The Stability of a Good Clustering

Marina Meila
University of Washington
Department of Statistics
mmp@stat.washington.edu

Thanks **Paul Tseng**, **Jim Burke, Linda Xu**

Optimizing these criteria is NP-hard

- Data

  similarities $S_{ij} > 0$

  $\{z_1, z_2, \ldots z_n\} \subseteq R^d$

- Objective

$$\mathcal{D}(Y) = \sum_{k=1}^{K} \sum_{i \in C_k} \frac{\sum_{j \notin C_k} S_{ij}}{\sum_j S_{ij}}$$

$$\mathcal{D}(Y) = \sum_k \sum_{i \in C_k} ||z_i - \mu_k||^2$$

- Algorithm

Spectral clustering

K-means

worst case

...but "spectral clustering, K-means work well when good clustering exists"

interesting case

Results summary

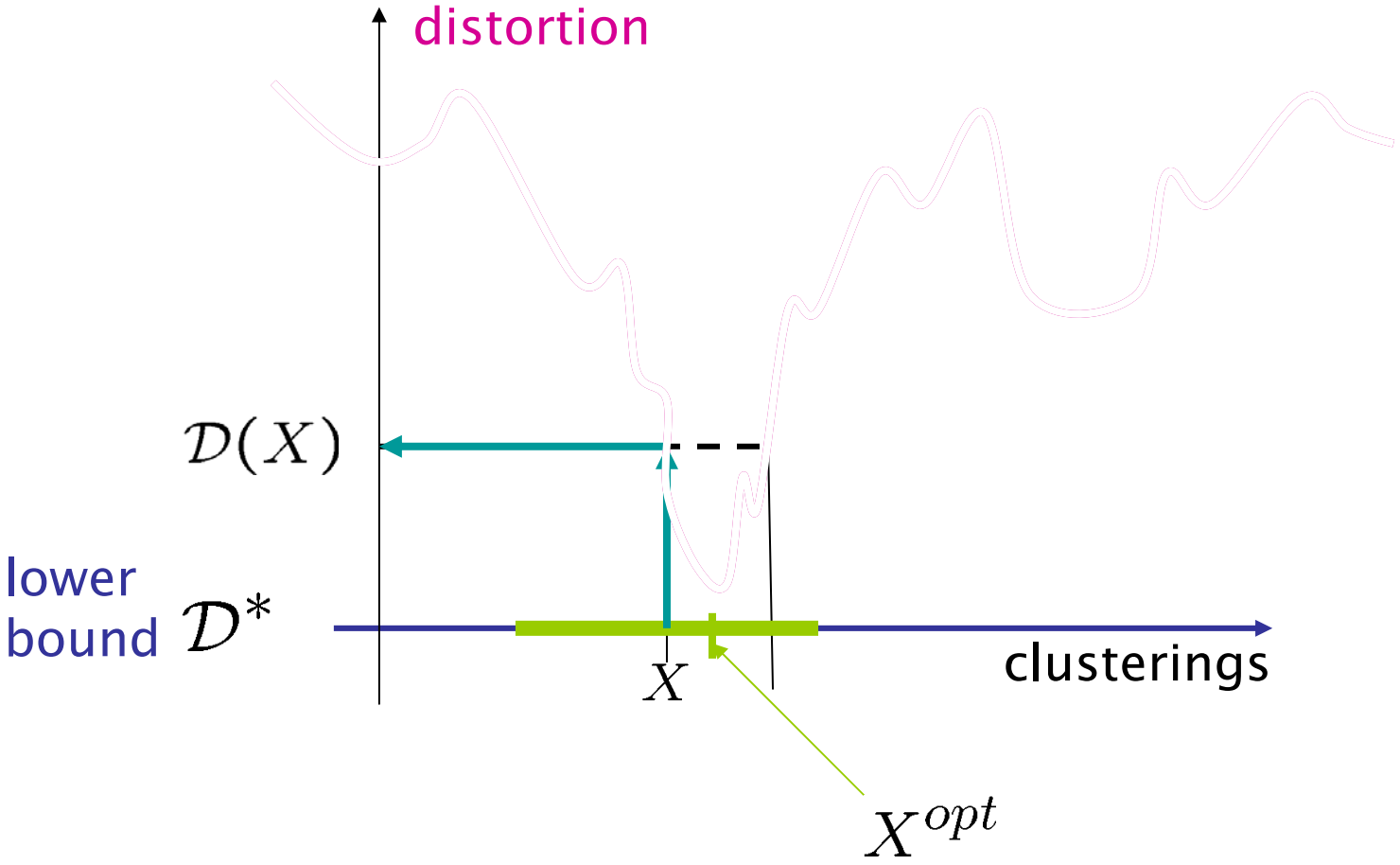
- Given
  - data
  - quadratic objective (distortion) $\mathcal{D}(X)$
  - clustering X with K clusters

- We have
  - Spectral lower bound $\mathcal{D}^*$ on distortion $\mathcal{D}(X)$
  - Theorem If $\mathcal{D}(X) - \mathcal{D}^*$ small then $d(X, X^{opt})$ small

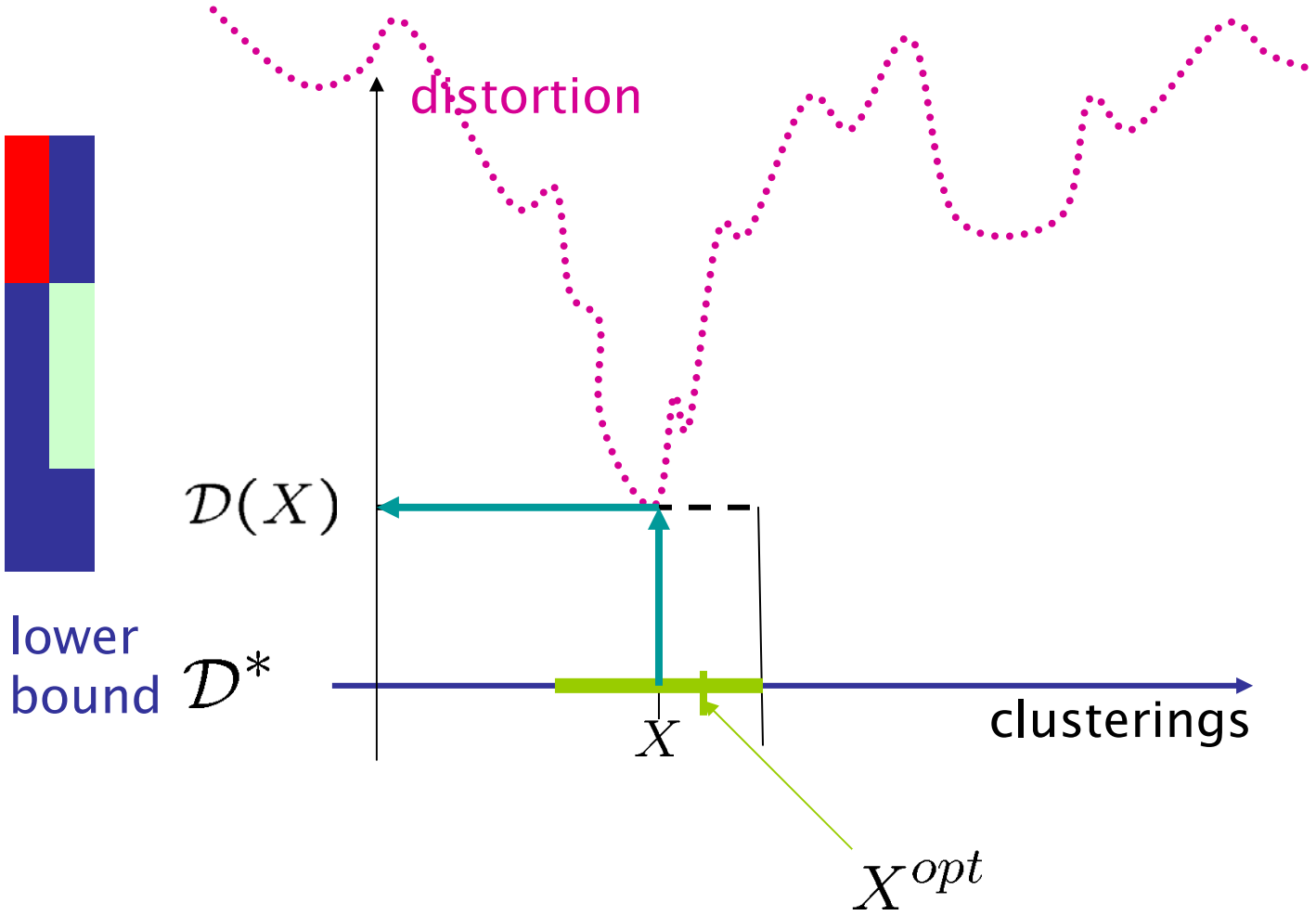
NEW

where $X^{opt} = \text{argmin}\mathcal{D}(X)$ = best K-clustering

A graphical view



A graphical view



Overview

- Representations
  - Distance between clusterings: the misclassification error
  - Subspace representation for clusterings
  - Quadratic representation for distortion

- The relaxed minimization problem
  - min quadratic function over subspaces
  - the eigengap

- The main theorem

- Experiments
- Extensions
- The χ^2 distance between clusterings
- Related work

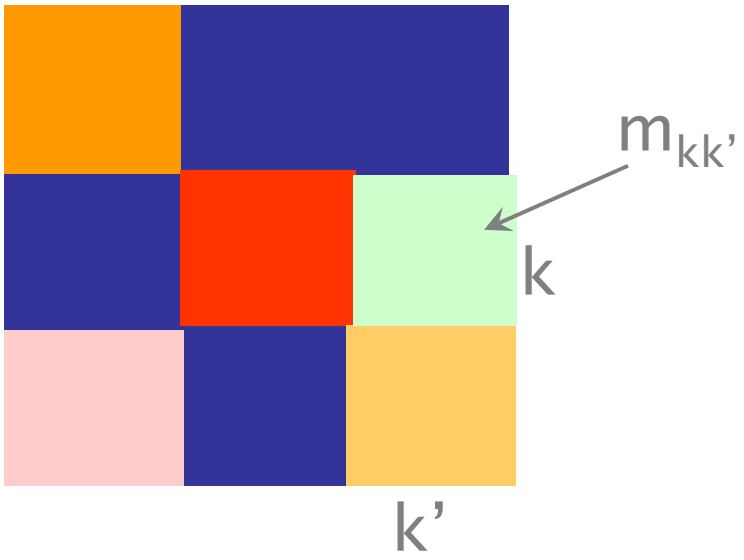
The Confusion Matrix

Two clusterings

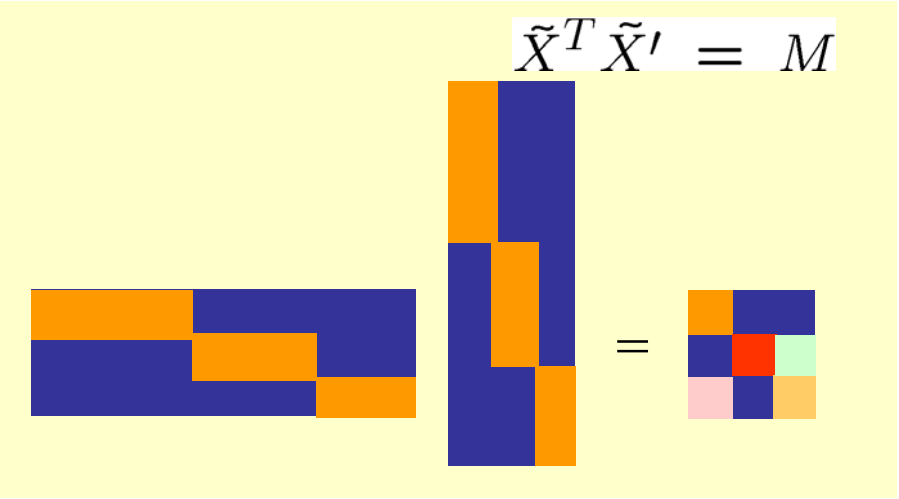
- $(C_1, C_2, \ldots C_K)$ with $|C_k| = n_k$
- $(C'_1, C'_2, \ldots C'_{K'})$ with $|C'_{k'}| = n'_{k'}$

$$m_{k,k'} = |C_k \cap C_{k'}|$$

- Confusion matrix $\tilde{M} = [m_{kk'}]$ (K x K')



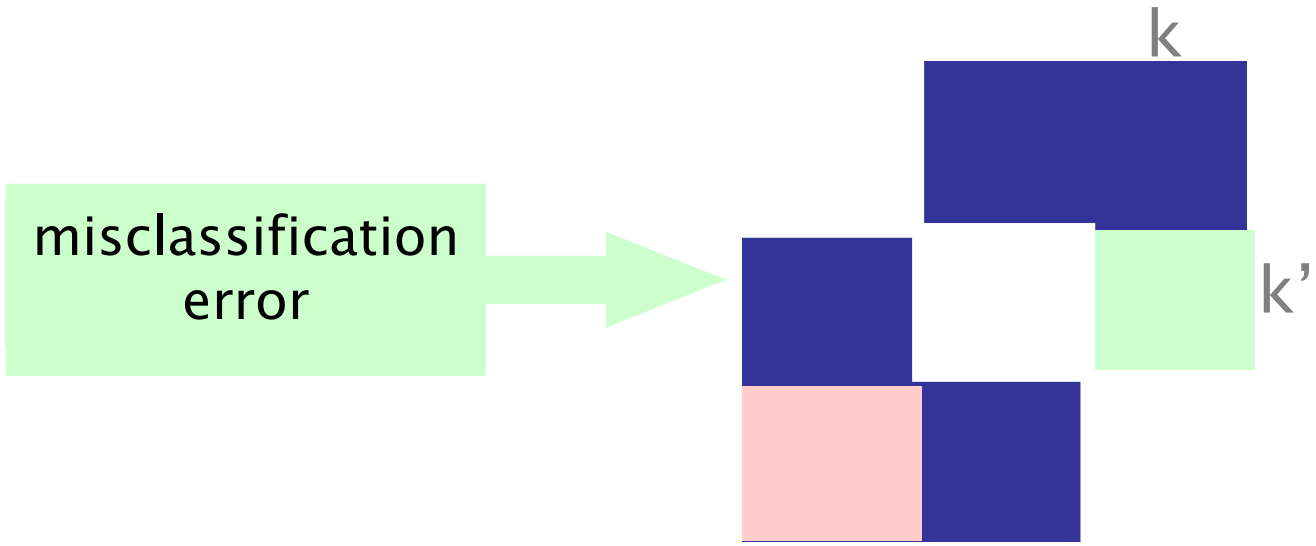
Matrix representation of M



The Misclassification Error distance

$$d_{ME}(X, X') = \min_{\text{label permutations}} \text{class error}(X, X') \in [0, 1]$$

- computed by the maximal bipartite matching algorithm between clusters



Representing clusterings as matrices

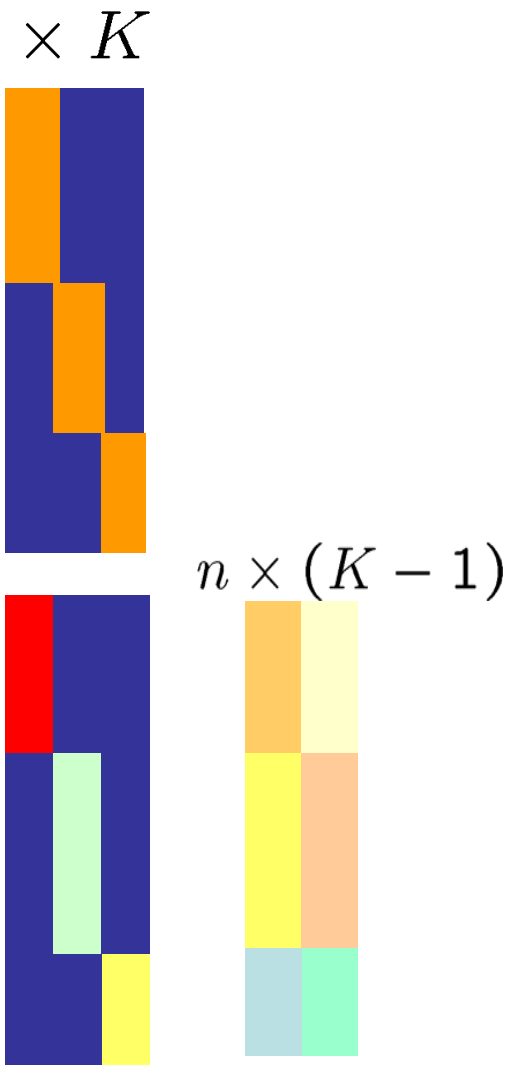
- Clustering of { 1,2,..., n } with K clusters (C_1, C_2,...C_K)
- Represented by n x K matrix
  - unnormalized

$$\tilde{X}_{ik} = \begin{cases} 1 & \text{if } i \in C_k \\ 0 & \text{otherwise} \end{cases}$$

  - normalized

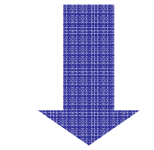
$$X_{ik} = \begin{cases} 1/\sqrt{n_k} & \text{if } i \in C_k \\ 0 & \text{otherwise} \end{cases}$$

- All matrices have orthogonal columns



Representing clusterings as subspaces

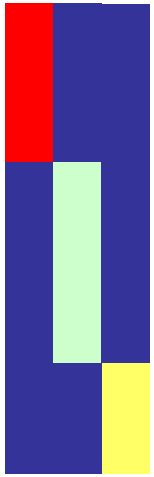
Orthogonal $n \times K$ matrix = basis for K-dimensional subspace



$n \times K$

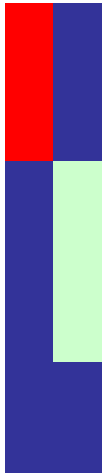
- one column in X redundant
  - removing it doesn't lose information

$n \times (K-1)$



X

drop last column

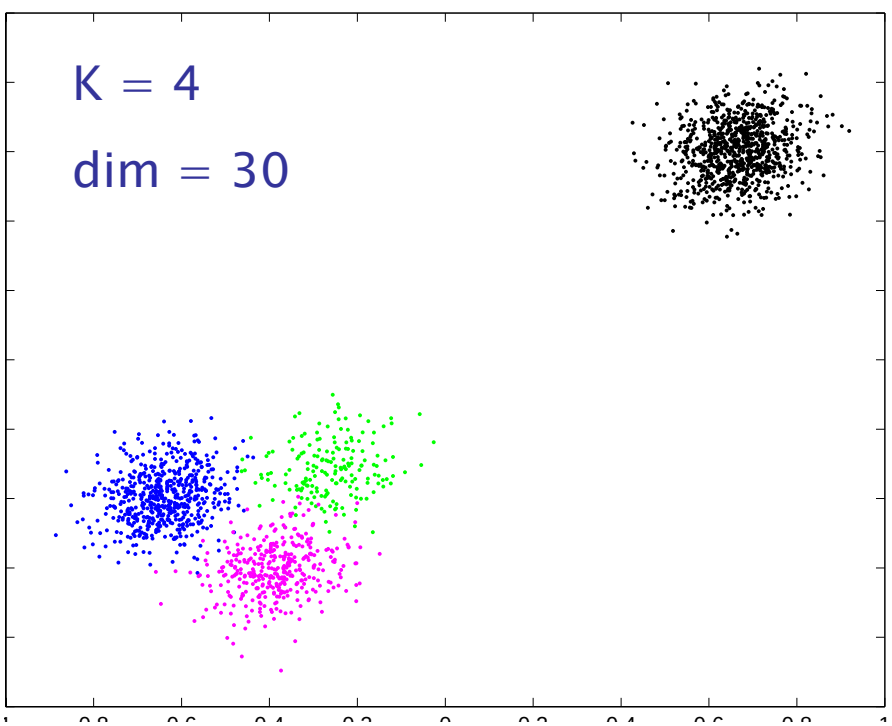


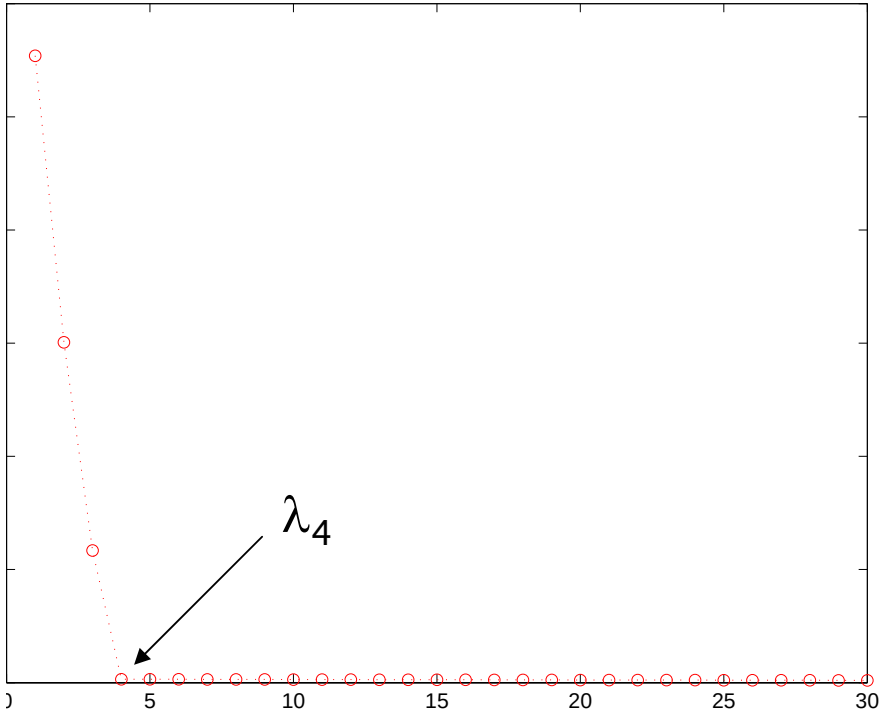
rotate

Y

basis for K-1 dimensional subspace

Why K-1 dimensions? Intuition





Distortion is quadratic in X

NCut

1,2,...n nodes in a graph

similarities $S_{ij} > 0$

$$\mathcal{D}(X) = \sum_{k=1}^{K} \sum_{i \in C_k} \frac{\sum_{j \notin C_k} S_{ij}}{\sum_j S_{ij}}$$

$$= K - \text{trace}\, X^T A X$$

$$A = D^{-1/2} S D^{-1/2}$$

$$D = \text{diag}\,(\ldots \sum_j S_{ij} \ldots)$$

K-means

$$\{z_1, z_2, \ldots z_n\} \subseteq R^d$$

$$\mathcal{D}(X) = \sum_k \sum_{i \in C_k} ||z_i - \mu_k||^2$$

$$= \text{trace}\, A - \text{trace}\, X^T A X$$

$$A = [z_I^T z_j]_{ij=1:n}$$

Why is distortion quadratic?

$$\mathcal{D}(X) = \sum_k \sum_{i \in C_k} ||z_i - \mu_k||^2$$

with

$$\mu_k = \frac{\sum_{i \in C_k} z_i}{|C_k|}$$

$$\mathcal{D}(X) = \sum_k \sum_{i,j \in C_k} ||z_i - z_j||^2$$

$$= \sum_k [2 \sum_{i \in C_k} ||z_i||^2 - 2 \sum_{i,j \in C_k} z_i^T z_j]$$

$$= 2\text{trace}\, A - 2\text{trace}\, X^T A X$$

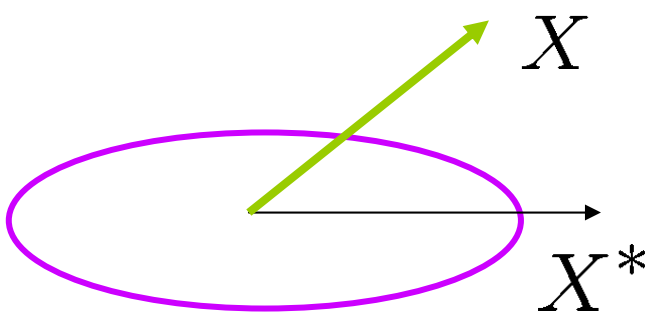
$$A = [z_i^T z_j]_{ij=1:n}$$

Quadratic functions and the eigengap

- X = n x K matrix with orthonormal columns
- quadratic function f $f(X) = \text{trace}\, X^T A X$
  - with A positive definite matrix
- f is maximized by
  - $X^* = [u_1\, u_2\, \ldots\, u_K]$ the leading eigenvectors of A
  - $f(X^*) = \lambda_1 + \lambda_2 + \ldots + \lambda_K$ (largest eigenvalues of A)

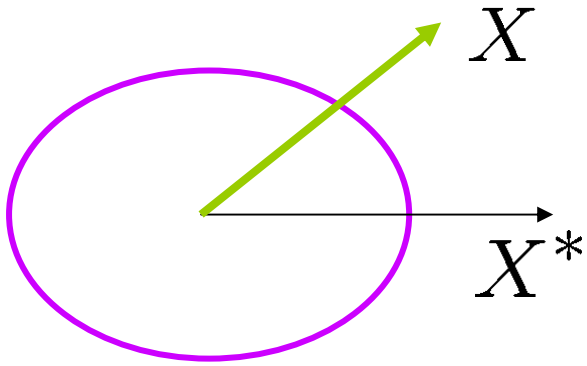
Quadratic functions and the eigengap

- Why does the eigengap matter?



large $\lambda_1 - \lambda_2$

f decreases fast around X^*



small $\lambda_1 - \lambda_2$

f decreases slowly around X^*

Eigengap Lemma

- If $\frac{f(X^*) - f(X)}{\lambda_K - \lambda_{K+1}}$ small, then X close to X^* (as subspaces)

Distortion is quadratic in X

$$\mathcal{D}(X) = \frac{1}{2}\sum_k \sum_{i \in C_k} ||z_i - \mu_k||^2$$

$$= \text{trace}\, A - \text{trace}\, X^T A X$$

$$A = [z_i^T z_j]_{ij=1:n}$$

- Relaxed minimization problem

$$\min_X \mathcal{D} \text{ s.t. } X \text{ orthonormal}$$

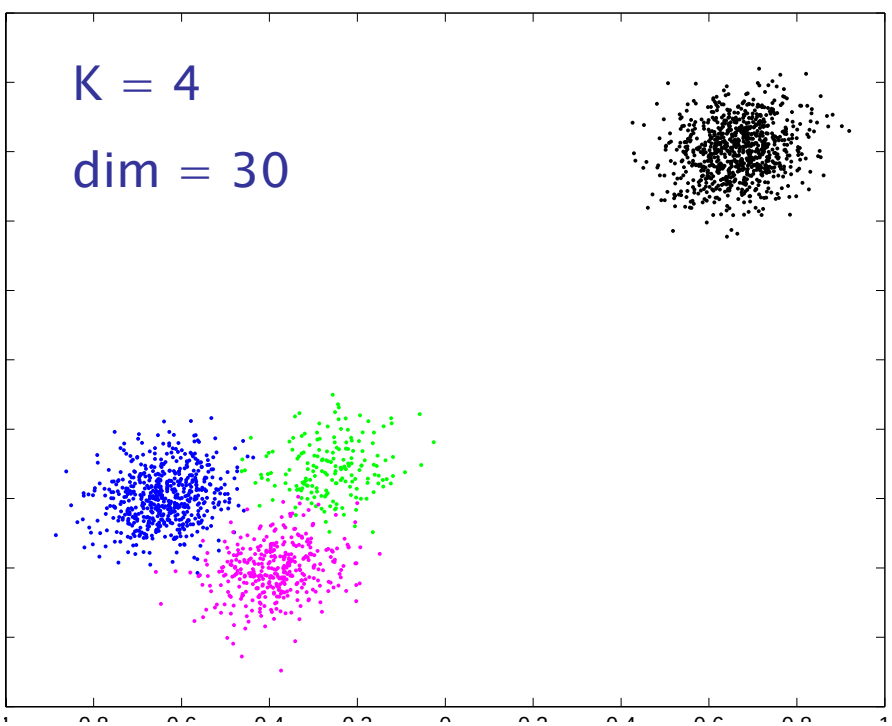
  - solution $X^* = [u_1\, u_2\, \ldots\, u_K]$
  - gives lower bound

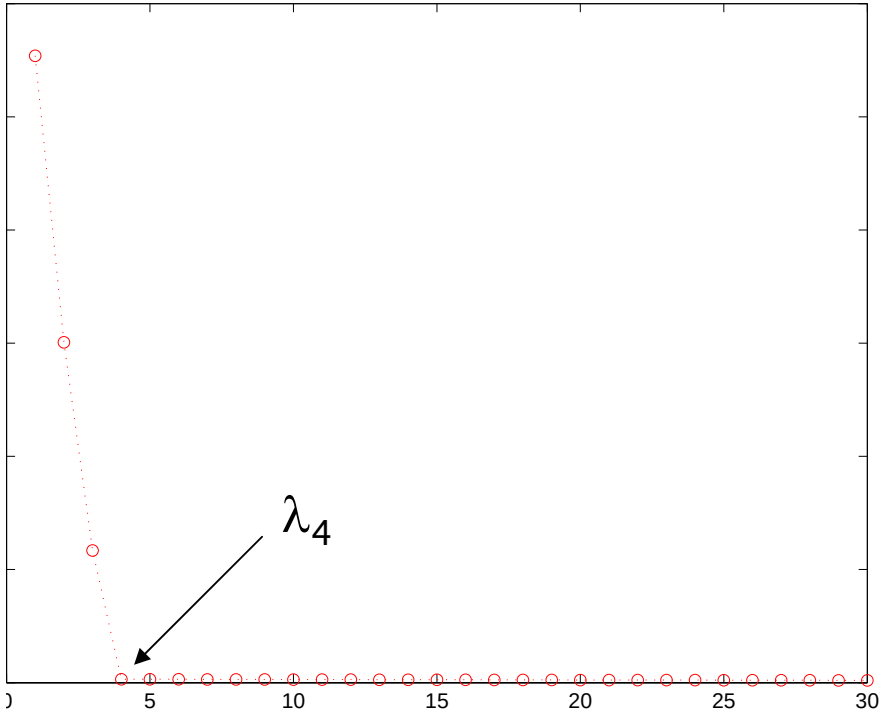
$$\mathcal{D}(X^*) = \text{trace}\, A - (\lambda_1 + \lambda_2 + \ldots + \lambda_K) = \mathcal{D}_K^*$$

  - if $\dfrac{\mathcal{D}(X) - \mathcal{D}_K^*}{\lambda_K - \lambda_{K+1}}$ small, X is close to X^* (as subspaces)

But this doesn't work...

- . K-th principal subspace typically not stable





n x (K-1) representation

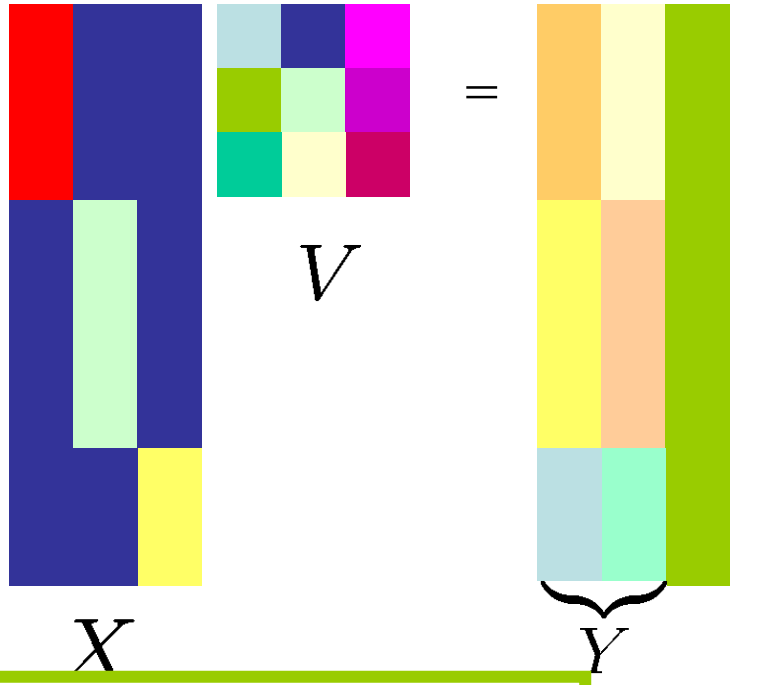
(Ding04)

- Non-redundant representation Y

$$w = \left[\sqrt{\frac{n_1}{n}}\ \sqrt{\frac{n_2}{n}} \cdots \sqrt{\frac{n_K}{n}}\right]^T$$

$$V = [\tilde{V}\ w] \text{ orthogonal}$$

$$XV = [Y\ 1/\sqrt{n}]$$



$n \times (K-1)$

- Distortion – new expression

$$\mathcal{D}(Y) = constant - \text{trace}\, Y^T A Y$$

... with lower bound $\mathcal{D}^* = ct - (\lambda_1 + \ldots + \lambda_{K-1})$ for

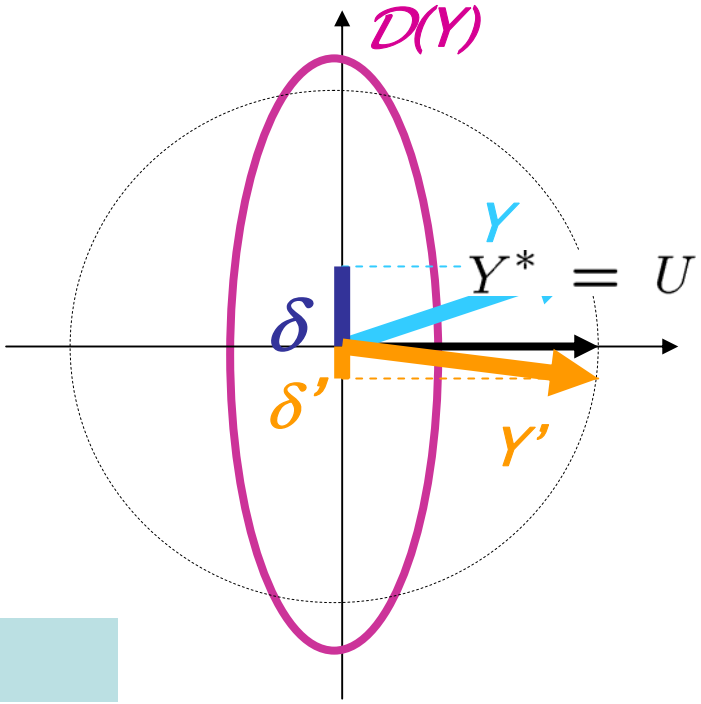
$$Y^* = [u_1 \ldots u_{K-1}]$$

Solve relaxed minimization

$$Y^* = U$$

$\mathcal{D}(Y) - \mathcal{D}^*$ small → Y close to Y*

$$\delta = \frac{\mathcal{D}(Y) - \mathcal{D}^*}{\lambda_{K-1} - \lambda_K} \geq \Pi_{U^\perp} Y$$



Y, Y' close to Y* → $||X^T X'||_F$ large

$$||X^T X'||_F^2 \geq K - \epsilon(\delta, \delta')$$

$$\epsilon(\delta, \delta') = 2\sqrt{\delta(1 - \frac{\delta}{K-1})}\sqrt{\delta'(1 - \frac{\delta'}{K-1})}$$

$||X^T X'||$ large → $d(X, X')$ small

$$d(Y, Y') \leq \epsilon p_{max}$$

Main Theorem

- Theorem

For any two clusterings X, X' with $\delta, \delta' \leq \frac{K-1}{2}$

$$d(X, X') \leq \epsilon(\delta, \delta') p_{max}$$

whenever $\epsilon < p_{min}$

$$\epsilon(\delta, \delta') = 2\sqrt{\delta(1 - \frac{\delta}{K-1})}\sqrt{\delta'(1 - \frac{\delta'}{K-1})}$$

$$p_{max} = \max \frac{C_k|}{n}$$

$$p_{min} = \min \frac{C_k|}{n}$$

Bound for d(X,Xopt)

- X' = X^{opt} $\quad \mathcal{D}(X^{opt}) \leq \mathcal{D}(X) \quad \Longrightarrow \delta_{Y^{opt}} \leq \delta$

Corollary: Whenever $\delta, \leq \frac{K-1}{2} \quad \epsilon(\delta, \delta) < p_{min}$

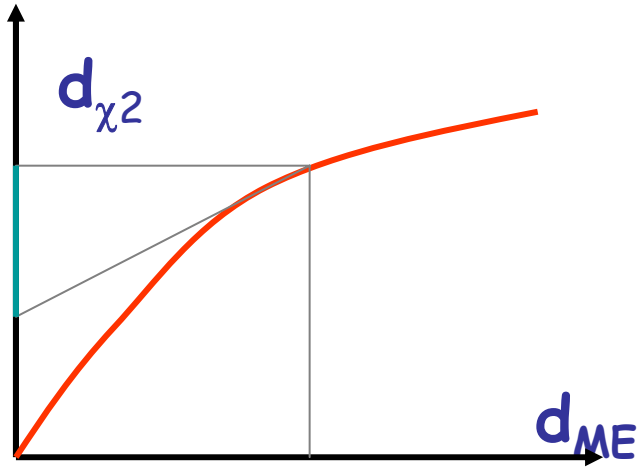
$$d(X, X^{opt}) \leq \epsilon(\delta, \delta) p_{max}$$

- the r.h.s depends only on A, X
- does not depend
  - on knowing the data distribution
  - on the clustering algorithm

The local equivalence of d_{ME} and $d_{\chi 2}$

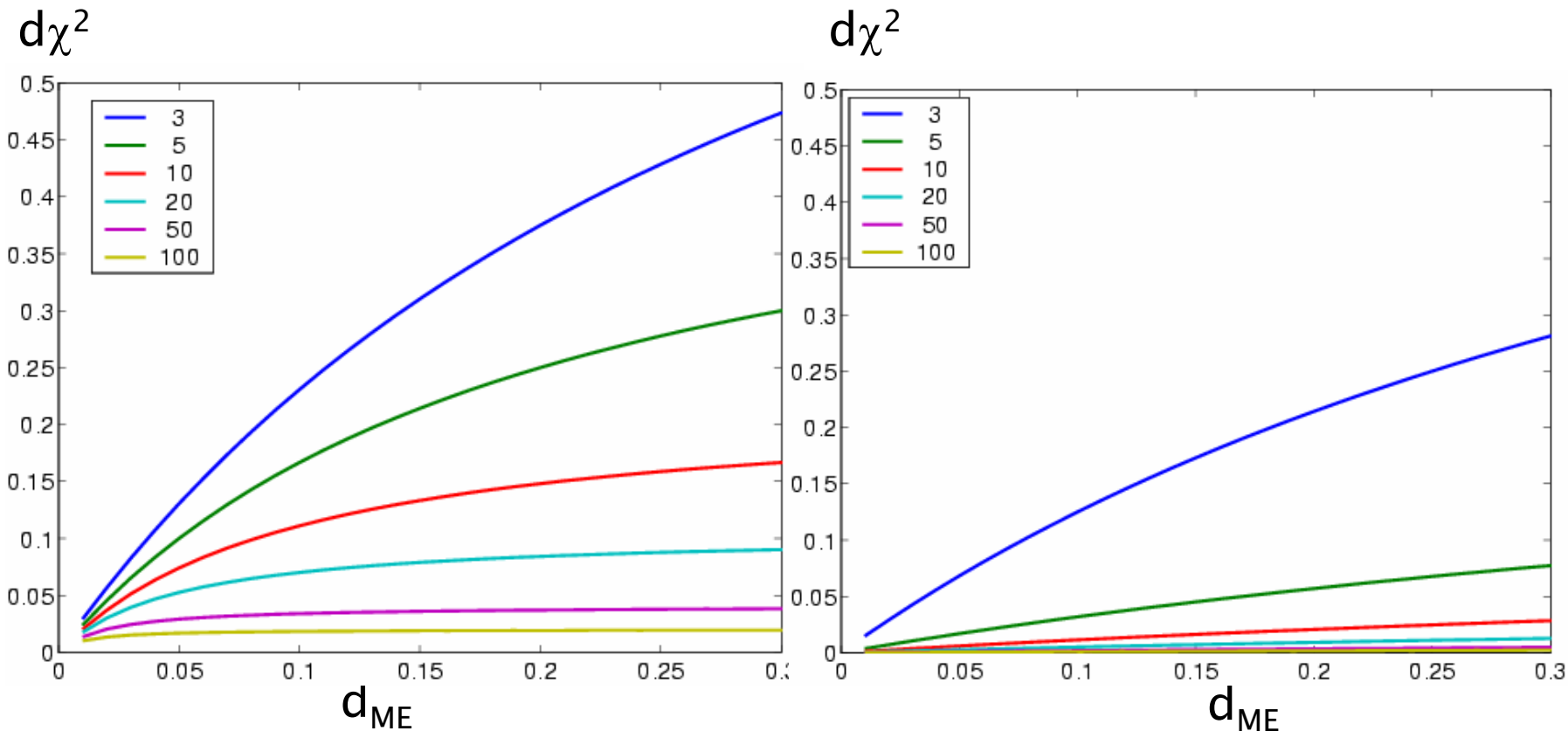
$$d_{\chi^2} \text{ small} \quad \Leftrightarrow \quad d_{ME} \text{ small}$$

- Proof based on convexity
  - both distances are concave functions
  - minimized when X=X'
  - look at extreme points of a probability set



  - Tighter bounds possible

Tighter bounds than ε p_{max}



C uniform

$$p_{max} = \frac{1}{K}$$

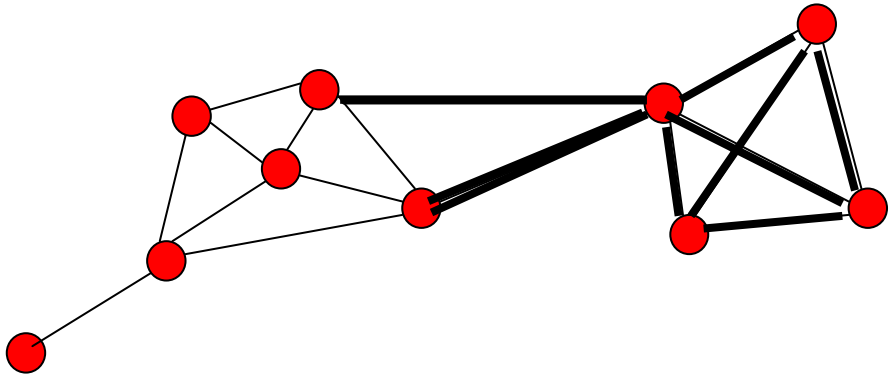
C non-uniform

$$p_{max} = 1 - \frac{1}{K}$$

Extensions

- Kernel K-means distortion
  - $A_{ij} = K(z_i, z_j)$
- Weighted data points
  - $w_i > 0$ weight of z_i
  - $A_{ij} = \sqrt{w_i w_j} z_i^T z_j$
  - $\tilde{X}_{ik} = \sqrt{w_i}$ if $i \in C_k$
  - p_{min}, p_{max}, d(X,X') computed w.r.t weighted data
- Normalized cut/average cut/weighted cut in graph

The normalized cut (Shi & Malik 99)



$$Cut(C, C') = \sum_{i \in C, j \in C'} S_{ij}$$

$$Vol\, C = \sum_{i \in C} D_i$$

$$NCut(C, C') = Cut(C, C') \left[\frac{1}{Vol\, C} + \frac{1}{Vol\, C'} \right]$$

- Minimizing the NCut = Finding a low weight cut that creates approximately equal clusters
- NCut for K=2
  - equivalent to isoperimetric number/Lovasz conductance
  - Is NP-hard to optimize (Shi & Malik 99)

The multiway normalized cut Meila & Xu, 03

- The Markov chain view of NCut
  - NCut(C, C') = $R_{CC'} + R_{C'C}$
  - $R_{CC'}$ = transition probability between clusters C, C'

    $R_{CC'}$ = Prob[C $\rightarrow$ C' | C]

    (Defined w.r.t a stationary distribution of the Markov chain)

- Generalization of NCut for K clusters
  - NCut(Δ) = $\sum_{C \neq C'} R_{CC'}$

    = sum of "off-diagonal" transition probabilities at the cluster level
- Remarks
  - R = [$R_{CC'}$] is a stochastic matrix
  - NCut(Δ) = K – trace (R)
  - (The transitions between clusters are generally NOT a Markov chain)

An improved bound for NCut

$$\delta = \frac{NCut(X) - [K - \sum_{k=1}^{K} \lambda_k]}{\lambda_K - \lambda_{K+1}}$$

- Theorem

For any two clusterings X, X' with $\delta, \delta' \leq \frac{K-1}{2}$

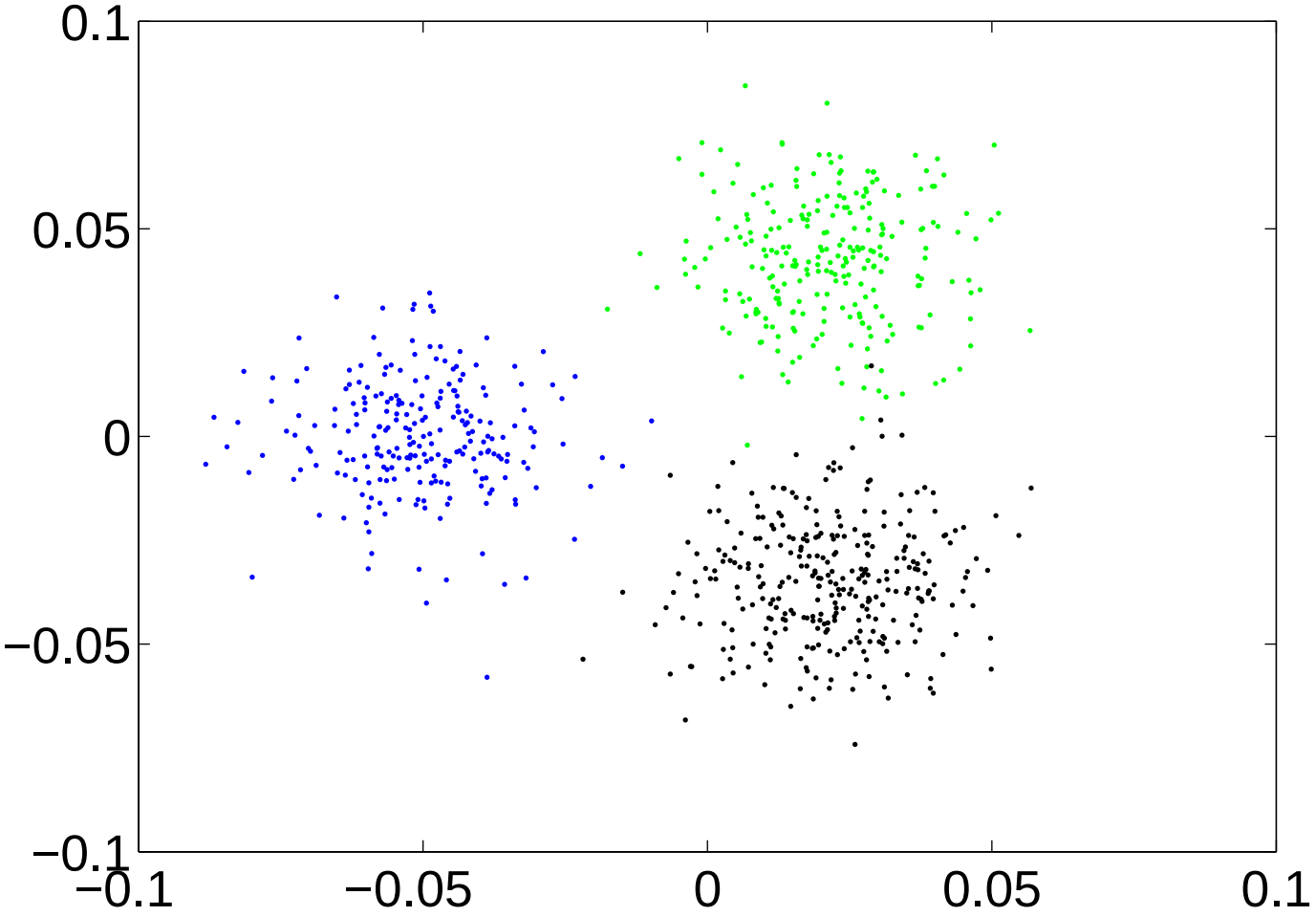
$$d(X, X') \leq \epsilon(\delta, \delta') p_{max}$$

whenever $\epsilon < p_{min}$

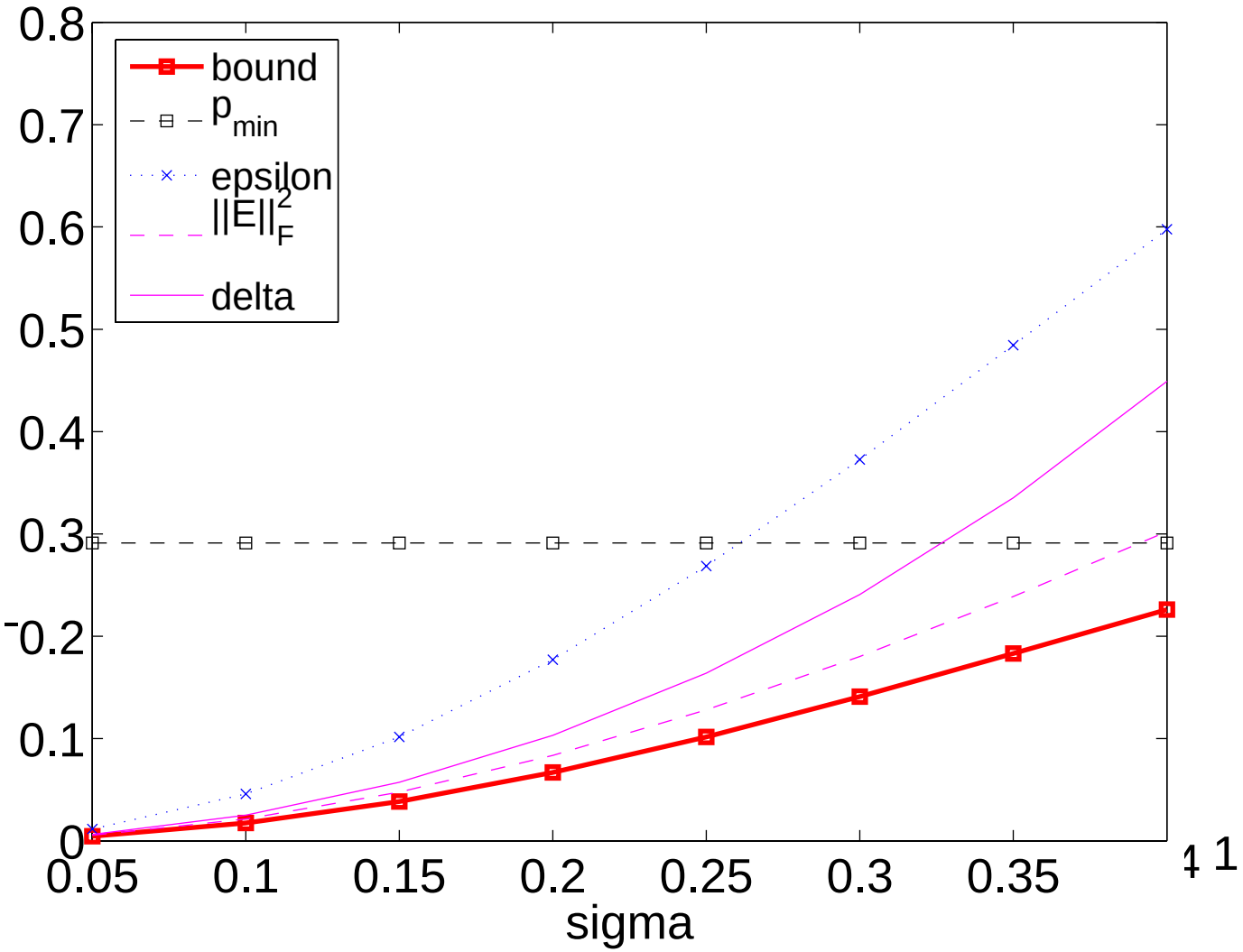
Lemma The new bound is an improvement on the bound of (M,Shortreed,Xu 05)

$$\epsilon^{old}(\delta, \delta) \geq \frac{K}{2} \epsilon(\delta, \delta)$$

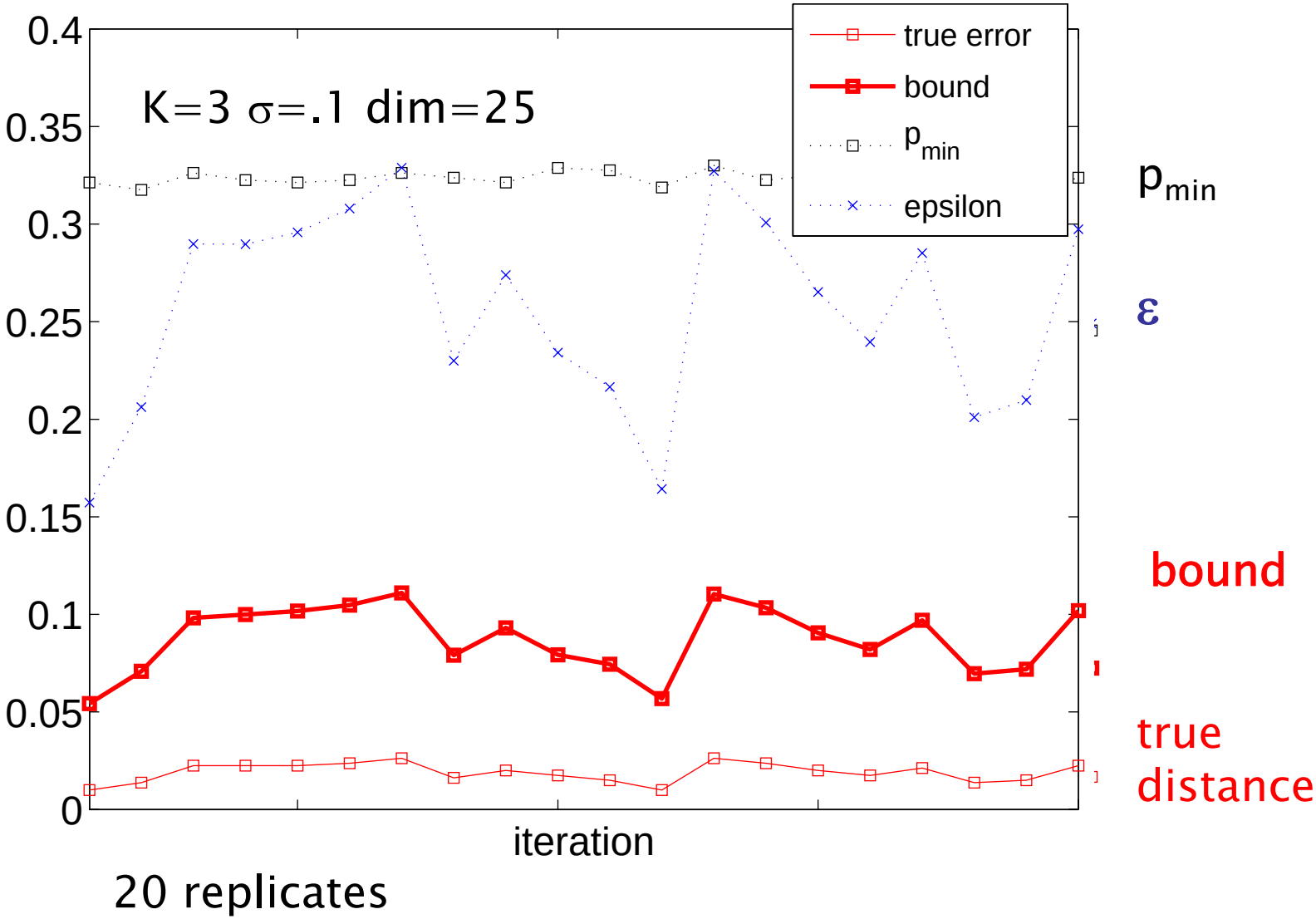
Experiments – bounds for best clustering



Experiments – bounds for best clustering



Experiments – random perturbation of X^{opt}



20 replicates

The χ^2 distance between clusterings

- Bounds exist when $\varepsilon \leq p_{min}$ and $\delta \leq (K-1)/2$

harder to satisfy

- Another distance
  - divergence from independence $\in [1,K]$

$$\chi^2(p_{XY}) = \sum_{xy} \frac{p_{xy}^2}{p_x p_y} - 1 = ||X^T X'||_F^2 - 1$$

a variant used by Bach & Jordan 03, Huber & Arabie 85

  - define "distance"

$$d_{\chi^2}(X, X') = \frac{K - ||X^T X'||_F^2}{K - 1} \in [0, 1]$$

Distances between clusterings

$$Vol\, A \propto Pr(A)$$

- The "χ^2" distance

$$\chi^2(\mathcal{C}, \mathcal{C}') = \sum_{k,k'} \frac{(Vol\, C_k \cap C'_{k'})^2}{Vol\, C_k . Vol\, C'_{k'}} \in [1, K]$$

a variant used by Bach & Jordan 03, Huber & Arabie 85c

- Pearson's χ^2 functional
  - $1 \leq \chi^2 \leq K$
  - $\chi 2(C, C') = K$ iff $\mathcal{C} = \mathcal{C}'$
  - minimum at independence

- define "distance" (not a metric)

$$d_{\chi^2}(\mathcal{C}, \mathcal{C}') = \frac{K - \chi^2(\mathcal{C}, \mathcal{C}')}{K - 1} \in [0, 1]$$

$f = p_{XY}$

$g = p_X p_Y$

$$
\begin{aligned}
\chi^2(f,g) &= \sum_{xy} \frac{(f-g)^2}{g} \\
&= \sum_{xy} \frac{(p_{xy} - p_x p_y)^2}{p_x p_y} \\
&= \sum_{xy} \frac{p_{xy}^2 - 2 p_{xy} p_x p_y + p_x^2 p_y^2}{p_x p_y} \\
&= \sum_{xy} \left[\frac{p_{xy}^2}{p_x p_y} - 2 p_{xy} + p_x p_y \right] \\
&= \sum_{xy} \frac{p_{xy}^2}{p_x p_y} - 2 + 1
\end{aligned}
$$

"Stability" of the best clustering

$$\chi^2(\Delta, \Delta') = \sum_{k,k'} \frac{(Vol\, C_k \cap C'_{k'})^2}{Vol\, C_k . Vol\, C'_{k'}} - 1$$

- χ^2 is Pearson's statistic
  - $0 \leq \chi^2 \leq K-1$
  - $\chi 2(\Delta,\Delta') = K-1$ iff $\Delta = \Delta'$
  - measures how "close" are two clusterings
  - define "distance" $d_{\chi^2} = 1 - (\chi^2 + 1)/K \quad \in [0, 1]$

- Theorem (M & Xu, 03)

  For any S and any clusterings Δ, Δ' with K clusters

$$gap(\Delta),\ gap(\Delta') \leq \varepsilon < \lambda_K - \lambda_{K+1}$$

$$\Rightarrow \quad d_{\chi^2}(\Delta, \Delta') \leq \frac{\varepsilon}{\lambda_K - \lambda_{K+1}} \cdot \frac{(\sqrt{K}+1)^2}{K} < \frac{3\varepsilon}{\lambda_K - \lambda_{K+1}}$$

- Theorem

For any two clusterings X, X' with $\delta, \delta' \leq \frac{K-1}{2}$

$$d_{\chi^2}(X, X') \leq \frac{\epsilon(\delta, \delta')}{K-1}$$

$$\epsilon(\delta, \delta') = 2\sqrt{\delta(1 - \frac{\delta}{K-1})}\sqrt{\delta'(1 - \frac{\delta'}{K-1})}$$

Corollary: $d_{\chi^2}(X, X^{opt}) \leq \frac{\epsilon(\delta, \delta')}{K-1}$

Conclusions

- First (?) model independent bounds on the clustering error
  - data dependent
    - hold when data well clustered (this is the case of interest)
  - depend explicitly on clustering (through p_{min},p_{max},...)
- Tight? – not yet...
- In addition
  - Showed local equivalence between "misclassification error" distance and "Frobenius norm distance" (also known as χ^2 distance)
- Related work
  - previous bound for NCut by (M,Shortreed,Xu)
  - Bounds for mixtures of Gaussians (Dasgupta, Vempala)
  - Nearest K-flat to n points (Tseng)