The Informativeness of k-Means for Learning Mixture Models

Vincent Y. F. Tan
(Joint work with Zhaoqiang Liu)

National University of Singapore

June 18, 2018

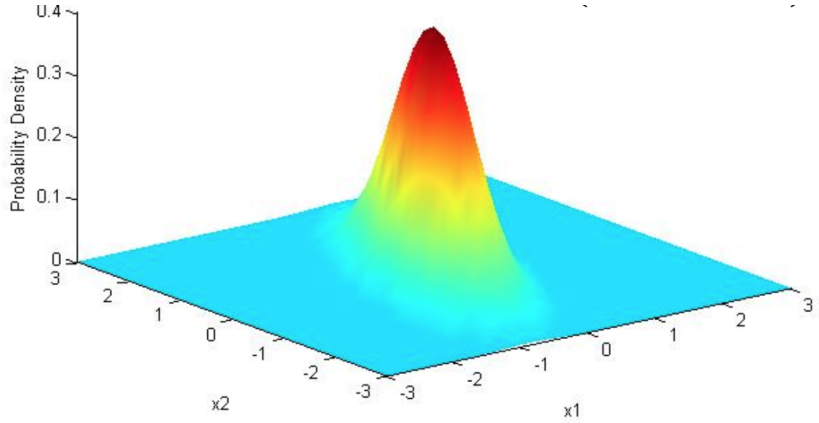
Gaussian distribution

For F dimensions, the Gaussian distribution of a vector $\mathbf{x} \in \mathbb{R}^F$ is defined by:

$$\mathcal{N}(\mathbf{x}|\mathbf{u}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{F/2}\sqrt{|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{u})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{u})\right),$$

where $\mathbf{u}$ is the mean vector, $\mathbf{\Sigma}$ is the covariance matrix of the Gaussian.

Example: Mean $\mathbf{u} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and Covariance matrix $\mathbf{\Sigma} = \begin{bmatrix} 0.25 & 0.3 \\ 0.3 & 1.0 \end{bmatrix}$



Gaussian mixture model (GMM)

$$\mathbb{P}(\mathbf{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x}|\mathbf{u}_k, \mathbf{\Sigma}_k).$$

- w_k: mixing weight
- $\mathbf{u}_k$: component mean vector
- $\mathbf{\Sigma}_k$: component covariance matrix; if $\mathbf{\Sigma}_k = \sigma_k^2 \mathbf{I}$, the GMM is said to be spherical

Learning GMM

Data samples independently generated from a GMM $\Rightarrow$ Correct target clustering of the samples according to which Gaussian distribution they are generated from

Learning GMM

Data samples independently generated from a GMM $\Rightarrow$
Correct target clustering of the samples according to which Gaussian distribution they are generated from

Definition 1 (correct target clustering)

Suppose

$$\mathbf{V} := [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$$

are samples independently generated from a K-component GMM.

Learning GMM

Data samples independently generated from a GMM $\Rightarrow$
Correct target clustering of the samples according to which Gaussian distribution they are generated from

Definition 1 (correct target clustering)

Suppose

$$\mathbf{V} := [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$$

are samples independently generated from a K-component GMM. The correct target clustering

$$\mathscr{I} := \{\mathscr{I}_1, \mathscr{I}_2, \ldots, \mathscr{I}_K\}$$

satisfies $n \in \mathscr{I}_k$ iff $\mathbf{v}_n$ comes from the k-th component.

Learning GMM

Data samples independently generated from a GMM $\Rightarrow$
Correct target clustering of the samples according to which Gaussian distribution they are generated from

Definition 1 (correct target clustering)

Suppose

$$\mathbf{V} := [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$$

are samples independently generated from a K-component GMM. The correct target clustering

$$\mathscr{I} := \{\mathscr{I}_1, \mathscr{I}_2, \ldots, \mathscr{I}_K\}$$

satisfies $n \in \mathscr{I}_k$ iff $\mathbf{v}_n$ comes from the k-th component.

Thereby inferring the important parameters of the GMM.

Algorithms for learning GMM

i) Expectation Maximization (EM)

- A local-search heuristic approach for maximum likelihood estimation in the presence of incomplete data;
- Cannot guarantee the convergence to global optima.

ii) Algorithms based on spectral decomposition and method of moments;

Definition 2 (non-degeneracy condition)

The mixture model is said to satisfy a non-degeneracy condition if the component mean vectors

$$\mathbf{u}_1, \ldots, \mathbf{u}_K$$

span a K-dimensional subspace, and the mixing weight $w_k > 0$, for $k \in \{1, 2, \ldots, K\}$.

iii) Algorithms proposed in theoretical computer science with guarantees;

Need to assume separability assumptions.

Vempala and Wang [2002]: for any $i, j \in [K]$, $i \neq j$,

$$\|\mathbf{u}_i - \mathbf{u}_j\|_2 > C \max\{\sigma_i, \sigma_j\} K^{\frac{1}{4}} \log^{\frac{1}{4}}\left(\frac{F}{w_{\min}}\right).$$

iii) Algorithms proposed in theoretical computer science with guarantees;

Need to assume separability assumptions.

Vempala and Wang [2002]: for any $i, j \in [K]$, $i \neq j$,

$$\|\mathbf{u}_i - \mathbf{u}_j\|_2 > C \max\{\sigma_i, \sigma_j\} K^{\frac{1}{4}} \log^{\frac{1}{4}}\left(\frac{F}{w_{\min}}\right).$$

A simple spectral algorithm with running time polynomial in both F and K works well for correctly clustering samples.

The k-means algorithm

Large number of algorithms for finding the (approximately) correct clustering of GMM;

The k-means algorithm

Large number of algorithms for finding the (approximately) correct clustering of GMM;

Many practitioners stick with k-means algorithm because of its simplicity and successful applications in various fields.

The objective function of k-means

Objective function: the so-called sum-of-squares distortion.

$$\mathcal{D}(\mathbf{V}, \mathscr{I}) := \sum_{k=1}^{K} \sum_{n \in \mathscr{I}_k} \|\mathbf{v}_n - \mathbf{c}_k\|_2^2,$$

where

- $\mathscr{I}_k$: the index set of k-th cluster;
- $\mathbf{c}_k := \frac{1}{|\mathscr{I}_k|} \sum_{n \in \mathscr{I}_k} \mathbf{v}_n$ is the centroid of the k-th cluster.

The objective function of k-means

Objective function: the so-called sum-of-squares distortion.

$$\mathcal{D}(\mathbf{V}, \mathscr{I}) := \sum_{k=1}^{K} \sum_{n \in \mathscr{I}_k} \|\mathbf{v}_n - \mathbf{c}_k\|_2^2,$$

where

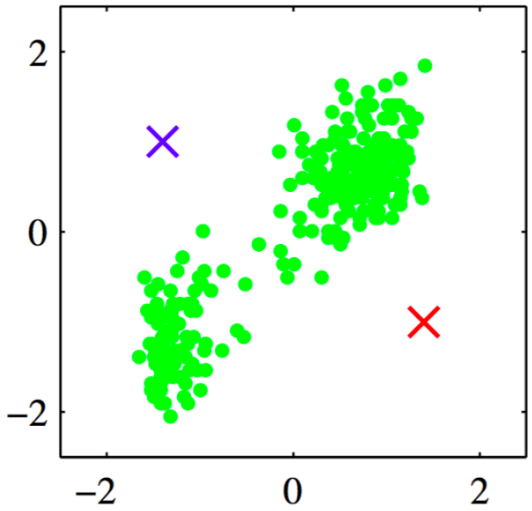
- $\mathscr{I}_k$: the index set of k-th cluster;
- $\mathbf{c}_k := \frac{1}{|\mathscr{I}_k|} \sum_{n \in \mathscr{I}_k} \mathbf{v}_n$ is the centroid of the k-th cluster.

Finding an optimal clustering $\mathscr{I}^{\mathrm{opt}}$ that satisfies

$$\mathcal{D}(\mathbf{V}, \mathscr{I}^{\mathrm{opt}}) = \min_{\mathscr{I}} \mathcal{D}(\mathbf{V}, \mathscr{I}) =: \mathcal{D}^*(\mathbf{V}).$$

k-Means: By Example

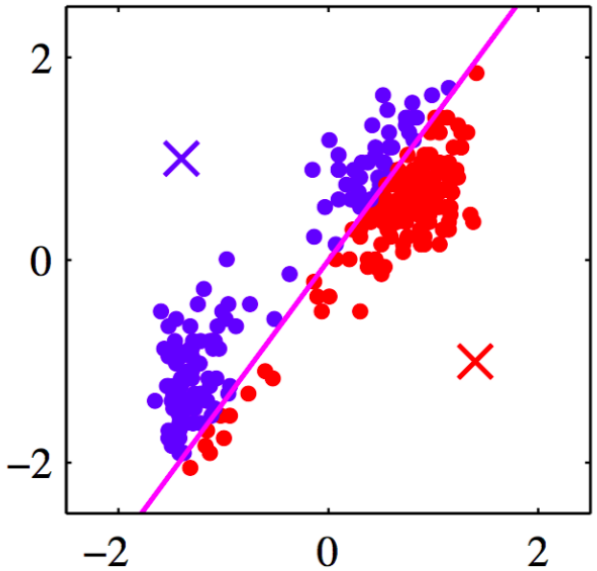
- Standardize the data.
- Choose two cluster centers.



From Bishop's *Pattern recognition and machine learning*, Figure 9.1(a).

k-means algorithm

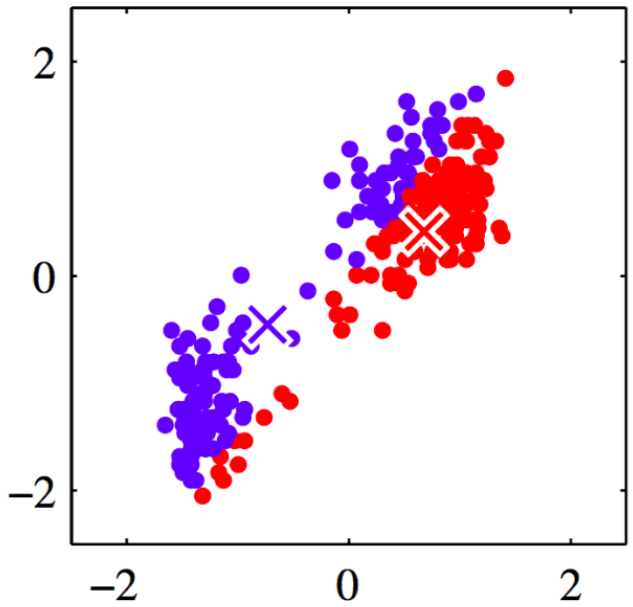
- Assign each point to closest center.



From Bishop's *Pattern recognition and machine learning*, Figure 9.1(b).

k-means algorithm

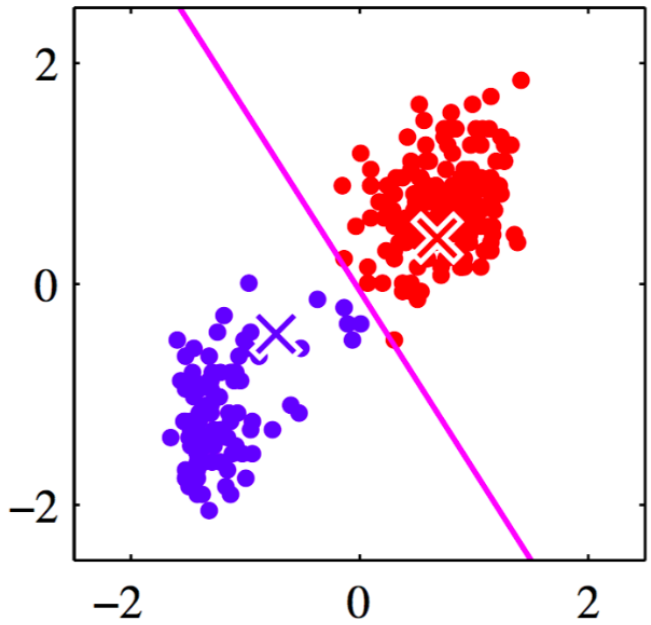
- Compute new class centers.



From Bishop's *Pattern recognition and machine learning*, Figure 9.1(c).

k-means algorithm

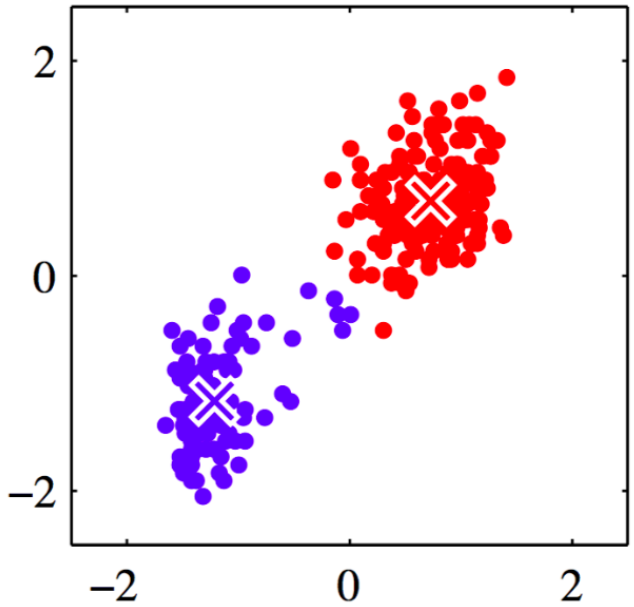
- Assign points to closest center.



From Bishop's *Pattern recognition and machine learning*, Figure 9.1(d).

k-means algorithm

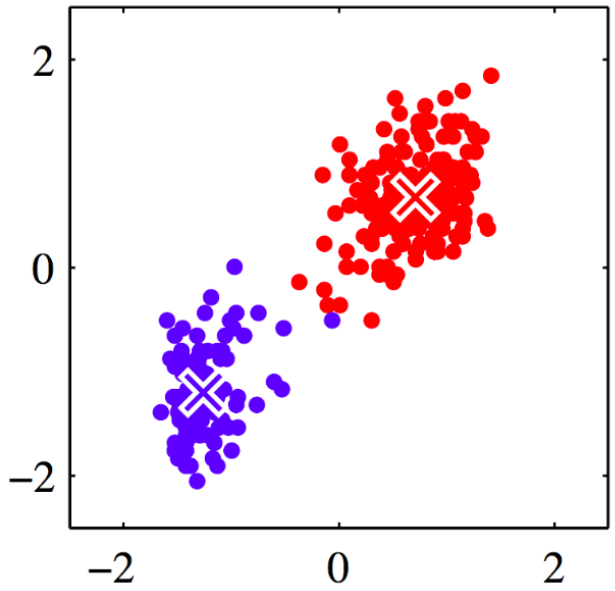
- Compute cluster centers.



From Bishop's *Pattern recognition and machine learning*, Figure 9.1(e).

k-means algorithm

- Iterate until convergence.



From Bishop's *Pattern recognition and machine learning*, Figure 9.1(i).

Using k-means to learn GMM?

Can we simply use k-means to learn the correct clustering of GMM?

Using k-means to learn GMM?

Can we simply use k-means to learn the correct clustering of GMM?

Yes!

Kumar and Kannan [2010] showed that if:

Data points satisfy a proximity condition, i.e., when they independently generated from a GMM with a certain separability assumption

$\Rightarrow$

Using k-means to learn GMM?

Can we simply use k-means to learn the correct clustering of GMM?

Yes!

Kumar and Kannan [2010] showed that if:

Data points satisfy a proximity condition, i.e., when they independently generated from a GMM with a certain separability assumption

$\Rightarrow$

k-means algorithm with a proper initialization can correctly cluster nearly all data points with high probability

Using k-means to learn GMM?

The key condition to be satisfied for performing k-means to learn the parameters of a GMM?

Using k-means to learn GMM?

The key condition to be satisfied for performing k-means to learn the parameters of a GMM?

The correct clustering $\approx$ Any optimal clustering

Main contributions

We prove if

- data points generated from a K-component spherical GMM;
- non-degeneracy condition and an separability assumption;

The correct clustering $\approx$ Any optimal clustering

Main contributions

We also prove if

- data points generated from a K-component spherical GMM;
- projected onto the low-dimensional space;
- non-degeneracy condition and an even weaker separability assumption;

The correct clustering $\approx$ Any optimal clustering for the dimensionality-reduced dataset

Advantages of dimensionality reduction

- Significantly faster running time
- Reduced memory usage
- Weaker separability assumption
- Other advantages

Lower bound of distortion

Let $\mathbf{Z}$ be the centralized data matrix of $\mathbf{V}$ and denote $\mathbf{S} = \mathbf{Z}^T\mathbf{Z}$. According to Ding and He [2004], for any K-clustering $\mathscr{I}$,

$$\mathcal{D}(\mathbf{V}, \mathscr{I}) \geq \mathcal{D}^*(\mathbf{V}) := \mathrm{tr}(\mathbf{S}) - \sum_{k=1}^{K-1} \lambda_k(\mathbf{S}),$$

where

$$\lambda_1(\mathbf{S}) \geq \lambda_2(\mathbf{S}) \geq \ldots \geq 0$$

are the sorted eigenvalues of $\mathbf{S}$.

Misclassification error (ME) distance

Definition 3 (ME distance)

The misclassification error distance of any two K-clusterings

$$\mathscr{I}^1 := \{\mathscr{I}_1^1, \mathscr{I}_2^1, \ldots, \mathscr{I}_K^1\}, \quad \text{and}$$
$$\mathscr{I}^2 := \{\mathscr{I}_1^2, \mathscr{I}_2^2, \ldots, \mathscr{I}_K^2\}$$

is defined as

$$\mathrm{d}(\mathscr{I}^1, \mathscr{I}^2) := 1 - \frac{1}{N} \max_{\pi \in \mathcal{P}_K} \sum_{k=1}^{K} |\mathscr{I}_k^1 \bigcap \mathscr{I}_{\pi(k)}^2|,$$

where $\pi \in \mathcal{P}_K$ represents that the distance is minimized over all permutations of the labels $\{1, 2, \ldots, K\}$.

Meilă [2005]: ME distance defined above is indeed a metric.

Important lemma

Lemma 1 (Meilă, 2006)

- *Given a partition* $\mathscr{I} := \{\mathscr{I}_1, \mathscr{I}_2, \ldots, \mathscr{I}_K\}$ *and a dataset* $\mathbf{V}$*;*

Important lemma

Lemma 1 (Meilă, 2006)

- *Given a partition* $\mathscr{I} := \{\mathscr{I}_1, \mathscr{I}_2, \ldots, \mathscr{I}_K\}$ *and a dataset* $\mathbf{V}$*;*
- *Let*

$$p_{\max} := \max_k \frac{1}{N}|\mathscr{I}_k|, \quad \textit{and} \quad p_{\min} := \min_k \frac{1}{N}|\mathscr{I}_k|$$

and

$$\delta := \frac{\mathcal{D}(\mathbf{V}, \mathscr{I}) - \mathcal{D}^*(\mathbf{V})}{\lambda_{K-1}(\mathbf{S}) - \lambda_K(\mathbf{S})}, \quad \textit{where} \quad \mathcal{D}^*(\mathbf{V}) := \min_{\mathscr{I}} \mathcal{D}(\mathbf{V}, \mathscr{I}).$$

Important lemma

Lemma 1 (Meilă, 2006)

- *Given a partition* $\mathscr{I} := \{\mathscr{I}_1, \mathscr{I}_2, \ldots, \mathscr{I}_K\}$ *and a dataset* $\mathbf{V}$*;*
- *Let*

$$p_{\max} := \max_k \frac{1}{N}|\mathscr{I}_k|, \quad \textit{and} \quad p_{\min} := \min_k \frac{1}{N}|\mathscr{I}_k|$$

and

$$\delta := \frac{\mathcal{D}(\mathbf{V}, \mathscr{I}) - \mathcal{D}^*(\mathbf{V})}{\lambda_{K-1}(\mathbf{S}) - \lambda_K(\mathbf{S})}, \quad \textit{where} \quad \mathcal{D}^*(\mathbf{V}) := \min_{\mathscr{I}} \mathcal{D}(\mathbf{V}, \mathscr{I}).$$

- *If*

$$\delta \leq \frac{K-1}{2} \quad \textit{and} \quad \tau(\delta) := 2\delta\left(1 - \frac{\delta}{K-1}\right) \leq p_{\min},$$

then

$$\mathrm{d}(\mathscr{I}, \textbf{optimal}) \leq p_{\max}\tau(\delta).$$

Definitions

Define the increasing function

$$\zeta(p) := \frac{p}{1 + \sqrt{1 - 2p/(K-1)}},$$

the average variances

$$\bar{\sigma}^2 := \sum_{k=1}^{K} w_k \sigma_k^2$$

and the minimum eigenvalue

$$\lambda_{\min} := \lambda_{K-1} \left(\sum_{k=1}^{K} w_k (\mathbf{u}_k - \bar{\mathbf{u}})(\mathbf{u}_k - \bar{\mathbf{u}})^T \right).$$

Theorem for original datasets

Theorem 1

- *Dataset* $\mathbf{V} \in \mathbb{R}^{F \times N}$ *consisting of samples generated from a* K*-component spherical GMM* ($N > F > K$);

Theorem for original datasets

Theorem 1

- *Dataset* $\mathbf{V} \in \mathbb{R}^{F \times N}$ *consisting of samples generated from a* K*-component spherical GMM* ($N > F > K$);
- *The non-degeneracy condition;*

Theorem for original datasets

Theorem 1

- *Dataset* $\mathbf{V} \in \mathbb{R}^{F \times N}$ *consisting of samples generated from a* K*-component spherical GMM* ($N > F > K$);
- *The non-degeneracy condition;*
- *Let*

$$w_{\min} := \min_k w_k, \quad \text{and} \quad w_{\max} := \max_k w_k$$

and assume

$$\delta_0 := \frac{(K-1)\bar{\sigma}^2}{\lambda_{\min}} < \zeta(w_{\min}).$$

Theorem for original datasets

Theorem 1

- *Dataset* $\mathbf{V} \in \mathbb{R}^{F \times N}$ *consisting of samples generated from a* K*-component spherical GMM* ($N > F > K$);
- *The non-degeneracy condition;*
- *Let*

$$w_{\min} := \min_k w_k, \quad \textit{and} \quad w_{\max} := \max_k w_k$$

and assume

$$\delta_0 := \frac{(K-1)\bar{\sigma}^2}{\lambda_{\min}} < \zeta(w_{\min}).$$

For sufficiently large N, *w.h.p.,*

$$\mathrm{d}(\textbf{correct}, \textbf{optimal}) \leq \tau\left(\delta_0\right) w_{\max}.$$

Remark for separability assumption

Remark 1

The condition $\delta_0 < \zeta(w_{\min})$ can be considered as a separability assumption. For example,

- *$K = 2$ implies that*

$$\lambda_{\min} = w_1 w_2 \|\mathbf{u}_1 - \mathbf{u}_2\|_2^2$$

and we have

$$\|\mathbf{u}_1 - \mathbf{u}_2\|_2 > \frac{\bar{\sigma}}{\sqrt{w_1 w_2 \zeta(w_{\min})}}.$$

Remark for non-degeneracy condition

Remark 2

The non-degeneracy condition is used to ensure that $\lambda_{\min} > 0$.

- *For* $K = 2$, *we have*

$$\lambda_{\min} = w_1 w_2 \|\mathbf{u}_1 - \mathbf{u}_2\|_2^2$$

and we only need the two component mean vectors are distinct and we do not need that they are linearly independent.

Theorem for dimensionality-reduced datasets

Theorem 2

- $\mathbf{V} \in \mathbb{R}^{F \times N}$: *generated under the same conditions given in Theorem 1;*
- *The separability assumption being modified to*

$$\delta_1 := \frac{(K-1)\bar{\sigma}^2}{\lambda_{\min} + \bar{\sigma}^2} < \zeta(w_{\min}).$$

- $\tilde{\mathbf{V}} \in \mathbb{R}^{(K-1) \times N}$: *the post-*$(K-1)$*-PCA dataset of* $\mathbf{V}$.

Theorem for dimensionality-reduced datasets

Theorem 2

- $\mathbf{V} \in \mathbb{R}^{F \times N}$: *generated under the same conditions given in Theorem 1;*
- *The separability assumption being modified to*

$$\delta_1 := \frac{(K-1)\bar{\sigma}^2}{\lambda_{\min} + \bar{\sigma}^2} < \zeta(w_{\min}).$$

- $\tilde{\mathbf{V}} \in \mathbb{R}^{(K-1) \times N}$: *the post-*$(K-1)$*-PCA dataset of* $\mathbf{V}$.

For sufficiently large N*, w.h.p.,*

$$\mathrm{d}(\mathbf{correct}, \tilde{\mathbf{optimal}}) \leq \tau\left(\delta_1\right) w_{\max}.$$

Upper bound for ME distance between optimal clusterings

Combining the results of Theorem 1 and Theorem 2, by the triangle inequality:

Corollary 1

- $\mathbf{V} \in \mathbb{R}^{F \times N}$: *generated under the same conditions given in Theorem 1;*
- $\tilde{\mathbf{V}}$: *the post-*$(K-1)$*-PCA dataset of* $\mathbf{V}$.

For sufficiently large N*, w.h.p.*

$$\mathrm{d}(\mathbf{optimal}, \tilde{\mathbf{optimal}}) \leq \left(\tau\left(\delta_0\right) + \tau\left(\delta_1\right)\right) w_{\max}.$$

Parameter settings

$K = 2$, for all $k = 1, 2$, we set

$$\sigma_k^2 = \frac{\lambda_{\min}\zeta(w_{\min} - \varepsilon)}{4(K-1)}, \text{ corr. to } \frac{\delta_0}{\zeta(w_{\min})} \approx \frac{1}{4},$$

or

$$\sigma_k^2 = \frac{\lambda_{\min}\zeta(w_{\min} - \varepsilon)}{K-1}, \text{ corr. to } \frac{\delta_0}{\zeta(w_{\min})} \approx 1,$$

where $\varepsilon = 10^{-6}$.

Parameter settings

$K = 2$, for all $k = 1, 2$, we set

$$\sigma_k^2 = \frac{\lambda_{\min}\zeta(w_{\min} - \varepsilon)}{4(K-1)}, \text{ corr. to } \frac{\delta_0}{\zeta(w_{\min})} \approx \frac{1}{4},$$

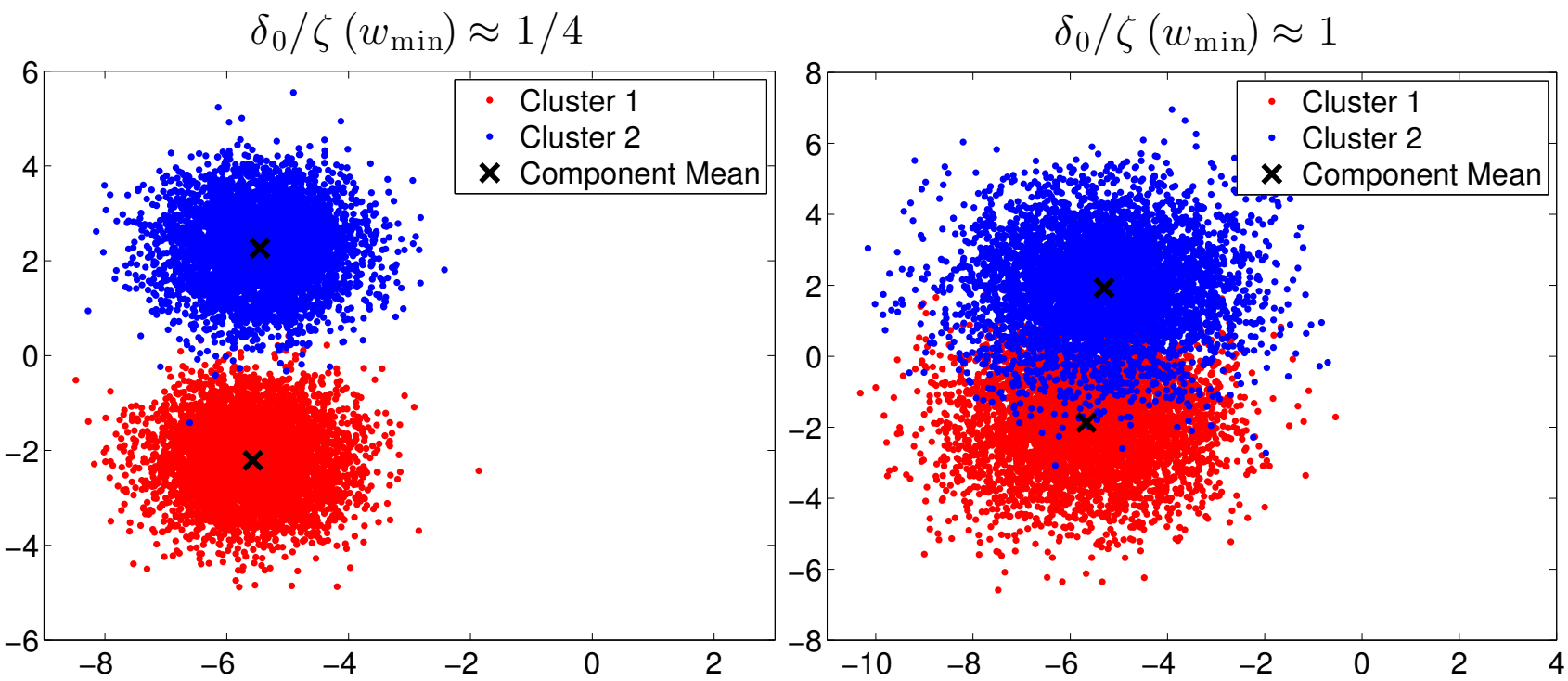
or

$$\sigma_k^2 = \frac{\lambda_{\min}\zeta(w_{\min} - \varepsilon)}{K-1}, \text{ corr. to } \frac{\delta_0}{\zeta(w_{\min})} \approx 1,$$

where $\varepsilon = 10^{-6}$.

The former corresponds to well-separated clusters; the latter corresponds to moderately well-separated clusters

Visualization of post-2-SVD datasets



Original datasets

$\mathrm{d}_{\mathrm{org}} := \mathrm{d}(\mathscr{I}, \mathscr{I}^{\mathrm{opt}})$,

$\bar{\mathrm{d}}_{\mathrm{org}} := \tau(\delta_0) w_{\max}$.

$\delta_0^{\mathrm{emp}} := \frac{\mathcal{D}(\mathbf{V},\mathscr{I}) - \mathcal{D}^*(\mathbf{V})}{\lambda_{K-1}(\mathbf{S}) - \lambda_K(\mathbf{S})}$ is an approximation of δ_0,

$\bar{\mathrm{d}}_{\mathrm{org}}^{\mathrm{emp}} := \tau(\delta_0^{\mathrm{emp}}) p_{\max}$ is an approximation of $\bar{\mathrm{d}}_{\mathrm{org}}$.

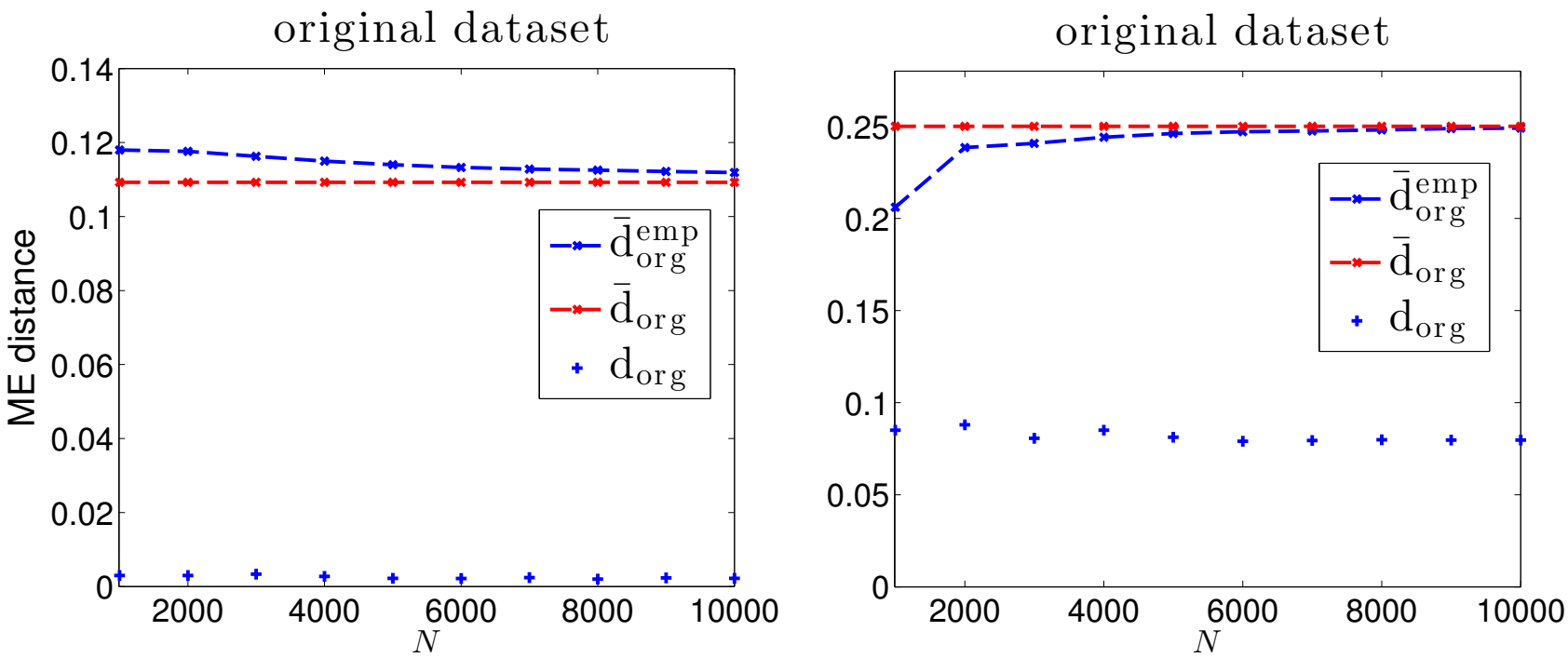


Figure: True distances and upper bounds for original datasets.

Dimensionality-reduced datasets

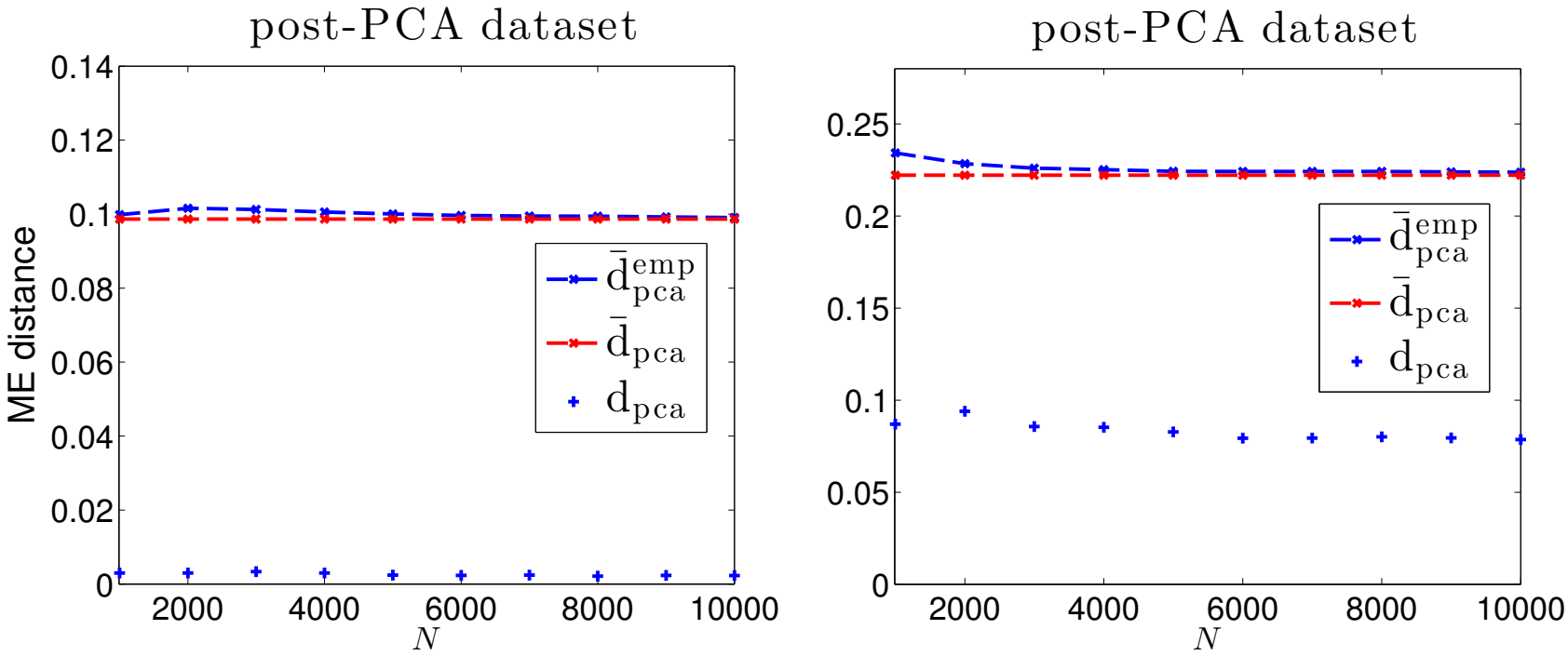
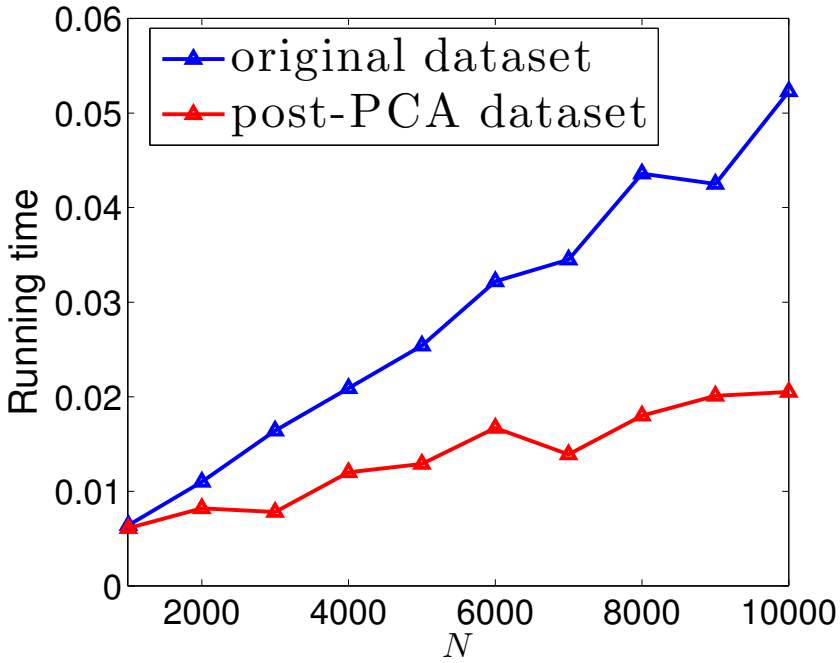
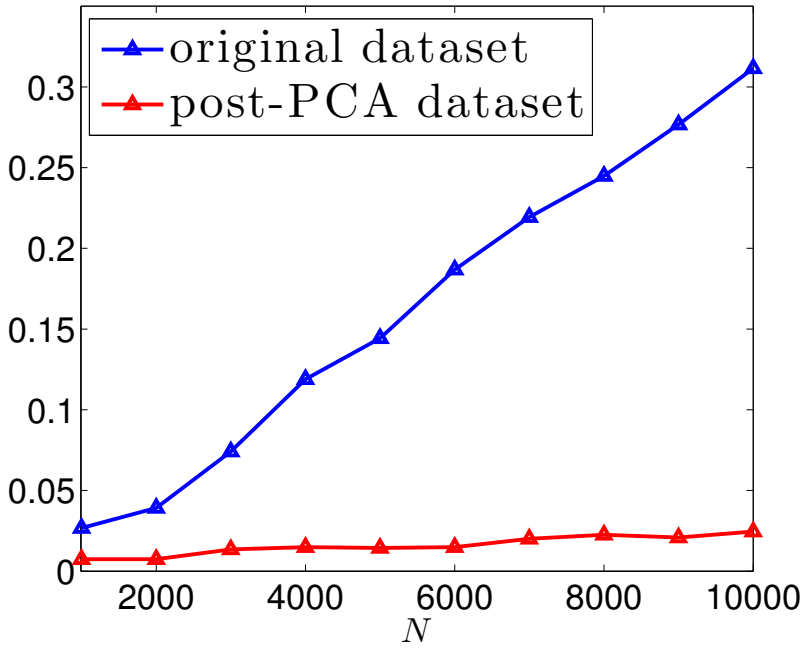


Figure: True distances and upper bounds for post-PCA datasets.

Comparisons of running time





Further extensions

- Randomized SVD instead of exact SVD;
- Random projection;
- Non-spherical Gaussian or even more general distributions, e.g., log-concave distributions;