

高维数据推断第 2 次作业

邵李翔 祖劭康 赵张弛 李子昊

目录

1	SIR	3
1.1	理论依据	3
1.2	样本计算流程	3
1.3	确定估计后降维维度	3
1.4	线性函数关系	3
1.5	生成 sin 函数关系	4
2	SAVE	5
2.1	理论依据	5
2.2	样本计算流程	5
2.3	线性回归模型	6
2.4	泊松回归模型	6
2.5	logistic 回归	7
2.6	cos 函数与 sin 函数生成的数据的降维计算	8
2.7	总结	9
3	IHT	9
3.1	理论依据	9
3.2	算法流程	9
3.3	线性回归模型	10
3.4	对数似然回归	10
3.5	cos 与 sin 函数关系	11
3.6	三角函数在 IHT 与 PHD 下降维效果的对比	12

1 SIR

1.1 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

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

1.2 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

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

4. 计算 \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 $v_k = \hat{\Sigma}_X^{-1/2} u_k, k = 1, \dots, q$

1.3 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 $\hat{q} = \arg \max \left\{ i \mid \frac{\lambda_{i+1} + C_n}{\lambda_i + C_n} \tau \right\}$ 估计得到的降维后维度为 1, 其中设置 $C_n = \frac{1}{n^{\frac{1}{3}}}$.

1.4 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

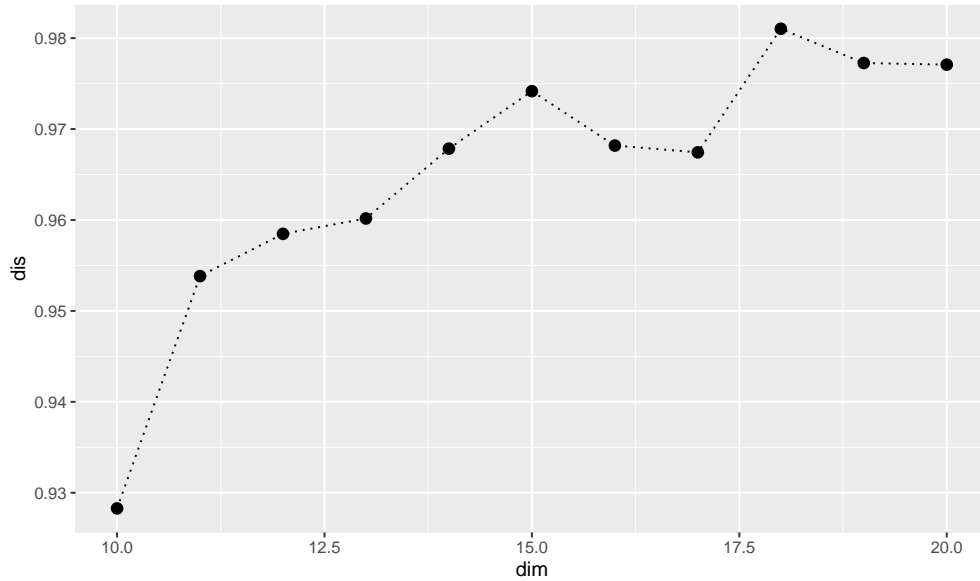


图 1: 线性函数关系

1.5 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

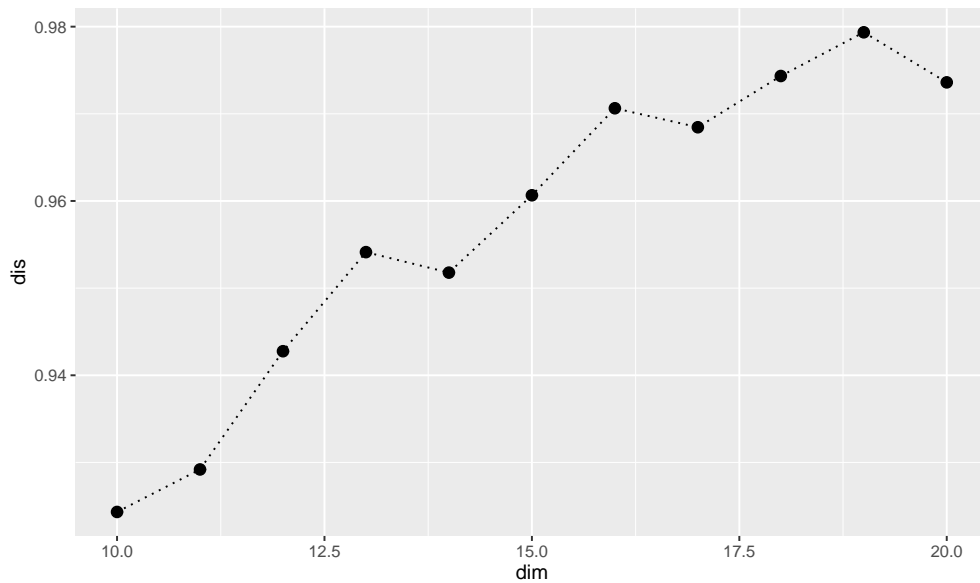


图 2: sin 函数生成的数据下维度对降维效果的影响 (二维)

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示

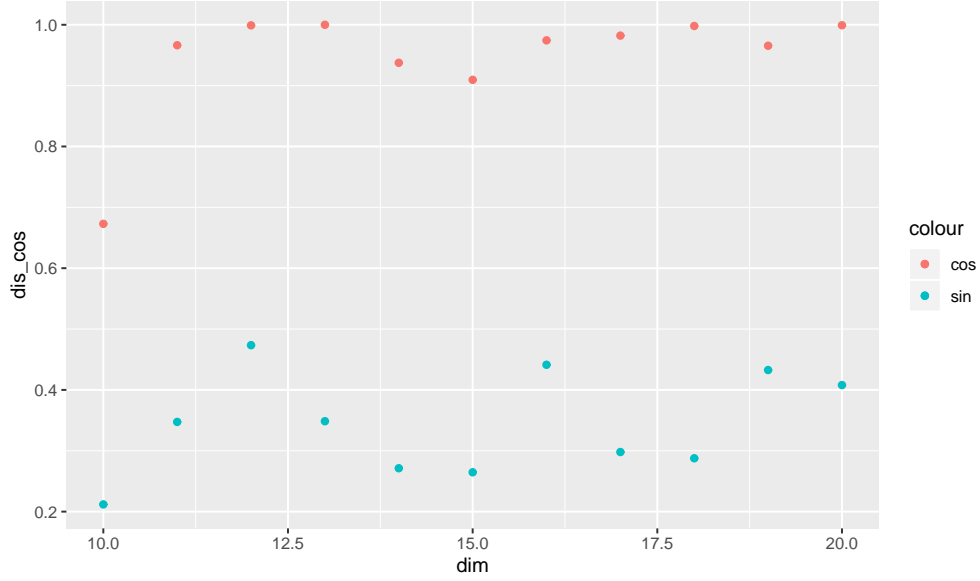


图 3: sin 函数生成的数据下维度对降维效果的影响 (一维)

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.

2 SAVE

2.1 理论依据

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

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. 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.

2.3 线性回归模型

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。设定样本量为 1000, 降维维数由 10 – 20, 判断其降维效果。

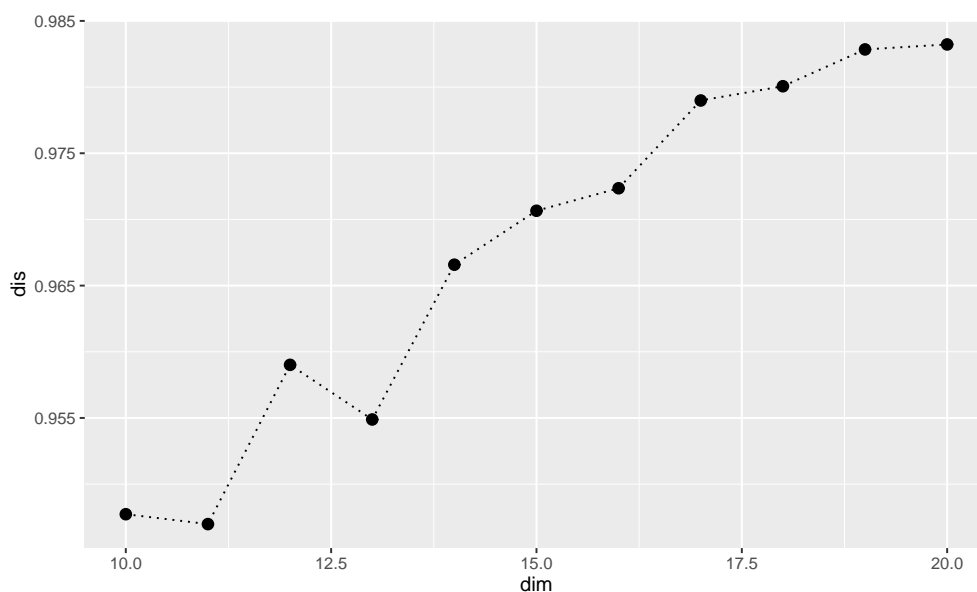


图 4: 线性回归下维度对降维效果的影响

2.4 泊松回归模型

按 $y = \exp(\beta^T x) + \varepsilon$ 生成样本 y , $\varepsilon \sim N(0, 0.1)$, 设定样本量为 1000, 降维维数由 10 – 20, 判断其降维效果。

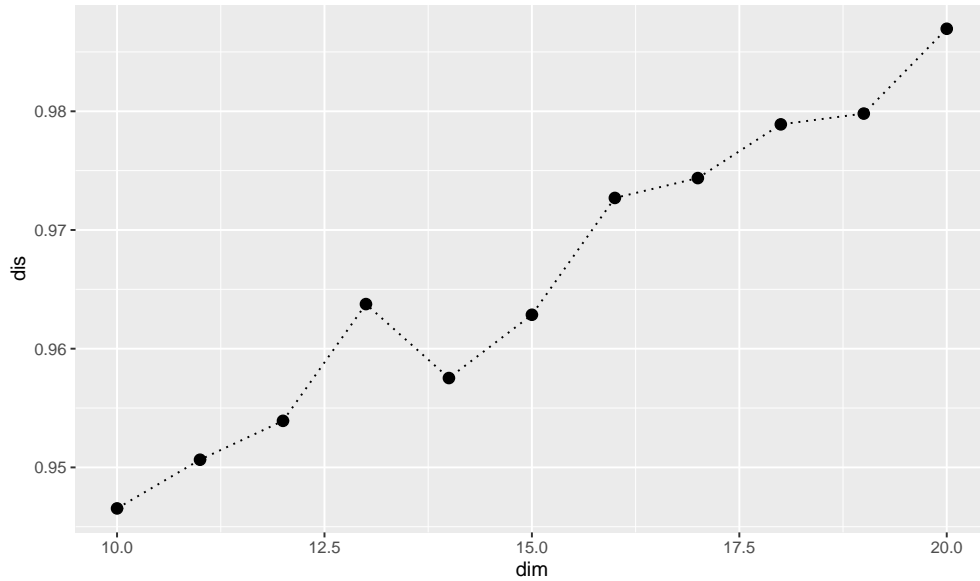


图 5: 泊松回归下维度对降维效果的影响

2.5 logistic 回归

按 $y = \frac{1}{1+\exp(-\beta^T x)} + \varepsilon$ 关系生成样本 $y, \varepsilon \sim N(0, 0.1)$, 设定样本量为 1000, 降维维数由 10 – 20, 判断其降维效果.

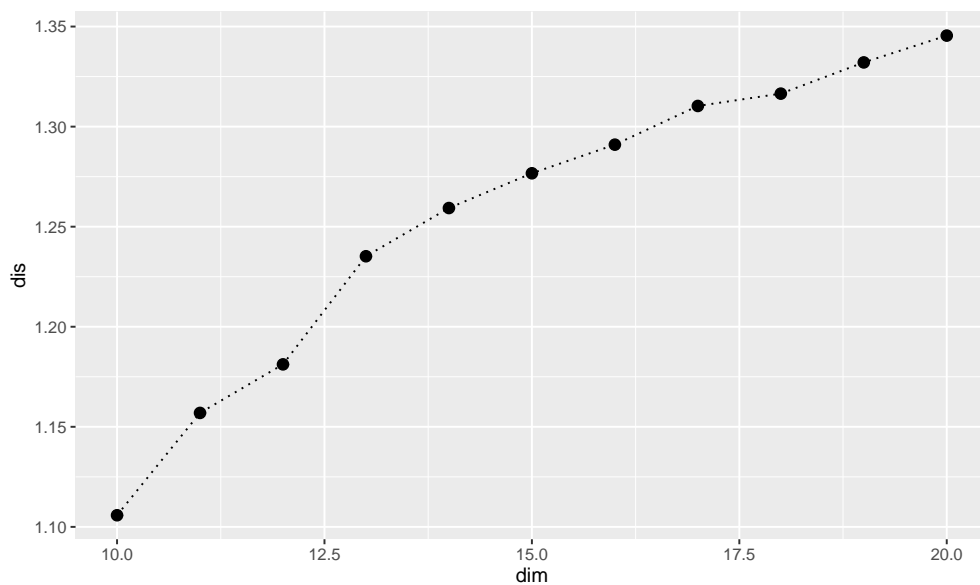


图 6: logistic 回归下维度对降维效果的影响

2.6 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 sin 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 。在一维情况下去掉 $\beta_2^T x$ 项。

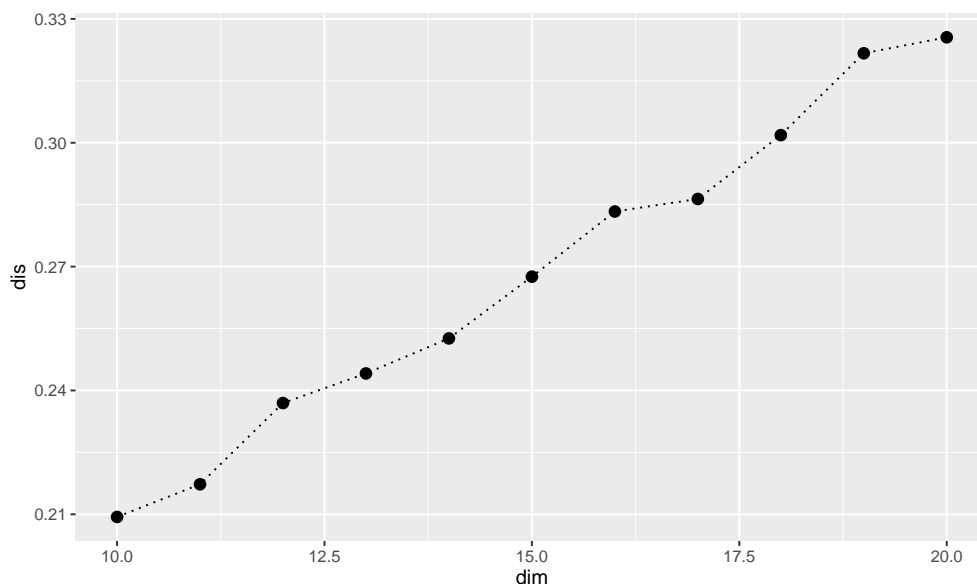


图 7: cos 函数生成的数据下维度对降维效果的影响 (二维)

而将目标降维维度设置为一维时,cos 函数与 sin 函数生成的数据下维度对降维效果的影响如图所示

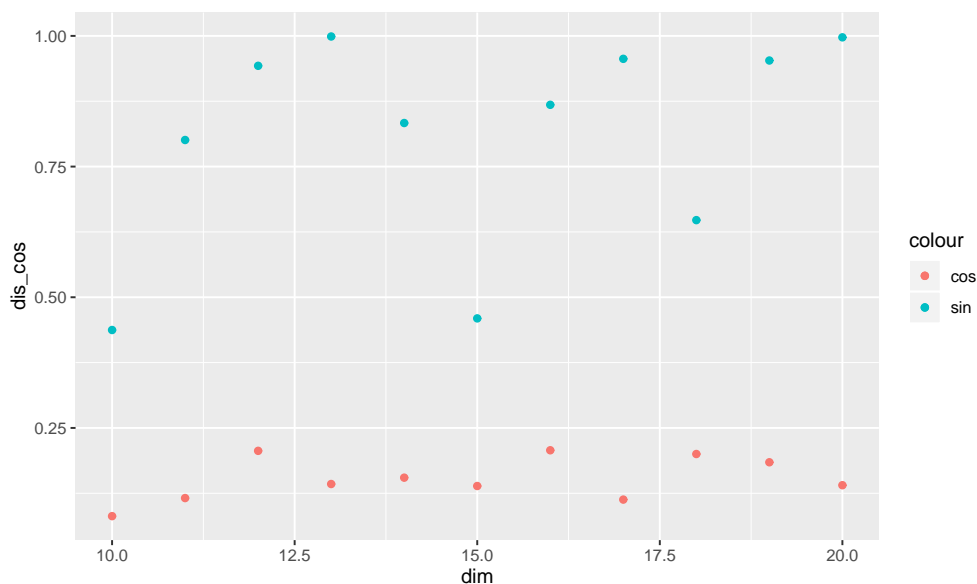


图 8: cos 函数与 sin 函数生成的数据下维度对降维效果的影响 (一维)

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

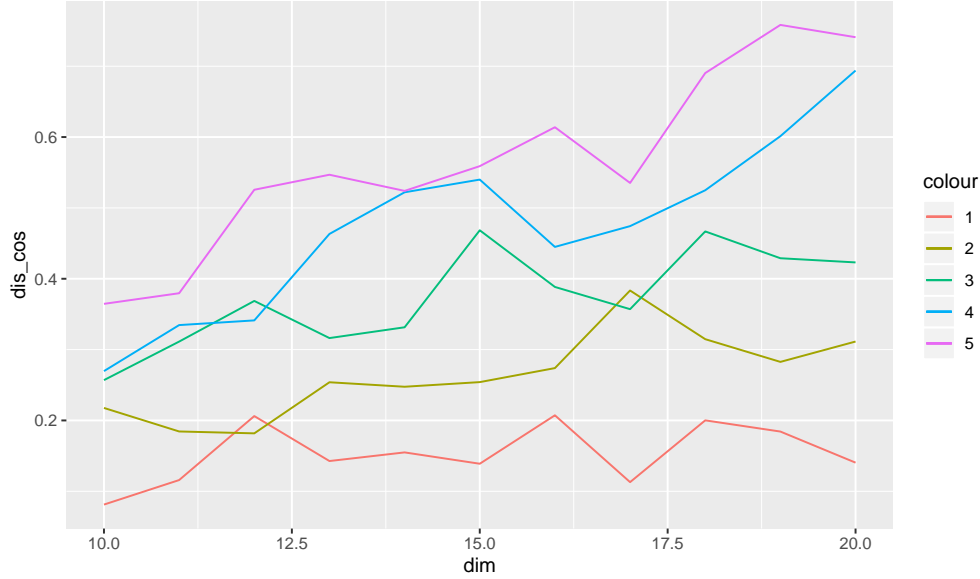


图 9: cos 函数生成的数据下降维后维度对降维效果的影响

2.7 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。

3 IHT

3.1 理论依据

$E(Z) = 0, \text{var}(Z) = I_p$, 则有 $S_{Y|Z}$ 是 Σ_{ZZY} 的不变子空间。

又有 $\Sigma_{ZY} \in S_{Y|Z}$, 由此可得 $\Sigma_{ZY}, \Sigma_{ZZY}\Sigma_{ZY}, \Sigma_{ZZY}^2\Sigma_{ZY}, \dots, \Sigma_{ZZY}^{p-1}\Sigma_{ZY} \in S_{Y|Z}$

3.2 算法流程

1. 将 X_1, \dots, X_n 标准化为 Z_1, \dots, Z_n
2. 计算 $\hat{\Sigma}_{ZY} = \text{Cov}_n(Z, Y), \hat{\Sigma}_{ZZY} = E_n(ZZ^TY)$
3. 计算 $p \times p$ 矩阵 $\hat{M} = (\hat{\Sigma}_{ZY}, \hat{\Sigma}_{ZZY}\hat{\Sigma}_{ZY}, \hat{\Sigma}_{ZZY}^2\hat{\Sigma}_{ZY}, \dots, \hat{\Sigma}_{ZZY}^{p-1}\hat{\Sigma}_{ZY})$ 并记 $\hat{\Lambda} = \hat{M}\hat{M}^T$
4. 计算 $\hat{\Lambda}$ 的前 q 个特征向量 u_1, \dots, u_q , 并令 $v_k = \hat{\Sigma}_{XX}^{-1/2}u_k, k = 1, \dots, q$
5. 则估计量为 $\{v_k^T[X_i - E_n(X)] : i = 1, \dots, n\}, k = 1, \dots, q$

3.3 线性回归模型

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。设定样本量为 1000, 降维维数由 10 – 20, 判断其降维效果。

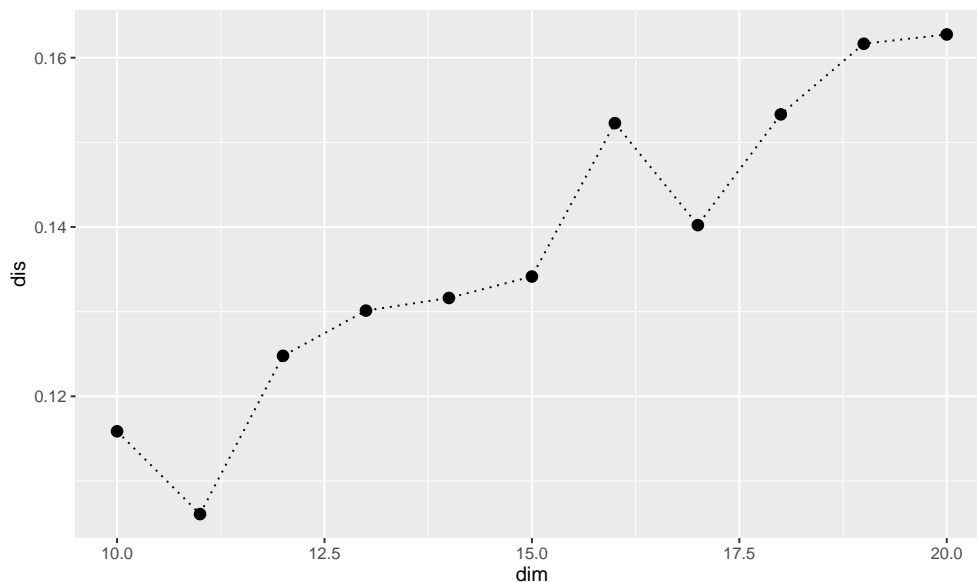


图 10: 线性回归下维度对降维效果的影响

3.4 对数似然回归

按 $y = \frac{1}{1+\exp(-\beta^T x)} + \varepsilon$ 关系生成样本 $y, \varepsilon \sim N(0, 0.1)$, 设定样本量为 1000, 降维维数由 10 – 20, 判断其降维效果。

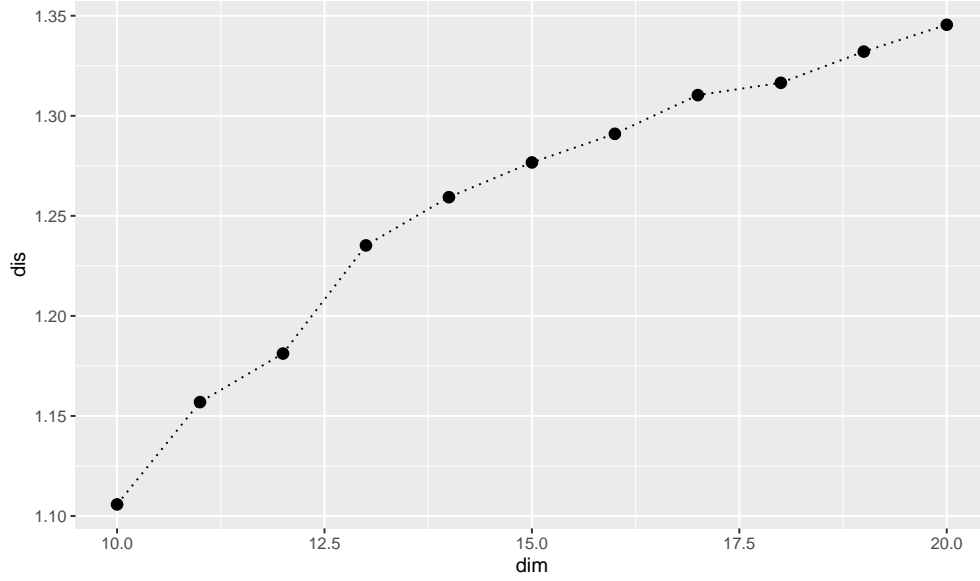


图 11: logistic 回归下维度对降维效果的影响

3.5 cos 与 sin 函数关系

按 $y = \sin(2\beta_1^T x) + \cos(\beta_2^T x) + \varepsilon$, 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 sin 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 。

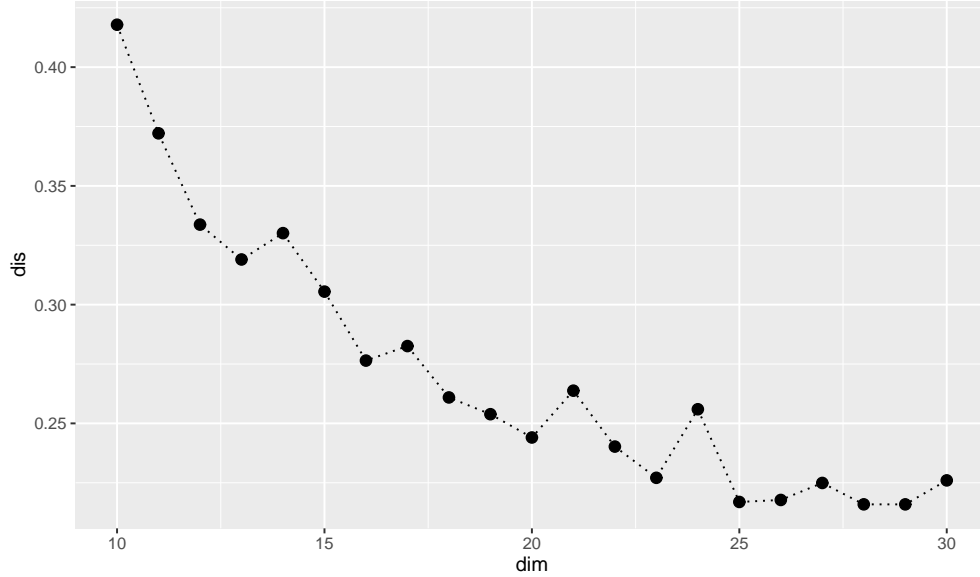


图 12: 三角函数生成的数据下维度对降维效果的影响

可以看出在降维子空间在 25 维之后降维效果有下降的趋势, 由于计算时间过长, 本次试验没有进行更高维度的判断。

3.6 三角函数在 IHT 与 PHD 下降维效果的对比

按 $y = \cos(2\beta_1^T x) + \sin(\beta_2^T x) + \varepsilon$, 生成样本 y . 分别使用 PHD 方法与 IHT 方法进行降维.

$$\beta_{phd} = \begin{pmatrix} 0.948449657 & 0.01971554 \\ -0.047997173 & 0.29983925 \\ 0.034641505 & -0.19857183 \\ 0.008065384 & 0.69594286 \\ -0.073057556 & -0.31804108 \\ 0.027100694 & 0.13619042 \\ -0.130396457 & -0.26034245 \\ 0.171602330 & -0.11436164 \\ 0.068274073 & -0.33846860 \\ 0.081761191 & -0.21252464 \end{pmatrix} \quad (1)$$

$$\beta_{iht} = \begin{pmatrix} 0.976733646 & -0.0647860186 \\ -0.019360699 & -0.9058644433 \\ 0.015809941 & 0.4508445520 \\ 0.006769810 & 0.1351203693 \\ -0.016888058 & -0.0002618049 \\ 0.007019351 & -0.0178304939 \\ -0.010093698 & 0.0408656636 \\ 0.008968595 & -0.0299744106 \\ 0.017278169 & 0.0001168891 \\ -0.003163035 & -0.0225223252 \end{pmatrix} \quad (2)$$

可以看出 IHT 方法处理此类函数关系时的降维效果要明显由于 PHD 方法, 即虽然在 H_1 方向下的降维效果接近真实值, 但是在 H_2 方向下降维效果 PHD 要劣于 IHT.