

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	3
2	pHd	3
2.1	原理	3
2.2	样本计算流程	3
2.3	度量两空间间的距离	3
2.4	线性回归	3
2.5	泊松回归	4
2.6	Logistic 回归	5
2.7	生成 cos 函数关系	5
2.8	总结	5
3	SIR	6
3.1	理论依据	6
3.2	样本计算流程	6
3.3	确定估计后降维维度	6
3.4	泊松回归	7
3.5	生成 sin 函数关系	7
4	SAVE	8
4.1	理论依据	8
4.2	样本计算流程	8
4.3	线性回归模型	9
4.4	cos 函数生成的数据的降维计算	9
4.5	总结	10

1 利用 OLS 求解广义线性模型参数

2 pHd

2.1 原理

$$H_1 = \Sigma_{XX}^{-1} \Sigma_{XXY} \Sigma_{XX}^{-1}$$

$$\text{span}(H_1) \subseteq S_{Y|X}$$

$$H_2 = E[(Y - E(XY)^T X) X X^T]$$

$$\text{span}(H_2) \subseteq S_{Y|X}$$

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 $\hat{\Sigma}_{ZZY} E_n(Z Z^T Y)$, $\hat{\Sigma}_{XX} \text{Var}_n(X)$
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 度量两空间间的距离

采用 $L_2 - Hausdorff$ 子空间距离度量 m 维子空间 U 与 n 维子空间 V 之间的距离:

$$d(U, V) = \max(\vec{d}(U, V), \vec{d}(V, U)) = \sqrt{\max(m, n) - \sum_{i=1}^m \sum_{j=1}^n (u_i^T v_j)^2}$$

2.4 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

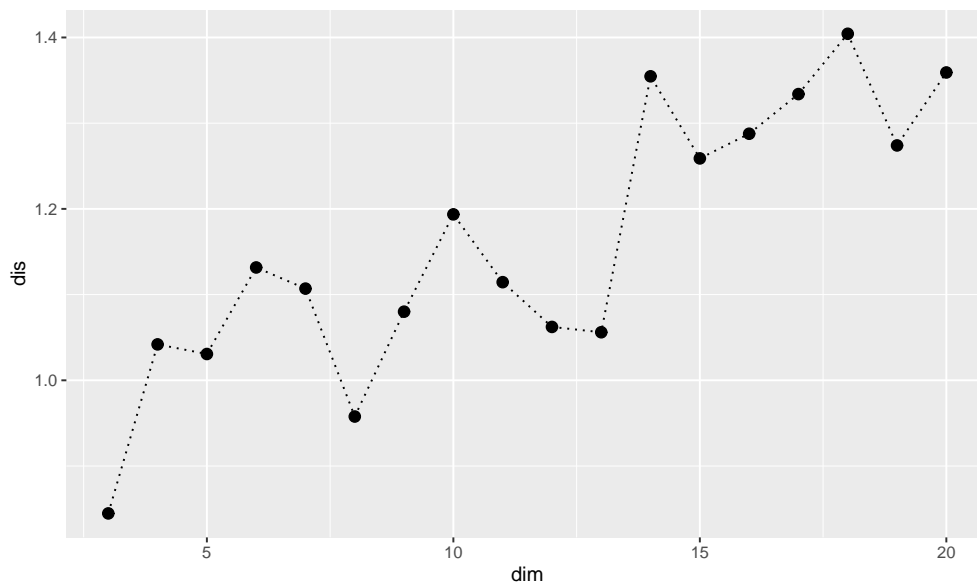


图 1: 线性回归下维度对降维效果的影响

2.5 泊松回归

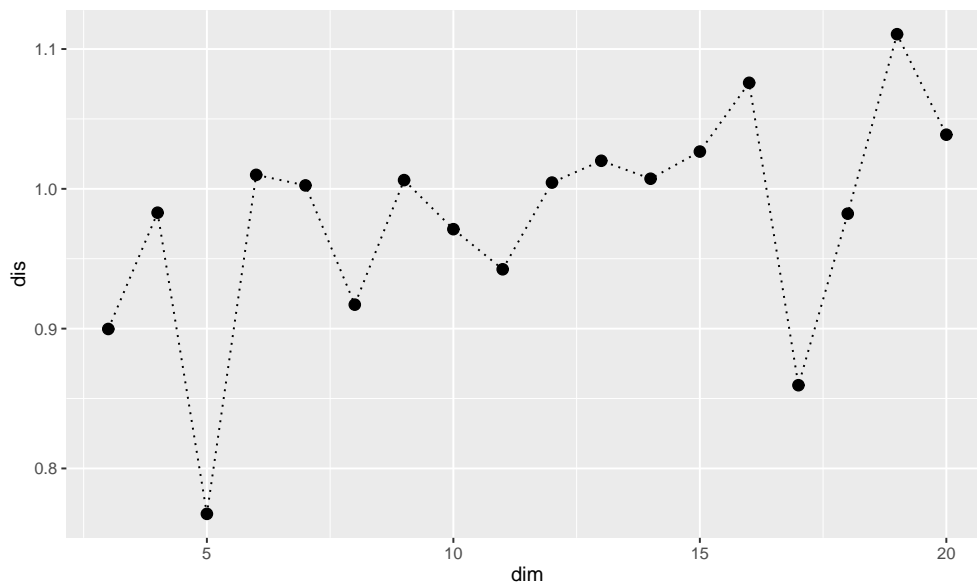


图 2: 泊松回归下维度对降维效果的影响

2.6 Logistic 回归

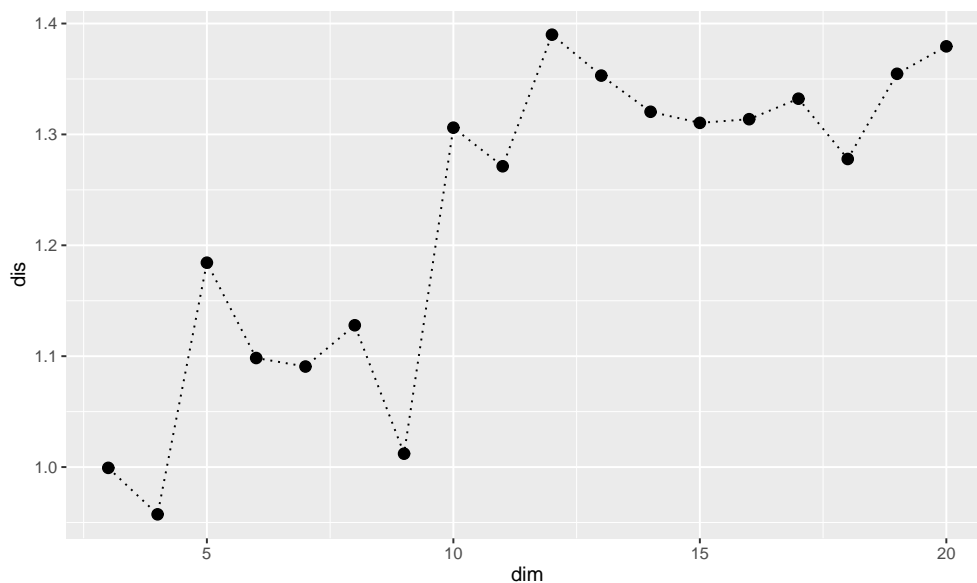


图 3: logistics 回归下维度对降维效果的影响

2.7 生成 cos 函数关系

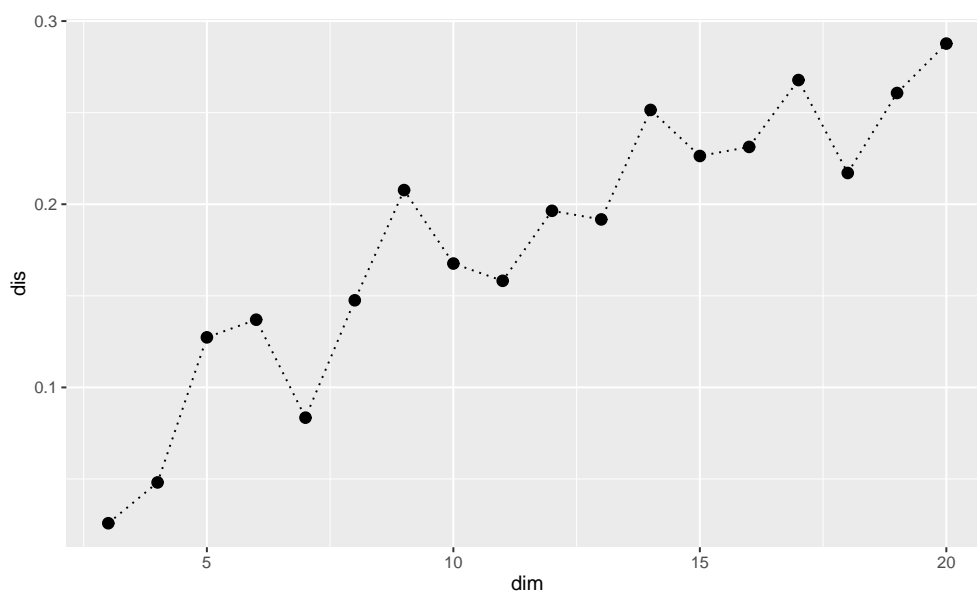


图 4: cos 函数生成的数据下维度对降维效果的影响

2.8 总结

样本量始终为 1000 的情况下，增大 X 的维数，估计空间与实际空间的差距呈上升趋势。但由于子空间间的距离可看出，这一方法在线性回归、Logistic 回归以及 Poisson 回归上的表现都不好，仅

在 Y 是 X 的 \cos 相关函数的情况下效果较为理想。

3 SIR

3.1 理论依据

$$\text{span}\{Cov(E(X|Y))\} \subseteq S_{Y|X}$$

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

$$\text{span}\{Cov(E(X|\tilde{Y}))\} \subseteq S_{Y|X}$$

3.2 样本计算流程

1. 将 X_1, \dots, X_n 标准化为 Z_1, \dots, Z_n
2. 将 $[a, b] = [\min Y_{i=1}^n, \max Y_{i=1}^n]$ 划分为 k 个区间, 得到 \tilde{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

$$\widetilde{M} = \sum_{i=1}^k E_n[I(Y \in I_i)] E_n(Z|\tilde{Y} = i) E_n(Z|\tilde{Y} = i)^T = \sum_{i=1}^k \frac{\#\{I_i\}}{n} \bar{\mu}_i \bar{\mu}_i^T$$

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

3.3 确定估计后降维维度

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 $\hat{q} = \arg \max \left\{ i \mid \frac{\lambda_{i+1} + C_n}{\hat{\lambda}_i + C_n} \tau \right\}$ 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

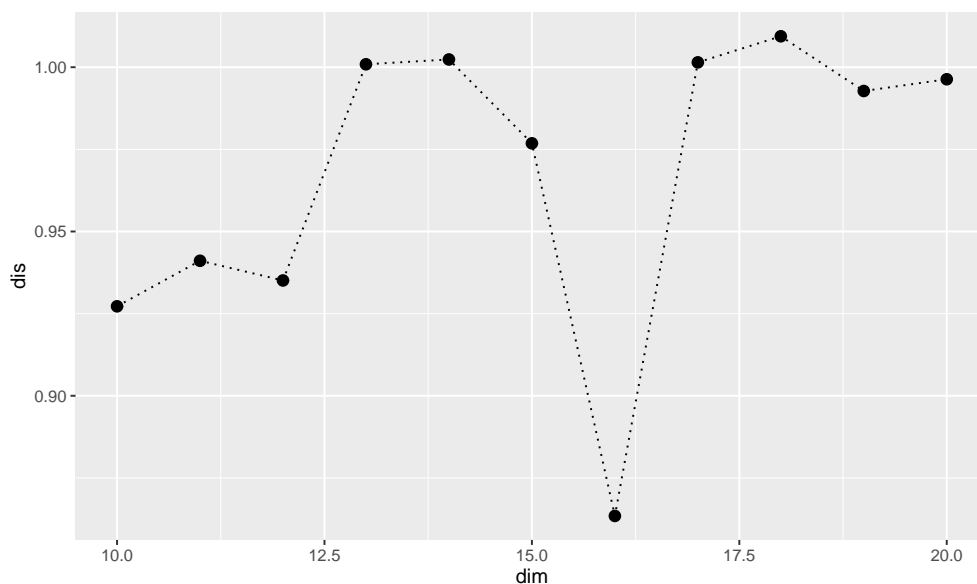


图 5: 泊松回归下维度对降维效果的影响

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

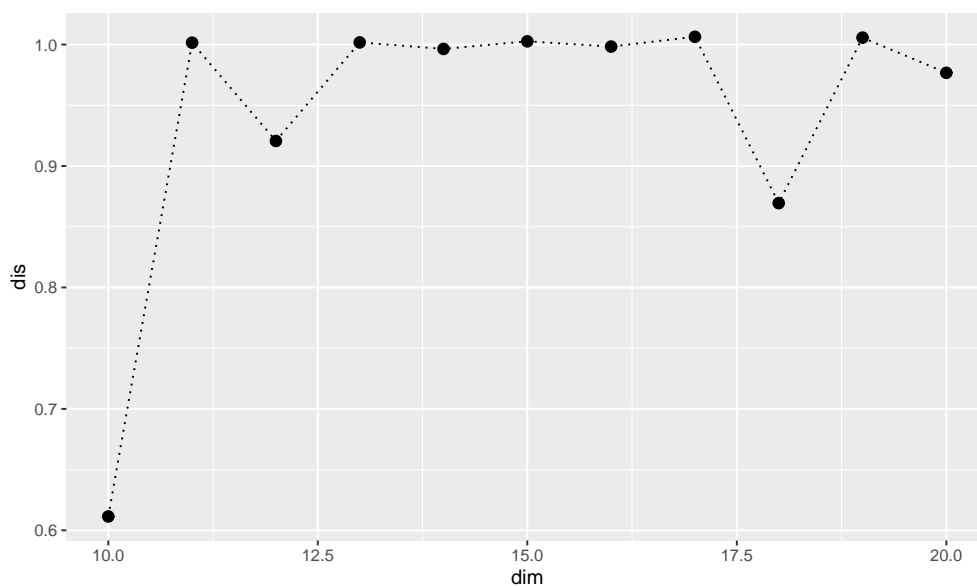


图 6: sin 函数生成的数据下维度对降维效果的影响 (二维)

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

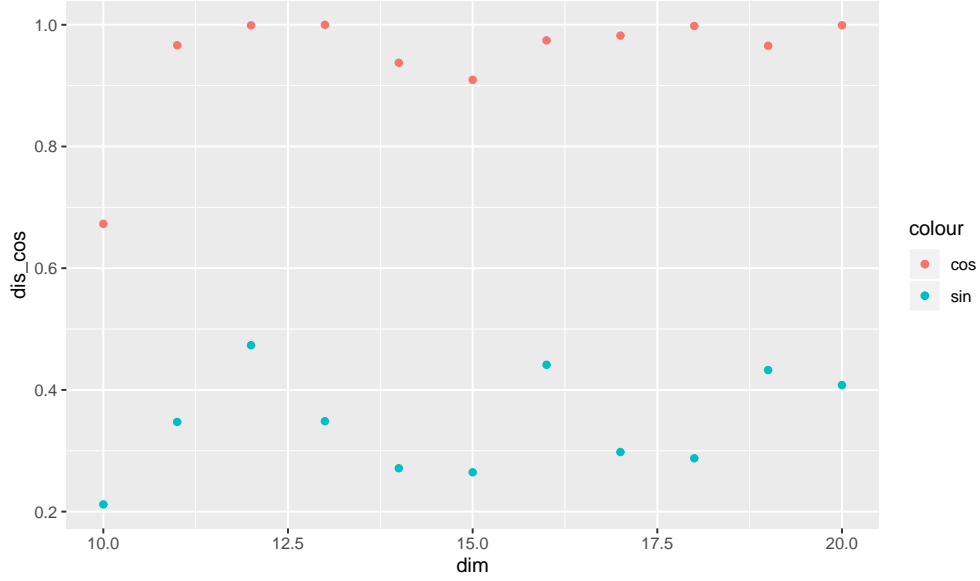


图 7: sin 函数生成的数据下维度对降维效果的影响 (一维)

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.

4 SAVE

4.1 理论依据

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

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 $[a, b] = [\min Y_{i=1}^n, \max Y_{i=1}^n]$ 划分为 k 个区间, 得到 \tilde{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 $\hat{\Sigma}_i = \sum_{j=1}^k \frac{1}{\#I_i} \sum_{Y_j \in I_i} (Z_j - \hat{\mu}_i)(Z_j - \hat{\mu}_i)^T$
4. 计算协方差矩阵估计量

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

5. 计算 \tilde{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.

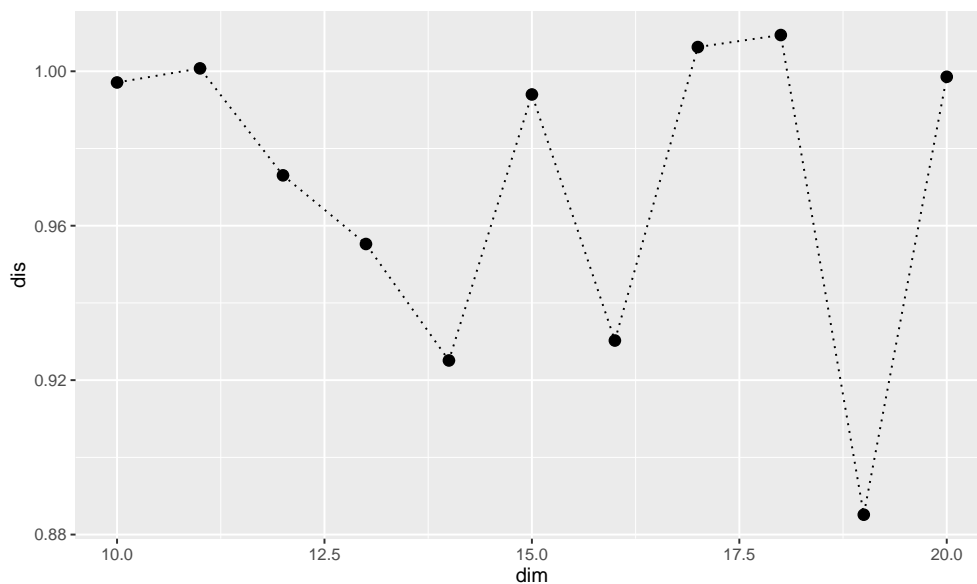


图 8: 线性回归下维度对降维效果的影响

4.4 cos 函数生成的数据的降维计算

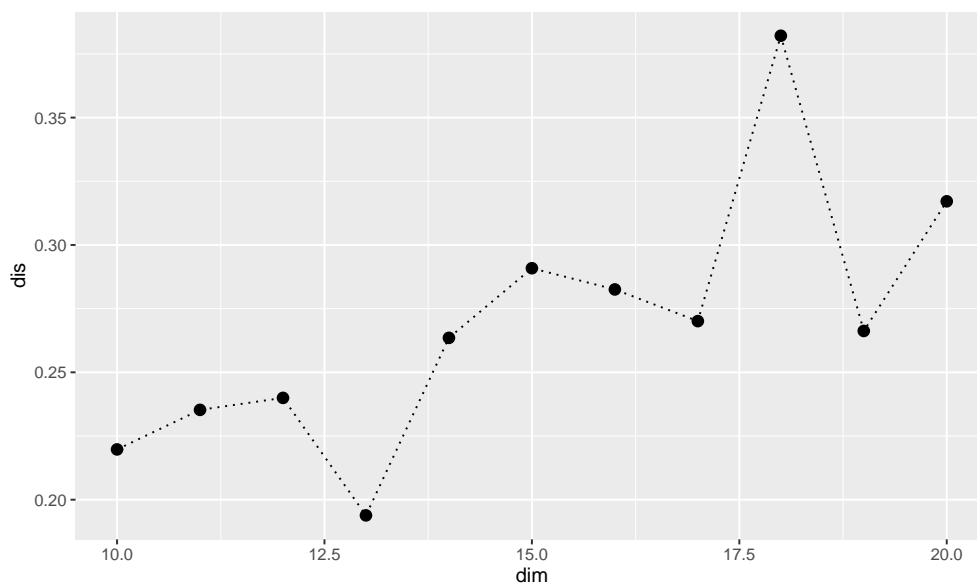


图 9: cos 函数生成的数据下维度对降维效果的影响 (二维)

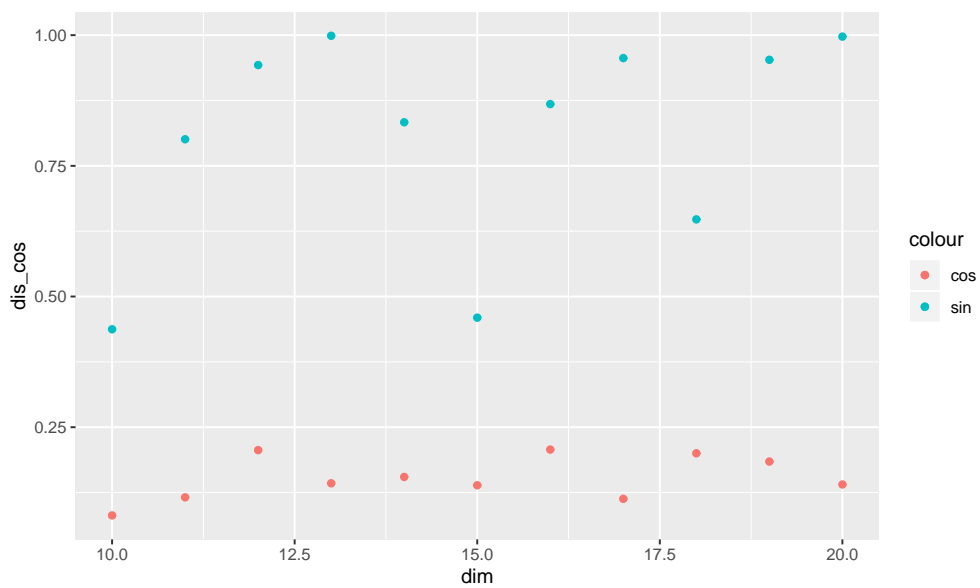


图 10: cos 函数与 sin 函数生成的数据下维度对降维效果的影响 (一维)

而在将降维后维数分别设定为 1,2,3,4,5 时, 其降维效果如图所示

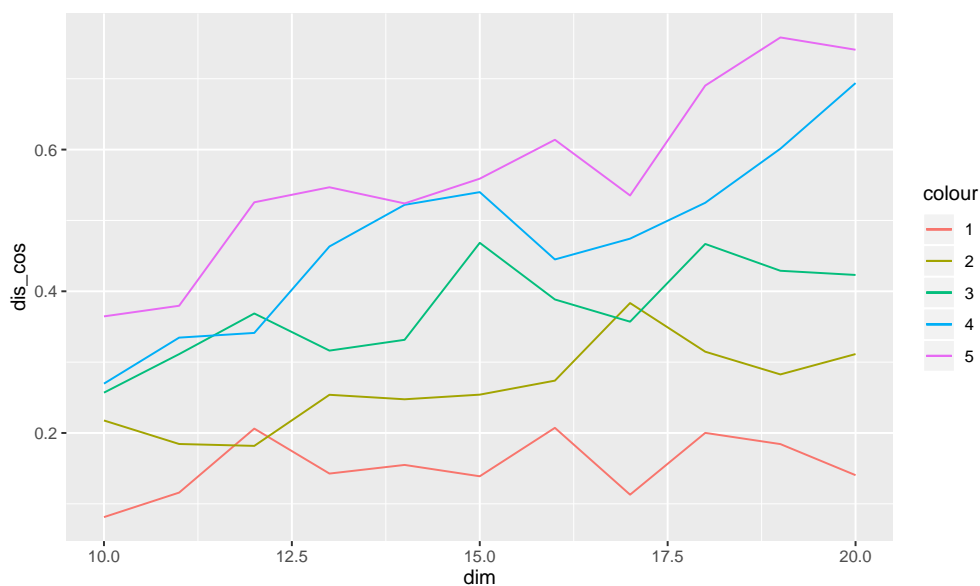


图 11: cos 函数生成的数据下降维后维度对降维效果的影响

4.5 总结

从图像可以看出, 样本量始终为 1000 的情况下, 增大 X 的维数, 估计空间与实际空间的差距呈上升趋势。线性回归下, 降维效果并不好, 且从其波动性可以看出, 模型结果首样本数据影响极大。由 cos 函数生成的 Y 的数据降维效果较好, 提高其生成函数实际空间的维数, 可见虽维数上升, 降维效果也有所下降。而由 sin 函数生成的 Y 的数据降维效果则非常不好, 与 cos 函数生成数据所

做结果进行比较，我们可一推测 SAVE 的降维效果可能与 Y 与 X 之间函数关系的奇偶性有关。