RaSE: A Variable Screening Framework via Random Subspace Ensembles

Key words: Ensemble learning; High-dimensional data; Random subspace method; Rank consistency;
Presenter: Fei Yang
2022/05/19

CONTENT

Background

随机子空间思想

- Random subspace method（RSM）又叫attribute bagging 或者 feature bagging，是集成学习的一种。随机子空间通过使用随机的部分特征而不是所有的特征来训练每个分类器，来降低每个分类器之间的相关性。
- It randomly selects a feature subset and grows each tree within the chosen subspace.
- 它类似bagging, bagging是随机使用部分训练数据，而Random subspace method是随机使用部分特征。因此，在非正式的情况下，随机子空间会导致个别学习器不会过度关注在训练集中表现出高度预测性/描述性的特征，但无法对训练集中以外的点进行预测。因此，对于特征数目远大于训练点数的问题，随机子空间是一个很有吸引力的选择，例如核磁共振、基因组序列、CSI（信道状态信息）。
- 实际上，**随机森林就是一个使用了RSM和bagging的decision tree**。同样的，RSM也可以用在SVM等其他分类器上。在训练出每个分类器之后进行预测，得到每个分类器对应的结果。根据多数投票或结合先验概率的方法获得最终结果。

Background

随机投影 Random Projection

- “Also, as Cannings and Samworth (2017) pointed out, the random subspace method can be regarded as the random projection ensemble classification method when the projection space is restricted to be axis-aligned.”

- J-L 引理

THEOREM 3.19 (Johnson-Lindenstrauss Lemma). *Let* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^D$ *for some* D. *Let* $\boldsymbol{A} \in \mathbb{R}^{m \times D}$ *be a random matrix whose entries are independent* $\mathcal{N}(0, 1/m)$ *random variables. Then for any* $\varepsilon \in (0, 1)$, *with probability at least* $1 - 1/n^2$, *the following holds:*

$$\forall\, i \neq j, \quad (1-\varepsilon)\left\|\boldsymbol{v}_i - \boldsymbol{v}_j\right\|_2^2 \;\leq\; \left\|\boldsymbol{A}\boldsymbol{v}_i - \boldsymbol{A}\boldsymbol{v}_j\right\|_2^2 \;\leq\; (1+\varepsilon)\left\|\boldsymbol{v}_i - \boldsymbol{v}_j\right\|_2^2, \quad (3.4.2)$$

provided $m > 32\frac{\log n}{\varepsilon^2}$.

Background

随机投影 Random Projection

J-L 引理：从N维到$\mathcal{O}\log(N)$维

高维背景空间中的两个低维子空间中的数据点，在经过高斯随机矩阵的压缩（降维）后，将在新的背景空间中形成两个新的低维子空间。这两个新子空间的距离（或理解为夹角）在很大概率上保持不变。

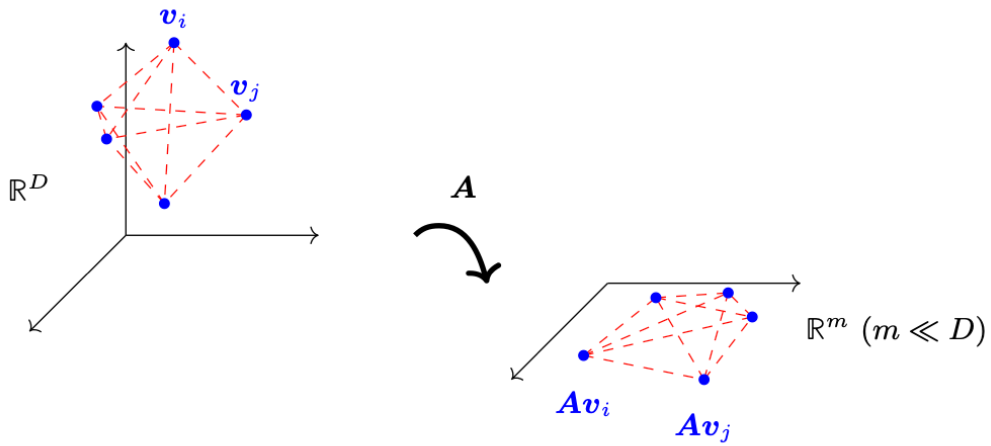


Figure 3.10 The Johnson-Lindenstrauss Lemma. Given a fixed collection of points $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ in a high-dimensional space $\mathbb{R}^D$, with high probability a random mapping into $m \sim \log n$ dimensions approximately preserves the distances between all pairs of points.

Motivation

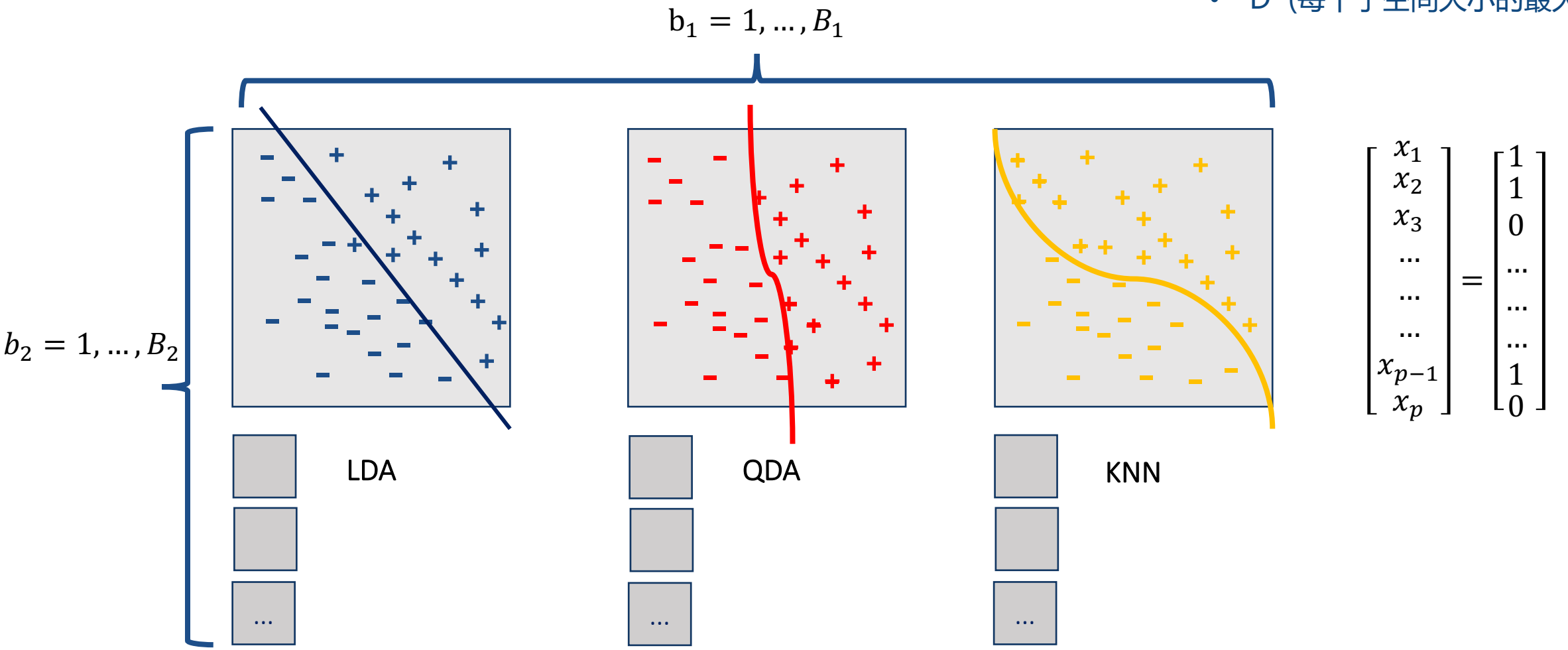
1. 边际不相关，联合相关的变量不易被基于边际扫描的方法选出；

2. 先有文献对于迭代式扫描框架没有理论支撑；

3. 与LASSO，SCAD等惩罚方法结合的扫描算法过度依赖于参数的选取

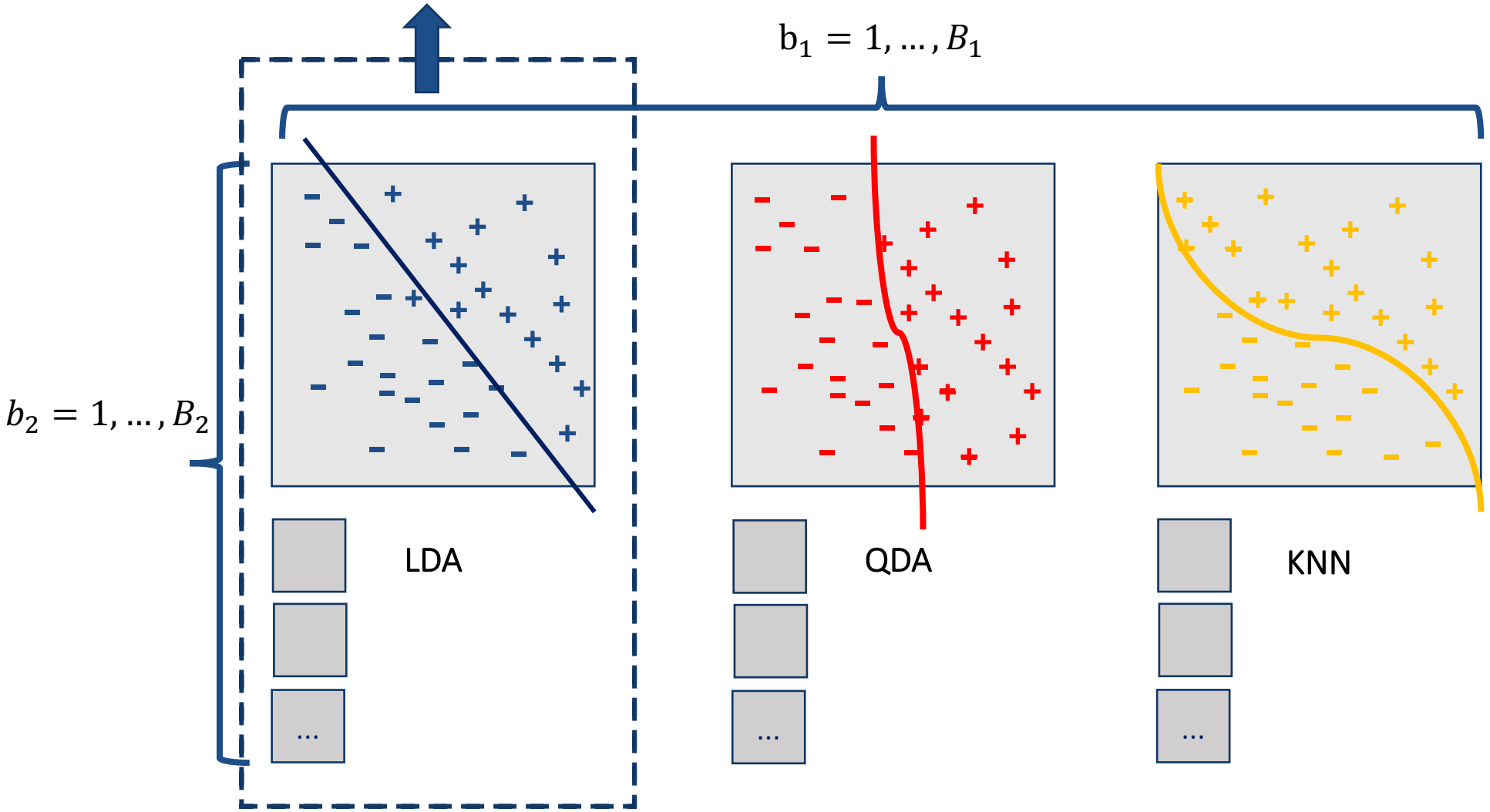
举例：S为子空间，$\{S_{b_1 b_2}, b_1 = 1 \ldots, B_1 \ ， \ b_2 = 1 \ldots, B_2\}$

- B_1 (子空间的组数)
- B_2 (每组子空间内的大小)
- D (每个子空间大小的最大值)

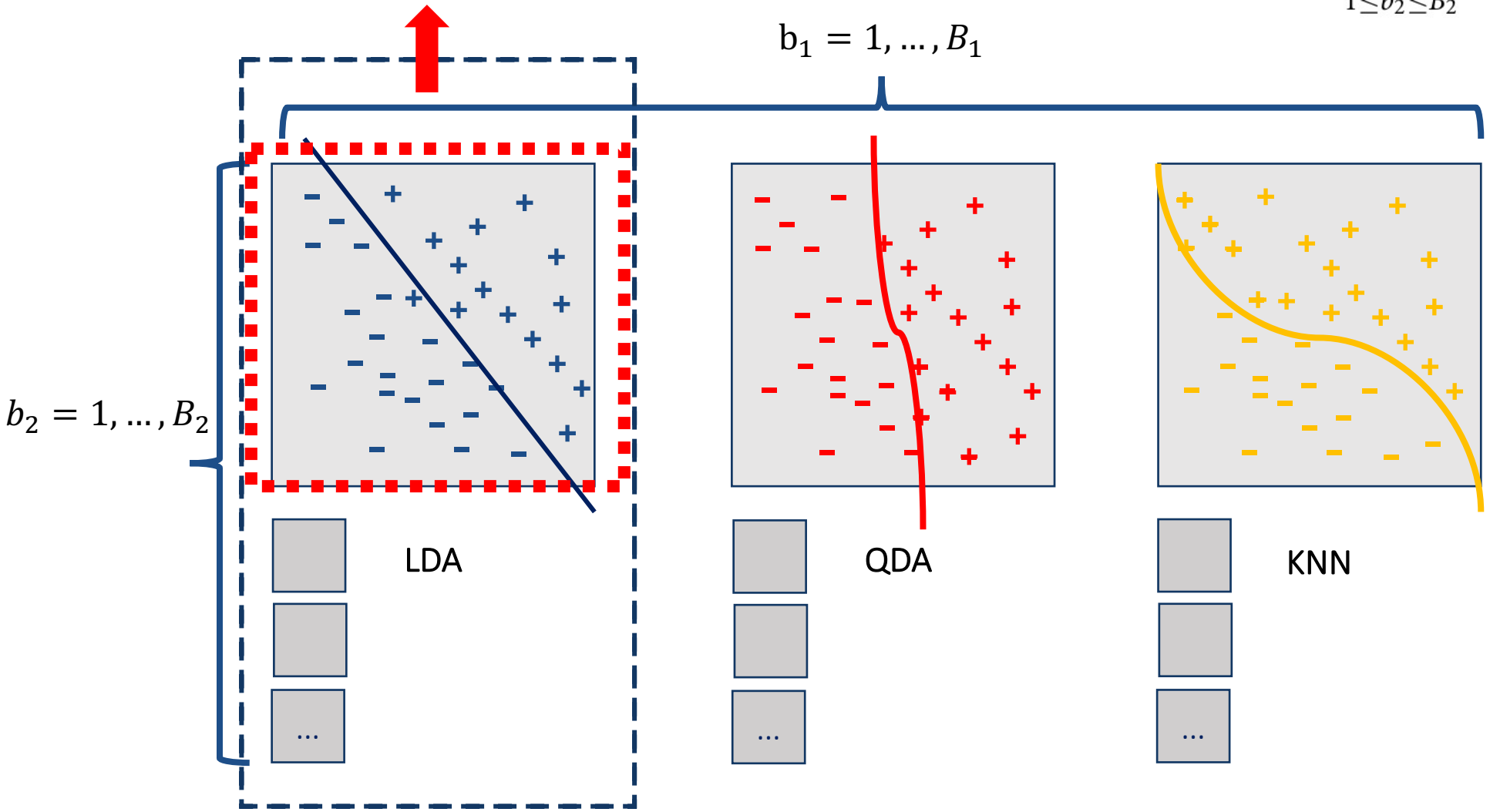


$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \ldots \\ \ldots \\ \ldots \\ x_{p-1} \\ x_p \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ \ldots \\ \ldots \\ \ldots \\ 1 \\ 0 \end{bmatrix}$$

虚线框内为$\mathrm{b}_1 th$ group of subspaces，$\{\mathrm{S}_{b_1 b_2}\}_{b_2=1}^{B_2}$



对该组$b_1 th$ group of subspaces，我们可以信息准测选择最优的一个S_{b_1*} Select the optimal subspace $S_{b_1*} = S_{b_1 b_2^*}$, where $b_2^* = \arg\min_{1 \le b_2 \le B_2} \mathrm{Cr}_n(S_{b_1 b_2})$



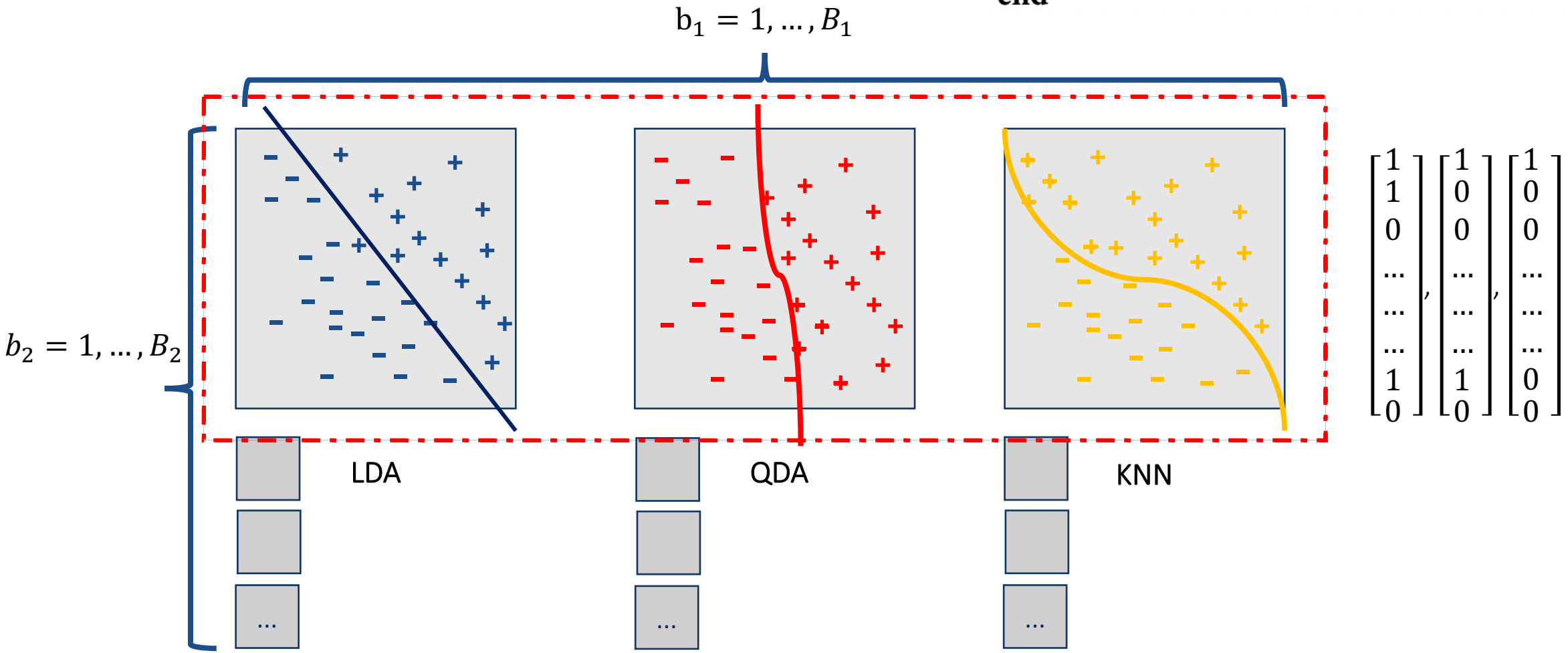
for $b_1 \leftarrow 1$ **to** B_1 **do**

Select the optimal subspace $S_{b_1*} = S_{b_1 b_2^*}$, where $b_2^* = \arg\min_{1 \le b_2 \le B_2} \mathrm{Cr}_n(S_{b_1 b_2})$

end

结合$B_1 = 3$个弱分类器的最优子空间为

The selected B_1 subspaces as $\{S_{b_1*}\}_{b_1=1}^{B_1}$



2.Methodology

Algorithm- Vanilla RaSE screening

Notation

- 子空间的分布为$\mathcal{D}$
- 子空间大小的最大值为D
- 初始步骤，从均匀分布$\{1, \ldots, D\}$生成子空间大小d
- E.g. $\{S_{11} \subseteq S_{full} \, |S_{11}| = d \}$

注意：

- D=1时，RaSE等同于marginal screening procedure
- 对于交互项的识别非常有效

Algorithm 1: Vanilla RaSE screening

Input: training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, subspace distribution $\mathcal{D}$, criterion function Cr_n, integers B_1 and B_2, number of variables N to select

Output: the selected proportion of each feature $\hat{\boldsymbol{\eta}}$, the selected subset $\hat{S}$

Independently generate random subspaces $S_{b_1 b_2} \sim \mathcal{D}, 1 \le b_1 \le B_1, 1 \le b_2 \le B_2$

for $b_1 \leftarrow 1$ **to** B_1 **do**

  Select the optimal subspace $S_{b_1 *} = S_{b_1 b_2^*}$, where $b_2^* = \arg \min_{1 \le b_2 \le B_2} \mathrm{Cr}_n(S_{b_1 b_2})$

end

Output the selected proportion of each feature $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_p)^T$ where $\hat{\eta}_j = B_1^{-1} \sum_{b_1=1}^{B_1} \mathbb{1}(j \in S_{b_1 *}), j = 1, \ldots, p$

Output $\hat{S} = \{1 \le j \le p : \hat{\eta}_j$ is among the N largest of all$\}$

2.Methodology

Algorithm- Vanilla RaSE screening

$The\ selected\ proportion\ of\ each\ feature\ \widehat{\eta_j}$

对各组子空间的optimal subset内含有的变量被挑选的次数做算术平均

/ 在全部分类器最佳分类结果下，第j个变量被包含在最佳子集下的概率

$$\begin{bmatrix}1\\1\\0\\\dots\\\dots\\\dots\\1\\0\end{bmatrix},\begin{bmatrix}1\\0\\0\\\dots\\\dots\\\dots\\1\\0\end{bmatrix},\begin{bmatrix}1\\0\\0\\\dots\\\dots\\\dots\\0\\0\end{bmatrix}$$

$$\widehat{\eta_1}=\frac{1}{3}\times(1+1+1)=1$$

$$\widehat{\eta_2}=\frac{1}{3}\times(1+0+0)=\frac{1}{3}$$

Algorithm 1: Vanilla RaSE screening

Input: training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, subspace distribution $\mathcal{D}$, criterion function Cr_n, integers B_1 and B_2, number of variables N to select

Output: the selected proportion of each feature $\hat{\boldsymbol{\eta}}$, the selected subset $\hat{S}$

1 Independently generate random subspaces $S_{b_1b_2} \sim \mathcal{D}, 1 \le b_1 \le B_1, 1 \le b_2 \le B_2$

2 **for** $b_1 \leftarrow 1$ **to** B_1 **do**

3 Select the optimal subspace $S_{b_1*} = S_{b_1b_2^*}$, where $b_2^* = \arg\min_{1\le b_2\le B_2} \mathrm{Cr}_n(S_{b_1b_2})$

4 **end**

5 Output the selected proportion of each feature $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \dots, \hat{\eta}_p)^T$ where $\hat{\eta}_j = B_1^{-1}\sum_{b_1=1}^{B_1} \mathbb{1}(j \in S_{b_1*}), j = 1, \dots, p$

6 Output $\hat{S} = \{1 \le j \le p : \hat{\eta}_j$ is among the N largest of all$\}$

2.Methodology

Algorithm- Vanilla RaSE screening

对$\hat{\eta}_j$进行排序，选择前$[\alpha D/c_{2n}]$个变量，作为扫描后的子集$\widehat{S_\alpha}$

$$\hat{S}_\alpha = \{1 \le j \le p : \hat{\eta}_j \text{ is among the } [\alpha D/c_{2n}] \text{ largest of all}\},$$

where c_{2n} is a constant (to be specified in the next section) depending on n, B_2, D, and the criterion Cr which is a population counterpart of Cr_n. Here, α can be any constant larger than 1, which will appear in the upper bound introduced in the sure screening theorem of Section 3.

2.Methodology

Algorithm-iterative RaSE screening

迭代式方法：

(1) 根据上一轮的选择概率 更新子空间分布$\mathcal{D}^{[t+1]}$

每个iteration t，根据$\tilde{\eta}_j^{[t]}$ (该值由$\hat{\eta}_j^{[t]}$构建)建立层次多项分布hierarchical restrictive multinomial distribution

$$\mathcal{R}(\mathcal{U}, p, \tilde{\boldsymbol{\eta}}), \text{ where } \sum_{j=1}^{p} \tilde{\eta}_j = 1 \text{ and } \tilde{\eta}_j \geq 0,$$

$$\tilde{\eta}_j^{[t]} \propto [\hat{\eta}_j^{[t]} \mathbb{1}(\hat{\eta}_j^{[t]} > C_0/\log p) + \frac{C_0}{p} \mathbb{1}(\hat{\eta}_j^{[t]} \leq C_0/\log p)]$$

(2) 在迭代结束后再进行变量扫描

Algorithm 2: Iterative RaSE screening (RaSE$_T$)

Input: training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, initial subspace distribution $\mathcal{D}^{[0]}$, criterion function Cr_n, integers B_1 and B_2, the number of iterations T, positive constant C_0, number of variables N to select

Output: the selected proportion of each feature $\hat{\boldsymbol{\eta}}^{[T]}$, the selected subset $\hat{S}$

for $t \leftarrow 0$ **to** T **do**
  Independently generate random subspaces $S_{b_1 b_2}^{[t]} \sim \mathcal{D}^{[t]}, 1 \leq b_1 \leq B_1, 1 \leq b_2 \leq B_2$
  for $b_1 \leftarrow 1$ **to** B_1 **do**
    Select the optimal subspace $S_{b_1 *}^{[t]} = S_{b_1 b_2^*}^{[t]}$, where $b_2^* = \arg\min_{1 \leq b_2 \leq B_2} \mathrm{Cr}_n(S_{b_1 b_2}^{[t]})$
  end
  Update $\hat{\boldsymbol{\eta}}^{[t]}$ where $\hat{\eta}_j^{[t]} = B_1^{-1} \sum_{b_1=1}^{B_1} \mathbb{1}(j \in S_{b_1 *}^{[t]}), j = 1, \ldots, p$
  Update $\mathcal{D}^{[t+1]} \leftarrow$ hierarchical restrictive multinomial distribution $\mathcal{R}(\mathcal{U}_0, p, \tilde{\boldsymbol{\eta}}^{[t]})$, where $\tilde{\eta}_j^{[t]} \propto [\hat{\eta}_j^{[t]} \mathbb{1}(\hat{\eta}_j^{[t]} > C_0/\log p) + \frac{C_0}{p} \mathbb{1}(\hat{\eta}_j^{[t]} \leq C_0/\log p)]$ and $\sum_{j=1}^{p} \tilde{\eta}_j^{[t]} = 1$
end
Output the selected proportion of each feature $\hat{\boldsymbol{\eta}}^{[T]}$
Output $\hat{S} = \{1 \leq j \leq p : \hat{\eta}_j^{[T]}$ is among the N largest of all$\}$

3.Theoretical Analysis

3.1 Sure Screening Property

Vanilla RaSE

Iterative RaSE

3.2 Rank Consistency

Vanilla RaSE

- **RaSE的效果依赖于对Cr (the criterion), B_1 (子空间的组数), B_2 (每组子空间内的大小), D(子空间大小的最大值)**
- **每次迭代的子空间分布可以通过，从被选中的B_1 子空间的大小的经验分布来决定D**

4.Numerical Studies

- **实验设定：**重复200次
- **评价指标：**
  - MMS：5%, 25%, 50%, 75%, 95% quantiles of the minimum model size (MMS) to include all signals
  - Predication performance: 样本外测试集的预测效力
- **比较方法：**
  - SIS, ISIS, DC-SIS,
  - MDC-SIS: Martingale Difference Correlation
  - MV-SIS: screening approach for discriminant analysis
  - HOLP: high-dimensional ordinary least-square projection
  - IPDC: interaction pursuit via distance correlation
  - CIS: covariate information number
- **实验环境：**R
- **实验预期：**
  - 变量筛选
  - 模型拟合：
    - 线性模型-最小化BIC / extened BIC (eBIC)
    - 非线性模型-最小化交叉验证误差
      - LOOCV MSE in KNN (k=5)
      - 5-fold CV MSE SVM with RBF kernel径向基函数核
- **参数设置：**
  - $\mathrm{B}_1 = 200$, $\mathrm{B}_1 = 20\times[\sqrt{n}]$, $\mathrm{D} = [\sqrt{n}]$

$$\mathrm{BIC}(s) = -2\log L_n\{\hat{\theta}(s)\} + \nu(s)\log n,$$

$$\mathrm{BIC}_\gamma(s) = -2\log L_n\{\hat{\theta}(s)\} + \nu(s)\log n + \boxed{2\gamma\log\tau(\mathcal{S}_j),} \quad 0 \le \gamma \le 1,$$

Size of partitioned subsets

高维数据中，由于维数较大导致计算量大、估计不稳定，因此，2008年Chen和Chen在考虑了高维数据中未知参数的个数和模型空间的复杂性在BIC基础上提出了EBIC（extended bayesian information criterion），在一定程度上控制了变量选择中的假阳率

- All the codes used in numerical experiments can be found on GitHub (https://github.com/ytstat/RaSE-screening-codes).

4.Numerical Studies

- **Example 1(from Fan and Lv 2008)**

$$y = 5x_1 + 5x_5 + 5x_3 - \frac{15}{\sqrt{2}}x_4 + \epsilon,$$

where $\boldsymbol{x} = (x_1, \ldots, x_p)^T \sim N(\mathbf{0}, \Sigma)$, $\Sigma = (\sigma_{ij})_{p\times p}$, $\sigma_{ij} = 0.5^{\mathbb{1}(i\neq j)}$, $\epsilon \sim N(0,1)$, and $\epsilon \perp\!\!\!\perp \boldsymbol{x}$. The signal set $S^* = \{1,2,3,4\}$. $n = 100$ and $p = 1000$.

模型构造

- X的协方差非对角线元素为0.5
- x_4 与余下$p-1$个变量有关系，y和x_4没有关系
- 将y投影至与 x_1 ，x_2 ，x_3中任一垂直的空间后，与x_4就有关系

结果解读

1、 其他基于度量边际重要性的方法无法识别出 x_4 ，如SIS

2、 RaSE-BIC, RaSE-eBIC 表现比SIS好

3、加入迭代后，$RaSE_1$-BIC, $RaSE_1$-eBIC改善更多; 但是对large quantiles MMS有恶化风险，如- $RaSE_1$-KNN

4、**最佳**- $RaSE_1$-eBIC

Table 1. Quantiles of MMS in Examples 1 and 2.

Method/MMS	Example 1				
	5%	25%	50%	75%	95%
SIS	227	317	397	647	922
ISIS	14	15	15	15	25
SIRS	87	370	594	762	949
DC-SIS	96	358	610	776	942
HOLP	912	949	969	986	999
IPDC	224	442	700	869	980
MDC-SIS	146	287	512	734	937
CIS	203	434	601	780	940
RaSE-BIC	5	12	37	126	650
$RaSE_1$-BIC	4	4	4	16	55
RaSE-eBIC	6	21	42	489	852
$RaSE_1$-eBIC	4	4	4	4	14
RaSE-*k*NN	22	88	233	312	883
$RaSE_1$-*k*NN	6	80	422	694	921
RaSE-SVM	13	59	150	336	842
$RaSE_1$-SVM	4	4	82	126	542

4.Numerical Studies

- **Example 1(from Fan and Lv 2008)**

$$y = 5x_1 + 5x_5 + 5x_3 - \frac{15}{\sqrt{2}}x_4 + \epsilon,$$

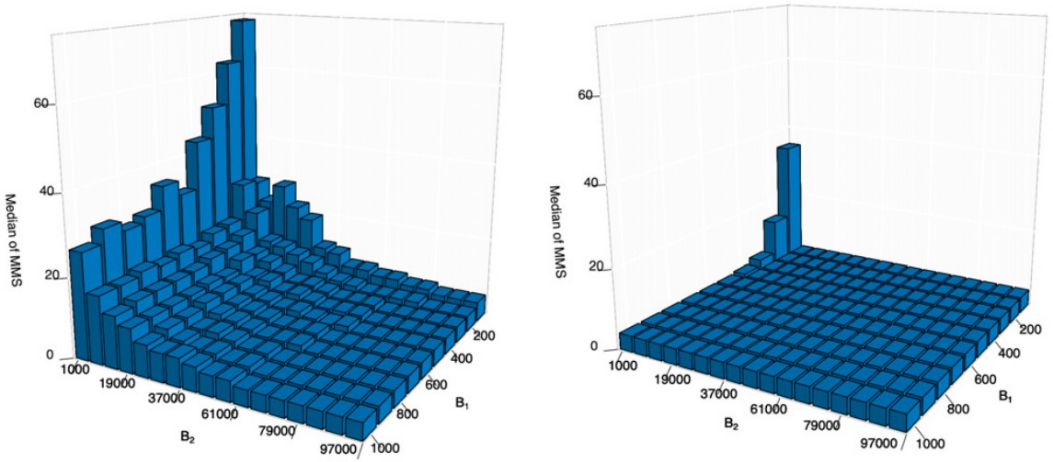
where $\boldsymbol{x} = (x_1, \ldots, x_p)^T \sim N(\mathbf{0}, \Sigma)$, $\Sigma = (\sigma_{ij})_{p\times p}$, $\sigma_{ij} = 0.5^{\mathbb{1}(i\neq j)}$, $\epsilon \sim N(0,1)$, and $\epsilon \perp\!\!\!\perp \boldsymbol{x}$. The signal set $S^* = \{1,2,3,4\}$. $n = 100$ and $p = 1000$.

为了解不同$B_1 B_2$,对结果的影响

- B_1：100 to 1000, 步长100
- B_2：1000 to 97，000, 步长6000

结果解读

- (B_1, B_2) 越大，效果越好
- B_2 较大时， B_1的效果稳定
- B_2增长时，效果连续变好--》足够的计算资源可以保证RaSE算法的效果
- 即便在B_2 较小时， $RaSE_1$-eBIC 表现仍然出色



(a) RaSE-BIC (b) $RaSE_1$-BIC

e 1. Median MMS to capture S^* ($|S^*| = 4$) as (B_1, B_2) varies for RaSE-BIC (a) and $RaSE_1$-BIC (b) in Example 1.

4.Numerical Studies

- **Example 1(from Fan and Lv 2008)**

$$y = 5x_1 + 5x_5 + 5x_3 - \frac{15}{\sqrt{2}}x_4 + \epsilon,$$

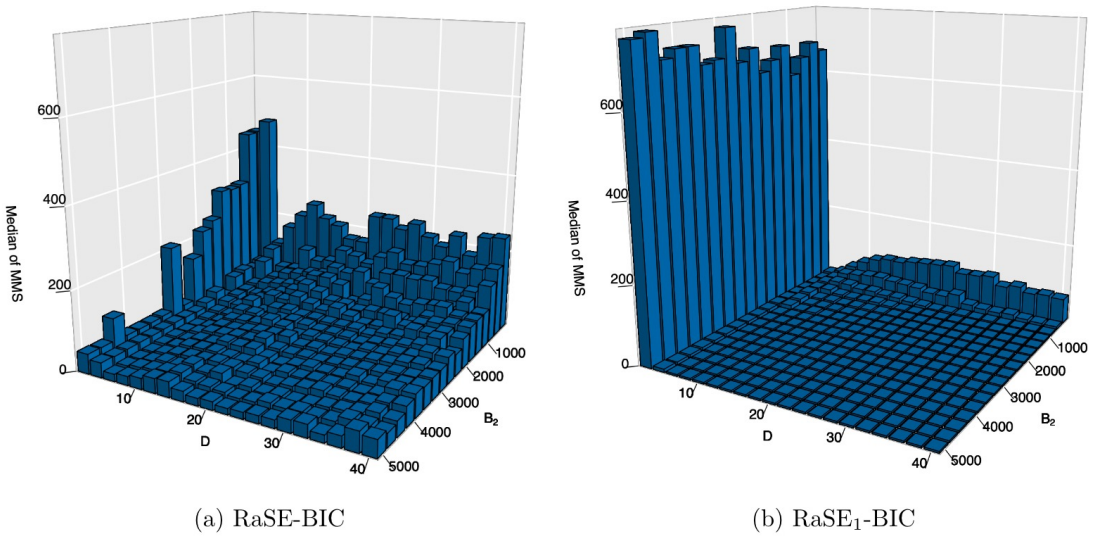
where $\boldsymbol{x} = (x_1, \ldots, x_p)^T \sim N(\mathbf{0}, \Sigma)$, $\Sigma = (\sigma_{ij})_{p\times p}$, $\sigma_{ij} = 0.5^{\mathbb{1}(i\neq j)}$, $\epsilon \sim N(0,1)$, and $\epsilon \perp\!\!\!\perp \boldsymbol{x}$. The signal set $S^* = \{1,2,3,4\}$. $n = 100$ and $p = 1000$.

为了解固定 B_1 不同D,B_2,对结果的影响

- B_1：200
- B_2：200 to 5000, 步长300
- D：2 to 40, 步长2

结果解读

- $D = \sqrt{n}$ 左右时，RaSE-BIC 表现稳定
- D， B_2 只要不太小， ， RaSE_1-BIC 非常稳健



(a) RaSE-BIC (b) RaSE_1-BIC

Figure 3: Median MMS to capture S^* ($|S^*| = 4$) as (D, B_2) varies for RaSE-BIC (a) and RaSE_1-BIC (b) in Example 1.

4.Numerical Studies

- **Example 1(from Fan and Lv 2008)**

$$y = 5x_1 + 5x_5 + 5x_3 - \frac{15}{\sqrt{2}}x_4 + \epsilon,$$

where $\boldsymbol{x} = (x_1, \ldots, x_p)^T \sim N(\mathbf{0}, \Sigma)$, $\Sigma = (\sigma_{ij})_{p\times p}$, $\sigma_{ij} = 0.5^{\mathbb{1}(i\neq j)}$, $\epsilon \sim N(0,1)$, and $\epsilon \perp\!\!\!\perp \boldsymbol{x}$. The signal set $S^* = \{1,2,3,4\}$. $n = 100$ and $p = 1000$.

计算时间比较

- B_1：200
- B_2：200 to 5000, 步长300
- D：2 to 40, 步长2

结果解读

- R$\boldsymbol{aSE}$ 由于要生成大量的子空间，消耗大量计算资源
- 可以通过并行计算和好的设备弥补缺点

Table 2 *Average (over 200 replications) computational time in seconds for various methods in Example 1. (Table view)*

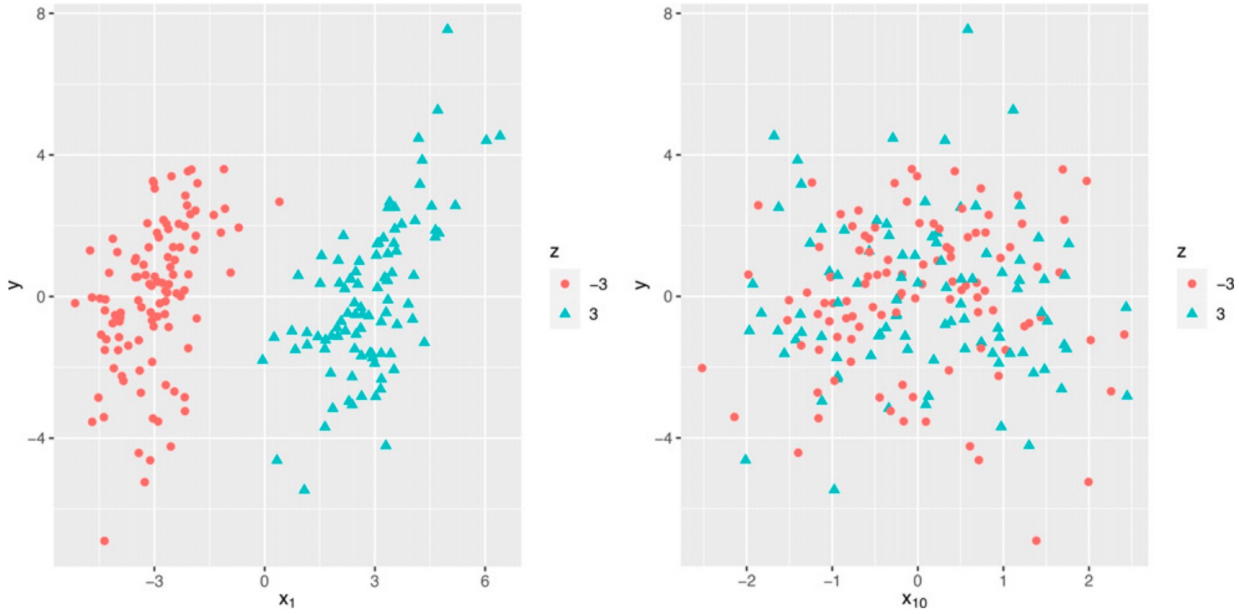
Other methods	*SIS*	*ISIS*	*SIRS*	*DC-SIS*	*HOLP*	*IPDC*	*MDC-SIS*	*CIS*
Time (s)	*0.01*	*0.67*	*0.28*	*1.30*	*0.02*	*0.49*	*0.28*	*1.18*
RaSE methods	*BIC*	*BIC_1*	*eBIC*	*$eBIC_1$*	*kNN*	*kNN_1*	*SVM*	*SVM_1*
Time (s)	*1.99*	*4.03*	*2.01*	*3.94*	*6.74*	*13.66*	*150.41*	*305.77*

NOTE: For simplicity, for RaSE methods, we use criteria to differentiate them and the subscript "1" denotes the one-step iterative version of the corresponding RaSE-based methods.

4.Numerical Studies

- **Example 2(Latent Cluster)** $y = 0.5(\tilde{x}_1 + \tilde{x}_2 + \tilde{x}_3 + \tilde{x}_4 + \tilde{x}_5 + \epsilon),$

where
$\tilde{\boldsymbol{x}} = (\tilde{x}_1, \ldots, \tilde{x}_p)^T \sim N(\mathbf{0}, \Sigma), \epsilon \sim t_2, \Sigma = (\sigma_{ij})_{p\times p} = (0.5^{|i-j|})_{p\times p}$, *and* $\epsilon \perp\!\!\!\perp \boldsymbol{x}$. *Generate* $z \sim \mathrm{Unif}(\{-3, 3\}) \perp\!\!\!\perp \tilde{\boldsymbol{x}}$ *and* $\boldsymbol{x} = \tilde{\boldsymbol{x}} + z\mathbf{1}_p$. *The signal set* $S^* = \{1, 2, 3, 4, 5\}$. $n = 200$ *and* $p = 2000$.



模型构造

- 不利于基于Pearson correlation **的方法**
- 如左图 y和$x_1$plot， 右图y和x_{10} plot

结果解读

- **整体，**KNN效果好
- 较小quantiles**时**（ 5%, 25% quantiles ）SIS, MDC-SIS表现好

4.Numerical Studies

- **Example 2(Latent Cluster)** $$y = 0.5(\tilde{x}_1 + \tilde{x}_2 + \tilde{x}_3 + \tilde{x}_4 + \tilde{x}_5 + \epsilon),$$

where
$\tilde{\boldsymbol{x}} = (\tilde{x}_1, \ldots, \tilde{x}_p)^T \sim N(\mathbf{0}, \Sigma), \epsilon \sim t_2, \Sigma = (\sigma_{ij})_{p\times p} = (0.5^{|i-j|})_{p\times p}$, *and* $\epsilon \perp\!\!\!\perp \boldsymbol{x}$. *Generate* $z \sim \mathrm{Unif}(\{-3, 3\}) \perp\!\!\!\perp \tilde{\boldsymbol{x}}$ *and* $\boldsymbol{x} = \tilde{\boldsymbol{x}} + z\mathbf{1}_p$. *The signal set* $S^* = \{1, 2, 3, 4, 5\}$. $n = 200$ *and* $p = 2000$.

模型构造

- 不利于基于Pearson correlation **的方法**

结果解读

- **整体，**KNN效果好
- 较小quantiles时（5%, 25% quantiles）SIS, MDC-SIS表现好

Table 1. Quantile

Method/MMS	Example 2				
	5%	25%	50%	75%	95%
SIS	6	28	105	592	1855
ISIS	172	861	1415	1825	1963
SIRS	6	1158	1492	1774	1964
DC-SIS	6	1083	1460	1752	1976
HOLP	45	196	576	1252	1906
IPDC	59	210	386	678	1517
MDC-SIS	6	20	93	999	1908
CIS	2000	2000	2000	2000	2000
RaSE-BIC	6	358	1514	1821	1956
RaSE_1-BIC	13	834	1507	1797	1969
RaSE-eBIC	8	26	1323	1789	1935
RaSE_1-eBIC	907	1485	1739	1878	1971
RaSE-*k*NN	5	5	6	76	1190
RaSE_1-*k*NN	5	5	5	13	1846
RaSE-SVM	5	5	5	6	68
RaSE_1-SVM	5	5	5	5	11

4.Numerical Studies

- **Example 3 (Li, Zhong, and Zhu 2012)** $y = 2\beta_1 x_1 x_2 + 3\beta_2 \mathbb{1}(x_{12} < 0) x_{22} + \epsilon,$

DC-SIS

where $\beta_j = (-1)^U(4\log n/\sqrt{n} + |Z|), j = 1, 2$, $U \sim$ Bernoulli(0.4), $Z \sim N(0, 1)$, $\epsilon \sim N(0, 1)$, $\boldsymbol{x} \sim N(\mathbf{0}, \Sigma)$ where $\Sigma = (\sigma_{ij})_{p\times p} = (0.8^{|i-j|})_{p\times p}$, $U \perp\!\!\!\perp Z$, $\epsilon \perp\!\!\!\perp \boldsymbol{x}$, and $(U, Z) \perp\!\!\!\perp (\epsilon, \boldsymbol{x})$. Note that we regenerate (U, Z) for each replication, so the results might differ from those in Li, Zhong, and Zhu (2012). The signal set $S^* = \{1, 2, 12, 22\}$. $n = 200$ and $p = 2000$.

模型构造

- **包含交互项x_1x_2**

结果解读

- 由于交互项和示性函数的存在，基于线性模型的方法，如SIS, ISIS, HOLP, RaSE with BIC, RaSE with eBIC 表现不佳
- CIS和RaSE$_1$-KNN 在5%，25%，50% quantiles 表现很好
- RaSE-KNN ， RaSE-SVM在小quantiles下表现好
- 最好- RaSE$_1$-SVM （except for 95% quantiles）

Table 3. Quantiles of MMS in Examples 3 and 4.

Method/MMS	Example 3				
	5%	25%	50%	75%	95%
SIS	184	810	1370	1732	1957
ISIS	362	1008	1482	1775	1945
SIRS	54	741	1294	1634	1920
DC-SIS	25	456	1222	1638	1923
HOLP	326	954	1475	1774	1975
IPDC	128	429	920	1397	1899
MDC-SIS	52	165	504	1331	1872
CIS	4	5	8	55	548
RaSE-BIC	637	1242	1619	1842	1959
RaSE$_1$-BIC	714	1196	1550	1839	1974
RaSE-eBIC	484	1137	1496	1794	1951
RaSE$_1$-eBIC	725	1330	1617	1806	1948
RaSE-*k*NN	5	33	168	1321	1855
RaSE$_1$-*k*NN	4	5	8	125	1528
RaSE-SVM	4	18	504	1282	1848
RaSE$_1$-SVM	4	4	5	14	1141

4.Numerical Studies

- **Example 4 (Interactions)**

$$y = 3\sqrt{|x_1|} + 2\sqrt{|x_1|}x_2^2 + 4\sin(x_1)\sin(x_2)\sin^2(x_3) + 12\sin(x_1)|x_2|\sin(x_3)x_4^2 + 0.5\epsilon,$$

where $x_1, \ldots, x_p \overset{\text{iid}}{\sim} N(0,1)$, $\epsilon \sim N(0,1)$, and $\epsilon \perp\!\!\!\perp \boldsymbol{x}$. The signal set $S^* = \{1,2,3,4\}$. $n = 300$ and $p = 1000$.

模型构造

- **甄别高阶交互项**

结果解读

- Lower quantiles: RaSE-BIC, Rase-KNN, Rase-SVM, RaSE_1-SVM , RaSE_1-KNN ， IPDC, CIS表现好
- RaSE framework可以有效识别高阶交互项

Table 3. Quar

Method/MMS	Example 4				
	5%	25%	50%	75%	95%
SIS	264	570	709	885	984
ISIS	293	626	810	911	978
SIRS	487	737	867	935	992
DC-SIS	44	304	603	814	949
HOLP	316	586	767	886	974
IPDC	7	19	68	158	528
MDC-SIS	189	482	736	889	979
CIS	5	33	136	352	789
RaSE-BIC	355	693	825	914	986
RaSE_1-BIC	424	661	824	918	981
RaSE-eBIC	302	553	784	913	987
RaSE_1-eBIC	480	686	860	930	986
RaSE-*k*NN	5	15	68	290	889
RaSE_1-*k*NN	4	8	51	446	910
RaSE-SVM	4	15	132	468	938
RaSE_1-SVM	4	30	232	645	898

4.Numerical Studies

- **Example 5 (Gaussian mixture, Cannings and Samworth 2017)**

Model 1: Here, $X|\{Y=0\} \sim \frac{1}{2}N_p(\mu_0, \Sigma)+\frac{1}{2}N_p(-\mu_0, \Sigma)$, and $X|\{Y=1\} \sim \frac{1}{2}N_p(\mu_1, \Sigma)+\frac{1}{2}N_p(-\mu_1, \Sigma)$, where, for $p = 100$, we set $\Sigma = I_{100\times 100}$, $\mu_0 = (2, -2, 0, \ldots, 0)^T$ and $\mu_1 = (2, 2, 0, \ldots, 0)^T$.

$$y \sim \text{Bernoulli}(0.5),\ \boldsymbol{x}|y = r \sim \frac{1}{2}N(\boldsymbol{\mu}_r, \Sigma) + \frac{1}{2}N(-\boldsymbol{\mu}_r, \Sigma), r = 0, 1,$$

where $\boldsymbol{\mu}_0 = (2, -2, 0, \ldots, 0)^T$, $\boldsymbol{\mu}_1 = (2, 2, 0, \ldots, 0)^T$, Σ is an identity matrix. The signal set $S^* = \{1, 2\}$. $n = 200$ and $p = 2000$.

模型构造

- 对基于边际相关性的方法不利；
- 唯一可以识别信号的方式是测量(x_1, x_2)的联合分布

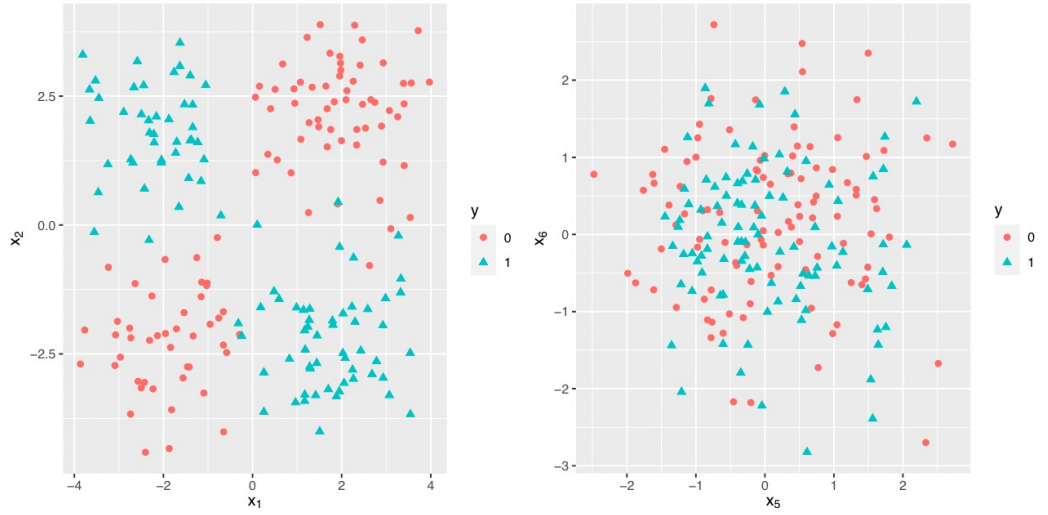


Figure 4: Scatterplots of x_2 vs. x_1 and x_6 vs. x_5 for Example 5 ($n = 200$).

4.Numerical Studies

- **Example 5 (Gaussian mixture, Cannings and Samworth 2017)**

$$y \sim \text{Bernoulli}(0.5),\ \boldsymbol{x}|y=r \sim \frac{1}{2}N(\boldsymbol{\mu}_r, \Sigma) + \frac{1}{2}N(-\boldsymbol{\mu}_r, \Sigma), r=0,1,$$

where $\boldsymbol{\mu}_0 = (2,-2,0,\ldots,0)^T$, $\boldsymbol{\mu}_1 = (2,2,0,\ldots,0)^T$, Σ is an identity matrix. The signal set $S^* = \{1,2\}$. $n = 200$ and $p = 2000$.

模型构造

- 对基于边际相关性的方法不利；
- 唯一可以识别信号的方式是测量(x_1, x_2)的联合分布

结果解读

- SIRS, , RaSE_1-SVM , RaSE_1-KNN 效果很好

Table 4. Quantiles of MMS in Examples 5 and 6.

Method/MMS	Example 5				
	5%	25%	50%	75%	95%
SIS	515	1090	1414	1746	1947
ISIS	445	1001	1470	1784	1967
SIRS	2	2	2	2	2
DC-SIS	451	960	1385	1706	1913
MV-SIS	379	957	1366	1692	1895
HOLP	495	1065	1381	1712	1936
IPDC	495	1010	1344	1673	1908
MDC-SIS	462	1038	1332	1708	1948
CIS	2000	2000	2000	2000	2000
RaSE-BIC	506	1081	1487	1804	1946
RaSE_1-BIC	464	968	1360	1692	1927
RaSE-eBIC	425	1045	1424	1705	1965
RaSE_1-eBIC	480	988	1370	1727	1938
RaSE-*k*NN	2	3	5	6	8
RaSE_1-*k*NN	2	2	2	2	2
RaSE-SVM	2	4	6	8	26
RaSE_1-SVM	2	2	2	2	2

4.Numerical Studies

- **Example 6 (Multinomial logistic regression, Fan, Samworth, and Wu 2009)**

We first generate $\tilde{x}_1, \ldots, \tilde{x}_4 \overset{\text{iid}}{\sim} \text{Unif}\left([-\sqrt{3}, \sqrt{3}]\right)$ *and* $\tilde{x}_5, \ldots, \tilde{x}_p \overset{\text{iid}}{\sim} N(0,1)$, *then let* $x_1 = \tilde{x}_1 - \sqrt{2}\tilde{x}_5, x_2 = \tilde{x}_2 + \sqrt{2}\tilde{x}_5, x_3 = \tilde{x}_3 - \sqrt{2}\tilde{x}_5, x_4 = \tilde{x}_4 + \sqrt{2}\tilde{x}_5$ *and* $x_j = \tilde{x}_j$ *for* $j = 5, \ldots, p$. *The response is generated from*

$$\mathrm{P}(y = r|\tilde{\boldsymbol{x}}) \propto \exp\{f_r(\tilde{\boldsymbol{x}})\}, r = 1, \ldots, 4,$$

where $f_1(\tilde{\boldsymbol{x}}) = -a\tilde{x}_1 + a\tilde{x}_4, f_2(\tilde{\boldsymbol{x}}) = a\tilde{x}_1 - a\tilde{x}_2, f_3(\tilde{\boldsymbol{x}}) = a\tilde{x}_2 - a\tilde{x}_3$ *and* $f_4(\tilde{\boldsymbol{x}}) = a\tilde{x}_3 - a\tilde{x}_4$ *with* $a = 5/\sqrt{3}$. *The signal set* $S^* = \{1, 2, 3, 4, 5\}$. $n = 200$ *and* $p = 2000$.

模型构造

- x_5 与y边际独立
- x_1 ，x_2, x_3, x_4 都与 x_5 有关
- 基于边际扫描的方法理应捕捉不到 x_5

4.Numerical Studies

- **Example 6 (Multinomial logistic regression, Fan, S**

We first generate $\tilde{x}_1, \ldots, \tilde{x}_4 \overset{\text{iid}}{\sim} \text{Unif}\left([-\sqrt{3}, \sqrt{3}]\right)$ *and* $\tilde{x}_5, \ldots, \tilde{x}$ $x_1 = \tilde{x}_1 - \sqrt{2}\tilde{x}_5, x_2 = \tilde{x}_2 + \sqrt{2}\tilde{x}_5, x_3 = \tilde{x}_3 - \sqrt{2}\tilde{x}_5, x_4 = \tilde{x}_4 + \sqrt{2}\tilde{x}_5$ $j = 5, \ldots, p$. *The response is generated from*

$$\mathrm{P}(y = r|\tilde{\boldsymbol{x}}) \propto \exp\{f_r(\tilde{\boldsymbol{x}})\}, r = 1, \ldots, 4,$$

where $f_1(\tilde{\boldsymbol{x}}) = -a\tilde{x}_1 + a\tilde{x}_4, f_2(\tilde{\boldsymbol{x}}) = a\tilde{x}_1 - a\tilde{x}_2, f_3(\tilde{\boldsymbol{x}}) = a\tilde{x}_2 - a\tilde{x}_3$ *with* $a = 5/\sqrt{3}$. *The signal set* $S^* = \{1, 2, 3, 4, 5\}$. $n = 200$ *and* $p = 2000$.

模型构造

- x_5 与y边际独立
- 基于边际扫描的方法理应捕捉不到 x_5

结果解读

- ISIS, , RaSE$_1$-BIC , RaSE$_1$-eBIC 效果很好
- 无迭代时,RaSE-BIC表现占优
- 与EX1相似的是，增加迭代可能导致large quantiles 的表现不好

Table 4. Quant

Method/MMS	Example 6				
	5%	25%	50%	75%	95%
SIS	170	471	910	1436	1932
ISIS	7	7	7	8	8
SIRS	821	1242	1551	1813	1966
DC-SIS	765	1155	1526	1775	1947
MV-SIS	199	706	1258	1660	1909
HOLP	—	—	—	—	—
IPDC	879	1425	1722	1884	1988
MDC-SIS	163	498	1064	1628	1917
CIS	229	736	1195	1652	1941
RaSE-BIC	8	14	20	26	1525
RaSE$_1$-BIC	5	5	5	6	14
RaSE-eBIC	26	346	894	1406	1919
RaSE$_1$-eBIC	5	7	10	14	1184
RaSE-*k*NN	38	202	294	1470	1925
RaSE$_1$-*k*NN	27	376	967	1486	1828
RaSE-SVM	11	39	118	343	1743
RaSE$_1$-SVM	5	5	118	1133	1792

5.Real Data Experiments

实验设计

- 变量筛选：以下为确定筛选个数的3种方法
  - $\mathrm{N} = [n/\log n]$ (Fan and Lv, 2008)
  - $\mathrm{N} = [\alpha D]$ for any α>1
  - Data Driven Strategies: use validation set and post-screening validation MSE/ classification error to determine N
- 模型拟合：
  - LASSO, kNN, SVM
  - Train: test = 9:1
  - 200 replications
  - standardization
- Benchmarks: 全变量进行拟合LASSO, kNN, SVM

5.Real Data Experiments

- 【Colon Cancer Dataset】

数据描述

- 2000 genes measured on 62 patients
- Class 1-colon cancer-40
- Class 2- healthy-22
- Y是连续性

结果解读

- 最佳-RaSE-BIC with LASSO
- 对于LASSO迭代后，效果都有所提升

Table 5. Average test classification error rate with standard deviations (in parentheses) for colon cancer dataset and average test mean square errors (MSEs) with standard deviations (in parentheses) for rat eye expression dataset.

Screening	Post-screening	Cancer	Eye
—		0.1792(0.1427)	0.0103(0.0091)
SIS		0.1633(0.1407)	*0.0091(0.0068)*
ISIS		0.1767(0.1444)	*0.0091(0.0068)*
SIRS		0.2800(0.1734)	0.0132(0.0123)
DC-SIS		0.3000(0.1998)	0.0124(0.0118)
MV-SIS		0.2958(0.1826)	—
HOLP	LASSO	0.1825(0.1491)	0.0228(0.0269)
IPDC		0.1917(0.1464)	0.0129(0.0132)
MDC-SIS		0.1600(0.1406)	0.0103(0.0071)
CIS		0.1550(0.1332)	0.0194(0.0231)
RaSE-BIC		**0.1192(0.1277)**	**0.0090(0.0066)**
$RaSE_1$-BIC		*0.1417(0.1324)*	0.0123(0.0104)
RaSE-eBIC		0.3083(0.2118)	0.0092(0.0069)
$RaSE_1$-eBIC		0.1458(0.1397)	0.0122(0.0098)
—		0.2258(0.1653)	0.0166(0.0206)
RaSE-*k*NN	*k*NN	0.1533(0.1340)	0.0131(0.0158)
$RaSE_1$-*k*NN		0.1867(0.1500)	0.0133(0.0161)
—		0.2025(0.1503)	0.0160(0.0243)
RaSE-SVM	SVM	*0.1375(0.1277)*	0.0158(0.0231)
$RaSE_1$-SVM		0.1858(0.1477)	0.0158(0.0232)

NOTE: We boldface the values corresponding to the best performances and italicize the values corresponding to the subsequent two best performances.

5.Real Data Experiments

- 【Colon Cancer Dataset】

比较筛选结果

- 选择TOP 10 选中的变量，计算选中概率
- 注意到在不同方法下，前几个变量的选中概率较高（100% or >50%）,说明结果的可靠与稳定
  - Gene1423
  - Gene377
  - gene1772

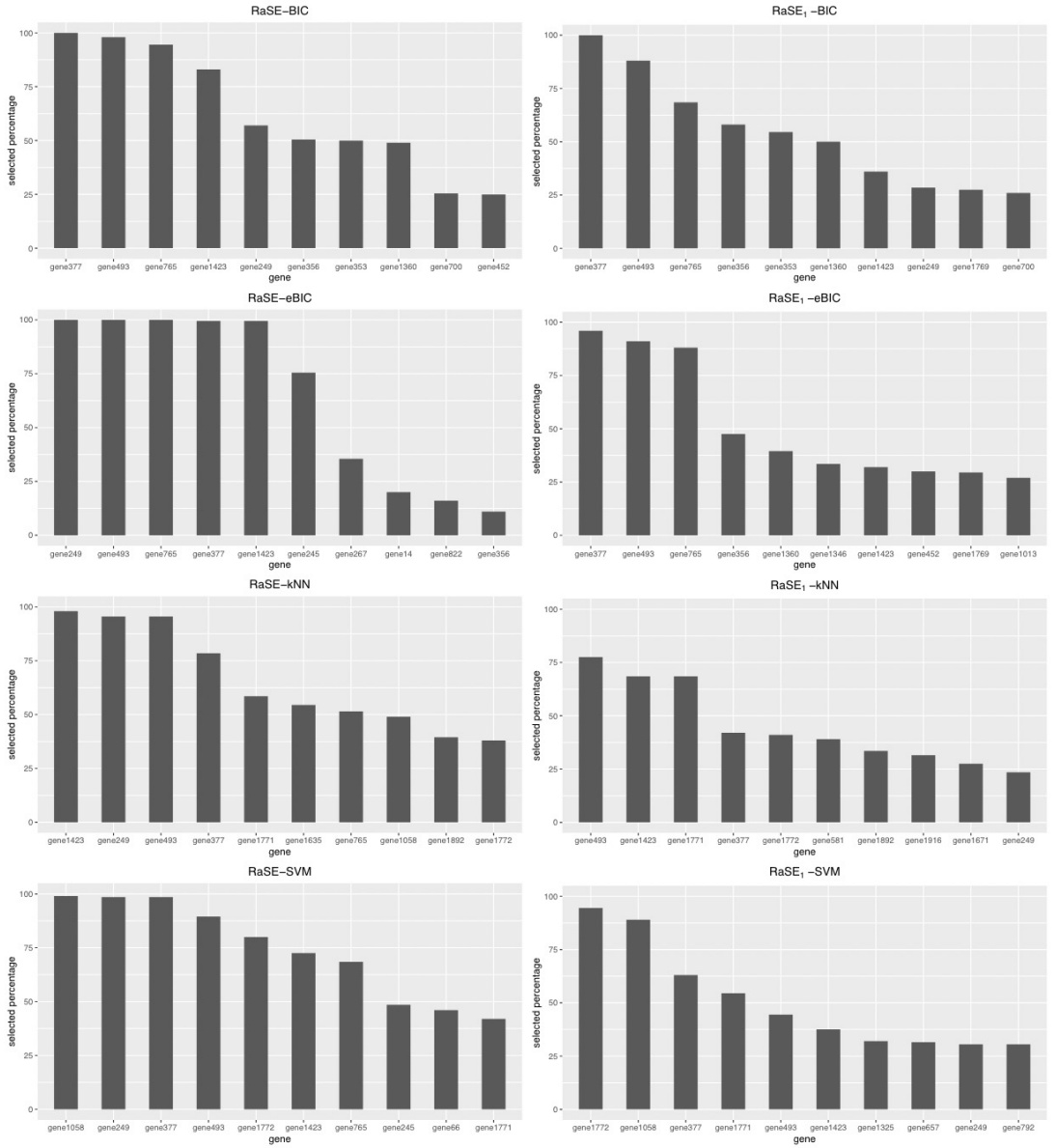


Figure 5: Features with the 10 highest selection rates (percentages in 200 replications) in the colon cancer data set.

5.Real Data Experiments

- 【Rat Eye Expression Dataset】

数据描述

- Gene expression values of 18,796 probes from 120 rats
- TRIM32 is the response, which is responsible to cause Bardet-Biedl syndrome
- 本次使用样本方差最大的前5000个基因(n=120, p=5000)

结果解读

- SIS, ISIS, RaSE-BIC, RaSE-eBIC with LASSO 的效果相当(优于vanilla LASSO)
- RaSE-KNN with KNN, RaSE_1-KNN with KNN 优于 vanilla KNN

Table 5. Average test classification error rate with standard deviations (in parentheses) for colon cancer dataset and average test mean square errors (MSEs) with standard deviations (in parentheses) for rat eye expression dataset.

Screening	Post-screening	Cancer	Eye
—		0.1792(0.1427)	0.0103(0.0091)
SIS		0.1633(0.1407)	*0.0091(0.0068)*
ISIS		0.1767(0.1444)	*0.0091(0.0068)*
SIRS		0.2800(0.1734)	0.0132(0.0123)
DC-SIS		0.3000(0.1998)	0.0124(0.0118)
MV-SIS		0.2958(0.1826)	—
HOLP	LASSO	0.1825(0.1491)	0.0228(0.0269)
IPDC		0.1917(0.1464)	0.0129(0.0132)
MDC-SIS		0.1600(0.1406)	0.0103(0.0071)
CIS		0.1550(0.1332)	0.0194(0.0231)
RaSE-BIC		**0.1192(0.1277)**	**0.0090(0.0066)**
RaSE_1-BIC		*0.1417(0.1324)*	0.0123(0.0104)
RaSE-eBIC		0.3083(0.2118)	0.0092(0.0069)
RaSE_1-eBIC		0.1458(0.1397)	0.0122(0.0098)
—		0.2258(0.1653)	0.0166(0.0206)
RaSE-*k*NN	*k*NN	0.1533(0.1340)	0.0131(0.0158)
RaSE_1-*k*NN		0.1867(0.1500)	0.0133(0.0161)
—		0.2025(0.1503)	0.0160(0.0243)
RaSE-SVM	SVM	*0.1375(0.1277)*	0.0158(0.0231)
RaSE_1-SVM		0.1858(0.1477)	0.0158(0.0232)

NOTE: We boldface the values corresponding to the best performances and italicize the values corresponding to the subsequent two best performances.

5.Real Data Experiments

- 【Rat Eye Expression Dataset】

比较筛选结果

- 选择TOP 10 选中的变量，计算选中概率
- 稳定筛选的几个变量
  - 1376747_at
  - 1390539_at
  - 1377791_at
  - 1383110_at

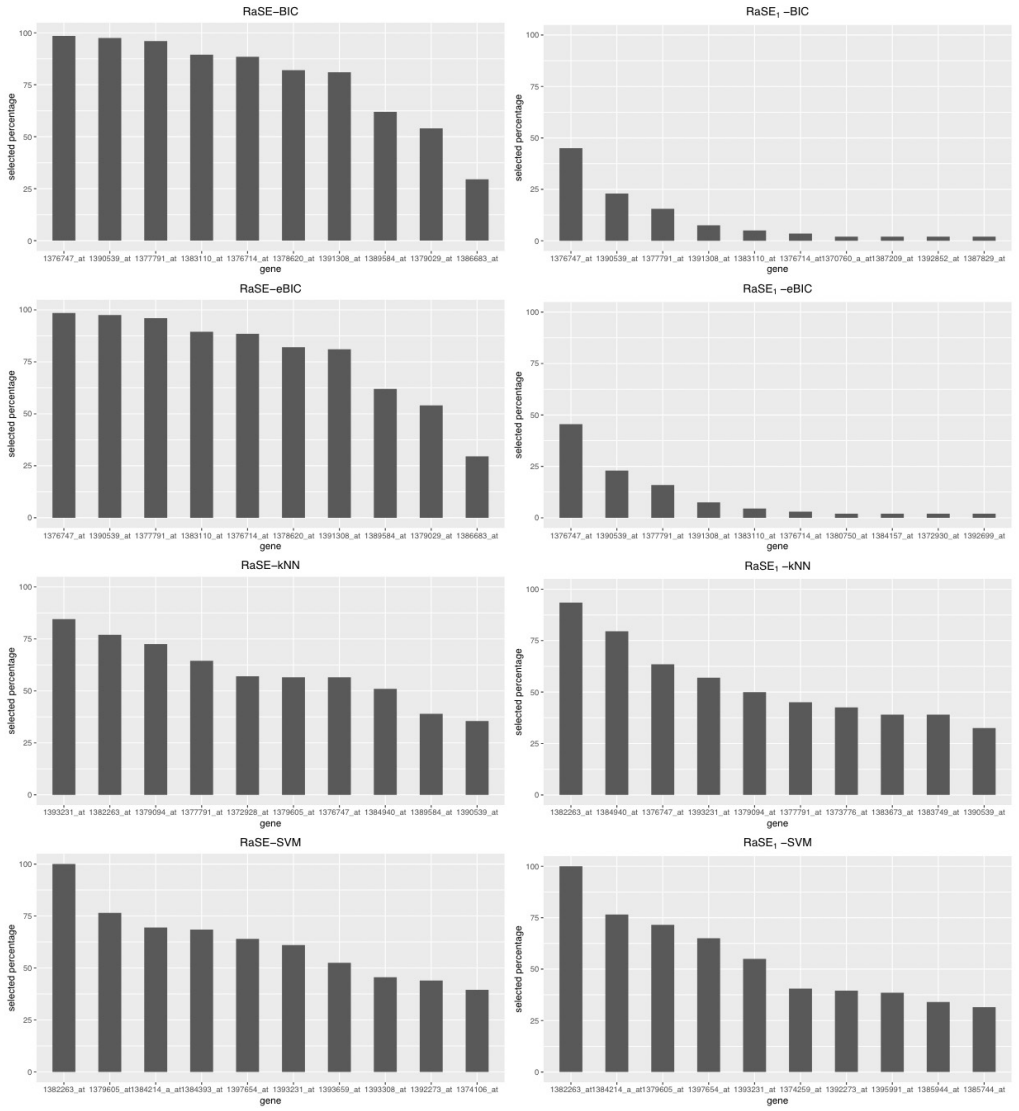


Figure 6: Features with the 10 highest selection rates (percentages in 200 replications) in the rat eye expression data set.

6.Discussion

SUMMARY

- 基于RaSE方法提出了一种变量扫描的框架，并结合BIC, eBIC等准则来比较子空间。注意到，与以往比较单一变量不同的是，我们比较的是子空间，因此可以捕捉到与相应变量没有边际关系的真实信号。

- 此外，迭代式的RaSE screening 方法可以一定程度改善效果，并放松对于B_2的假设条件。在理论上，我们对这两类框架都建立了sure screening property性质的证明。并证实vanilla RaSE具有筛选一致性。在信号较弱的情况下，我们需要增大B_2来保证效果。

- 数值模拟和实证都验证了方法的可靠性。

METHOD

- **RaSE的效果依赖于对Cr (the criterion), B_1 (子空间的组数), B_2 (每组子空间内的大小), D(每个子空间大小的最大值)**
- **每次迭代的子空间分布可以通过，从被选中的B_1 子空间的大小的经验分布来决定D**

FUTURE WORK

- 探寻适应性的方法自主选择迭代次数T
- 在迭代中选择不同的 B_2 值，以加速算法

Reference

[1] Tian, Y., & Feng, Y. (2021). RaSE: Random Subspace Ensemble Classification. J. Mach. Learn. Res., 22, 45-1.
[2] Tian, Y., & Feng, Y. (2021). RaSE: A variable screening framework via random subspace ensembles. Journal of the American Statistical Association, (just-accepted), 1-30.
[3] Zhu, J., & Feng, Y. (2021). Super RaSE: Super Random Subspace Ensemble Classification. Manuscript.
[4] R Package RaSEn: https://cran.r-project.org/web/packages/RaSEn/index.html

感谢聆听！请大家批评指正!

THANK YOU FOR YOUR CRITICISM

presenter: Fei Yang 2022/05/19

Random Subspace Ensembles

Algorithm-RaSE-classification

Algorithm 1: Random subspace ensemble classification (RaSE)

Input: training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, new data $\boldsymbol{x}$, subspace distribution $\mathcal{D}$, criterion $\mathcal{C}$, integers B_1 and B_2, type of base classifier $\mathcal{T}$

Output: predicted label $C_n^{RaSE}(\boldsymbol{x})$, the selected proportion of each feature $\boldsymbol{\eta}$

Independently generate random subspaces $S_{jk} \sim \mathcal{D}, 1 \le j \le B_1, 1 \le k \le B_2$

for $j \leftarrow 1$ **to** B_1 **do**

  Select the optimal subspace S_{j*} from $\{S_{jk}\}_{k=1}^{B_2}$ according to $\mathcal{C}$ and $\mathcal{T}$

end

Construct the ensemble decision function $\nu_n(\boldsymbol{x}) = \frac{1}{B_1}\sum_{j=1}^{B_1} C_n^{S_{j*}-\mathcal{T}}(\boldsymbol{x})$

Set the threshold $\hat{\alpha}$ according to (2)

Output the predicted label $C_n^{RaSE}(\boldsymbol{x}) = \mathbb{1}(\nu_n(\boldsymbol{x}) > \hat{\alpha})$, the selected proportion of each feature $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^T$ where $\eta_l = B_1^{-1}\sum_{j=1}^{B_1} \mathbb{1}(l \in S_{j*}), l = 1, \ldots, p$

Random Subspace Ensembles

Algorithm-iterative RaSE

Algorithm 2: Iterative RaSE (RaSE_T)

Input: training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, new data $\boldsymbol{x}$, initial subspace distribution $\mathcal{D}^{(0)}$, criterion $\mathcal{C}$, integers B_1 and B_2, the type of base classifier $\mathcal{T}$, the number of iterations T

Output: predicted label $C_n^{RaSE}(\boldsymbol{x})$, the proportion of each feature $\boldsymbol{\eta}^{(T)}$

for $t \leftarrow 0$ **to** T **do**

  Independently generate random subspaces $S_{jk}^{(t)} \sim \mathcal{D}^{(t)}, 1 \le j \le B_1, 1 \le k \le B_2$

  for $j \leftarrow 1$ **to** B_1 **do**

    Select the optimal subspace $S_{j*}^{(t)}$ from $\{S_{jk}^{(t)}\}_{k=1}^{B_2}$ according to $\mathcal{C}$ and $\mathcal{T}$

  end

  Update $\boldsymbol{\eta}^{(t)}$ where $\eta_l^{(t)} = B_1^{-1} \sum_{j=1}^{B_1} \mathbb{1}(l \in S_{j*}^{(t)}), l = 1, \ldots, p$

  Update $\mathcal{D}^{(t)} \leftarrow$ restrictive multinomial distribution with parameter $(p, d, \tilde{\boldsymbol{\eta}}^{(t)})$, where $\tilde{\eta}_l^{(t)} = \eta_l^{(t)} \mathbb{1}(\eta_l^{(t)} > C_0/\log p) + \frac{C_0}{p} \mathbb{1}(\eta_l^{(t)} \le C_0/\log p)$ and d is sampled from the uniform distribution over $\{1, \ldots, D\}$

end

Set the threshold $\hat{\alpha}$ according to (2)

Construct the ensemble decision function $\nu_n(\boldsymbol{x}) = \frac{1}{B_1} \sum_{j=1}^{B_1} C_n^{S_{j*}^{(T)} - \mathcal{T}}(\boldsymbol{x})$

Output the predicted label $C_n^{RaSE}(\boldsymbol{x}) = \mathbb{1}(\nu_n(\boldsymbol{x}) > \hat{\alpha})$ and $\boldsymbol{\eta}^{(T)}$

```
#knn伪代码
import numpy as np

##给出训练数据以及对应的类别
def createDataSet():
    group = np.array([[1.3,1.1],[0,0],[1.0,2.0],[1.2,0.1],[3,1.4],[3.4,3.5],[3.2,2.2],[3.5,2.7],[4,2.4]])
    labels = ['A','A','A','A','B','B','B','B','B']
    return group,labels

#计算欧氏距离
def get_distance(X,Y):
 return np.sum((X-Y)**2)**0.5

def knn(x_test,x_train,y_train,k):
 distances = []
 y_kind={}
 #计算点到每个训练集样本的距离
 for i in x_train:
  distances.append(get_distance(x_test,i))
 tmp=list(enumerate(distances))
 #距离进行排序，取前k个距离最近的
 tmp.sort(key=lambda x:x[1])
 min_k_dis=tmp[:k]
 #前k个的y标签进行字典统计
 for j in min_k_dis:
  t_key = y_train[j[0]] #标签 j[0]是索引下标
  if t_key in y_kind.keys():
   y_kind[t_key] += 1
  else:
   y_kind.setdefault(t_key,1)
 #标签结果进行排序
 t=sorted(y_kind.items(),key=lambda x:x[1],reverse=True)
 #返回标签最多的一个
 return t[0][0]

x_train,y_train = createDataSet()
x_test = np.array([4,3.4])
n_neighbors = 3
output = knn(x_test,x_train,y_train,n_neighbors)
print("测试数据为:",x_test,"分类结果为：",output)
```