高维数据推断作业

邵李翔　祖劭康　赵张弛　李子昊

目录

1 利用 OLS 求解广义线性模型参数

1.1 模型表示

利用改进的计算方法得到的改进模型为

$$Y|X \sim G(\beta^T X) = G(E(XY)^T Xc)$$

利用最小线性二乘作为目标函数，牛顿迭代法计算此模型参数, 即用牛顿迭代法计算 $\min_c(Y - G(E(XY)^T Xc))^2$。得到参数 c, 从而得到参数 β 的估计。

1.2 模型计算

接下来通过广义线性模型具体计算估计参数的 MSE 来判断高维数据下的计算效率. 由于在泊松回归与 logistics 回归中需要计算回归平方和对参数 c 的一阶导与二阶导, 这会导致计算时间大大增加. 因此只罗列出其在 10 维,15 维,20 维数据中的降维效果.

将样本量设置为 1000,X 的每一维度均从 $[-1, 1]$ 上的均匀分布取样。设定参数 β 是每一维均为 1 的列向量. 在参数维度为 10 时, 三种模型其所估计的参数的各个分量如下表1所示

表 1: 三种回归模型在样本量 1000, 维度 10 时估计的参数

线性回归	泊松回归	logistics 回归
0.9975943	1.0448115	1.3493012
0.9994737	1.1038849	1.0676927
0.9972221	0.8499783	0.7865872
1.0039932	0.7242002	0.8772376
0.9909994	0.7985460	1.3469553
0.9998097	1.0124794	0.9293493
0.9988749	1.2205892	1.1112117
0.9948789	0.9288008	1.3919609
0.9938299	0.9899140	1.4916090
0.9933786	0.8456531	1.1946454

而三种模型在 10,15,20 维时的估计 MSE 如下表2所示 (由于泊松回归与 logistics 回归在样本量为 1000 时参数估计的效果不是很好, 所以在这两个模型所用的样本量均为 2000)

表 2: 三种回归模型在三种维度下的估计参数 MSE(重复 100 次)

模型	线性回归 (样本量 1000)	泊松回归 (样本量 2000)	logistics 回归 (样本量 2000)
10 维	0.0176	0.3403	0.3325
15 维	0.0209	1.2782	1.2826
20 维	0.0250	1.5676	1.6912

可以看出在维度增加的同时, 对于线性回归模型的参数估计精度并未太大影响（在每个维度上的变化程度并未明显改变，在 β 的每一维上的估计只在 1 上下 0.01 左右浮动）. 而在泊松回归与 logistics 回归模型中不仅会增加计算时间, 而且对精度也会产生很大的影响。因为随着维度增加，logistics 回归和泊松回归的 MSE 的变化幅度要大于线性回归的变化幅度。

2 pHd

2.1 原理

在课上的假设 2（线性条件）和假设 3（X 的条件方差是常矩阵）条件成立下，有：

$$
\begin{aligned}
&H_1 = \Sigma_{XXY} \\
&span(H_1) \subseteq S_{Y|X} \\
&H_2 = E[(Y - E(XY)^T X)XX^T] \\
&span(H_2) \subseteq S_{Y|X}
\end{aligned}
$$

2.2 样本计算流程

1. 将 $X_1, \ldots, X_n$ 标准化为 $Z_1, \ldots, Z_n$
2. 计算 $\hat{\Sigma}_{ZZY} E_n(ZZ^T Y), \quad \hat{\Sigma}_{XX} Var_n(X)$
3. 计算 $\hat{\Sigma}_{ZZY}$ 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 $v_k = \hat{\Sigma}_{XX}^{-1/2} u_k, k = 1, \ldots, q$

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：

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

在 pHd 方法中，选取了 y 对 x 的线性回归模型、泊松回归模型、logistics 回归模型与 cos 函数关系来观察 pHd 方法的降维效果。结果也可以看处，

2.4 线性回归

设定样本量为 1000, 样本 X 维度为 5,X 每一维度上的值均从标准正态分布中取样。β 作用 X 降维后的维度为 2. 即 β 是 10*2 的矩阵。一次的计算结果 H_1, H_2 如下表所示。设定的真实的 β 如下

$$\beta = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \tag{1}$$

表 3: 线性回归结果

H_1		H_2	
$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
-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

表3可以看出，与真实的参数 β 相比，phd 用 H1 方法得到的结果，并不是很接近。而且在降维后的维度不变情况下，随着 X 本身维度增加，估计效果越不好。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, 采用 H1 方法，估计 β，重复实验 100 次，计算距离的均值。判断依据为两个空间的距离（$L_2 - Hausdorff$ 子空间距离，越接近 0 表示两个空间越接近）. 如下图所示

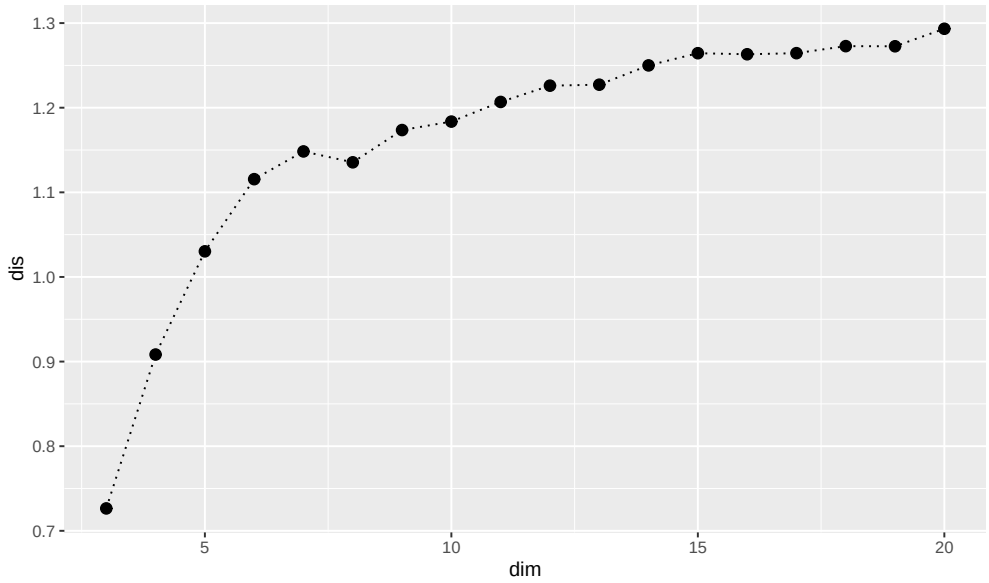

图 1: 线性回归下维度对降维效果的影响

2.5 泊松回归

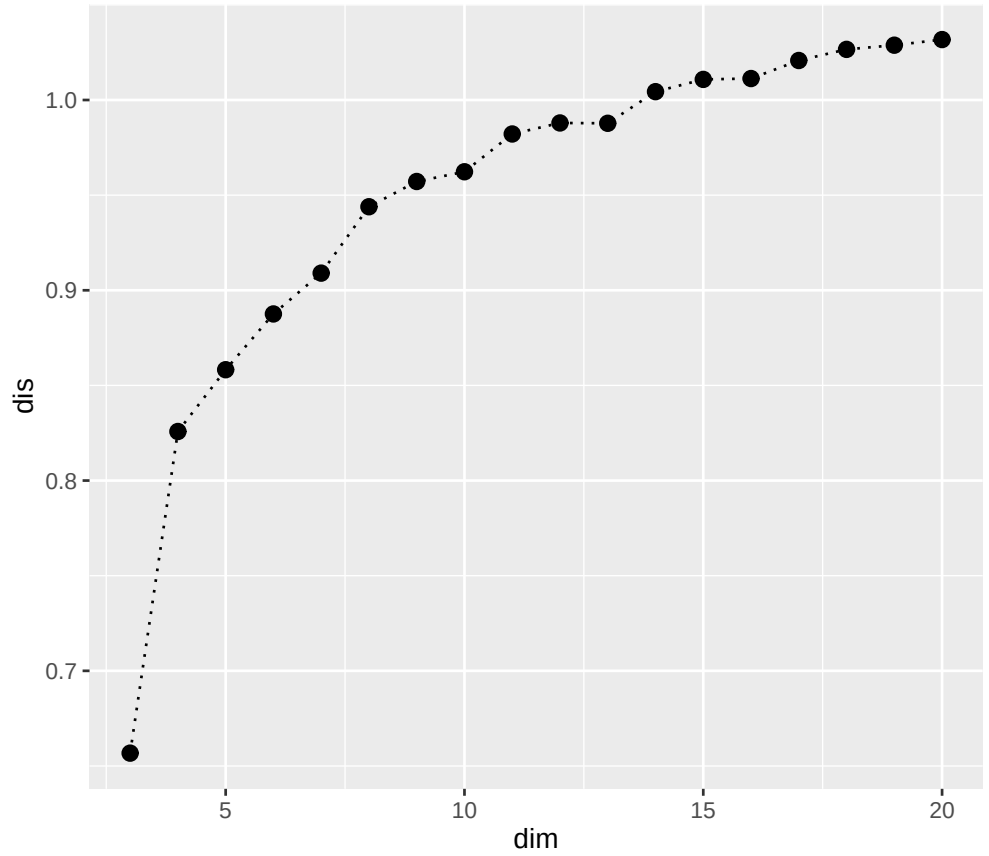

图 2: 泊松回归下维度对降维效果的影响

2.6 Logistic 回归

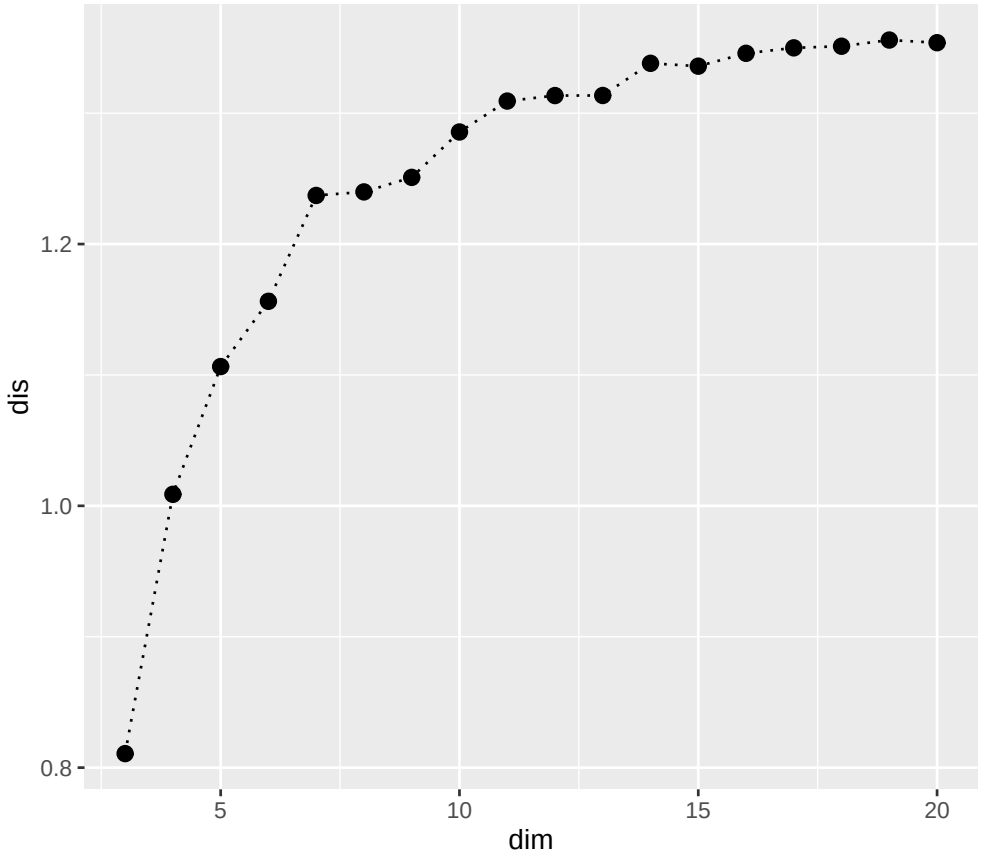

图 3: logistics 回归下维度对降维效果的影响

2.7 生成 cos 函数关系

按 $y = \cos(2\beta_1^T x) + \cos(\beta_2^T x) + \varepsilon$, 生成样本 y。其中 β_1 是参数矩阵 β 的第一列向量，同理 β_2，而 ε 服从均值为 0，方差 0.1 的正态分布。

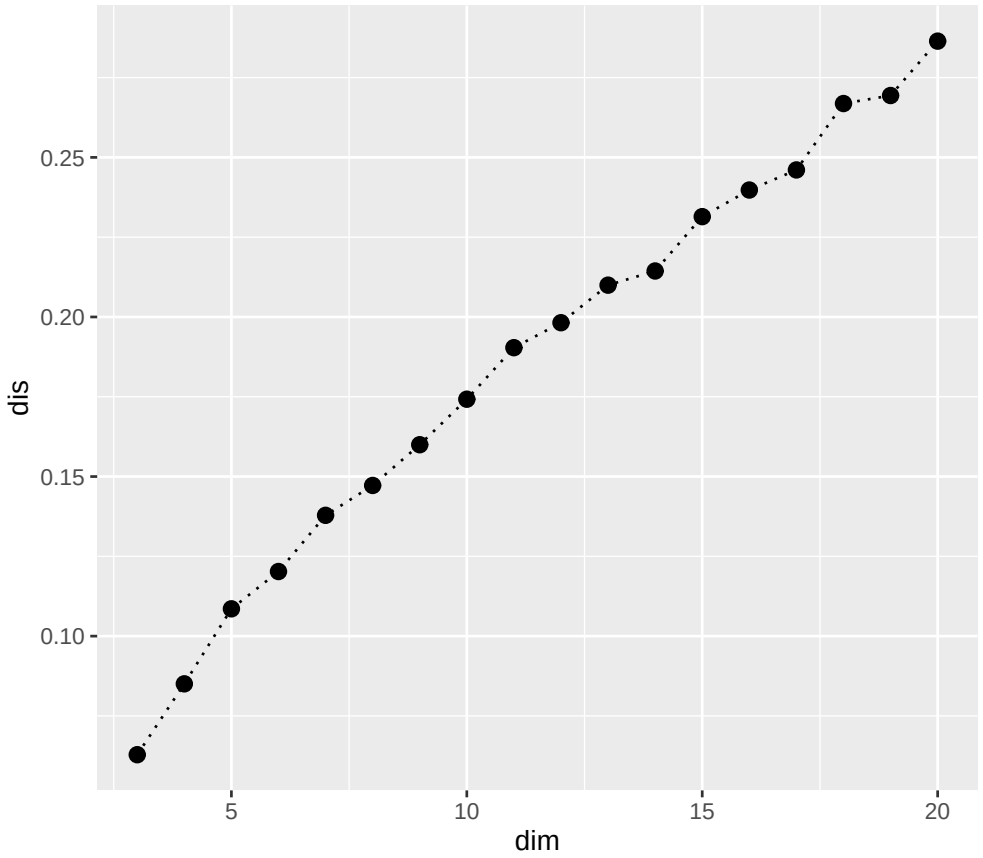

图 4: cos 函数生成的数据下维度对降维效果的影响

2.8 总结

样本量始终为 1000 的情况下，且降维后的维度不变条件下，增大 X 的维数，估计空间与实际空间的差距呈上升趋势。而且实际上 $L_2 - Hausdorff$ 子空间距离最大值为子空间 U 和 V 的较大的维度的开根号，在实验中，即两个空间的距离上限为 $\sqrt{2}$, 所以很明显这一方法在线性回归、Logistic 回归以及 Possion 回归上的表现都不好，因为计算的 β 张成的空间的距离与 pHd 方法得到的估计张成的空间的距离是快达到上限的，仅在 Y 是 X 的 cos 相关函数的情况下效果较为理想。到 20 维距离也才为 0.3，与 0 相近。

3 SIR

3.1 理论依据

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

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

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

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 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 $\hat{q}=arg\max\left\{i|\frac{\hat{\lambda_{i+1}}+C_n}{\hat{\lambda_i}+C_n}\tau\right\}$ 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 线性函数关系

设置降维维度为二维, 即按 $y=\beta_1^T x+\beta_2^T x+\varepsilon$, 生成样本 y。其中 β_1 是参数矩阵 β 的第一列向量，同理 β_2，而 ε 服从均值为 0，方差 0.1 的正态分布。得到的结果如下图所示

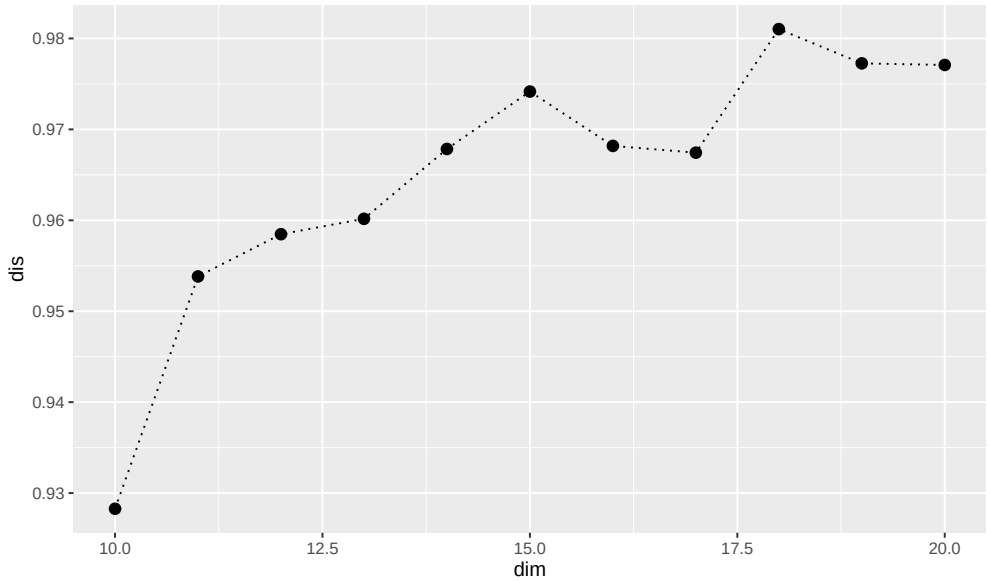

图 5: 线性函数关系

3.5 生成 sin 函数关系

首先设置降维维度为二维, 即按 $y = \sin(2\beta_1^T x) + \sin(\beta_2^T x) + \varepsilon$, 生成样本 y。其中 β_1 是参数矩阵 β 的第一列向量，同理 β_2，而 ε 服从均值为 0，方差 0.1 的正态分布。得到的结果如下图所示

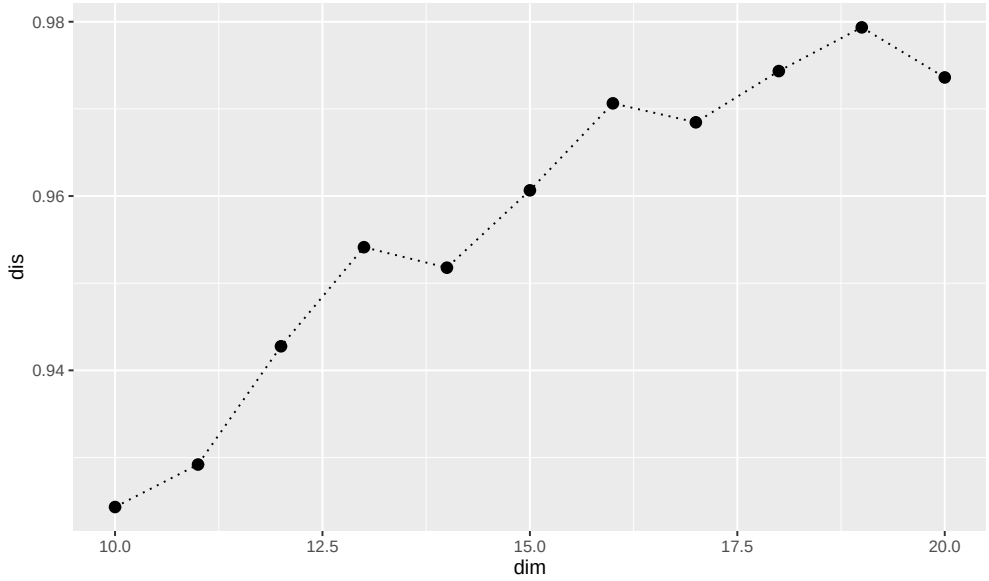

图 6: sin 函数生成的数据下维度对降维效果的影响 (二维)

而降维结果为一维时, 即按 $y = \sin(2\beta_1^T x) + \varepsilon$, 生成样本 y。与 cos 函数生成的数据结果比较如图所示

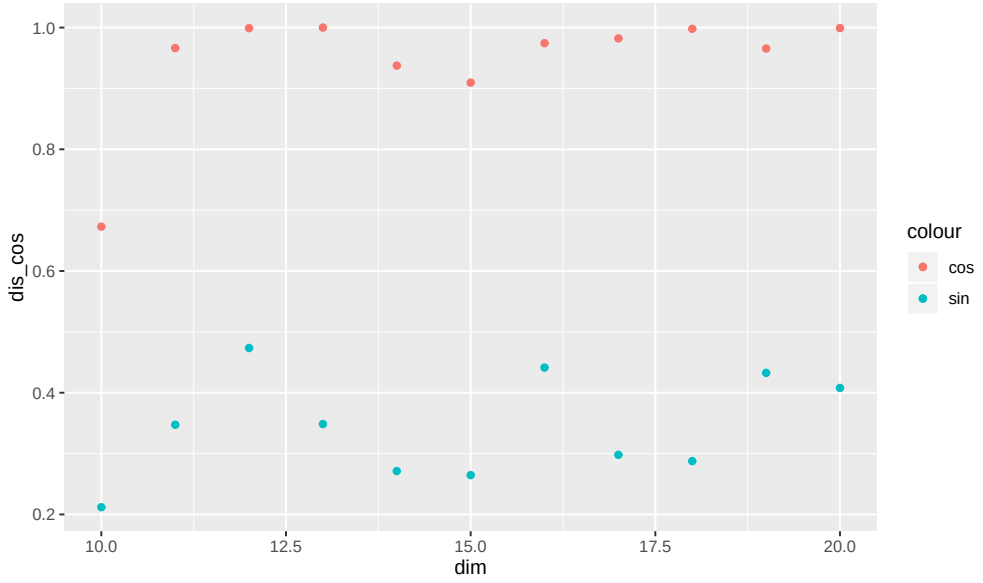

图 7: sin 函数生成的数据下维度对降维效果的影响 (一维)

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

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

4.2 样本计算流程

1. 将 $X_1, \ldots, X_n$ 标准化为 $Z_1, \ldots, Z_n$
2. 将 $[a, b] = [minY_{i i=1}^{n}, maxY_{i i=1}^{n}]$ 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 $u_1, \ldots, u_n$，则对 $S_{Y|X}$ 中向量的估计为 $v_k = \hat{\Sigma}_X^{-1/2} u_k, k = 1, \ldots, q$

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.

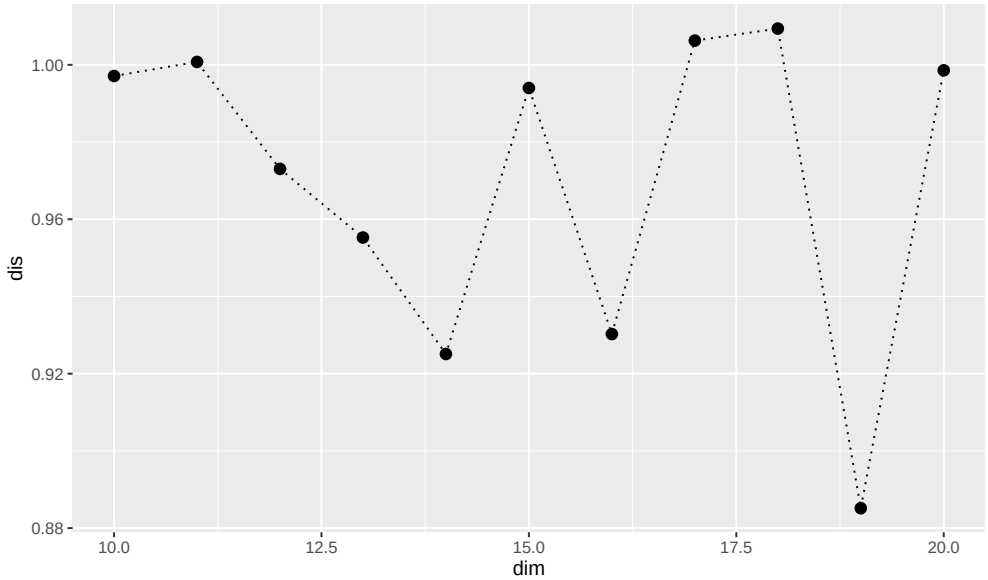

图 8: 线性回归下维度对降维效果的影响

4.4 cos 函数与 sin 函数生成的数据的降维计算

按 $y = \cos(2\beta_1^T x) + \cos(\beta_2^T x) + \varepsilon$, 生成样本 y。其中 β_1 是参数矩阵 β 的第一列向量，同理 β_2，而 ε 服从均值为 0，方差 0.1 的正态分布。对于 sinh 函数关系，我们按 $y = \sin(2\beta_1^T x) + \sin(\beta_2^T x) + \varepsilon$, 生成样本 y。一维下就去掉 $\beta_2^T x$ 项。

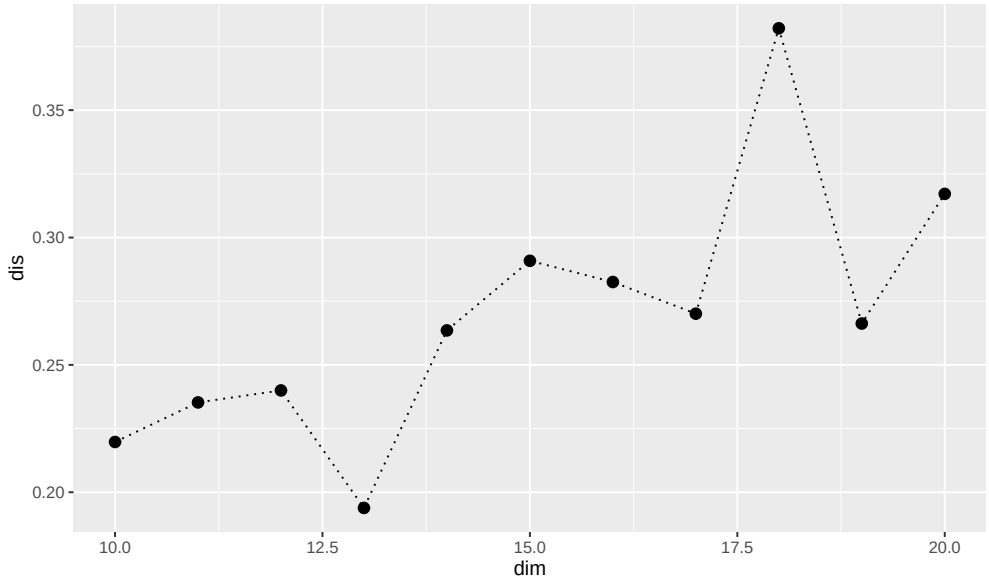

图 9: cos 函数生成的数据下维度对降维效果的影响 (二维)

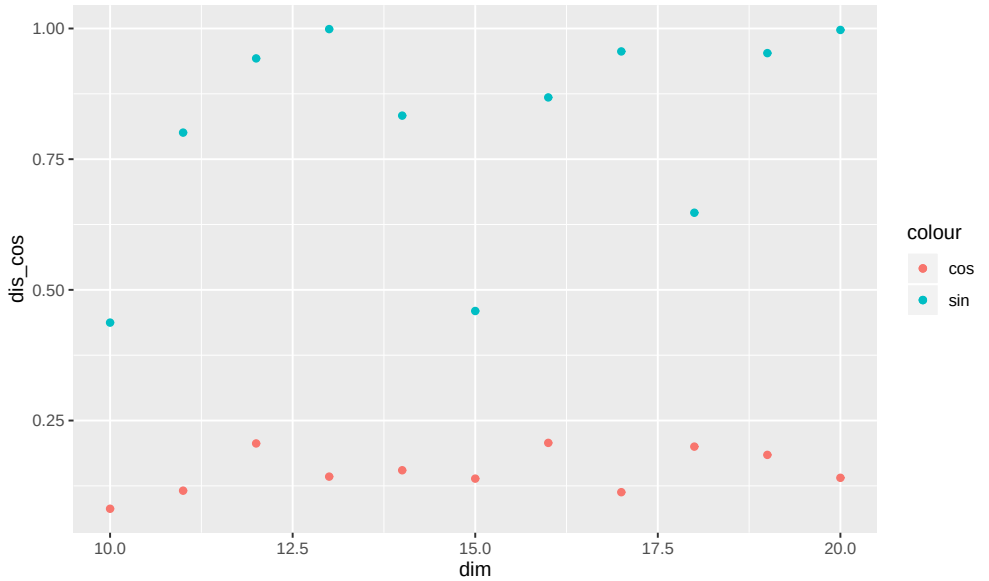

图 10: cos 函数与 sin 函数生成的数据下维度对降维效果的影响 (一维)

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

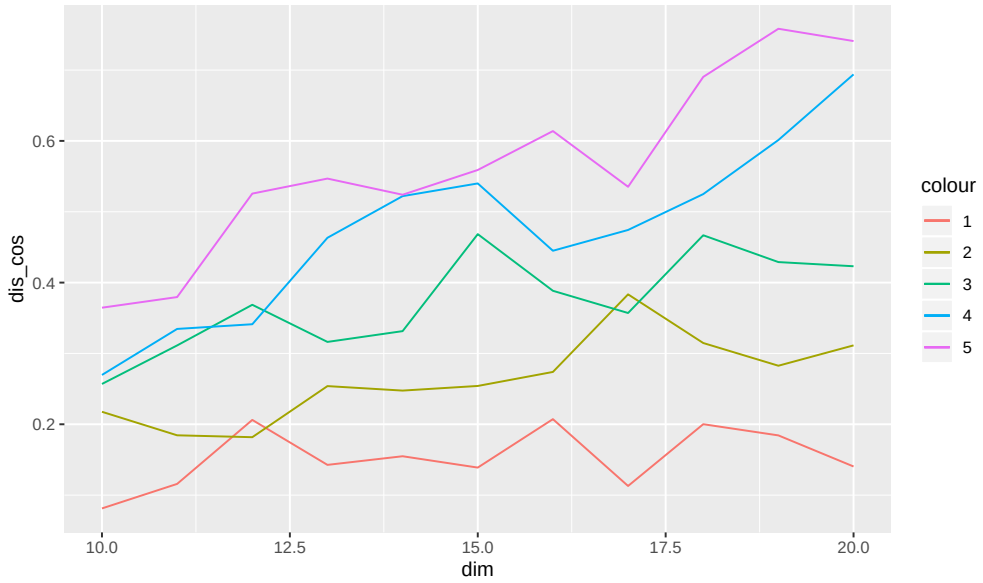

图 11: cos 函数生成的数据下降维后维度对降维效果的影响

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。