Mixture Models and SVD

Monday, August 30, 2021 6:57 AM

Gaussian Mixture Model: $N(\mu_1, \Sigma_1)$, $N(\mu_2, \Sigma_2)$ . . . .

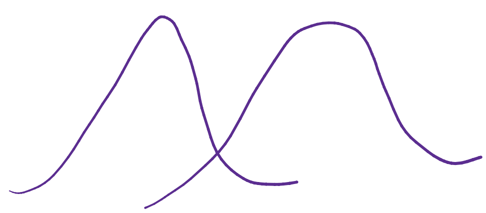
$\omega_1 \geq 0 \qquad \omega_2 \geq 0 \quad \ldots \omega_k$

$$\sum_{i=1}^{k} \omega_i = 1.$$

Problem: Given random samples from an unknown k-GMM estimate its parameters.

$k=1$: $\omega_1 = 1$, $\mu =$ Sample mean $\Sigma =$ sample covariance

$k=2$?



special case: Separable GMMs.

The components are pairwise separated.

How to measure separation.

1-dim $|\mu_i - \mu_j| > ?? \max\{\sigma_i, \sigma_j\}$

d-dim. (Geometric)

(probabilistic) $d_{TV}(F_i, F_j) \geq 1-\varepsilon$.

$$\boxed{d_{TV}(p,q) = \frac{1}{2}\int |p(x) - q(x)|\, dx}$$

d_{TV} "large"

Lemma. In 1-dim, d_{TV} "large"

$\Rightarrow$ either $\|\mu_i - \mu_j\|$ is large

or $\max\left\{ \frac{\sigma_i}{\sigma_j}, \frac{\sigma_j}{\sigma_i} \right\}$ is large.

Thm. (Concentration) $X \sim N(\mu, \sigma^2 I)$

$$\Pr\left(|X-\mu| > t\sigma\sqrt{d} \right) \le e^{-\frac{dt^2}{2}}$$

$$\Pr\left(\left| |X-\mu|^2 - d\sigma^2 \right| > t\sigma^2\sqrt{d} \right) \le 2e^{-t^2/8}$$

How to solve P1?
In the separable case, cluster and estimate each component separately.

P2. Cluster a sample from a K-GMM by component of origin.

How? Suppose mean separated

$X, Y \in F_i$: $\mathbb{E}(\|X-Y\|^2) = \mathbb{E}(\|X-\mu_i - (Y-\mu_i)\|^2)$

$= \mathbb{E}(\|X-\mu_i\|^2) + \mathbb{E}(\|Y-\mu_i\|^2) + 0.$

$x, y \in F_i$

$$= \mathbb{E}(\|x-\mu_i\|^2) + \mathbb{E}(\|y-\mu_i\|^2) + 0.$$
$$= 2d\sigma^2.$$

$x \in F_i,\ y \in F_j$

$$\mathbb{E}(\|x-y\|^2) = \mathbb{E}(\|x-\mu_i - (y-\mu_j) + \mu_i - \mu_j\|^2)$$
$$= \mathbb{E}(\|x-\mu_i\|^2) + \mathbb{E}(\|y-\mu_j\|^2) + \|\mu_i-\mu_j\|^2$$
$$= 2d\sigma^2 + \|\mu_i-\mu_j\|^2.$$

By conc. with prob $\geq 1-2e^{-t^2/8}$

$$\|x-y\|^2 \leq 2d\sigma^2 + 2t\sqrt{d}\,\sigma^2 \qquad x, y \in F_i$$

$$\|x-y\|^2 > 2d\sigma^2 + \|\mu_i-\mu_j\|^2 - 2t\sqrt{d}\,\sigma^2. \qquad x \in F_i,\ y \in F_j$$

$\therefore$ it suffices to have

$$\|\mu_i-\mu_j\|^2 > 4t\sqrt{d}\,\sigma^2$$

to ensure that pairs from same Gaussian are closer.

<u>Cluster using distances</u>

- put nearest pair in same cluster
- Repeat till only k clusters.

<u>Thm</u>. with prob $1-\delta$, random sample with m points and $\|\mu_i-\mu_j\| > C\left(\log\frac{m}{\delta}\cdot d\right)^{\frac{1}{4}} \max\{\sigma_i, \sigma_j\}$

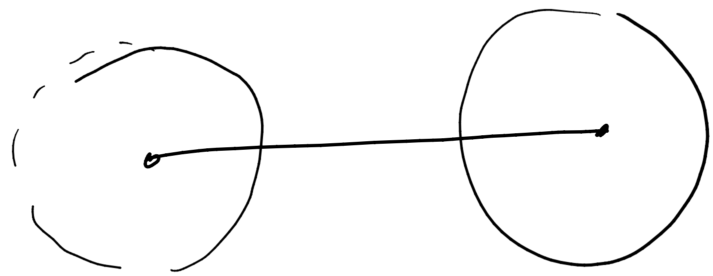
can be clustered using pairwise distances in

can be clustered using [illegible]
polynomial time.

Pf. Set $t = \sqrt{c \log \frac{m}{\delta}}$.

Is this the right answer?
Separation grows with $d^{1/4}$.

No!



Project to line joining μ_i, μ_j.
Separation needed is $O(\sigma)$. Not $d^{1/4}\sigma$.

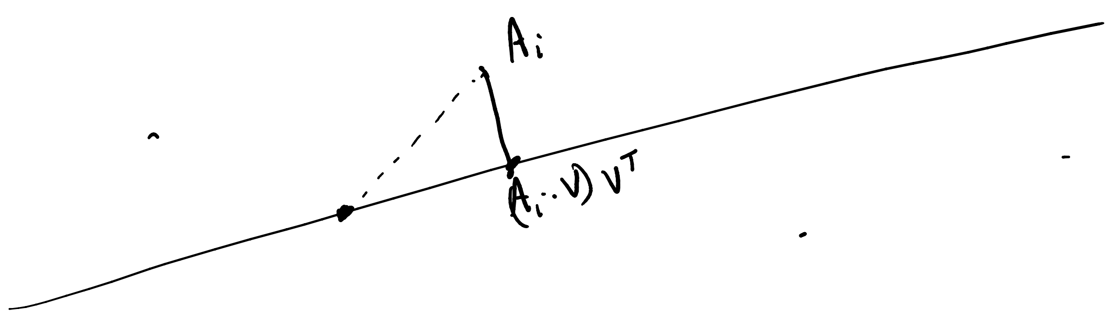
Q. But how to find line joining means?

Best fit line: $v = \underset{\|v\|=1}{\text{argmax}} \|Av\|^2$
maximizes sum of squared projections of rows of A.

For each row A_i

$$\|A_i\|^2 = (A_i v)^2 + \|A_i - (A_i v) v^T\|^2$$

$\|A_i\| = (A_i \cdot v) + \|A_i - (A_i \cdot v) v$



$$\arg\max \|Av\|^2 = \arg\min \|A - (Av)v^T\|^2$$

least squared error.

$A \in \mathbb{R}^{n \times d}$

u, v : left, right singular vectors of A

$Av = \sigma u \qquad \sigma \geq 0.$

$A^T u = \sigma v$

<u>Lema</u> $v = \arg\max_{\|v\|=1} \|Av\|^2$ is a right singular vector of A with largest singular value.

<u>Pf.</u> $A^T A v = A^T(\sigma u) = \sigma^2 v$

v is an eigen vector of $A^T A$.

goes both ways: $A^T A v = \sigma^2 v$

define $u = \frac{1}{\sigma} Av$

define $\quad u = \frac{1}{\sigma} A v$

Then $\quad A^T u = \frac{1}{\sigma} A^T A v = \sigma v.$

Now consider $f(v) = \|Av\|^2 = v^T A^T A v.$

$$\nabla_v f(v) = 2A^T A v$$

at any local max/min $\nabla_v f(v) = \lambda v$

$$\Rightarrow \quad A^T A v = \lambda v.$$

Hence the maximizer is an eigenvector.

$$\left.\begin{aligned} &\|Av\|^2 + \lambda(1 - \|v\|^2) \\ &2A^T A v + 2\lambda v = 0 \\ &\|v\|^2 = 1 \end{aligned}\right]$$

SVD.

$v_1 = \arg\max \|Av_1\| \qquad \sigma_1 = \|Av_1\|$

$v_2 = \underset{v_2 \perp v_1}{\arg\max} \|Av_1\|$

$\vdots$

$v_k = \underset{v_k \perp v_1 \ldots v_{k-1}}{\arg\max} \|Av_k\| \qquad \sigma_k.$

$v_k = \quad\quad k$
$v_k \perp v_1 \ldots v_{k-1}$

Thm (SVD) $V_k = \text{span}\{v_1 \ldots v_k\}$ satisfies

$$V_k = \underset{V:\, \dim(V) = k}{\arg\min} \sum_i d(A_{(i)}, V)^2$$

minimizes sum of squared distances among all k-dim subspaces. $v_1, \ldots v_k$ are singular vectors with singular values $\sigma_1 \ldots \sigma_k \ldots$

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T$$

Pf. By induction on k.

V_1 ✓

Suppose true for V_{k-1}.

Consider V_k' OPT k-dim subspace.

Let $w_1 \ldots w_k$ be an orthonormal basis of V_k' with $w_k \perp V_{k-1}$. ($\exists$ such a w_k).

$\therefore V_k'$ maximizes

$$\|Aw_1\|^2 + \|Aw_2\|^2 + \ldots \|Aw_k\|^2$$

$$\|Aw_1\|^2 + \|Aw_2\|^2 + \dots \|Aw_k\|^2$$

$$\leq \sum_{i=1}^{n} d(A_{(i)}, V_{k-1})^2 + \|Aw_k\|^2$$

$$= \|Av_1\|^2 + \dots + \|Av_{k-1}\|^2 + \|Av_k\|^2$$

Hence $V_k' = \text{Span } V_{k-1} \cup \{w_k\}$
WLOG

must be a maximizer.

But $w_k \perp V_{k-1}$ and

$\therefore \quad V_k = V_k'$.

$\left(\sum \sigma u_i v_i^T\right) v_j$ is the same as Av_j

Hence also for any $x \ (= \sum \alpha_j v_j)$. $\blacksquare$

Back to k-GMMs.

Thm (Mean subspace). V_k for a mixture of spherical Gaussians $\supseteq \text{Span}\{\mu_1 \dots \mu_k\}$.

Algorithm. - Project sample to top k-dim SVD subspace.
- Cluster according to distances in $\mathbb{R}^k$.

— cluster according to ...

Thm (SVD + Cluster). $|\mu_i - \mu_j| > C \left(\log \frac{m}{\delta} \cdot k\right)^{\frac{1}{4}} \max\{\sigma_i, \sigma_j\}$

suffices!

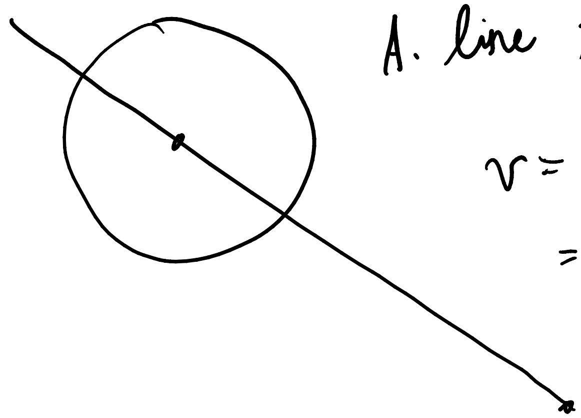
---

k instead of d.

Pf (of Thm. [Mean Subspace])

Suppose $k=1$. Q. What is the best subspace?

A. line through μ!



$$v = \arg\min \mathbb{E}\left(\|X - (X\cdot v)v\|^2\right)$$

$$= \arg\max \mathbb{E}\left(\|X\cdot v\|^2\right)$$

$$= \mathbb{E}\left[\|(X-\mu)\cdot v + \mu\cdot v\|^2\right]$$

$$= \sigma^2 + (\mu\cdot v)^2 + 0.$$

to maximize set $v = \frac{\mu}{\|\mu\|}$.

For 1 Gaussian

best k-d subspace is any subspace containing μ.

Eigenvalues of $\mathbb{E}(XX^T)$ are $\sigma^2 + \|\mu\|^2$

σ^2

$\vdots$

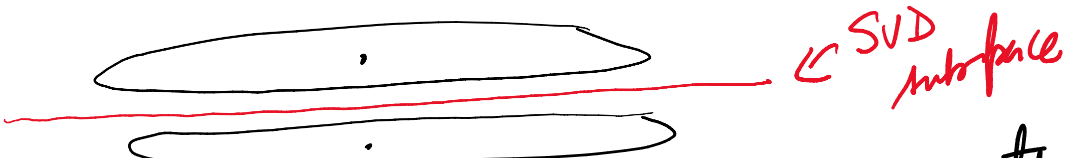
$\sigma^2.$

So for k Gaussians, best k-dim subspace is any subspace containing all their means!

Q. Does SVD work for general Gaussians

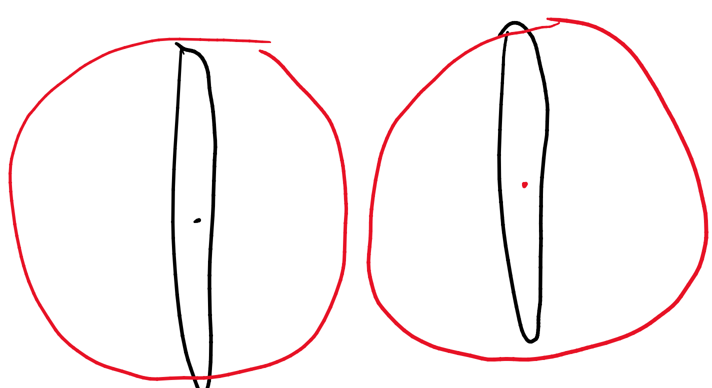
A. Only if the pairwise separation grows with the largest variance of each component.

E.g.



Projecting to SVD subspace simply merges the two "pancakes".

Need:



Two questions:

1) smaller separation?

2) general Gaussians?

1) [illegible]

2) general Gaussians?

Idea: SVD is about second moments. $\left(\mathbb{E}((X\cdot v)^2)\right)$

$\boxed{\text{Q. Would higher moments help?}}$

Mixture of k spherical Gaussians

$w_i \quad N(\mu_i, \sigma_i^2 I)$. $\boxed{\text{Assume } \{\mu_1 \dots \mu_k\} \text{ are lin. ind.}}$

① Make $\boxed{\text{isotropic}}$ i.e. $\mathbb{E}_F(X) = 0$

$$\mathbb{E}_F(XX^T) = I.$$

Thm [isotropic transformation]

Any distribution with bounded second moments can be made isotropic via an affine transformation.

Pf. Suppose $\mathbb{E}(X) = \mu$ $\quad \mathbb{E}((X-\mu)(X-\mu)^T) = A.$

Then $Y = X - \mu$ has $\mathbb{E}(Y) = 0$.

and $Y = A^{-1/2}(X-\mu)$ is isotropic!

$\mathbb{E}(Y) = 0 \qquad \mathbb{E}(YY^T) = A^{-1/2}\,\mathbb{E}\left((X-\mu)(X-\mu)^T\right)\left(A^{-\frac{1}{2}}\right)^T$

$$= A^{-\frac{1}{2}} A \left(A^{-\frac{1}{2}}\right)^T = I$$

$$= A^{-\frac{1}{2}} A (A^{-\frac{1}{2}})' = I$$

What is $A^{-1/2}$?

Note $\mathbb{E}((x-\mu)(x-\mu)^T) \succ 0$

Positive Definite matrix.

$\therefore A = BB^T$ for some B.

$A^{-1/2} = B^{-1}$

So $A^{-\frac{1}{2}} A (A^{-\frac{1}{2}})^T$

$B^{-1} BB^T (B^{-1})^T = I$.

So assume that $F = \sum_i w_i F_i$ is isotropic.

Next consider $\mathbb{E}(x \otimes x \otimes x) = T$

this is a 3-dim array, a tensor of size $d \times d \times d$.

$T_{ijk} = \mathbb{E}(x_i x_j x_k)$

$T = \sum_i w_i \mathbb{E}_{F_i}(x \otimes x \otimes x)$

So we need

$\mathbb{E}(x \otimes x \otimes x) \quad x \sim N(\mu, \sigma^2 I)$

$= \mathbb{E}((x-\mu+\mu) \otimes (x-\mu+\mu) \otimes (x-\mu+\mu))$

$= \mathbb{E}(\otimes^3 (x-\mu)) + \mathbb{E}((x-\mu) \otimes (x-\mu) \otimes \mu)$

$+ \mathbb{E}((x-\mu) \otimes \mu \otimes (x-\mu))$

$$+ \mathbb{E}\left((X-\mu) \otimes \mu \otimes (X-\mu)\right)$$

$$+ \mathbb{E}\left(\mu \otimes (X-\mu) \otimes (X-\mu)\right)$$

$$\left.\begin{array}{l} + \mathbb{E}\left((X-\mu) \otimes \mu \otimes \mu\right) \\ + \\ \vdots \\ + \end{array}\right\} = 0$$

$$+ \mu \times \mu \times \mu .$$

$$\mathbb{E}\left((X-\mu)_i (X-\mu)_j (X-\mu)_k\right) = \begin{cases} 0 & \text{if } i,j,k \text{ are not all equal} \\ \mathbb{E}((X-\mu_i)^3) & \text{if } i=j=k \\ = 0 . & \end{cases}$$

So we are left with

$$\mathbb{E}(X \otimes X \otimes X) = \sigma^2 I \otimes \mu + \mu \otimes \sigma^2 I$$

$$+ \mathbb{E}\left((X-\mu) \otimes \mu \otimes (X-\mu)\right)$$

$$+ \mu \otimes \mu \otimes \mu .$$

$$\mathbb{E}\left((X-\mu)_i \mu_j (X-\mu)_k\right) = \begin{cases} \sigma^2 \mu_j & \text{if } i=k \\ 0 & \text{o.w.} \end{cases}$$

$$= \left(\sigma^2 \sum^{d} e_i \otimes \mu \otimes e_i\right) \quad = \begin{cases} \sigma^2 \mu_j & \text{if } i=k=l \end{cases}$$

$$= \left(\sigma^2 \sum_{l=1}^{d} e_l \otimes \mu \otimes e_l \right)_{ijk} = \begin{cases} \sigma^2 \mu_j & \text{if } i=k=l \\ 0 & \text{o.w.} \end{cases}$$

Note $I = \sum_{l=1}^{d} e_l \otimes e_l$.

So,

<u>Lema</u> - (a) $\mathbb{E}(X \otimes X \otimes X) = \mu \times \mu \times \mu$

$X \sim N(\mu, \sigma^2 I)$

$$+ \sigma^2 \sum_{i=1}^{d} e_i \times e_i \times \mu + e_i \times \mu \times e_i + \mu \times e_i \times e_i .$$

(b) $X \sim \sum w_i N(\mu_i, \sigma_i^2 I)$

$$\mathbb{E}(X \otimes X \otimes X) = \sum_i w_i \mu_i \otimes \mu_i \times \mu_i$$

$$+ \underbrace{\sum_i w_i \sigma_i^2 \sum_{j=1}^{d} e_j \times e_j \times \mu_i + e_j \times \mu_i \times e_j + \mu_i \times e_j \times e_j}_{\text{Can we estimate this ?!}}$$

For any tensor $T = (T_{ijk})$ $x, y, z \in \mathbb{R}^d$

$T(x, y, z) = \sum_{i,j,k} T_{ijk} x_i y_j z_k$ ← scalar

$T(\cdot, y, z) = \sum_{j,k} T_{ijk} y_j z_j$ ← vector

matrix

$$T(\cdot, y, z) = \sum_{j,k} T_{ijk}\, y_j\, z_j$$

$$T(\cdot,\cdot,z) = \sum_{k} T_{ijk}\, z_j$$ ← matrix.

Consider $\mathbb{E}\big(X \otimes (X-\mu) \otimes (X-\mu)\big)[\cdot, v, v]$

$$= \mathbb{E}(X \otimes X \otimes X)[\cdot, v, v]$$
$$+ \ \mathbb{E}\big(X \otimes -\mu \otimes (X-\mu)\big)[\cdot, v, v]$$
$$+ \ \mathbb{E}\big(X \otimes (X-\mu) \otimes -\mu\big)[\cdot, v, v]$$

$\boxed{\text{We choose } \underset{\|v\|=1}{v} \perp \mu_i}$

$$= \mathbb{E}(X \otimes X \otimes X)[\cdot, v, v]$$

use Lemma (b):

$$= 0 + \underbrace{\sum_i w_i \sigma_i^2 \mu_i}_{\text{call this vector } u.}$$

<u>Then</u>

$$\mathbb{E}(X \otimes X \otimes X) = \sum w_i\, \mu_i \otimes \mu_i \otimes \mu_i$$
$$+ \underbrace{\sum_{j=1}^{d} u \otimes e_j \otimes e_j + e_j \otimes u \otimes e_j + e_j \otimes e_j \otimes u}_{\text{We know this!}}$$

and this!

So, we have $\quad T = \sum w_i\, \mu_i \otimes \mu_i \otimes \mu_i$

So, we have $\quad T = \sum w_i \mu_i \otimes \mu_i \otimes \mu_i$

and $\quad \mathbb{E}(xx^T) = \sum_i w_i \mu_i \otimes \mu_i + \sum_i w_i \sigma_i^2 I$

$\mathbb{E}((x-\mu)\cdot v)^2 = \sum w_i \sigma_i^2 = \hat{\sigma}^2$

$\therefore$ we also have $\sum w_i \mu_i \times \mu_i$

and by linear trasformation we have $= I$.

<u>Claim</u>. $\sum_{i=1}^{K} w_i \mu_i \times \mu_i = I \Rightarrow \mu_i$ are orthogonal

$\sqrt{w_i}\,\mu_i$ are orthonormal.

$\|\mu_i\|^2 = \frac{1}{w_i}$

$\sum a_i a_i^T = I$

$AA^T = I \quad \Rightarrow \|a_i\|^2 = 1 \quad a_i^T a_j = 0 \quad i \neq j$.

μ_i are orthogonal and we know $\sum w_i \mu_i \otimes \mu_i \otimes \mu_i$

Is this enough?

<u>Thm</u>. [Tensor Decomposition] Given $T = \sum_i \alpha_i u_i \otimes u_i \otimes u_i \quad \{u_i\}$ orthogonal

there is a polytime algorithm to recover α_i, u_i

<u>Pf.</u> Consider the iteration

Pf. Consider the iteration

$$x = \frac{T(\cdot, x, x)}{\|T(\cdot, x, x)\|}$$ starting with x_0 random.

assume $\|u_i\| = 1$

$x = \sum \beta_i u_i$

$$x^{(1)} \propto T(\cdot, x, x) = \sum_i \alpha_i \beta_i^2 u_i$$

$$x^{(2)} \propto T(\cdot, x^{(1)}, x^{(1)}) = \sum_i \alpha_i^3 \beta_i^4 u_i$$

$$= \sum_i \alpha_i^7 \beta_i^8 u_i$$

$$= \sum (\alpha_i \beta_i)^{2^k - 1} \cdot \beta_i \cdot u_i$$

So i with largest $\alpha_i \beta_i$ will quickly dominate!

$$x^{(t)} \longrightarrow u_i$$

Peel off and repeat.

(*) need to be careful about error accumulation.

So now we have an algorithm!.

$F = \sum w_i N(\mu_i, \sigma_i^2 I)$ $\{\mu_1 \dots \mu_k\}$ lin. ind

$F = \sum_i w_i N(\mu_i, \sigma_i^2 I)$ $\{\mu_1 \dots \mu_k\}$ lin. ind.

① $M = \mathbb{E}_S(X \otimes X)$ find top k eigenvectors. $v_1, v_2 \dots v_k$.

$\hat{\sigma}^2 = (k+1)^{st}$ eigenvalue

② $(M - \hat{\sigma}^2 I) = WW^T$

compute $\hat{S} = W^{-1}S \leftarrow$ (sample)

③ $v \perp \{W^{-1}v_1, \dots W^{-1}v_k\}$

$$u = \mathbb{E}_{\hat{S}}\left(X\left((X-\mu)\cdot v\right)^2\right)$$

$$T = \mathbb{E}_{\hat{S}}(X \otimes X \otimes X) - \left(\sum_j u \otimes e_j \otimes e_j + e_j \otimes u \otimes e_j + e_j \otimes e_j \otimes u\right)$$

④ Decompose T using tensor decomposition

for vector y set $\hat{\mu}_i = T(y,y,y)\,y$

and $w_i = \frac{1}{\|\hat{\mu}_i\|^2}$

and finally $\sigma_i^2 = u \cdot \hat{\mu}_i = w_i \sigma_i^2 \|\hat{\mu}_i\|^2 = \sigma_i^2$.

Note: complexity depends on the condition number

Note: complexity depends on the condition number

of $\begin{pmatrix} - \mu_1 - \\ - \mu_k - \end{pmatrix}$.