Achieving near-perfect clustering for high dimension, low sample size data

Yoshikazu Terada

Graduate School of Engineering Science, Osaka University, Japan

Visiting PhD student of Machine Learning Group (Prof. Dr. Ulrike von Luxburg)

at the Department of Informatics of the University of Hamburg

AG DANK/BCS Meeting 2013 in London

November 08 - 09, 2013, Department of Statistical Science , University College London, London, the United Kingdom

Room: the Galton Lecture Theatre, Room 115, 1-19 Torrington Place,

11:15 – 11:35, 09/11/2013 (Sat.)

This work is supported by Grant-in-Aid for JSPS Fellows.

0. Contents

1. Introduction

1.1 Geometric representation of HDLSS data

(Hall et al., 2005; JRSS B)

1.2 Difficulty of clustering HDLSS data

2. Previous study

2.1 Clustering with MDP distance (Ahn et al, 2013; Statist. Sinica)

3. Clustering with distance (or inner product) vectors

3.1 Main idea

3.2 Proposal method

3.3 Theoretical result of the proposal methods

4. Conclusion

1.1 Geometric representation of HDLSS data

- **<u>Notation and General setting</u>**
  - K : Number of clusters, N : Sample size, p : dimensions,
  - n_k : Sample size of Cluster k, $N = \sum_{k=1}^{K} n_k$,
  - $\boldsymbol{X}_k^{(p)}$: p -dimensional random vector of Cluster k ,
  - $\boldsymbol{X}_k^{(p)} := (X_{k1}, \ldots, X_{kp})^T$, $d_{ij}^{(p)} := \left\| \boldsymbol{X}_i^{(p)} - \boldsymbol{X}_j^{(p)} \right\|$
  - $\boldsymbol{X}_k^{(p)}$ $(k = 1, \ldots, K)$ are independent.

1.1 Geometric representation of HDLSS data

- **Notation and General setting** $(k = 1, \ldots, K)$

(a) $\exists M > 0;\ \forall s \in \mathbb{N};\ \mathbb{E}\left[|X_{ks}|^4\right] < M$

(b) $$\frac{1}{p}\sum_{s=1}^{p} \mathbb{E}[X_{ks}]^2 \to \mu_k^2 \quad \text{as} \quad p \to \infty$$

(c) $$\frac{1}{p}\sum_{s=1}^{p} \mathrm{Var}(X_{ks}) \to \sigma_k^2 \quad \text{as} \quad p \to \infty$$

(d) $$\frac{1}{p}\sum_{s=1}^{p} \{\mathbb{E}[X_{ks}] - \mathbb{E}[X_{ls}]\}^2 \to \delta_{kl}^2 \quad \text{as} \quad p \to \infty$$

1.1 Geometric representation of HDLSS data

- **Notation and General setting** $(k = 1, \ldots, K)$

  (e) $\displaystyle \frac{1}{p}\sum_{s=1}^{p} \mathbb{E}[X_{ks}]\mathbb{E}[X_{ls}] \to \eta_{kl} \quad \text{as} \quad p \to \infty$

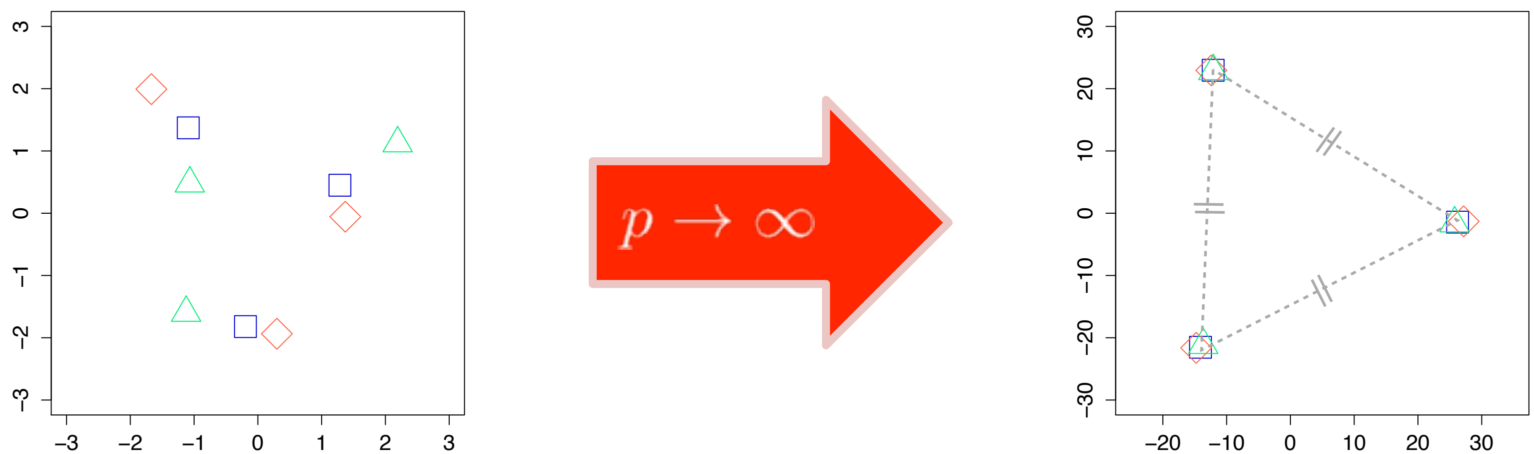
  (f) There is some permutation of $\boldsymbol{X}_k^{(\infty)}$,

  which is ρ -mixing*.

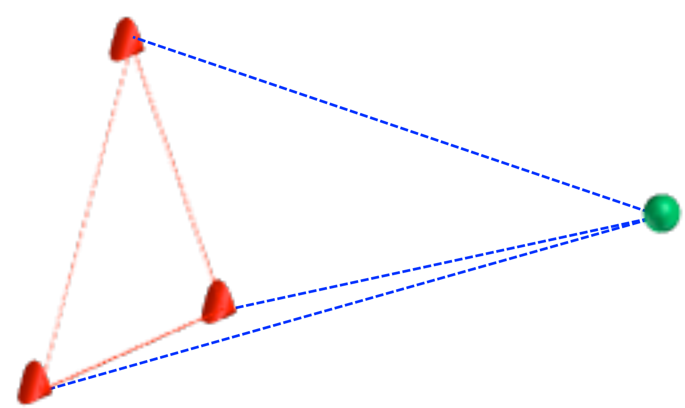
  *The concepts of ρ -mixing is useful as a mild condition for the development of laws of large number.

1.1 Geometric representation of HDLSS data

- **<u>Hall et al. (2005; JRSS B)</u>**
  - The distance between data vectors from a same cluster is approximately-constant after scaled by $\sqrt{p}$!

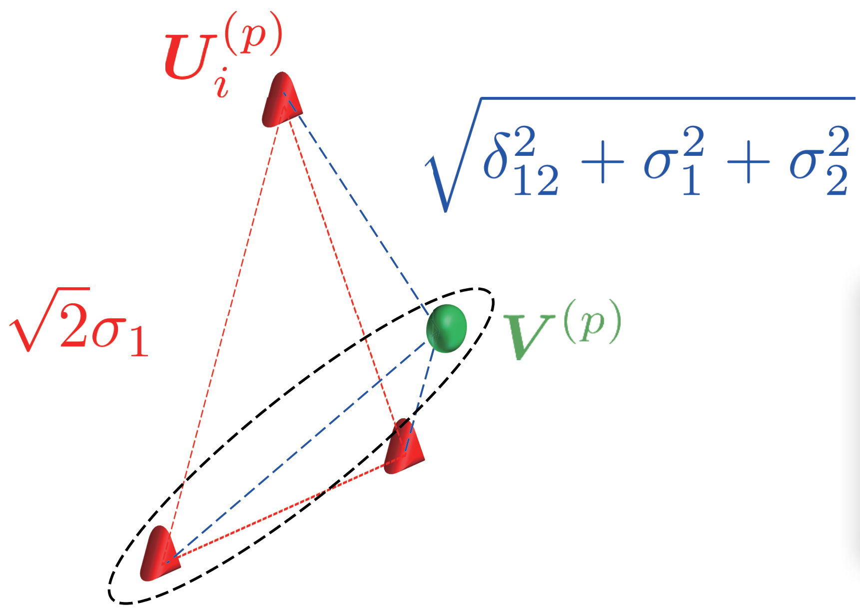


  - The distance between data vectors from different clusters is also approximately constant after scaled by $\sqrt{p}$!



1.2. Difficulty of clustering for HDLSS data

- **Hierarchical clustering in HDLSS contexts**
  - $\boldsymbol{U}_1^{(p)}, \boldsymbol{U}_2^{(p)}, \boldsymbol{U}_3^{(p)} \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}_1^{(p)}$; $\boldsymbol{V}^{(p)} \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}_2^{(p)}$



$$\sqrt{2}\sigma_1 \geq \sqrt{\delta_{12}^2 + \sigma_1^2 + \sigma_2^2}$$

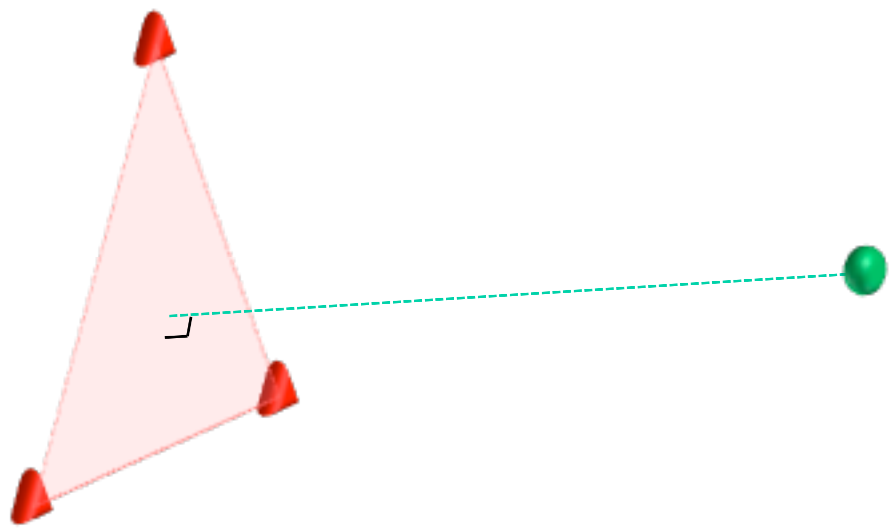
Condition for label consistency

$$2\sigma_k^2 < \mu_{kl}^2 + \sigma_k^2 + \sigma_l^2$$

  - In some cases, classical methods do not work well…

2. Previous study (MDP clustering)

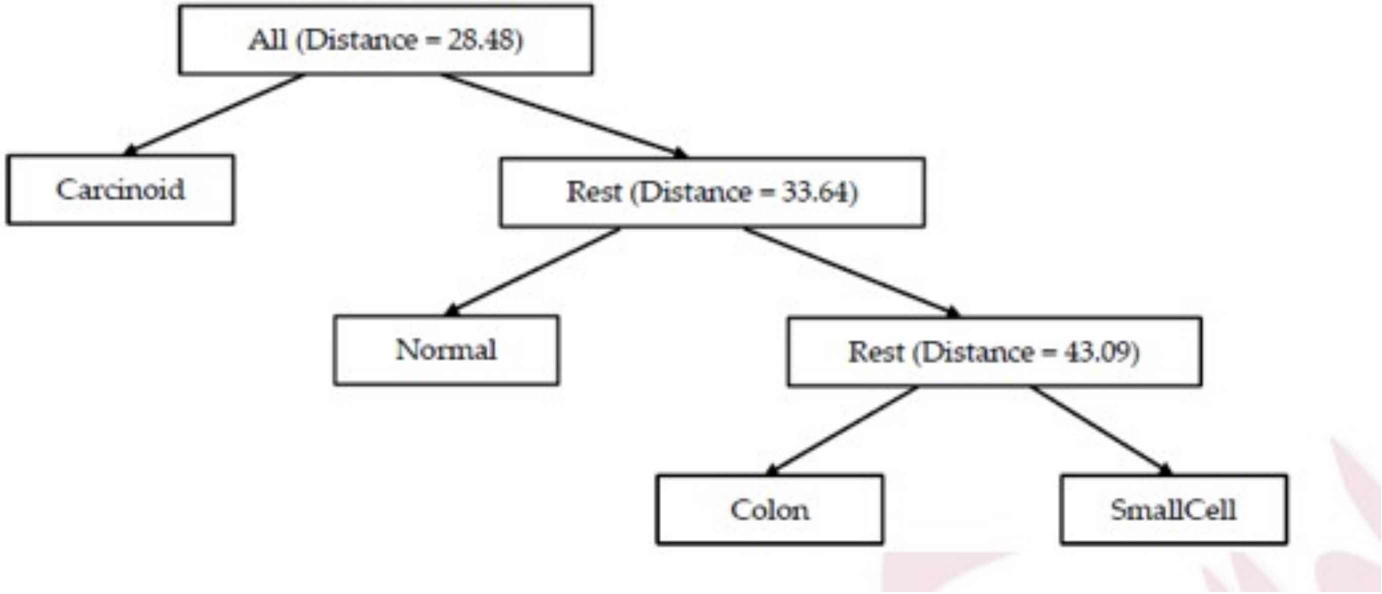
- **Maximal data piling (MDP) distance (Ahn and Marron, 2007)**
  - The orthogonal distance between the affine subspaces generated by the data vectors in each cluster.



$$\boldsymbol{U}_1^{(p)}, \boldsymbol{U}_2^{(p)}, \boldsymbol{U}_3^{(p)} \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}_1^{(p)};\ \boldsymbol{V}^{(p)} \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}_2^{(p)}$$

2. Previous study (MDP clustering)

- **Clustering with MDP distance (Ahn, et al., 2013)**
  - Find successive binary split, each of which creates two clusters in such a way that the MDP distance between them is as large as possible.



2. Previous study (MDP clustering)

- **MDP distance Clustering (Ahn, et al., 2013)**
  - A sufficient condition for the label consistency

$$\delta_{12}^2 + \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} > \max\left\{\frac{n_1+G}{n_1 G}\sigma_1^2, \frac{n_2+G}{n_2 G}\sigma_2^2\right\}$$

  - where $G \leq \max\{n_1, n_2\}$.
  - If $\delta_{12}^2 > 0$ is sufficient large, the label consistency holds.

2. Previous study (MDP clustering)

- **Some problems of MDP clustering**
  - The sufficient condition depends on variances (and sample sizes.
  - Cannot discover differences between variances in each cluster.

Avoiding stereotypes of clustering method,
we can conduct simple and effective methods based on
a distance matrix or an inner product matrix.

3. Clustering with distance vectors

3.1 Main idea

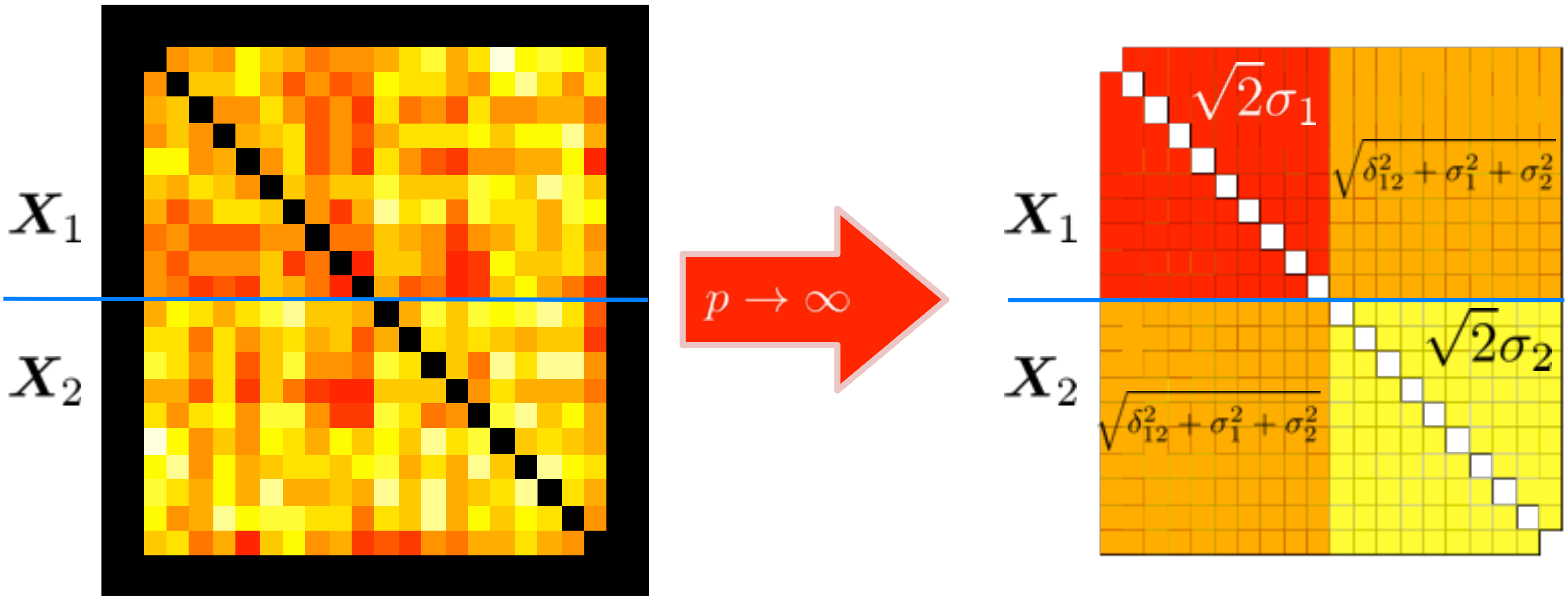
– Toy example:

- $\boldsymbol{X}_1 \sim N_p(\mathbf{0},\ I_p)$
- For $c \neq 1$, $\boldsymbol{X}_2 \sim N_p(\mathbf{0},\ cI_p)$,
- The condition of Ahn et al. (2013) dose not hold.

$$\xcancel{\delta_{12}^2 + \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} > \max\left\{\frac{n_1+G}{n_1 G}\sigma_1^2, \frac{n_2+G}{n_2 G}\sigma_2^2\right\}}$$

3. Clustering with distance vectors

3.1 Main idea



Standardized distances converge to some constants in prob.

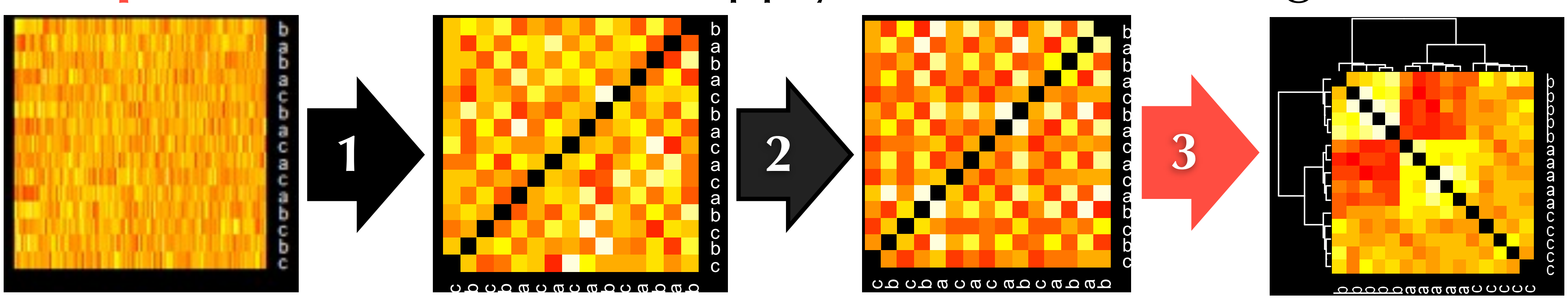
Distance "*vectors*" have the cluster information!

3. Clustering with distance vectors

- **Proposed method**
  - Step 1. Compute the distance matrix D from the data matrix X (or the inner product matrix $S := XX^T$).
  - Step 2. Calculate the following distances ($\Xi := (\xi_{ij})_{n \times n}$),

$$\xi_{ij} = \sqrt{\sum_{s \neq i,\, s \neq j} (d_{is} - d_{js})^2} \quad \left(\text{or } := \sqrt{\sum_{t \neq i,\, t \neq j} (s_{it} - s_{jt})^2} \right).$$

  - <u>**Step 3.**</u> For the matrix Ξ, apply a usual clustering method.



Data matrix X — Inner product S — Distance matrix Ξ — Inner product S

3. Clustering with distance vectors

- ***K*-means Type**

$$Q_p(\mathscr{C} \mid K) := \sum_{i=1}^{N} \min_k \sum_{j \neq i} \left(d_{ij}^{(p)} - \bar{d}_{kj}^{(p)} \right)^2 ,$$

where $\bar{d}_{kj}^{(p)} := \frac{1}{n_k - 1} \sum_{i \neq j} d_{ij}^{(p)}.$

  – We can optimize this by the usual *k*-means algorithm.

- **Important property**

  – Under the assumptions a) $\sim$ f), for all $K^* \geq K$,

$$\min_{\mathscr{C}} Q_p(\mathscr{C} \mid K^*) \xrightarrow{\mathbb{P}} 0 \quad \text{as} \quad p \to \infty.$$

3. Clustering with distance vectors

- **Theoretical results of the *k*-means type**
  - **In the case of using a distance matrix**

Assume a) $\sim$ f) .

If $\forall k,\ l\ (k \neq l);\ \sigma_k \neq \sigma_l \text{ or } \delta_{kl}^2 > 0,$

then the estimate label vector converges to the true label vector in probability as $p \to \infty$.

- Ahn et al., 2013

$$\delta_{12}^2 + \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} > \max\left\{\frac{n_1+G}{n_1 G}\sigma_1^2, \frac{n_2+G}{n_2 G}\sigma_2^2\right\}$$

3. Clustering with distance vectors

- **Theoretical results of the *k*-means type**
  - **In the case of using an inner product matrix**

Assume a) $\sim$ f) .

If $\delta_{12}^2 > 0,$

then the estimate label vector converges to the true label vector in probability as $p \to \infty$.

– Ahn et al., 2013

$$\delta_{12}^2 + \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} > \max\left\{ \frac{n_1 + G}{n_1 G}\sigma_1^2, \frac{n_2 + G}{n_2 G}\sigma_2^2 \right\}$$

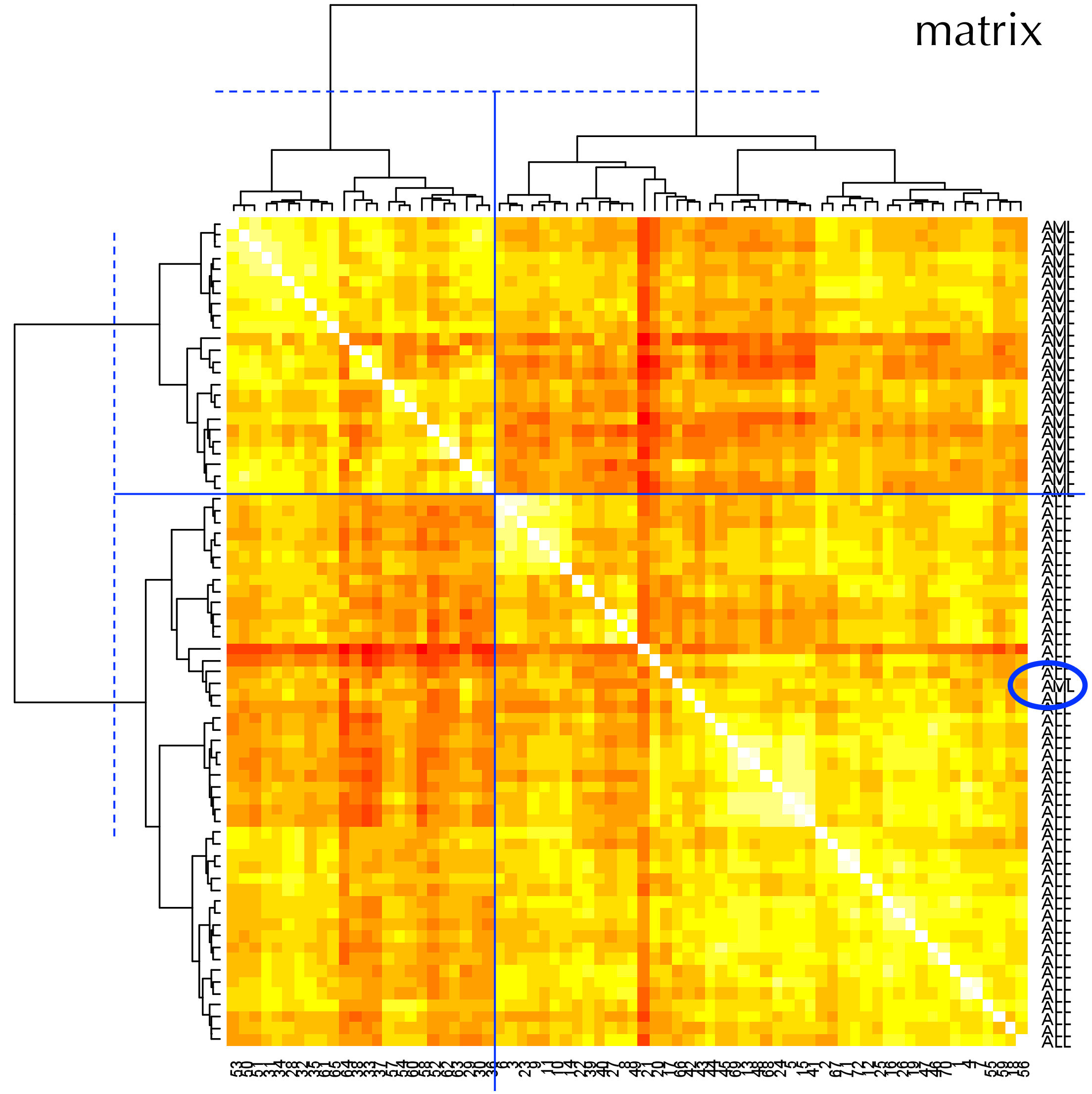
3. Clustering with distance vectors

Application to Cancer Microarray data (Leukemia data)

- **Summary:**
  - The number of labels：2
  - Sample size：72
  - Dimensions：3571

- **Comparison**
  - Reg. *k*-means: 2/72 (2)
  - MDP: 2/72 (2)
  - Proposal method：1/72 (2)

Inner product matrix



4. Conclusion

Thank you for your attention!

- In this presentaion,
  - Introduce geometric representations of HDLSS data,
  - Propose a new efficient clustering method for HDLSS data.
- Remark:
  - In HDLSS contexts,

    the closeness between data points may not be meaningful,

    but "*vectors*" of distances have the cluster information!

Reference

[01] Ahn, J. , Lee, M.H., and Yoon, Y.J. (2013). Clustering high dimension, low sample size data using the maximal data piling distance, *Statist. Sinica*.

[02] Ahn, J., Marron, J.S., Muller, K.M., and Chi, Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild condition, *Biometrika*, **94** (3), 760 –766.

[03] Hall, P., Marron, J.S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data, *J. R. Statist. Soc. B*, **67** (3), 427 – 444.

[04] Kolmogorov, A.N. and Rozanov, Y.A. (1960). On strong mixing conditions for stationary Gaussian processes. *Theor. Probab. Appl.* **5**, 204-208.

[05] Sun, W., Wang, J., and Fang, Y. (2012). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electron. J. Statist*. **6**, 148–167.

A. Definition of ρ-mixing

- **ρ-mixing** (Kolmogorov and Rozanov, 1960; Theor. Probab. Appl.)
  - For $-\infty \leq J \leq L \leq \infty$,

    $\mathcal{F}_J^L$: the σ-field of events generated by the r.v.s $(Z_i,\ J \leq i \leq L)$.
  - For any σ-field $\mathcal{A}$,

    $L_2(\mathcal{A})$: the space of square-integrable, $\mathcal{A}$-measurable r.v.s.
  - For each $m \geq 1$, define the maximal correlation coefficient

    $$\rho(m) := \sup |\mathrm{Corr}(f,\ g)|, \quad f \in L_2(\mathcal{F}_{-\infty}^{j}),\ g \in L_2(\mathcal{F}_{j+m}^{\infty}),$$

    where $j \in \mathbb{Z}$.
  - The sequence $\{Z_i\}$ is said to be ρ**-mixing** if

    $$\rho(m) \to 0 \quad \text{as} \quad m \to \infty.$$