehat{W}_d^\top B_d^{(u)}O_2\} \\ &\le 2\|\sin\Theta\big(\widehat{W}_d, B_d^{(u)}\big)\|_F^2.\end{aligned}$$

回顾一下，$\widehat{W}_d$ 和 $B_d^{(u)}$ 分别是 d 和 $\widehat{\Sigma}$的前 $BB^\top$ 个特征向量。Davis-Kahan 定理（Davis 和 Kahan，1970年；Yu 等，2015）给出：

$$\|\sin\Theta\big(\widehat{W}_d, B_d^{(u)}\big)\|_F \le \frac{\|\widehat{\Sigma} - BB^\top\|_F}{\|\boldsymbol{\beta}_d\|_2^2 - \|\boldsymbol{\beta}_{d+1}\|_2^2}.$$

在 (D4') 条件下，$\|\boldsymbol{\beta}_d\|_2^2 - \|\boldsymbol{\beta}_{d+1}\|_2^2 \ge C\cdot p$ 对某个常数 $C > 0$ 成立。在 (D1)–(D3) 条件下，

标准论证（例如参见 Fan 等，2013年；Fan 和 Gu，2024年）可得

$$\|\widehat{\Sigma} - BB^\top\|_F = O_{\mathbb{P}}(p\delta),$$

其中 $\delta = n^{-1/2} + (n/\log p)^{-1/2} + p^{-1/2}$。因此，

$$\|\sin\Theta(\widehat{W}_d, B_d^{(u)})\|_F = O_{\mathbb{P}}(\delta),$$

这意味着 $\|W_d^\top B_{-d}\|_2 = O_{\mathbb{P}}(p\delta)$。结合前述界可得

$$\|p^{-1}H_d^{-1}W_d^\top B_{-d} f_{-[d]}\|_{L_2(\mathbb{P})} = O_{\mathbb{P}}(\delta).$$

接下来，我们对（B.4）右侧的第二项进行界定。根据 Fan 和 Gu （2024年）中引理 2 的证明，在（D1）–（D3）条件下，有 $\|W_d^\top u\|_{L_2(\mathbb{P})}^2 = O(p)$。结合 $\|H_d^{-1}\|_2 = O_{\mathbb{P}}(1)$，可得 $\|p^{-1}H_d^{-1}W_d^\top u\|_{L_2(\mathbb{P})} = O_{\mathbb{P}}(p^{-1/2})$。

代入这些界可证明，当 $O = H_d^{-1}$，$\|O\widehat{R}_{[d]} - f_{[d]}\|_{L_2(\mathbb{P})} = O_{\mathbb{P}}(\delta)$ 时，结论成立。

□

特别地，命题 B.9 意味着，只要满足 $(\log p)/\sqrt{n} = o(1)$ 和 $\sqrt{n/p} = o(1)$，要求 $\|OR_{[d]} - f_{[d]}\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}(n^{-1/4})$ 就以 $O = H_d^{-1}$ 的形式成立。

B.4.2 充分降维

我们将证明，在适当条件下，当 $d < d^*$ 时，对于某个依赖于数据的旋转矩阵 O，表示误差 $\|O\widehat{R}_{[d]} - B_d^\top X\| = o_{\mathbb{P}}(n^{-1/4})$ 满足相应结论。我们聚焦于基于核矩阵的充分降维方法（如 B.3.2 节所述的“逆向回归”类方法）。设 $M \in \mathbb{R}^{p\times p}$ 为总体核矩阵，其 d^* 个非零特征值决定了充分降维子空间。令 $B \in \mathbb{R}^{p\times d^*}$ 收集相应的特征向量，并按特征值递减顺序排列。对于 $1 \le d \le d^*$，令 B_d 表示 B 的前 d 列。将相同的充分降维方法应用于 $\{(X_i, Y_i) : i \in I_{-k}\}$，得到经验核矩阵 $\widehat{M}$，其前 $d_{\max}$ 个特征向量构成 $\widehat{B}_{d_{\max}}$。对于任意 $d \le d_{\max}$，用 $\widehat{B}_d$ 表示 $\widehat{B}_{d_{\max}}$ 的前 d 列。下面的结果与 $d < d^*$ 情形下的命题 B.8 类似。

命题 B.10。令 $1 \le d < d^*$。存在 $O = B_d^\top \widehat{B}_d \in \mathbb{R}^{d\times d}$ 使得

$$\|O\widehat{R}_{[d]} - B_d^\top X\| \le \|P_{B_d} - P_{\widehat{B}_d}\|_2 \|X\|.$$

证明。参见命题 B.8的证明。 □

利用恒等式 $\|P_{\widehat{B}_d} - P_{B_d}\|_F = 2^{1/2}\|\sin\Theta(\widehat{B}_d, B_d)\|_F$，Davis-Kahan 定理

（Davis 和 Kahan，1970年；Yu 等，2015）表明

$$\|\sin\Theta(\widehat{B}_d, B_d)\|_F \leq \frac{\|\widehat{M} - M\|_F}{\lambda_d - \lambda_{d+1}},$$

其中，λ_1、……、λ_{d^*} 是 M的非零特征值，按降序排列。当 $p < n$固定时，已有研究（Hsing 和 Carroll，1992；Zhu 和 Ng，1995；Li和Wang，2007）证明了 $\|\widehat{M} - M\|_F = O_{\mathbb{P}}(n^{-1/2})$。进一步假设特征间隙条件 $\lambda_j - \lambda_{j+1} > c > 0$ 对所有 $1 \leq j < d^*$ 均成立，如 Zhu 和 Ng（1995）和 Yu 等（2025）所述。那么，对于任意 $d < d^*$，$\|\sin\Theta(\widehat{B}_d, B_d)\|_F = O_{\mathbb{P}}(n^{-1/2})$，$\|P_{B_d} - P_{\widehat{B}_d}\|_2 = O_{\mathbb{P}}(n^{-1/2})$。结合命题 B.10，并假设 $\|X\|$ 有界，我们即可得出所需结论：$\|O\widehat{R}_{[d]} - B_d^\top X\| = o_{\mathbb{P}}(n^{-1/4})$。

C 补充实验与实现细节

C.1 降维映射的任务特定构造

我们使用示例 1–3 来说明如何在不同的降维范式和预测任务下构造降维映射 $\widehat{\varphi}$。

因子回归。 大量文献提出了构造 $\widehat{\varphi}$ 的方法，包括主成分分析（Jolliffe，2002年）、极大似然方法（Bai 和 Li，2012年；Doz 等，2012年）以及基于协方差的方法（Fan 等，2013年）。投影主成分分析（Fan 等，2016）利用协变量信息改进因子估计。Fan 和 Liao（2022年）进一步提出了针对特定结构的构造方法，包括基于特征的权重、移动窗口估计量、初始变换和 Hadamard 投影。方法的选择应依据数据生成情境；例如，当存在有信息量的协变量时，基于特征的加权方案或投影主成分分析可能更为合适。

充分降维。 应根据预测任务（等价地说，损失函数）选择降维方法。正如第B.1节所述，不同的损失函数对应不同的目标充分降维子空间，因此应据此选择估计程序。例如，平方损失的目标是中心均值子空间，适用于主海森方向（Li，1992）、最小平均方差估计量（Xia 等人，2002年）、迭代海森变换（Cook 和 Li，2002年）等方法。负对数似然和交叉熵损失的目标是中心子空间，可采用多种逆回归和正向回归方法（Li，1991年；Cook 和 Weisberg，1991年；Li 和 Wang，2007；Xia，2007；Wang 和 Xia，2008；Yin 和 Li，2011年）。在 0-1 损失下，目标是中心判别子空间，可以通过

使用诸如 Cook 和 Yin （2001年）等方法进行估计。

降秩回归。 拟合低秩$\widehat{}$系数矩阵 $A \in \mathbb{R}^{p\times q}$ 并施加秩约束（Reinsel 等，2022年；Yuan 等，2007；Bunea 等，2011年；Zou 等，2022年）有多种可用的估计量。任何此类方法均可用于获得$\widehat{}$ $\widehat{}$秩-$d_{\max}$ 估计值 $A_{d_{\max}}$。对 $A_{d_{\max}}$ 进行奇异值分解，可得到其前$d_{\max}$ 个左奇异向量，记为 $\widehat{B}_{d_{\max}} \in \mathbb{R}^{p\times d_{\max}}$。在此框架下，可以取$\widehat{\varphi} = B_{d_{\max}}$。

C.2 信息准则基线的实现细节

我们将我们的方法与信息准则（IC）进行比较，具体而言是 IC_{p1}，由Bai 和 Ng （2002年）提出。令$\mathbf{X} = [X_1, \ldots, X_n] \in \mathbb{R}^{p\times n}$ 表示数据矩阵，并令$\mathbf{F}^k = [f_1, \ldots, f_n] \in \mathbb{R}^{k\times n}$ 为一个由 k 潜在因子组成的矩阵。定义

$$V(k) = \min_{B^k \in \mathbb{R}^{p\times k}, \mathbf{F}^k} \frac{\|\mathbf{X} - B^k \mathbf{F}^k\|_F^2}{np}.$$

该准则为

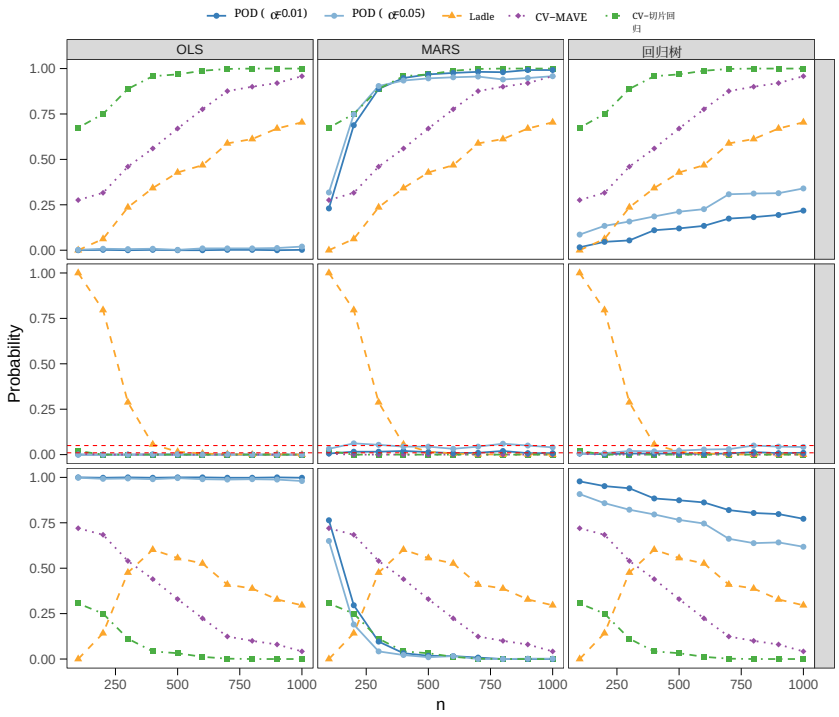
$$IC_{p1}(k) = \ln\left(V(k)\right) - k\frac{n+p}{np}\ln\left(\frac{n+p}{np}\right). \tag{C.1}$$

等价地，$V(k)$ 可以计算为从（$k+1$）到 p 的 $\mathbf{X}\mathbf{X}^\top/(np)$ 的特征值之和。

C.3 对学习器类别选择的敏感性

我们使用第 4.1.2 节中的模型 4，$\mathcal{F}$ 以考察学习器类别的选择如何影响 POD。在保持其他所有设定与正文相同的情况下，我们将 $\mathcal{F}$ 限制为以下学习器类别之一：普通最小二乘法、多元自适应回归样条或回归树。每种方法的最终表现见图 S.2。

在该模型中，OLS 和回归树无法充分近似 Y与 (X_1, X_2) 之间的关系，反映出这些学习器类别在捕捉潜在函数形式方面的局限性；这种模型设定错误进而降低了 POD 的性能。相比之下，多元自适应回归样条提供了更准确的近似，并取得了令人满意的结果。这些发现凸显了选择足够丰富的学习器类别 $\mathcal{F}$的重要性，即该类别应包含能够表示潜在结构的灵活模型。



^图 S.2：SDR 的正确估计（$d = d^*_{\mathrm{CMS}}$）、高估^ ^（$d > d^*_{\mathrm{CMS}}$）和低估（$d < d^*_{\mathrm{CMS}}$）的经验概率，跨越 n 。第二行中的红色虚线表示名义显著性水平 $\alpha = 0.01$ 和 $\alpha = 0.05$。

C.4 响应变量为分类变量的充分降维

对于分类响应，令 $X \sim \mathcal{N}(0,\ I_p)$，其中 $p = 10$，并考虑

模型 6：$Y \sim$ 二项分布(2,逻辑函数(X_1)),

模型 7：$Y = \mathbf{1}\{X_1 + \cdots + X_5 + \sigma\varepsilon > 1\} + 2 \cdot \mathbf{1}\{X_6 + \cdots + X_{10} + \sigma\varepsilon > 0\}$,

其中 $\varepsilon \sim \mathcal{N}(0, 1)$ 与 X和 $\sigma = 0.5$相互独立。结构维数为$d^*_{\mathrm{CS}} = 1$ （模型 6）和 $d^*_{\mathrm{CS}} = 2$ （模型 7，Zhu 和 Zeng (2006)）；在两个模型中，$d^*_{\mathrm{CS}} = d^*_{\mathrm{CMS}}$。POD 使用交叉熵损失函数评估预测性，因此 $d^*(\mathcal{L}) = d^*_{\mathrm{CS}}$。对于 POD，我们使用切片逆回归估计模型 6–7 的降维映射，模型 6 使用 3 个切片，模型 7 使用 4 个切片。通过二折交叉验证，在支持向量机和分类树之间选择下游分类器。

我们将 POD 与三种备择方法进行比较：基于切片逆回归的 ladle 估计量，以及基于 MAVE 和切片回归的交叉验证准则。所有方法的目标分别是模型 6 的 $d^*_{\mathrm{CS}} = 1$ 和模型 7 的 $d^*_{\mathrm{CS}} = 2$ 。

图S.3报告了d^*_{CS} 随着 n 增加时正确估计、高估和低估的经验概率。在这两种模型中，POD 都比其他方法收敛得更快。其高估概率趋近于名义水平，而

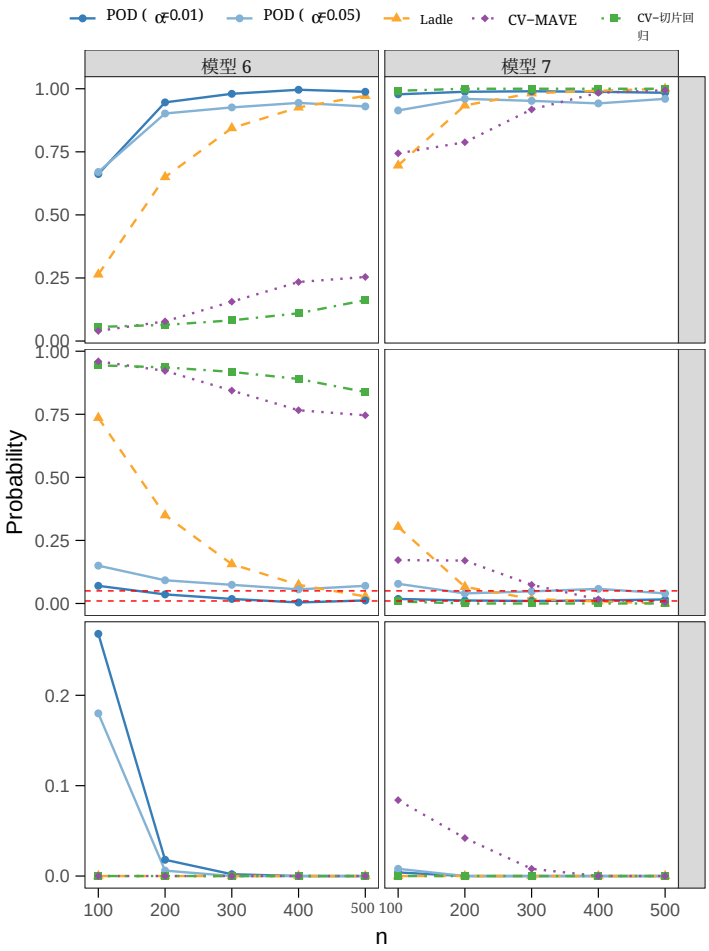


图 S.3：SDR 模型 6–7 在 n 下正确估计 ($\hat{d} = d^*_{\mathrm{CS}}$)、高估 ($\hat{d} > d^*_{\mathrm{CS}}$) 和低估 ($\hat{d} < d^*_{\mathrm{CS}}$) 的经验概率。第二行中的红色虚线水平线标示名义显著性水平 $\alpha = 0.01$ 和 $\alpha = 0.05$。

低估概率则趋于零。

C.5 真实数据分析中的选择频率

由于我们的方法采用 K 折交叉拟合，ladle 估计量依赖自助法，而加权-χ^2 检验的临界值通过模拟获得，因此即使应用于同一数据集，所有方法都存在随机性。因此，我们在表 S.1–S.2 中报告了重复运行 100 次时各估计维数的选择频率，这表明 POD 能够稳定地进行维数选择。

ˆ 表 S.1： d 在$d_{\max} = 8$时，100 次重复中估计维数的选择频率。

	POD										Ladle	BY，加权-χ^2		
	0-1 损失					交叉熵								
	$\alpha = 1\%$			$\alpha = 5\%$		$\alpha = 1\%$			$\alpha = 5\%$			$\alpha = 1\%$		$\alpha = 5\%$
$\widehat{d}$	1	2	3	2	≥ 3	1	2	3	2	≥ 3	3	14	15	16
频率	1	93	6	89	11	1	97	2	94	6	100	81	19	100

ˆ 表 S.2：当$d_{\max} = 16$时， d 在 100 次重复中的估计维数选择频率。

	POD								Ladle		BY，加权-χ^2		
	0-1 损失				交叉熵								
	$\alpha = 1\%$		$\alpha = 5\%$		$\alpha = 1\%$		$\alpha = 5\%$				$\alpha = 1\%$		$\alpha = 5\%$
$\widehat{d}$	2	3	2	≥ 3	2	3	2	≥ 3	≤ 13	14	14	15	16
频率	93	7	84	16	95	5	94	6	13	87	81	19	100