

# 几类基于重建的子空间聚类问题研究

姓 名：李宝华  
专 业：计算数学

指导教师：卢湖川 教授  
吴微 教授

大连理工大学

October 14, 2017

## 主要内容：

### ① 子空间聚类问题

- 概述
- 子空间聚类问题的分类
- 原理示例
- 以前相关工作

### ② 主要工作

- $k$ -support 范数正则化子空间聚类
- 多元混合高斯回归子空间聚类
- 乘性噪音子空间聚类

### ③ 工作总结

### ④ 展望

### ⑤ 博士学位期间论文情况

## 子空间聚类问题概述

聚类是指对于给定的取自不同类、不同簇的数据，通过算法自动的进行分类，使得同一类的数据之间的差异尽量小，不同类数据之间的差异尽量大。随着数据的维度不断升高，一方面出现维数灾难问题为传统方法带来了压力；另一方面收集到的高维数据可以建模为来自不同子空间的元素而构成，子空间聚类技术能很好对高维数据进行解释、建模，近年来得到广泛关注与发展。

## 子空间聚类的应用与分类

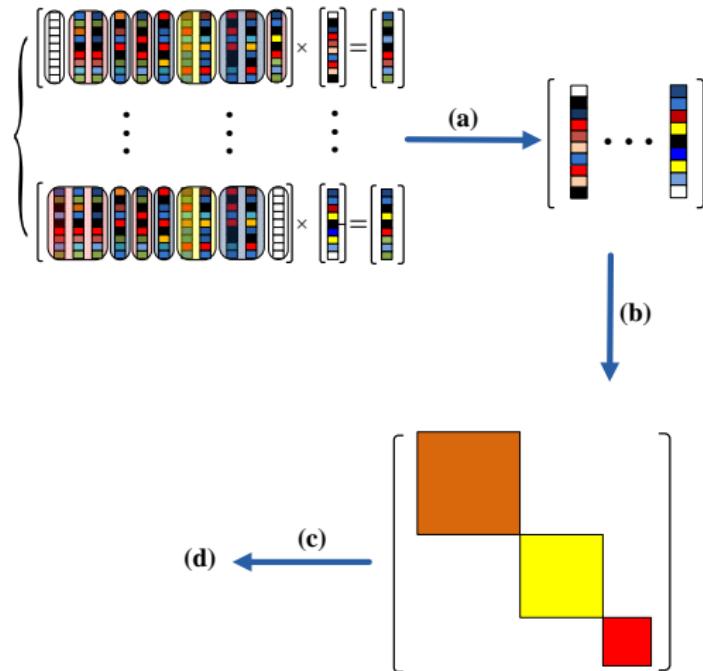
由于子空间聚类在计算机视觉、机器学习、图像处理、系统识别、医疗诊断、以及涉及到子空间聚类技术的其它领域的应用，很多有效的模型被提了出来。这些模型主要分为如下四类：

## 模型的分类

- ① 基于迭代的方法
- ② 基于代数的方法
- ③ 基于统计的方法
- ④ 基于谱聚类的方法

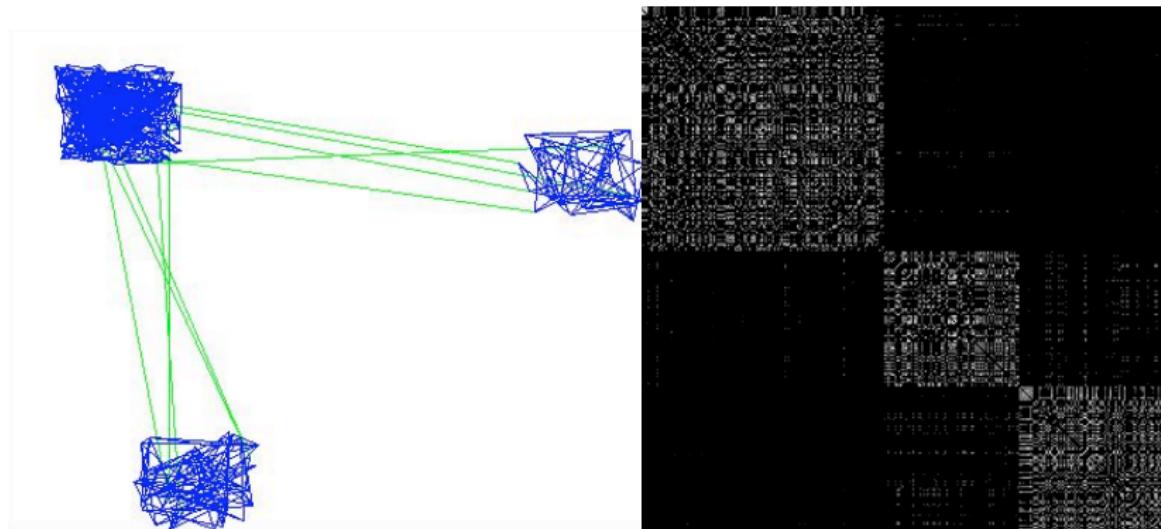
上面列出的这几种主要方法中，基于谱聚类的方法操作简单、聚类精度高。在这其中，利用数据重建策略来建立数据点之间全局关联性的方法表现尤为突出。

## 原理示例图：



# 子空间聚类问题

## 关联矩阵示例图：



## 相关的模型：

稀疏子空间聚类(SSC)模型

- 无噪音情形

$$\begin{aligned} & \min_{\mathbf{Z}} \|\mathbf{Z}\|_1 \\ & \mathbf{X} = \mathbf{XZ}, \text{ s.t. } \text{diag}(\mathbf{Z}) = 0 \end{aligned} \tag{1}$$

- 有噪音情形

$$\begin{aligned} & \min_{\mathbf{Z}, \mathbf{E}} \|\mathbf{Z}\|_1 + \lambda_1 \|\mathbf{E}\|_1 + \lambda_2 \|\mathbf{N}\|_F^2 \\ & \mathbf{X} = \mathbf{XZ} + \mathbf{E} + \mathbf{N}, \text{ s.t. } \text{diag}(\mathbf{Z}) = 0 \end{aligned} \tag{2}$$

## 相关的模型：

低秩子空间聚类(LRR)模型

- 无噪音情形

$$\begin{aligned} & \min_{\mathbf{Z}} \| \mathbf{Z} \|_* \\ & s.t. \mathbf{X} = \mathbf{XZ} \end{aligned} \tag{3}$$

- 有噪音情形

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{E}} \| \mathbf{X} \|_* + \lambda \| \mathbf{E} \|_{1,2} \\ & \mathbf{A} = \mathbf{AX} + \mathbf{E} \end{aligned} \tag{4}$$

## 相关的模型：

稳健最小二乘子空间聚类(LSR)模型

- 无噪音情形

$$\begin{aligned} & \min_{\mathbf{Z}} \| \mathbf{Z} \|_F \\ & s.t. \mathbf{X} = \mathbf{XZ}, \text{diag}(\mathbf{X}) = 0 \end{aligned} \tag{5}$$

- 有噪音情形

$$\begin{aligned} & \min_{\mathbf{Z}} \| \mathbf{XZ} - \mathbf{X} \|_F^2 + \lambda \| \mathbf{Z} \|_F^2 \\ & s.t. \text{diag}(\mathbf{Z}) = 0 \end{aligned} \tag{6}$$

## 存在问题的概述

前面提到的算法成功的关键在于关联矩阵

$$\mathbf{C} = |\mathbf{Z}^\top| + |\mathbf{Z}|$$

的建立。然而通过SSC得到的  $\mathbf{C}$  过于稀疏、通过 LRR 以及 LSR 得到的  $\mathbf{C}$  过于稠密，从而影响数据点之间正确的关联；另一方面，从以上含噪音模型可以看出，对噪音项的建模假设噪音服从某种分布，实际场景的噪音复杂，这种过强的假设不能很好的除去噪音，影响重构效果，降低聚类精度。

# 主要工作(一): $k$ -support 范数介绍

## 定义

令向量  $\mathbf{x} \in \mathbb{R}^n$ , 对任意的  $k \in \{1, 2, \dots, n\}$ , 则  $k$ -support 范数  $\|\cdot\|_k^{sp}$  定义为

$$\|\mathbf{x}\|_k^{sp} \triangleq \min \left\{ \sum_{I \in \mathcal{G}_k} \|v_I\|_2 : \text{supp}(v_I) \subseteq I, \sum_{I \in \mathcal{G}_k} v_I = \mathbf{x} \right\} \quad (7)$$

这里  $\mathcal{G}_k$  代表集合  $\{1, 2, \dots, n\}$  的所有元素个数不超过  $k$  的子集。

- ① Argyriou A, Foygel R, Srebro N. Sparse Prediction with the  $k$ -Support Norm. Advances in Neural Information Processing Systems. 2012:1457 – 1465.

## 性质

对每一个  $\mathbf{x} \in \mathbb{R}^n$ , 它的  $k$ -support 范数可以表示成:

$$\|\mathbf{x}\|_k^{sp} = \left( \sum_{i=1}^{k-r-1} \left( |\mathbf{x}|_i^\downarrow \right)^2 + \frac{1}{1+r} \left( \sum_{i=k-r}^n |\mathbf{x}|_i^\downarrow \right)^2 \right)^{\frac{1}{2}} \quad (8)$$

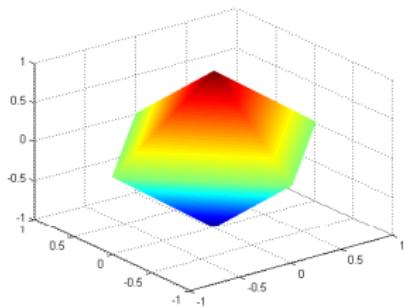
这里我们令  $|\mathbf{x}|_0^\downarrow$  表示正无穷  $+\infty$ ,  $r$  是集合  $\{0, 1, \dots, k-1\}$  中的唯一一个满足如下关系

$$|\mathbf{x}|_{k-r-1}^\downarrow > \frac{1}{r+1} \sum_{i=k+r}^n |\mathbf{x}|_i^\downarrow \leq |\mathbf{x}|_{k-r}^\downarrow \quad (9)$$

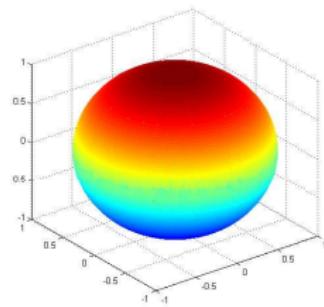
的整数。

# 主要工作：建模

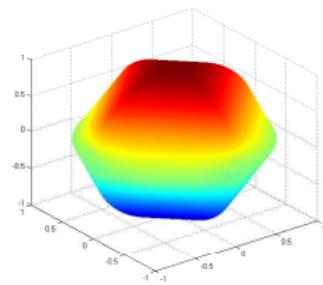
$k$ -support 范数单位球



(a)



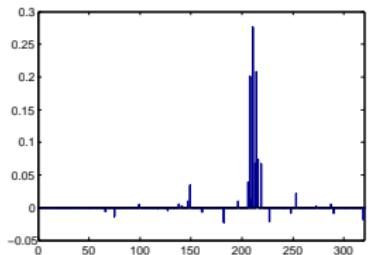
(b)



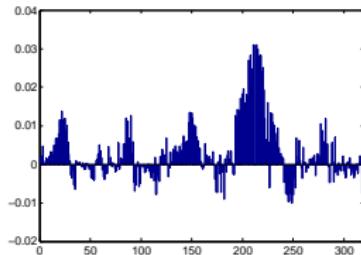
(c)

# 主要工作：建模

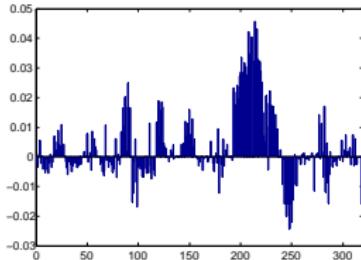
## $k$ -support 重构系数可视化



(a)



(b)



(c)

# 主要工作：建模

## $k$ -support 范数正则化子空间聚类

当数据是干净的，对于数据矩阵的每一列  $\mathbf{x}_i, i = 1, 2, \dots, n$  利用如下模型

$$\begin{aligned} & \min(\|\mathbf{z}_i\|_k^{sp})^2 \\ s.t. \quad & \mathbf{Xz}_i = \mathbf{x}_i, \quad \mathbf{z}_{ii} = 0 \end{aligned} \tag{10}$$

重建。

当数据受到噪声污染，则采用模型

$$\begin{aligned} & \min_{\mathbf{Z}} \frac{1}{2} \|\mathbf{XZ} - \mathbf{X}\|_F^2 + \frac{\lambda}{2} \sum_{i=1}^n (\|\mathbf{z}_i\|_k^{sp})^2 \\ s.t. \quad & \text{diag}(\mathbf{Z}) = 0 \end{aligned} \tag{11}$$

# 主要工作：建模

## 一个例子

$$\mathbf{x} = \begin{pmatrix} 1 & 2 & 3 & 6 & 0 & 0.2 & 0 & 0 \\ 0.1 & 0 & 0.3 & 0 & 1 & 2 & 3 & 4 \end{pmatrix}$$



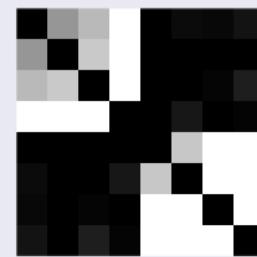
LSR



SSC



LRR



KSP

method	LSR	SSC	LRR	KSC
Accuracy	25	75	25	100

## 关于EBD 条件的定理

令  $\bar{\mathbf{X}}_{-i} = \mathbf{X}_{-i}\mathbf{P} = (\mathbf{X}_1, \dots, \mathbf{X}_L)$ ,  $j = 1, \dots, L$ ,  $\mathbf{X}_j$  属于子空间  $\mathcal{S}_j$ , 并且诸  $\mathcal{S}_j$  相互独立。对  $i = 1, \dots, n$  如果优化问题 (12) 的最优解  $\mathbf{h}_i$  和可行解  $\hat{\mathbf{h}}_i$  满足  $k$ EBD 条件。那么由优化问题 (12) 求得的系数矩阵具有块对角形式。

$$\begin{aligned} & \min \left( \| \mathbf{P}^{-1} z_i \|_k^{sp} \right)^2 \\ s.t. \quad & \bar{\mathbf{X}}_{-i} \mathbf{P}^{-1} z_i = \mathbf{x}_i, \quad i = 1, \dots, n \end{aligned} \tag{12}$$

注：问题(10)与问题 (12) 等价。

$$\mathbf{X}_{-i} = (\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{0}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_n)$$

## 关于Grouping Effect性质的定理

给定一个列归一化的数据矩阵  $\mathbf{X} \in \mathbb{R}^{m \times n}$ ，正则化参数  $\lambda > 0$ ，令  $\mathbf{z}^*$  是优化问题 (14) 的最优解。任取数据矩阵  $\mathbf{X}$  两列，就有如下关系式

$$|z_{il}^* - z_{lj}^*| \leq \frac{1}{\lambda} \sqrt{2(1 - \mathbf{x}_j^\top \mathbf{x}_l)} \quad (13)$$

成立。其中  $z_j^*$  和  $z_l^*$  分别表示最优解  $\mathbf{z}^*$  的第  $j$  和  $l$  个元素，并且  $j, l \geq k - r - 1$ 。

注：问题 (14) 如下所示：

$$\min_{\mathbf{z}} \frac{1}{2} \|\mathbf{Xz} - \mathbf{x}\|_F^2 + \frac{\lambda}{2} (\|\mathbf{z}\|_k^{sp})^2 \quad (14)$$

## 统计恢复理论(一)

令数据矩阵  $\mathbf{X} \in \mathbb{R}^{m \times n}$  的元素服从独立同分布的标准正态分布,  $\Omega$  表示集合  $\mathcal{T}_{k^{sp}}(\mathbf{z}) \cap \mathbb{S}^{n-1}$ 。对于某个常数  $C$  如果  $m > C + 1$ , 那么优化问题(16)的最优值等于真实值  $\hat{\mathbf{z}}$  的概率不小于  $1 - \exp\left(-\frac{1}{2}\left(\frac{m}{\sqrt{m+1}} - \sqrt{C}\right)^2\right)$ 。其中常数  $C$  由下式给出

$$C = \left( \sqrt{2k \ln \left( \left( m - k - \lceil \frac{s}{k} \rceil + 2 \right) \right)} + \sqrt{k} \right)^2 \lceil \frac{s}{k} \rceil + s \quad (15)$$

$s$  等于满足定义条件 11 的集合  $\mathbb{I}$  的并的个数。

注: 问题 (16) 如下所示:

$$\begin{aligned} \min_{\mathbf{z}} & (\|\mathbf{z}\|_k^{sp})^2 \\ \text{s.t. } & \mathbf{Xz} = \mathbf{x}, \end{aligned} \quad (16)$$

## 统计恢复理论(二)

令数据矩阵  $\mathbf{X} \in \mathbb{R}^{m \times n}$  的元素服从标准正态分布，并且  $\Omega$  表示  $\mathcal{T}_{k^{sp}}(\mathbf{z}) \cap \mathbb{S}^{n-1}$ . 含有噪声的观测数据表示为  $\mathbf{x} = \mathbf{X}\mathbf{z} + \boldsymbol{\varphi}$ ，这里噪声项满足关系式  $\|\boldsymbol{\varphi}\|_2 \leq \alpha$ 。那么，存在一个正数  $\epsilon$ ，只要  $m$  满足

$$m > \frac{C + \sqrt{2C^2 + 4}}{2(1 - \epsilon)^2} \quad (17)$$

就有式子  $\|\hat{\mathbf{d}}\|_2 = \|\hat{\mathbf{z}} - \mathbf{z}^*\|_2 \leq \frac{2\alpha}{\epsilon}$  以不小  
于  $1 - \exp\left(-\frac{1}{2}\left(\frac{m}{\sqrt{m+1}} - \sqrt{C} - \epsilon\right)^2\right)$  的概率成立。这里的常  
数  $C$  同上个定理一致。

注：本定理考虑的模型为：

$$\min_{\mathbf{z}} (\|\mathbf{z}\|_k^{sp})^2 \quad s.t. \quad \|\mathbf{X}\mathbf{z} - \mathbf{x}\|_2 \leq \alpha \quad (18)$$

## 交替方向乘子法

$$\begin{aligned} \min_{\mathbf{z}} \quad & \frac{1}{2} \|\mathbf{X}_{-\mathbf{y}} \mathbf{z} - \mathbf{y}\|^2 + \frac{\lambda}{2} (\|\mathbf{w}\|_k^{sp})^2 \\ \text{s.t.} \quad & \mathbf{z} = \mathbf{w} \end{aligned} \quad (19)$$

进一步，我们可得：

$$\min_{\mathbf{z}, \mathbf{w}} \frac{1}{2} \|\mathbf{X}_{-\mathbf{y}} \mathbf{z} - \mathbf{y}\|^2 + \frac{\lambda}{2} (\|\mathbf{w}\|_k^{sp})^2 + \frac{\beta}{2} \|\mathbf{z} - \mathbf{w} - \frac{\Gamma}{\beta}\|^2 \quad (20)$$

把这个优化问题拆分成如下两个子问题：

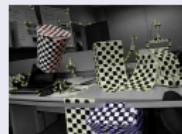
$$\min_{\mathbf{z}} \frac{1}{2} \|\mathbf{X}_{-\mathbf{y}} \mathbf{z} - \mathbf{y}\|^2 + \frac{\beta}{2} \|\mathbf{z} - \mathbf{w} - \frac{\Gamma}{\beta}\|^2 \quad (21)$$

和

$$\min_{\mathbf{w}} \frac{\beta}{2} \|\mathbf{w} - \mathbf{z} + \frac{\Gamma}{\beta}\|^2 + \frac{\lambda}{2} (\|\mathbf{w}\|_k^{sp})^2 \quad (22)$$

# $k$ -support 范数子空间聚类实验结果

## 用于评测的数据库



—N<sub>4</sub> L<sub>5</sub>

1 2 3 4 5

1 2 3 4

# $k$ -support 范数子空间聚类实验

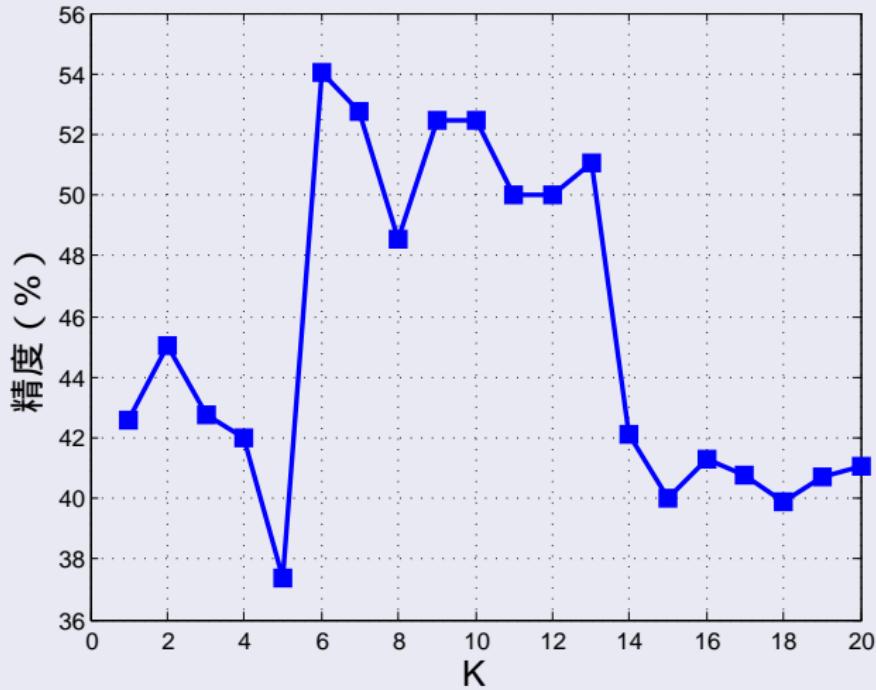
## 不同数据集上的实验结果

	LRR	SSC	LSR	CASS	LS3C	WBLRR	$S^3C$	FaLRR	KSC
2 motions	98.71	98.68	98.61	98.71	98.61	98.70	98.81	98.76	98.80
3 motions	97.25	97.02	97.23	97.43	97.10	97.23	97.51	97.32	97.56

	LRR	SSC	LSR	CASS	LS3C	WBLRR	$S^3C$	FaLRR	KSC
5 objects	86.56	80.31	92.19	94.03	93.10	87.03	96.59	87.54	94.58
8 objects	78.91	62.90	80.66	91.41	85.66	80.03	95.85	81.77	91.50
10 objects	65.00	52.19	73.59	81.88	77.08	68.73	94.82	79.22	82.04

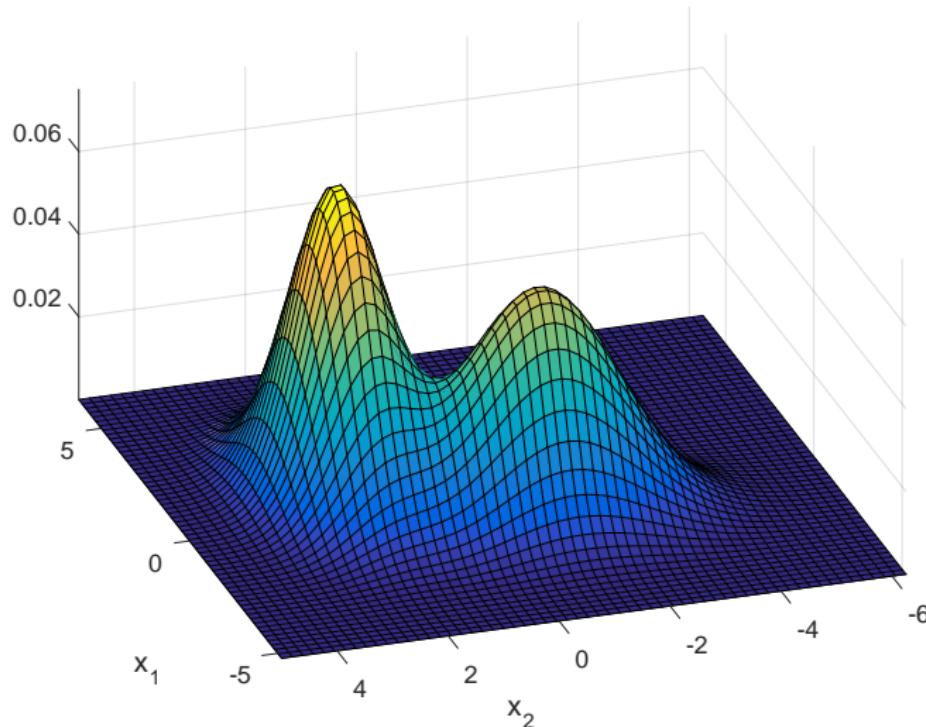
	LRR	SSC	LSR	CASS	LS3C	WBLRR	$S^3C$	FaLRR	KSC
MR	34.65	35.73	31.76	32.54	36.04	35.46	36.01	35.61	39.76
MRB	25.54	26.01	21.00	24.63	27.62	25.58	28.36	25.93	32.00
MBR	40.76	42.6	41.05	42.33	45.36	41.33	47.12	42.41	54.06

## $K$ 的变化



# 主要工作(二): 高斯混合回归子空间聚类

## 直观认识



## 简单推导

$$\min_{\mathbf{Z}} \mathcal{F}(\mathbf{XZ} - \mathbf{X}) + \mathcal{R}(\mathbf{Z}) \quad (23)$$

$$\begin{aligned} p(\mathbf{Z} | \mathbf{X}, \mathbf{XZ}) &= \frac{p(\mathbf{Z}, \mathbf{XZ} | \mathbf{X})}{\int p(\mathbf{Z}, \mathbf{XZ} | \mathbf{X}) d\mathbf{Z}} \\ &= \frac{p(\mathbf{XZ} | \mathbf{X}, \mathbf{Z}) p(\mathbf{Z})}{\int p(\mathbf{Z}, \mathbf{XZ} | \mathbf{X}) d\mathbf{Z}} \end{aligned} \quad (24)$$

$$p(\mathbf{Z} | \mathbf{X}, \mathbf{XZ}) \propto p(\mathbf{XZ} | \mathbf{X}, \mathbf{Z}) p(\mathbf{Z}) \quad (25)$$

$$\min_{\mathbf{Z}} -\ln(p(\mathbf{XZ} | \mathbf{X}, \mathbf{Z})) - \ln(p(\mathbf{Z})) \quad (26)$$

## 一个例子

$$p(\mathbf{XZ} \mid \mathbf{X}, \mathbf{Z}) = \prod_{i=1}^n \mathcal{N}(\mathbf{Xz}_i \mid \mathbf{x}_i, \beta \mathbf{I}) \quad (27)$$

以及

$$p(\mathbf{Z}) = \prod_{i=1}^n \mathcal{N}(\mathbf{z}_i \mid \mathbf{0}, \alpha \mathbf{I}) \quad (28)$$

因此我们得到如下优化问题

$$\min_{\mathbf{Z}} \frac{1}{2} \| \mathbf{XZ} - \mathbf{X} \|_F^2 + \frac{\alpha}{2\beta} \| \mathbf{Z} \|_F^2 \quad (29)$$

## 模型的引出

$$p(\mathbf{XZ} \mid \mathbf{X}, \mathbf{Z}) = \prod_{i=1}^n \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{Xz}_i - \mathbf{x}_i \mid \mu_k, \Sigma_k) \quad (30)$$

$$-\sum_{i=1}^n \ln \left( -\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{Xz}_i - \mathbf{x}_i \mid \mu_k, \Sigma_k) \right) + \frac{\alpha}{\beta} \|\mathbf{Z}\|_F^2 \quad (31)$$

这里， $\mu_k$ 和 $\Sigma_k$ ,  $k = 1, 2, \dots, K$ 分别表示均值和方差，  
 $\pi_k > 0$ ,  $k = 1, 2, \dots, K$ 是权重，并且满足 $\sum_{k=1}^K \pi_k = 1$ 。

## 具体模型

$$\begin{aligned} & - \sum_{i=1}^n \ln \left( - \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{Xz}_i - \mathbf{x}_i \mid \mathbf{0}, \boldsymbol{\Sigma}_k) \right) + \lambda \|\mathbf{Z}\|_F^2 \\ & s.t. \quad \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \text{diag}(\mathbf{Z}) = \mathbf{0}, \end{aligned} \tag{32}$$

$$\pi_k \geq 0, \boldsymbol{\Sigma}_k \in \mathbb{S}^+, k = 1, \dots, K, \sum_{k=1}^K \pi_k = 1,$$

## EM 迭代

$$\gamma_{i,k} = \frac{\pi_k \mathcal{N}(\tilde{\mathbf{e}}_i | \mathbf{0}, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\tilde{\mathbf{e}}_i | \mathbf{0}, \Sigma_j)}, \quad (33)$$

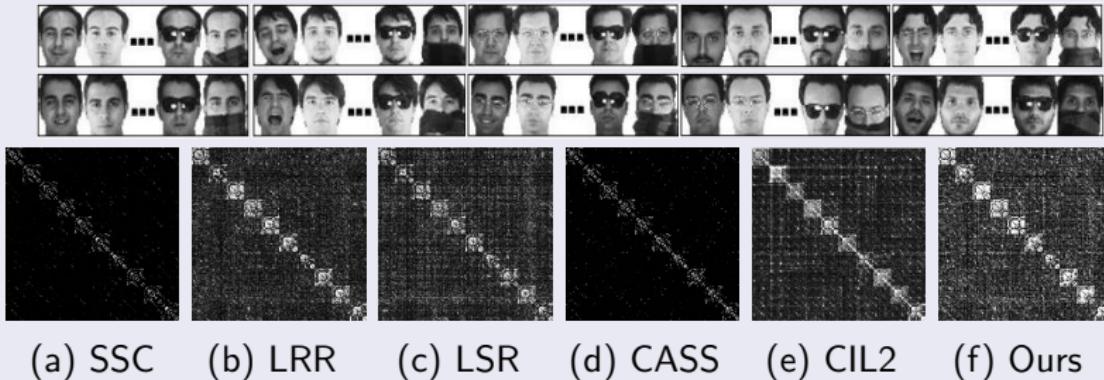
$$\begin{aligned} & \min_{\Sigma_k} - \sum_{i=1}^n \ln \left( \sum_{k=1}^K \pi_k \mathcal{N}(\tilde{\mathbf{e}}_i | \mathbf{0}, \Sigma_k) \right) \\ & \text{s.t. } \Sigma_k \in \mathbb{S}^+. \end{aligned} \quad (34)$$

$$\min_{\pi_k \geq 0} - \sum_{i=1}^n \ln \left( \sum_{k=1}^K \pi_k \mathcal{N}(\tilde{\mathbf{e}}_i | \mathbf{0}, \Sigma_k) \right) + \beta \left( \sum_{k=1}^K \pi_k - 1 \right), \quad (35)$$

$$\min_{\mathbf{z}_i} - \sum_{i=1}^n \ln \left( \sum_{k=1}^K \pi_k \mathcal{N}(\tilde{\mathbf{e}}_i | \mathbf{0}, \Sigma_k) \right) + \lambda \| \mathbf{z}_i \|_F^2. \quad (36)$$

# 高斯混合回归子空间聚类

## AR数据集评测



	SSC	LRR	LSR	CASS	CIL2	Ours
	73.51	75.41	52.10	75.18	76.35	<b>80.32</b>

	SSC	LRR	LSR	CASS	CIL2	Ours
5 subjects	83.05	84.41	87.69	78.46	85.38	<b>93.85</b>
10 subjects	75.06	78.54	63.07	77.69	80.39	<b>88.85</b>

## Grouping Effect

已知数据点  $\mathbf{x} \in \mathbb{R}^m$ , 列归一化数据矩阵  $\mathbf{X}$  以及正则化参数  $\lambda > 0$ ,  
不失一般性, 令  $\hat{\mathbf{z}}$  是如下问题的最优解

$$\min_{\mathbf{z}} -\ln \left( \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{X}\mathbf{z} - \mathbf{x} \mid \mathbf{0}, \Sigma_k) \right) + \lambda \|\mathbf{z}\|^2, \quad (37)$$

那么存在一个常数  $a$  使得

$$|\hat{z}^i - \hat{z}^j| \leq \frac{a}{\lambda} \sqrt{\frac{1-\rho}{2}}, \quad (38)$$

其中  $\rho = \cos\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ . 令  $\hat{z}^i$  和  $\hat{z}^j$  分别表示向量  $\hat{\mathbf{z}}$  的第  $i$  个和第  $j$  个元素, 而  $\mathbf{x}_i$  和  $\mathbf{x}_j$  分别表示数据矩阵  $\mathbf{X}$  的第  $i$  列和第  $j$  列。

在数据集中，对于一个给定的数据点为了合理的反映这个数据点和同一个子空间其它的数据点之间的相关程度，我们加了一个约束  $\text{diag}(\mathbf{Z}) = 0$ ，促使这个给定的数据点尽可能表示成和它同处一个子空间的其它的数据点的线性表示。由于混合模型的特殊性下面给出了证明

## 对角约束的证明

给定一个数据矩阵  $\mathbf{X} \in \mathbb{R}^{m \times n}$ , 对于  $i = 1, \dots, n$ ,

令  $\mathbf{X}_{-i} = (\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{0}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_n)$ 。那么优化问题(39)的解满足  $\mathbf{z}_{ii} = 0$ ，进一步可知问题 (32) 的解满足  $\text{diag}(\mathbf{Z}) = \mathbf{0}$ 。

$$\min_{\mathbf{z}} -\ln \left( \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{X}_{-i}\mathbf{z} - \mathbf{x} | \mathbf{0}, \Sigma_k) \right) + \lambda \|\mathbf{z}\|^2, \quad (39)$$

## 渐进性质

令矩阵误差矩阵  $\mathbf{E} \in \mathbb{R}^{m \times n}$  的列是独立同分布的，并且令模型 (32) 中的高斯元的个数为  $K$ 。如果  $\frac{\lambda}{\sqrt{n}} = o(1)$  并且  $\theta_{\mathbf{z}}^{ini} - \theta_{\mathbf{z}}^t = \mathcal{O}_p\left(n^{-\frac{1}{2}}\right)$  并且在模型 (36) 满足正则条件 (Jianqing fan, Runze Li, 2001) (A) – (C) 的前提下，我们可知模型 (36) 的局部极小值  $\theta_{\mathbf{z}}^{lm}$  满足如下的关系式

$$\sqrt{n} \left( \mathbf{z}_i^{lm} - \mathbf{z}_i^t \right) \xrightarrow{d} \mathcal{N} \left( 0, I_s (\mathbf{z}_i^t)^{-1} \right), \quad i = 1, \dots, n \quad (40)$$

## 估计高斯元个数

$$\Theta = (\pi_1, \dots, \pi_K, \Sigma_1, \dots, \Sigma_K) \quad (41)$$

$$length(\mathbf{E}, \Theta) = length(\mathbf{E} | \Theta) + length(\Theta) \quad (42)$$

$$\begin{aligned} \hat{\Theta} = \arg \min_{\Theta} & \left\{ -\ln p(\Theta) - \ln p(\mathbf{XZ} | \Theta) + \frac{1}{2} \ln |\mathbf{F}_c(\Theta)| \right. \\ & \left. + \frac{D(\Theta)}{2} \left( 1 + \ln \frac{1}{12} \right) \right\} \end{aligned} \quad (43)$$

- ① M. A. Figueiredo, A. K. Jain,. Unsupervised learning of finite mixture models. IEEE TPAMI 2002:381-396.
- ② I. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud, EM Algorithms for Weighted-data Clustering with Application to Audio-visual Scene Analysis. IEEE TPAMI, 2016,(38) 2402 – 2415.

## 修改的模型

综上，我们把模型 (32) 改造成如下最小描述长度形式：

$$\begin{aligned} (\hat{\Theta}, \mathbf{Z}^*) &= \arg \min_{\Theta, \mathbf{Z}} \mathcal{L}(\Theta, \mathbf{Z}) \\ &= \arg \min_{\Theta, \mathbf{Z}} \left\{ \frac{D}{2} \sum_{k: \pi_k > 0} \ln \left( \frac{n\pi_k}{12} \right) + \frac{K}{2} \ln \left( \frac{n}{12} \right) - \lambda \|\mathbf{Z}\|_F^2 \right. \\ &\quad \left. + \frac{K(D+1)}{2} - \sum_{i=1}^n \ln \left( \sum_{k=1}^K \pi_k \mathcal{N}(\tilde{\mathbf{e}}_i | \mathbf{0}, \Sigma_k) \right) \right\} \\ s.t. \quad \pi_k &\geq 0, \Sigma_k \in \mathbb{S}^+, k = 1, \dots, K, \sum_{k=1}^K \pi_k = 1 \end{aligned} \tag{44}$$

## 修改的EM迭代

$$\pi_k^{mnew} = \frac{\max\{0, \sum_{i=1}^n \gamma_{i,k} - \frac{D}{2}\}}{\sum_{j=1}^K \max\{0, \sum_{i=1}^n \gamma_{i,j} - \frac{D}{2}\}} \quad (45)$$

$$\gamma_{i,k} = \frac{\pi_k \mathcal{N}(\tilde{\mathbf{e}}_i | \mathbf{0}, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\tilde{\mathbf{e}}_i | \mathbf{0}, \Sigma_j)}, \quad (46)$$

$$\Sigma_k^{mnew} = \frac{1}{\gamma_{n,k}} \left( \sum_{n=1}^N \frac{\pi_k^{mnew} \mathcal{N}(\tilde{\mathbf{e}}_n | \mathbf{0}, \Sigma_k)}{\sum_{j=1}^{K_{\max}} \pi_j^{mnew} \mathcal{N}(\tilde{\mathbf{e}}_n | \mathbf{0}, \Sigma_j)} \tilde{\mathbf{e}}_n^{old} (\tilde{\mathbf{e}}_n^{old})^\top + \epsilon \mathbf{I} \right)$$

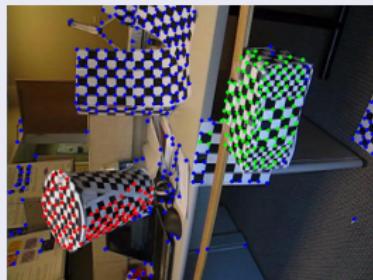
$$\mathbf{z}_n^{new} = \left( \frac{\sum_{k=1}^{K_{\max}} \xi_k \tilde{\mathbf{X}}_n^\top (\Sigma_k^{mnew})^{-1} \tilde{\mathbf{X}}_n}{\sum_{j=1}^{K_{\max}} \xi_j} + 2\lambda \mathbf{I} \right)^{-1} \mathbf{b}_n,$$

其中  $\mathbf{b}_n = \frac{\sum_{k=1}^{K_{\max}} \xi_k \tilde{\mathbf{X}}_n (\Sigma_k^{mnew})^{-1}}{\sum_{j=1}^{K_{\max}} \xi_j} \mathbf{x}_n$ ,  $\xi_k = \pi_k^{mnew} \mathcal{N}(\tilde{\mathbf{e}}_n^{old} | \mathbf{0}, \Sigma_k^{new})$ ,

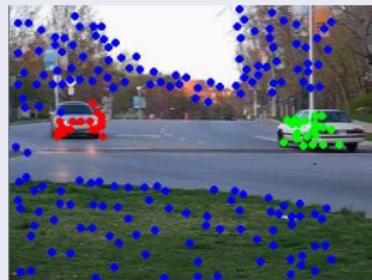
$$\xi_j = \pi_j^{mnew} \mathcal{N}(\tilde{\mathbf{e}}_n^{old} | \mathbf{0}, \Sigma_j^{mnew}).$$



## Hopkins 155 聚类可视化



(a)



(b)

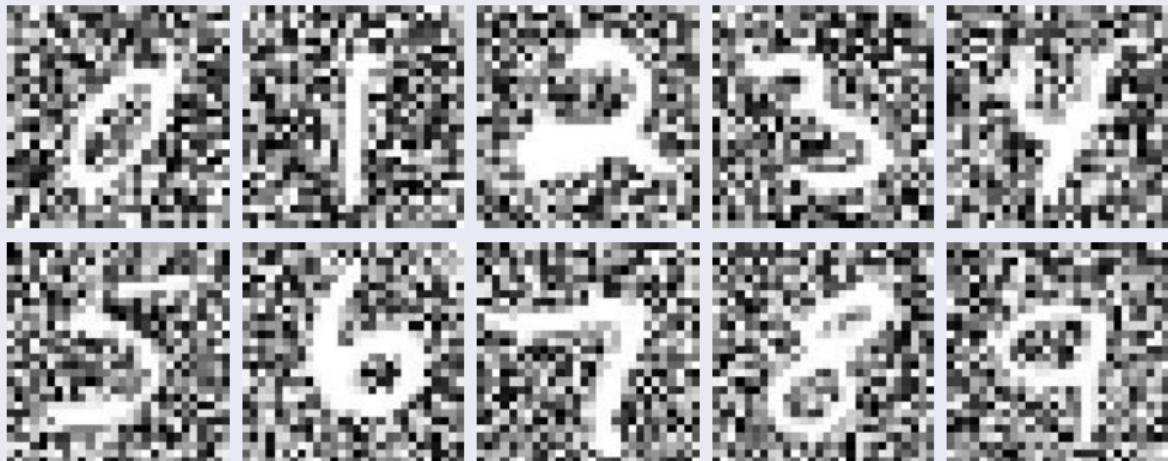


(c)

	SSC	LRR	LSR	CASS	CIL2	Ours
2 motions	95.69	96.43	97.48	97.01	97.63	<b>98.76</b>
3 motions	91.97	92.35	93.21	94.06	94.34	<b>95.03</b>

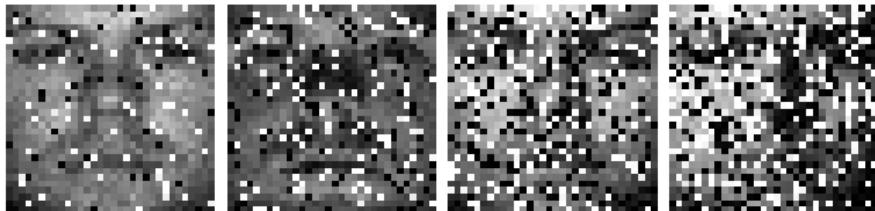
# 实验结果

数据库样本

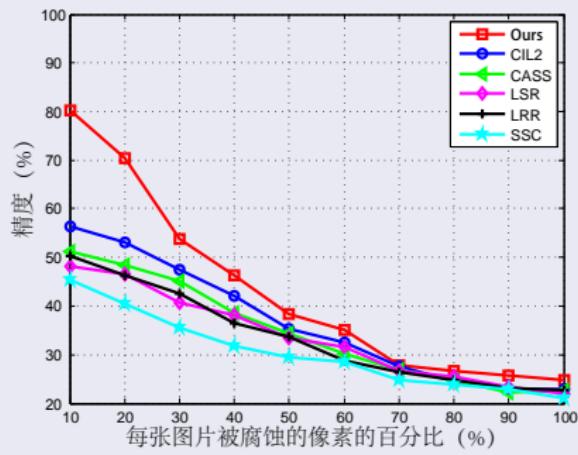


SSC	LRR	LSR	CASS	CIL2	Ours
33.56	22.85	20.55	29.05	36.50	<b>51.98</b>

# 高斯混合回归子空间聚类

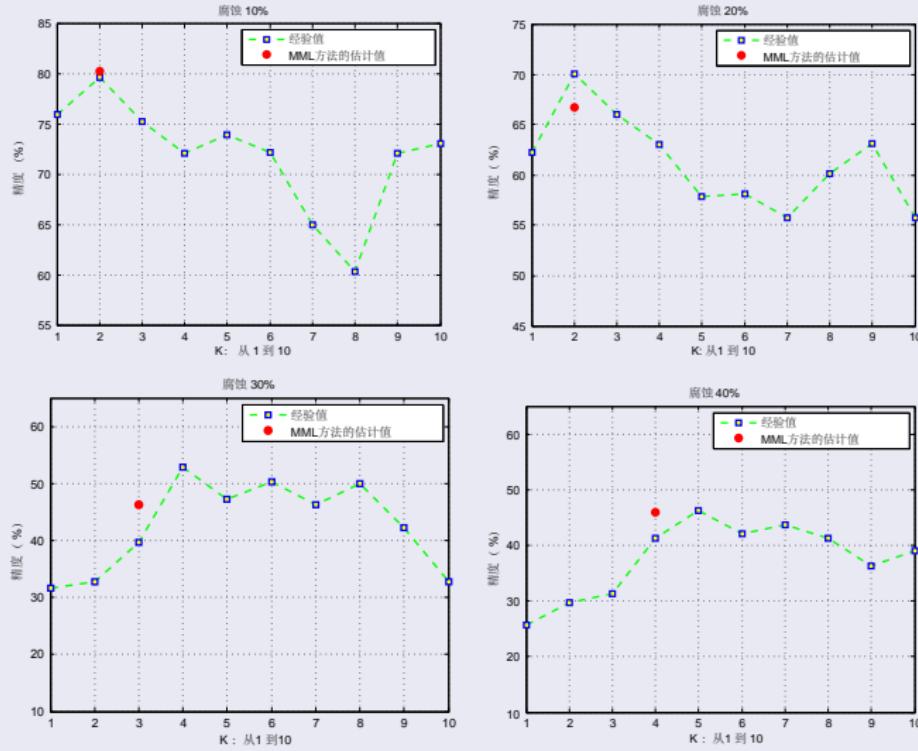


## 多元混合高斯



# 高斯混合回归子空间聚类

## 多元混合高斯



## 主要思想

干净数据 $\mathbf{Y}$ 和污染的数据 $\mathbf{X}$ 以及噪音 $\varepsilon$ 这三者之间可以表示成如下的线性关系：

$$\mathbf{X} = \mathbf{Y} + \varepsilon \quad (47)$$

$$\mathbf{X} = \mathbf{XZ} + \varepsilon \quad (48)$$

这里 $\mathbf{X}$ 表示数据矩阵， $\mathbf{Z}$ 表示重构系数矩阵， $\varepsilon$ 表示噪音。乘性噪音对数据的污染机理可表示如下

$$\mathbf{X} = \mathbf{B} \odot \varepsilon \quad (49)$$

其中 $\odot$ 表示哈达马积

## 模型的提出

$$\begin{aligned} \min_{\mathbf{H}, \mathbf{Z}, \mathbf{B}} & \frac{1}{2} \|\mathbf{H} - \alpha \mathbf{E}\|_F^2 + \lambda_1 \|\mathbf{B} - \mathbf{X} \odot \mathbf{H}\|_1 \\ & + \frac{1}{2} \|\mathbf{BZ} - \mathbf{B}\|_F^2 + \lambda_2 \|\mathbf{Z}\|_* \end{aligned} \quad (50)$$

这里  $\mathbf{X}$ 、 $\mathbf{B}$  和  $\mathbf{Z}$  分别表示污染的数据矩阵、待估计的干净数据矩阵以及重构系数矩阵。另外  $\mathbf{H}$  中的每一项  $\mathbf{H}_{i,j}$  表示相应噪音的倒数， $\mathbf{E}$  中的每个元素都是 1。 $\alpha$  表示噪音的期望， $\lambda_1$ 、 $\lambda_2$  是正则化参数。常见的有乘性 *Gamma* 噪音、乘性 *Rayleigh* 噪音以及乘性 *Gaussian* 噪音这三种。

- ① Xi-Le Zhao, Fan Wang, and Michael K. Ng, A New Convex Optimization Model for Multiplicative Noise and Blur Removal. SIAM Journal on Imaging Sciences 2014,(7): 456-475

## ADMM方法求解

更新  $B$ :

$$\min_{\mathbf{B}} \frac{1}{2} \| \mathbf{B}\mathbf{W}^{(n)} - \mathbf{B} \|_F^2 + \frac{\beta}{2} \| \mathbf{B} - \mathbf{A} \odot \mathbf{H}^{(n)} - \mathbf{Z}^{(n)} + \frac{\Lambda_1^{(n)}}{\beta} \|_F^2 \quad (51)$$

更新  $\mathbf{W}$

$$\min_{\mathbf{W}} \frac{1}{2} \| \mathbf{B}^{(n+1)}\mathbf{W} - \mathbf{B}^{(n+1)} \|_F^2 + \frac{\beta}{2} \| \mathbf{W} - \mathbf{X}^{(n)} + \frac{\Lambda_2^{(n)}}{\beta} \|_F^2 \quad (52)$$

更新  $\mathbf{Z}$

$$\min_{\mathbf{Z}} \frac{1}{2} \| \mathbf{B}^{(n+1)} - \mathbf{A} \odot \mathbf{H}^{(n)} - \mathbf{Z} + \frac{\Lambda_2^{(n)}}{\beta} \|_F^2 + \| \mathbf{Z} \|_1 \quad (53)$$

更新  $\mathbf{X}$

$$\min_{\mathbf{X}} \frac{1}{2} \| \mathbf{W}^{(n+1)} - \mathbf{X} + \frac{\Lambda_2^{(n)}}{\beta} \|_F^2 + \frac{\lambda_2}{\beta} \| \mathbf{X} \|_* \quad (54)$$

更新  $\mathbf{H}$

$$\min_{\mathbf{H}} \frac{1}{2} \| \mathbf{H} - \alpha \mathbf{E} \|_F^2 + \frac{\beta}{2} \| \mathbf{B}^{(n+1)} - \mathbf{A} \odot \mathbf{H} - \mathbf{Z}^{(n+1)} + \frac{\Lambda_2^{(n)}}{\beta} \|_F^2 \quad (55)$$

# 乘性噪音子空间聚类

## 实验



(a) (b) (c)

图 1: (a) 干净图片, (b) 从上到下表示受到乘性Rayleigh、Gamma 以及 Gaussian 噪音污染的图片. (c) 相应的去噪结果.

# 乘性噪音子空间聚类

## 实验

	5 Objects		
	Gamma	Gaussian	Rayleigh
Our Method	64.06%	61.75%	59.07%
SSC	52.64%	53.08%	50.43%
LRR	52.73%	51.77%	47.33%
LSR	50.35%	48.04%	45.17%
CASS	56.97%	55.04	52.72%

	8 Objects		
	Gamma	Gaussian	Rayleigh
Our Method	53.05%	51.83%	49.31%
SSC	45.69%	43.83%	40.76%
LRR	46.74%	44.01%	42.33%
LSR	41.63%	39.78%	40.02%
CASS	48.76%	46.55%	43.75%

	10 Objects		
	Gamma	Gaussian	Rayleigh
Our Method	41.96%	40.32%	39.31%
SSC	37.56%	34.85%	33.61%
LRR	34.07%	31.86%	32.63%
LSR	33.73%	29.06%	28.61%
CASS	38.77%	36.01%	33.88%

	MINST data		
	Gamma	Gaussian	Rayleigh
Our Method	39.59%	39.04%	38.88%
SSC	35.06%	33.67%	30.96%
LRR	34.66%	31.07%	28.99%
LSR	32.86%	29.43%	27.96%
CASS	36.11%	35.01%	31.96%

## 工作总结

我们针对以前方法中体现出的不足相继提出了基于 $k$ -support范数正则化的子空间聚类模型、 $\ell_2$ 范数惩罚下的高斯回归子空间聚类模型，另外针对成型噪音污染数据的机理的特殊性，我们提出了两阶段乘性噪音子空间聚类模型。我们对上述模型即给出了理论上的合理性，也通过实验验证了模型的有效性。

## 对未来工作的展望

事实上，我们提出的 $k$ -support范数策略下的子空间聚类模型中， $K$ 的确定、在混合高斯回归中高斯源个数的确定，目前都没有完善的理论保证，另外如何提出凸的乘性噪音子空间聚类模型并给出相应理论分析，这些将是未来工作的方向与重点。

- ① **Baohua Li**, Huchuan Lu, Fu Li, Wei Wu, *Subspace Clustering with K-Support Norm*, IEEE Transactions on Circuits and Systems for Video Technology, 2016, (录用).
- ② **Baohua Li**, Ying Zhang, Zhouchen Lin, Huchuan Lu, *Subspace clustering by Mixture of Gaussian Regression*, IEEE Conference on Computer Vision and Pattern Recognition , 2015, 2094 - 2102.(发表)
- ③ **Baohua Li**, Ying Zhang, Zhouchen Lin, Huchuan Lu, Wei Wu, *Subspace Clustering under Complex Noise*, IEEE Transactions on Circuits and Systems for Video Technology, (Reversion being Processing)
- ④ **Baohua Li**, Wei Wu, *Subspace Clustering Under Multiplicative Noise Corruption*, International Conference on Intelligent Science and Big Data Engineering, 2017, 461-470.(发表)

谢谢！