高维数据推断作业

邵李翔　祖劭康　赵张弛

目录

1 利用 OLS 求解广义线性模型参数

2 pHd

2.1 原理

$$
\begin{aligned}
&H_1 = \Sigma_{XX}^{-1}\Sigma_{XXY}\Sigma_{XX}^{-1} \\
&span(H_1) \subseteq S_{Y|X}
\end{aligned}
$$

2.2 样本计算流程

1. 将 $X_1,\dots,X_n$ 标准化为 $Z_1,\dots,Z_n$
2. 计算 $\hat{\Sigma}_{ZZY}E_n(ZZ^TY), \quad \hat{\Sigma}_{XX}Var_n(X)$
3. 计算 $\hat{\Sigma}_{ZZY}$ 的前 q 个特征向量 $u_1,\dots,u_n$，则对 $S_{Y|X}$ 中向量的估计为 $v_k = \hat{\Sigma}_{XX}^{-1/2}u_k, k = 1,\dots,q$

2.3 线性回归

设定样本量为 1000, 样本 X 维度为 5, 降维后维度为 2. 一次的计算结果 H_1, H_2 如下表所示

表 1: 线性回归结果

H_1		H_2	
-0.7193995	0.07794245	0.6920215	0.4112301
-0.2187786	-0.24974545	-0.1080706	-0.7261988
-0.5618548	0.28103448	0.6192357	-0.3741050
-0.1909770	-0.92664516	-0.2790291	0.2171582
-0.3313467	-0.03092594	-0.2530621	0.2974159

接下来判断维度的增大对降维效果的影响, 设定 X 维度依次为 $3-20$, 统一降维至二维, 判断依据为两个空间的距离. 如下图所示

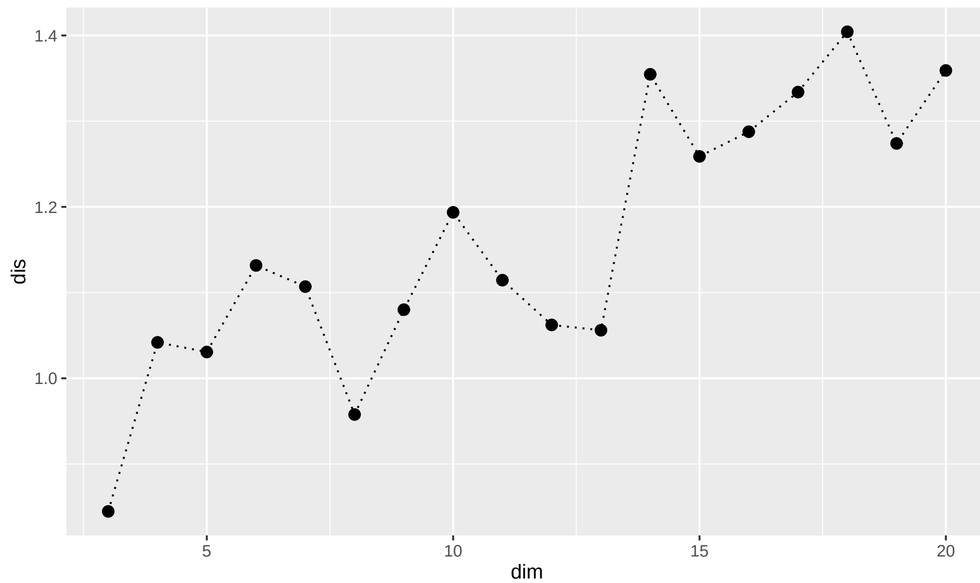

图 1: 线性回归下维度对降维效果的影响

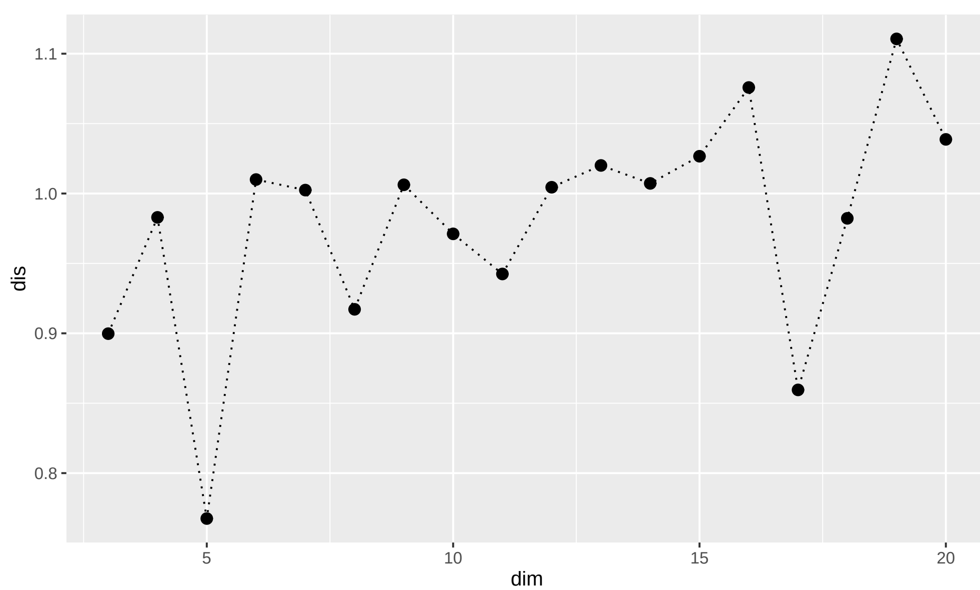

图 2: 泊松回归下维度对降维效果的影响

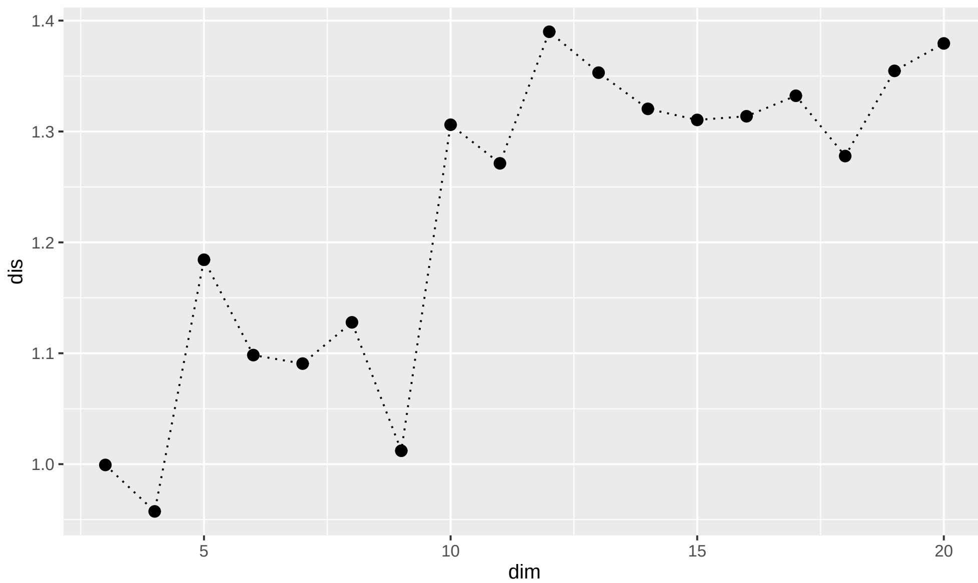

图 **3:** logistics 回归下维度对降维效果的影响

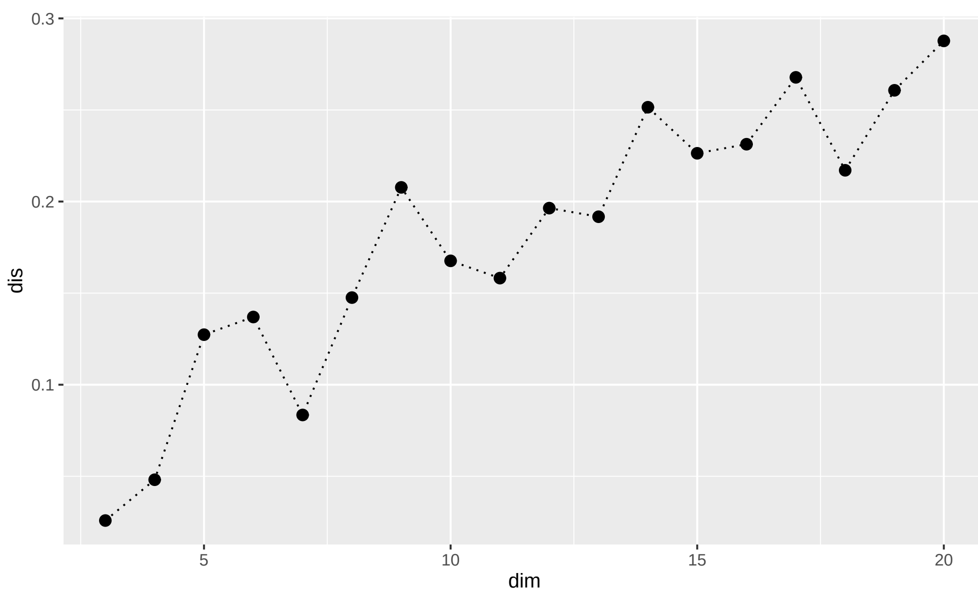

图 **4:** cos 函数生成的数据下维度对降维效果的影响

2.4 泊松回归

2.5 Logistic 回归

2.6 生成 cos 函数关系

3 SIR

3.1 理论依据

$$spanCov(E(X|Y)) \subseteq S_{Y|X}$$

将区间 $(-\infty, +\infty)$ 划分为 k 个区间 $I_i, i = 1, \ldots, k$, 定义 $\widetilde{Y} = \sum_{i=1}^{k} iI\{Y \in I_i\}$，则有

$$spanCov(E(X|\widetilde{Y})) \subseteq S_{Y|X}$$

3.2 样本计算流程

1. 将 $X_1, \ldots, X_n$ 标准化为 $Z_1, \ldots, Z_n$
2. 将 $[a, b] = [minY_{ii=1}^n, maxY_{ii=1}^n]$ 划分为 k 个区间，得到 $\widetilde{Y}_i$；由此计算 $\bar{\mu}_i = \frac{1}{\#\{I_i\}} \sum_{Y_j \in I_i} Z_j$
3. 计算协方差矩阵估计量
$$\widetilde{M} = \sum_{i=1}^{k} E_n[I(Y \in I_i)]E_n(Z|\widetilde{Y} = i)E_n(Z|\widetilde{Y} = i)^T = \sum_{i=1}^{k} \frac{\#\{I_i\}}{n} \bar{\mu}_i \bar{\mu}_i{}^T$$
4. 计算 $\widetilde{M}$ 的前 q 个特征向量 $u_1, \ldots, u_n$，则对 $S_{Y|X}$ 中向量的估计为 $v_k = \hat{\Sigma}_X^{-1/2} u_k, k = 1, \ldots, q$

3.3 线性回归

设定样本量为 1000, 样本 X 维度为 10, 降维后维度为 2.

根据岭比率阈值准则 $\hat{q} = arg\max\left\{i|\frac{\hat{\lambda_{i+1}} + C_n}{\hat{\lambda_i} + C_n}\tau\right\}$ 估计得到的降维后维度为 1, 其中设置 $C_n = \frac{1}{n^{\frac{1}{3}}}$.

3.4 泊松回归

3.5 生成 sin 函数关系

首先设置降维维度为二维, 得到的结果如下图所示

而降维结果为一维时, 与 cos 函数生成的数据结果比较如图所示 可以看出 SIR 对奇函数的降维效果相较于偶函数有比较不错的改进.

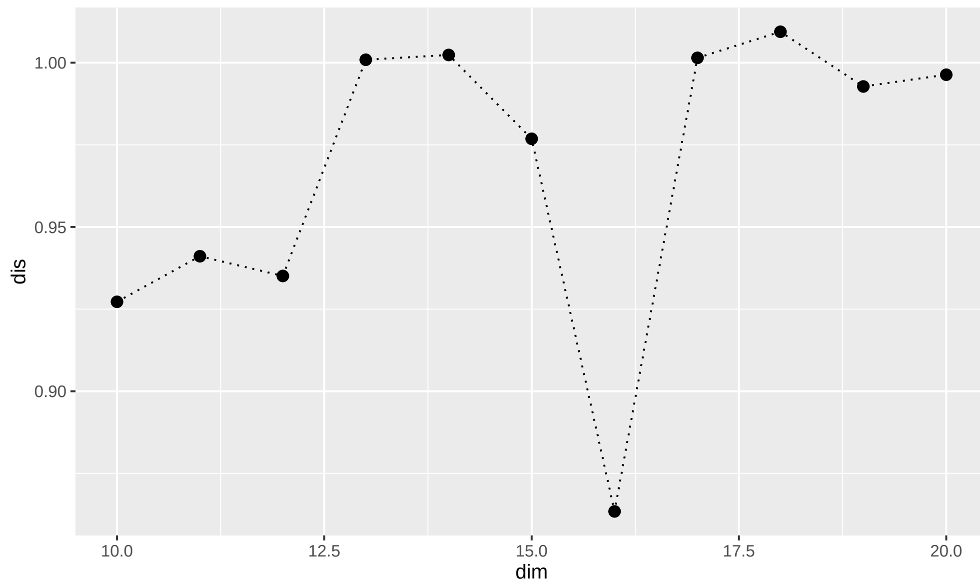

图 5: 泊松回归下维度对降维效果的影响

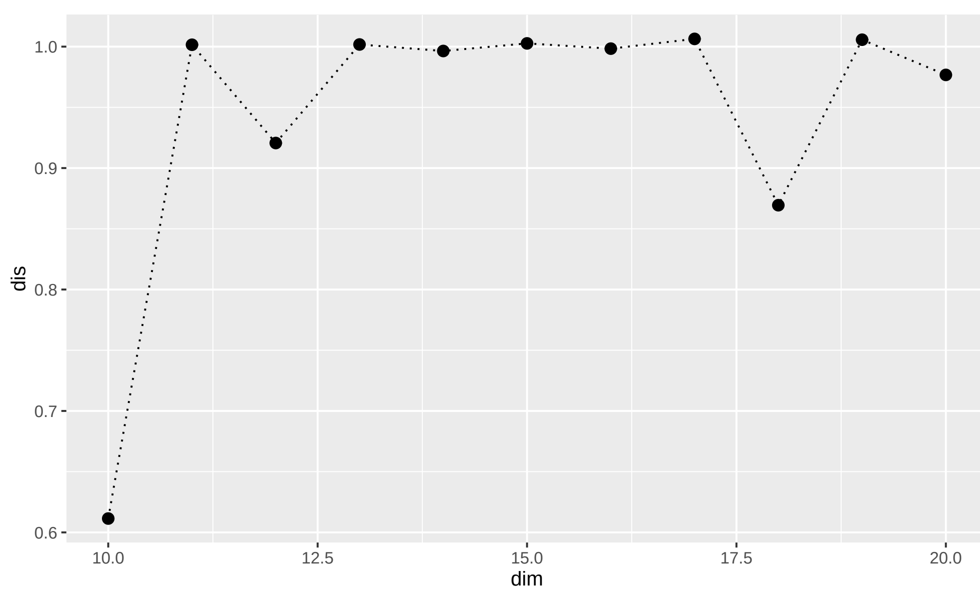

图 6: sin 函数生成的数据下维度对降维效果的影响 (二维)

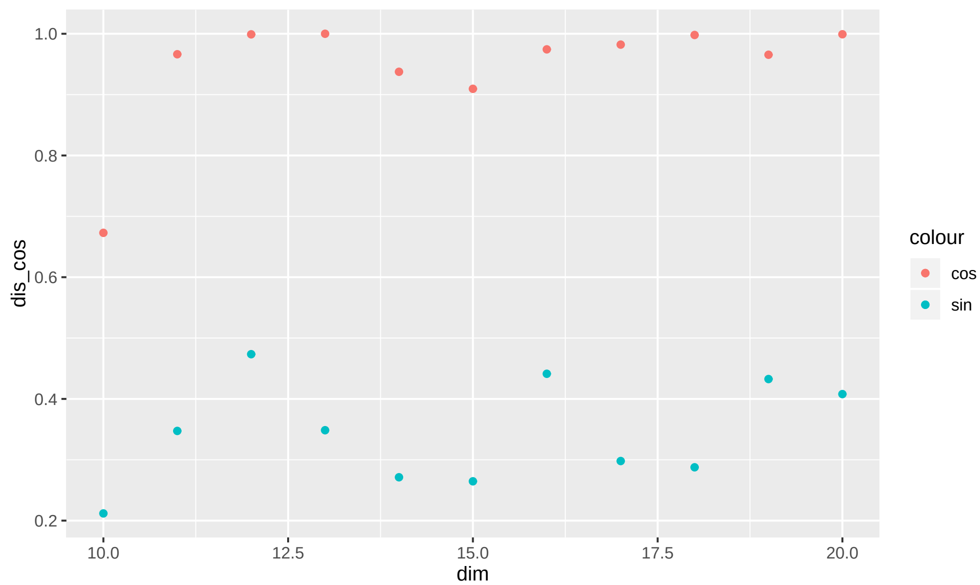

图 7: sin 函数生成的数据下维度对降维效果的影响 (一维)

4 SAVE

4.1 理论依据

$$span(E[I_p - Var(X|Y)]^2) \subseteq S_{Y|X}$$

4.2 样本计算流程

1. 将 $X_1, \dots, X_n$ 标准化为 $Z_1, \dots, Z_n$
2. 将 $[a,b] = [minY_{ii=1}^n, maxY_{ii=1}^n]$ 划分为 k 个区间，得到 $\widetilde{Y}_i$；由此计算 $\bar{\mu}_i = \frac{1}{\#\{I_i\}} \sum_{Y_j \in I_i} Z_j$
3. 计算条件方差估计量 $\hat{\Sigma}_i = \sum_{i=1}^k \frac{1}{\#I_i} \sum_{Y_j \in I_i} (Z_j - \hat{\mu}_i)(Z_j - \hat{\mu}_i)^T$
4. 计算协方差矩阵估计量

$$\begin{aligned}\widetilde{M} &= \sum_{i=1}^k E_n[I(Y \in I_i)][I_{p\times p} - Var(Z|\widetilde{Y} = i)]^2 \\ &= \sum_{i=1}^k \frac{\#\{I_i\}}{n}(I_{p\times p} - \hat{\Sigma}_i)^2\end{aligned}$$

5. 计算 $\widetilde{M}$ 的前 q 个特征向量 $u_1, \dots, u_n$，则对 $S_{Y|X}$ 中向量的估计为 $v_k = \hat{\Sigma}_X^{-1/2} u_k, k = 1, \dots, q$

在实际计算中, 将 $[Y_{(1), Y_{(n)}}]$ 划分为 10 个区间.

4.3 线性回归模型

设定样本量为 1000, 降维维数由 $10 - 20$, 判断其降维效果.

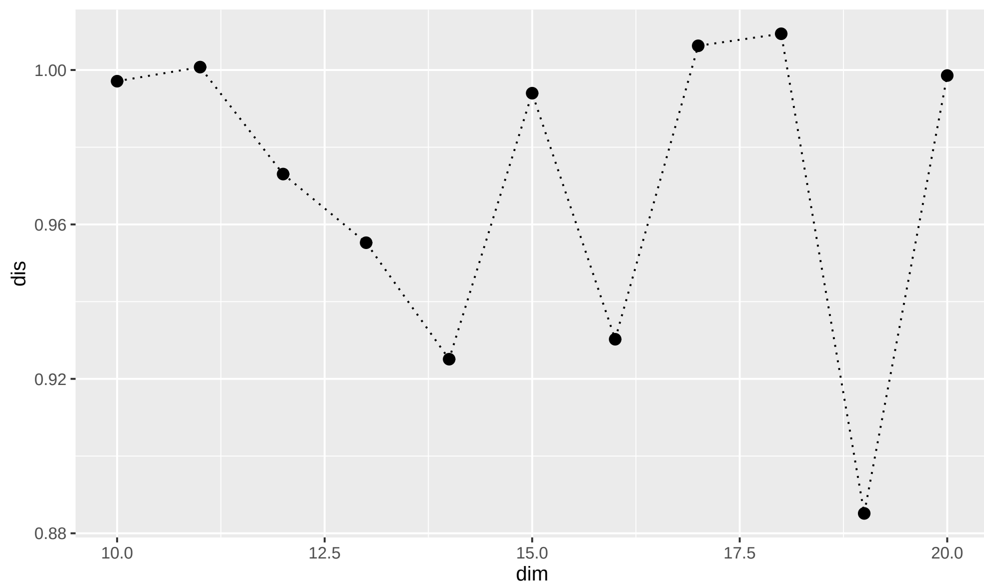

图 **8:** 线性回归下维度对降维效果的影响

4.4 cos 函数生成的数据的降维计算

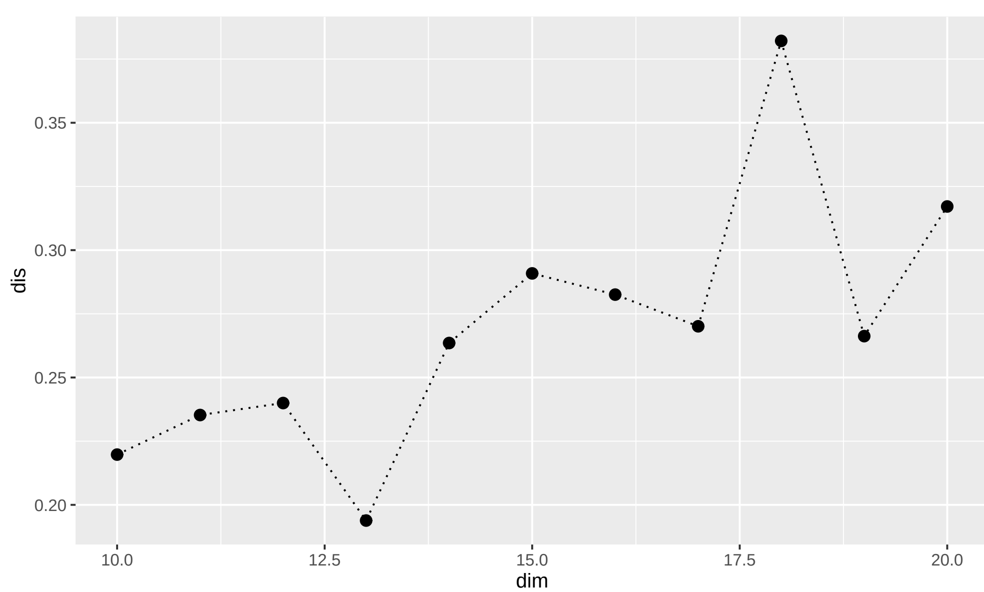

图 **9:** cos 函数生成的数据下维度对降维效果的影响 (二维)

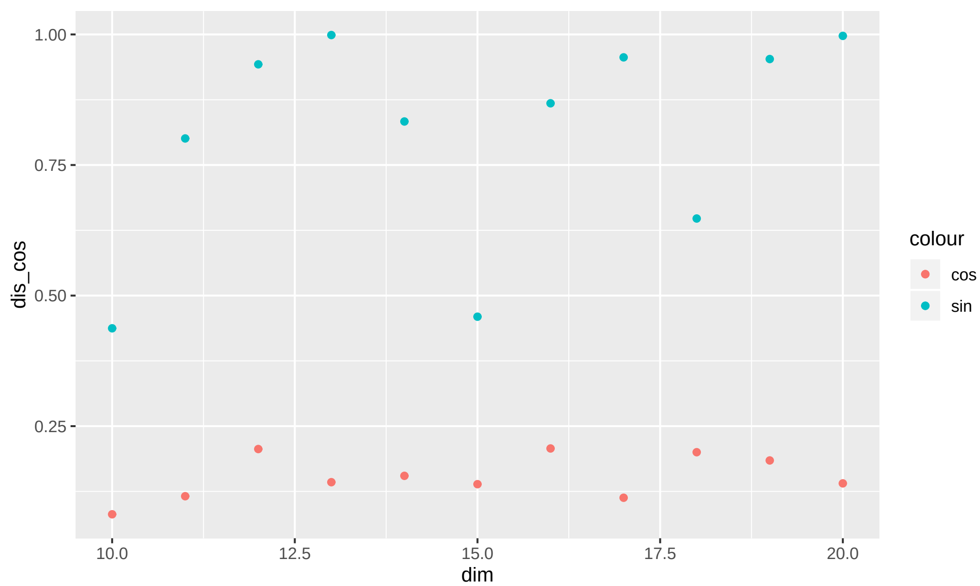

图 10: cos 函数与 sin 函数生成的数据下维度对降维效果的影响 (一维)