Theoretical Foundations of Clustering- Some Progress, Many Challenges

Shai Ben-David
U. of Waterloo

Bertinoro, 2007

The Theory-Practice Gap

Clustering is one of the most widely used tool for exploratory data analysis.

Social Sciences

Biology

Astronomy

Computer Science

.

.

All apply clustering to gain a first understanding of the structure of large data sets.

Yet, there exist distressingly little theoretical understanding of clustering

Inherent Obstacles

Clustering is not well defined.

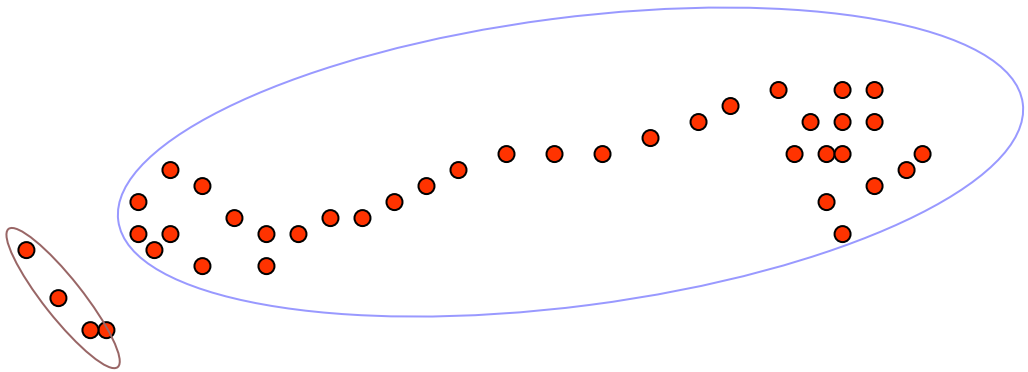
There is a wide variety of different clustering tasks, with different (often implicit) measures of quality.

In most practical clustering tasks there is no clear ***ground truth*** to evaluate your solution by.

There are Many Clustering Tasks

"Clustering" is an ill defined problem

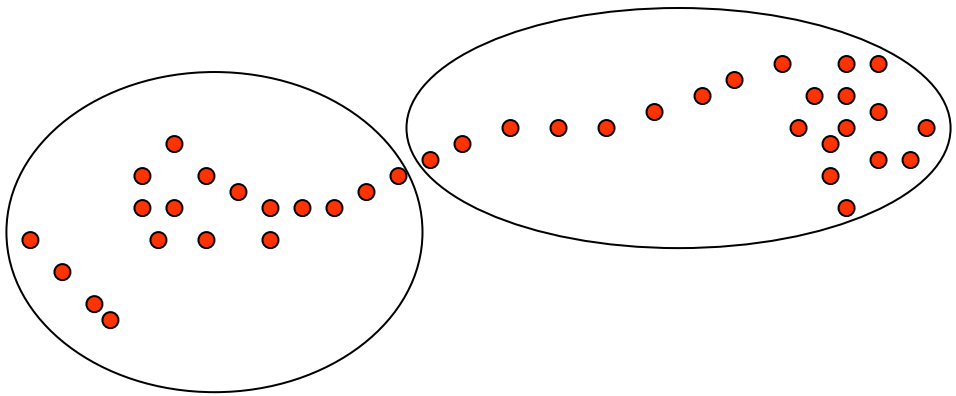
There are many different clustering tasks, leading to different clustering paradigms:



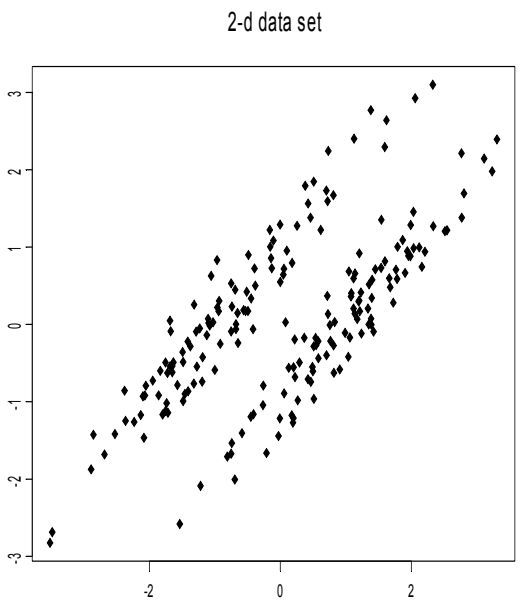
There are Many Clustering Tasks

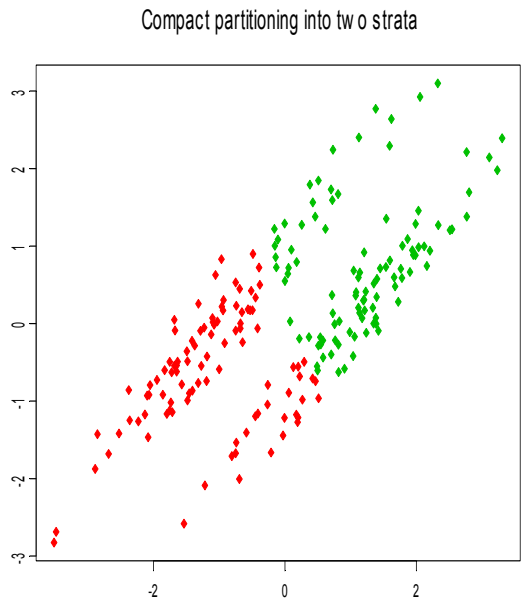
"Clustering" is an ill defined problem

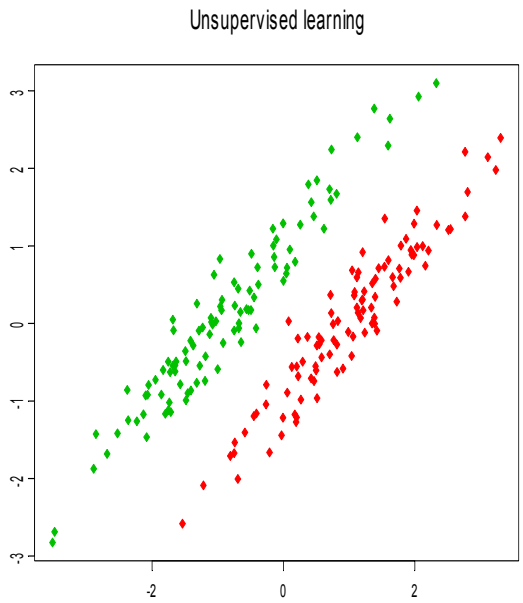
There are many different clustering tasks, leading to different clustering paradigms:



Some more examples







Common Solutions

Axiomatic approach:

Postulate ***'clustering axioms'***
that, ideally, every clustering approach should satisfy - usually conclude negative results
(e.g. [Hartigan 1975], [Puzicha, Hofmann, Buhmann '00], [Kleinberg '03]).

Add structure:

"Relevant Information" –
Information Bottleneck approach [Tishby, Pereira, Bialek '99]

Objective utility functions – sum of in-cluster distances, average distances to center points, cut weight, etc.
(Shmoys, Charikar, Meyerson …)

Common Solutions (2)

Consider a restricted set of distributions:

E., g, Mixtures of Gaussians

[Dasgupta ‘99], [Vempala,, ’03], [Kannan et al ‘04], [Achlitopas, McSherry ‘05].

Focus on specific algorithmic paradigms:

Projections based clustering (random/spectral)

all the above papers

Spectral-based representations – (Meila and Shi, Belkin, von Luxburg …)

Many more

Quest for a general theory

What can we say *independently* of any

particular algorithm,

particular objective function

or specific generative data model

?

What questions should research address?

- What is clustering?
- What is a “good” clustering?
- Can clustering be carried out efficiently?
- Can we distinguish “clusterable” from “structureless” data?
- Many more …

The Basic Setting

- For a finite domain set ***S***, a *dissimilarity function* (*DF*) is a mapping, $d:S\times S \rightarrow R^{+}$, such that:

  d is symmetric,

  and

  $d(x,y)=0$ iff $x=y$.

- A *clustering function* takes a dissimilarity function on ***S*** and returns a partition of ***S***.

- *We wish to define the properties that distinguish clustering functions from other functions that output domain partitions.*

Kleinberg's Axioms

- *Scale Invariance*

  $F(\lambda d)=F(d)$ for all d and all strictly positive λ.

- *Richness*

  For any finite domain S,

  $\{F(d): d$ *is a DF over* $S\}=\{P:P$ *a partition of* $S\}$

- *Consistency*

  d' equals d except for shrinking distances within clusters of $F(d)$ or stretching between-cluster distances (w.r.t. $F(d)$), then $F(d)=F(d')$.

Note that any pair is realizable

Consider Single-Linkage with different stopping criteria:

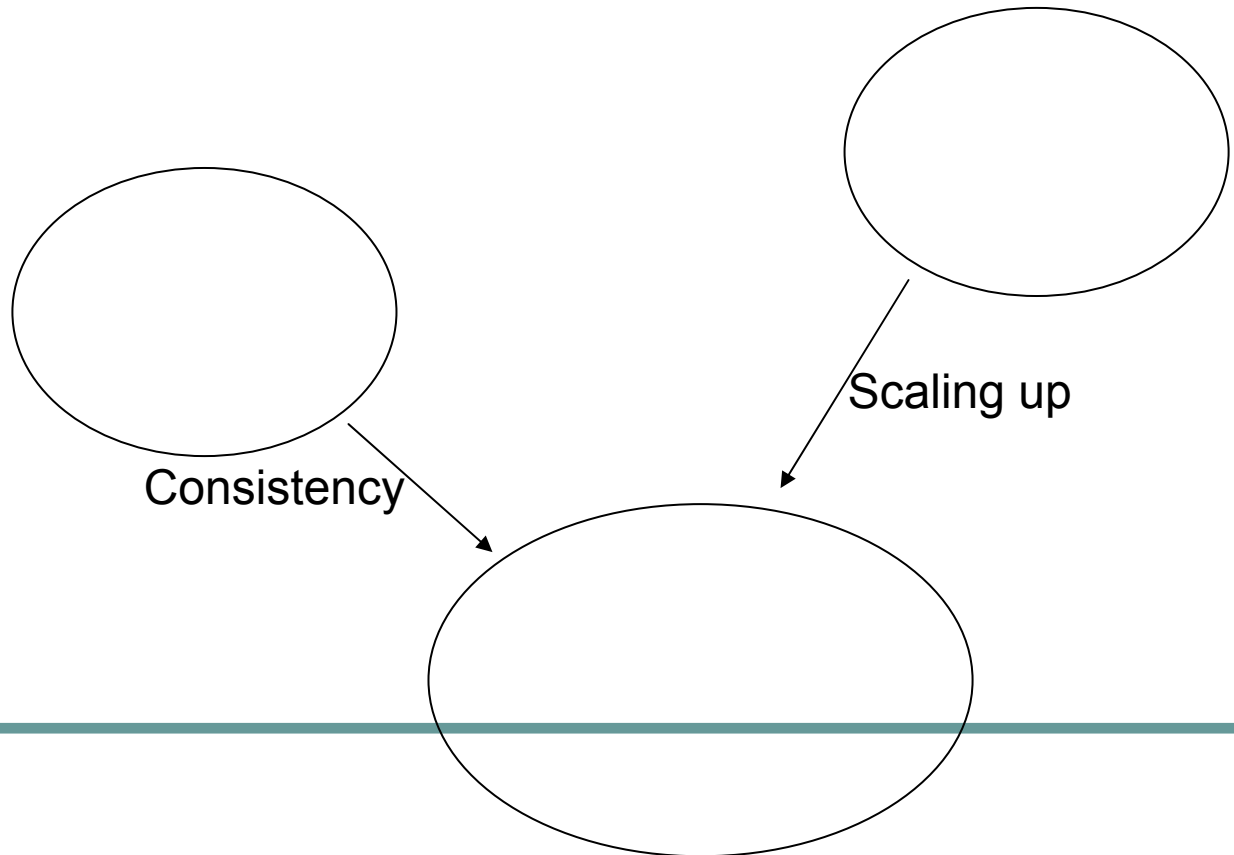
- *k* connected components.
- Distance *r* stopping.
- Scale α stopping:

  add edges as long as their length is at most α(max-distance)

Kleinberg's Impossibility result

There exist no clustering function

Proof:



Ideal Theory

- We would like the ***axioms*** to be such that:

  1. *Any clustering* method satisfies *all* the axioms,

  and

  2. *Any function* that is clearly not a clustering fails to satisfy at least one of the axioms.

  (this is probably too much to hope for).

- We would like to have a list of *simple properties*

  so that major clustering methods are distinguishable from each other using these properties.

Axioms as a tool for a taxonomy of clustering paradigms

- The goal is to generate a variety of axioms (or properties) over a fixed framework, so that different clustering approaches could be classified by the different subsets of axioms they satisfy.

	"Axioms"			*"Properties"*		
	Scale Invariance	**Antichain Richness**	**Local Consistency**	**Full Consistency**	**Richness**	
Single Linkage	+	+	+	+	-	
Center Based	+	+	+	-	+	
Sum of Distances	+	+	+	+	-	
Spectral	+	+	+	+	-	
Silly F	+	+	-	-	+	

Types of Axioms/Properties

- *Richness requirements.*

  E.g., relaxations of Kelinberg's richness, such as

  K-Richness -

  {F(d): d** is a DF over **S}={P:P** a partition of **S** into **k** sets**}

- *Invariance/Robustness/Stability* requirements.

  E.g., Scale-Invariance, Consistency, robustness to perturbations of ***d*** ("smoothness" of ***F***) or stability w.r.t. sampling of ***S***.

Relaxations of Consistency

Local Consistency –

Let $C_1, \ldots C_k$ be the clusters of $F(d)$.

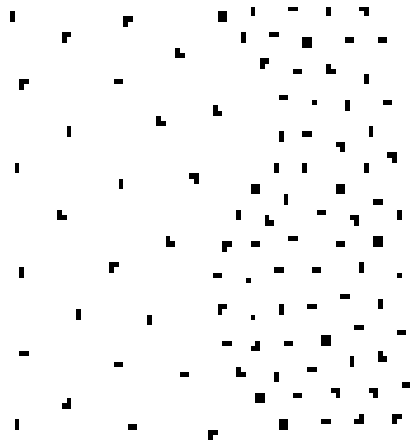
For every $\lambda_0 \geq 1$ and positive $\lambda_1, \ldots \lambda_k \leq 1$, if d' is defined by:

$$d'(a,b) = \begin{cases} \lambda_i d(a,b) & \text{if } a \text{ and } b \text{ are in } C_i \\ \lambda_0 d(a,b) & \text{if } a,b \text{ are not in the same } F(d)\text{-cluster,} \end{cases}$$

then $F(d)=F(d')$.

Other types of clustering

- *Edge-Detection* (advantage to smooth contours)
- *Texture clustering*



- The professors example.

Some Open Questions

- What do we want from a set of clustering axioms? (Meta axiomatization …)
- Can we define a notion of "completeness" of a set of axioms? How can we show that certain subsets are just too weak?
- Are there basic properties that distinguish different clustering paradigms, say, *Center-based* from *Linkage-based* clustering?

A Different Approach

Formulate conditions that should be satisfied by any conceivable clustering function.

(Sidestepping the issue of “what is clustering?”)

In other words –
find ***necessary*** conditions for good clustering

Stability basic idea

1) *Cluster independent samples of the data.*

2) *Compare the resulting clusterings.*

Meaningful clusterings should not change much

from one independent sample to another.

This idea has been employed as a tool for choosing the number of clusters in several empirical studies ([Ben-Hur et al'02], [Lange, Brown, Roth, Buhmann '03] and many more).

However, currently there is very limited theoretical support.

A Different Perspective – Replication

S*tability* can be viewed as the fundamental issue of replication --

to what extent are the results of an experiment (*sample-based clustering*) reproducible?

Replication has been investigated in many applications of clustering - mostly by ***visual inspection*** of the results of cluster analysis on two samples.

Stability - a formal definition

Given,

- Probability dist. P over some domain X.
- Clustering function A defined on $\{S : S \subseteq X\}$.
- Similarity measure over clusterings, D.
- Sample size m.

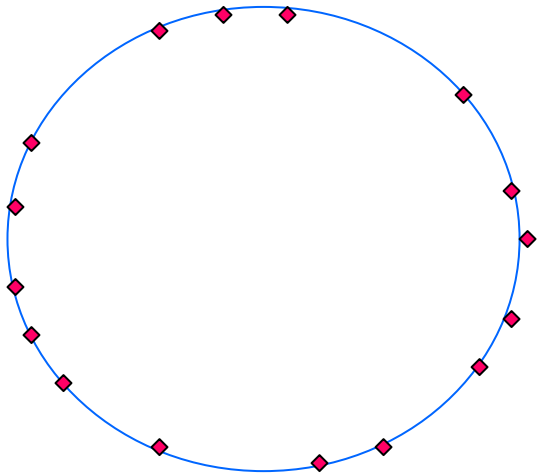
$$\mathrm{InStab}_m(A, P) = \mathrm{E}_{S, S' \in P^m} D(A(S), A(S'))$$

Namely, the expected distance between the clusterings generated by two P-random i.i.d. samples of size m.

Negative Observation:

There is no distribution-free stability guarantee.

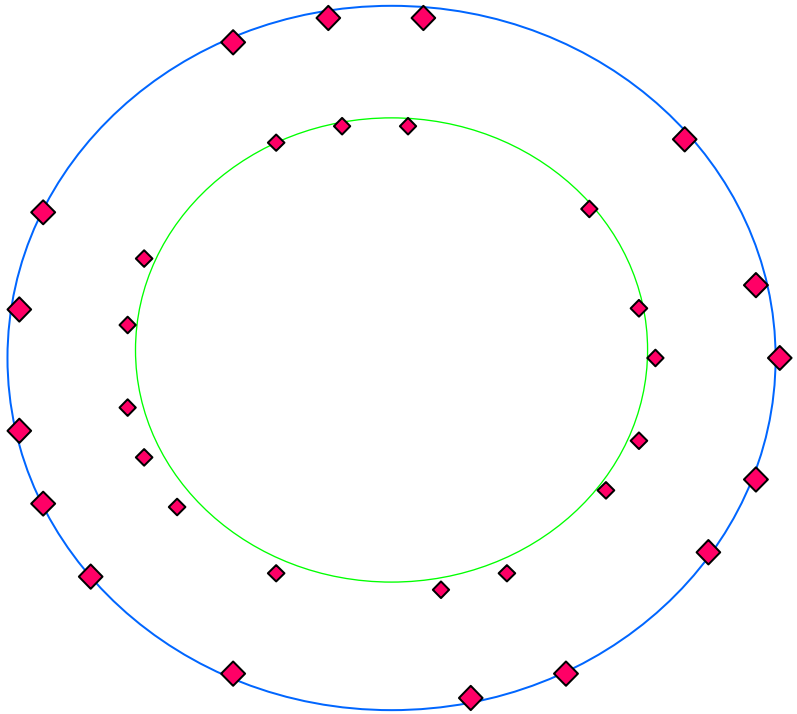
Example 1: The uniform distribution over a circle



InStab(C,P) will be large for any non-trivial clustering function.

Another ‘unstable’ example:

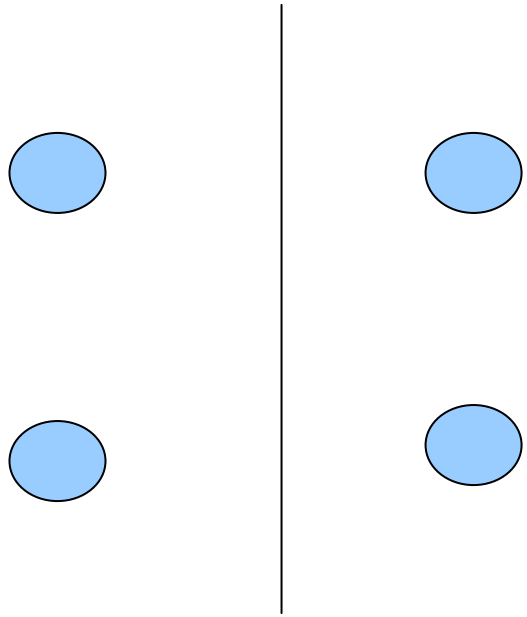
Example 2: A mixture of two uniform distribution over circles



InStab(C,P) will be large for any center-based clustering function.

Yet another ‘unstable’ example:

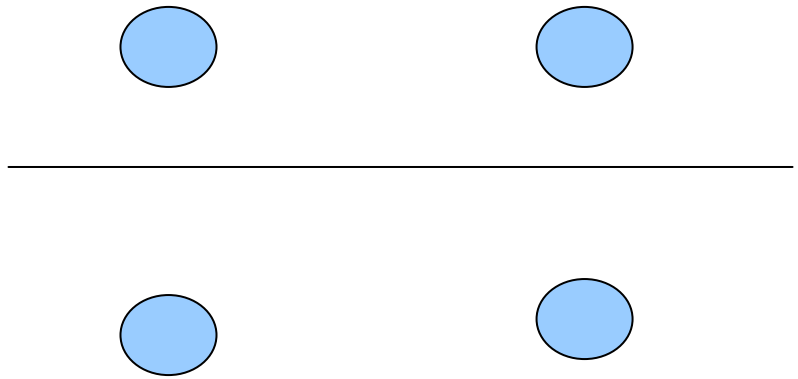
Example 3: A mismatch of # of clusters



UnStab(C,P) will be large for any center-based 2-cluster function.

Yet another ‘unstable’ example:

Example 3: A mismatch of # of clusters



UnStab(C,P) will be large for any center-based 2-cluster function.

An Optimistic view of the "negative " observations:

We may view stability of a measure of fit between a probability distribution and a clustering model.

The previous examples can be interpreted as saying:

stability fails when the clustering model is not aligned with the input data.

Stability as a Model-selection Tool:

On a given input data distribution, P,
for every candidate clustering algorithm, C,

- *Etimate* InStab(C,P)
- *Choose algorithm/parameters for which this measure is small.*

Examples of stability success as a model-selection tool:

We considered two types of clustering models:

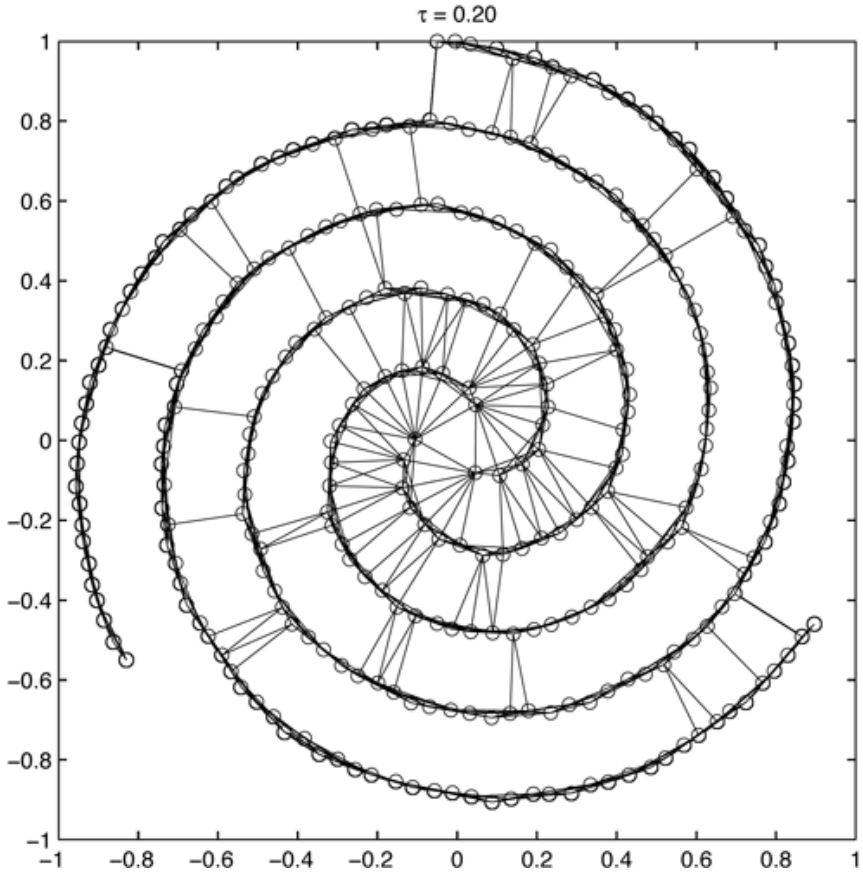
- *Center Based (both k-median and k-means)*
  -- Varying the number of clusters, k.
- *Linkage Based*
  -- Varying "stopping distance".

We analyzed their stability on two types of distributions:

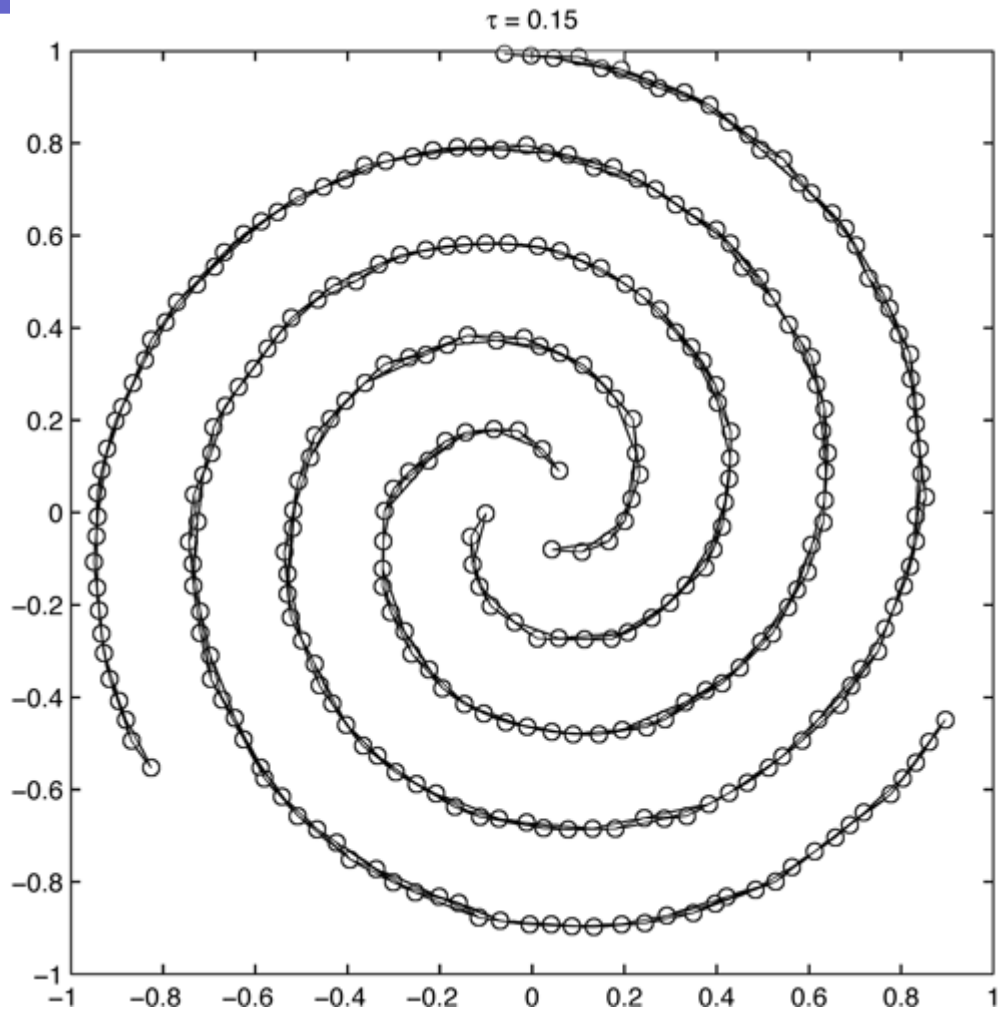
- Mixtures of shifted spherical, equal *Gaussians*
- Mixture of *disjoint co-centric spherical distributions*.

In all these cases, stability *provably* picks the correct model and correct model parameter.

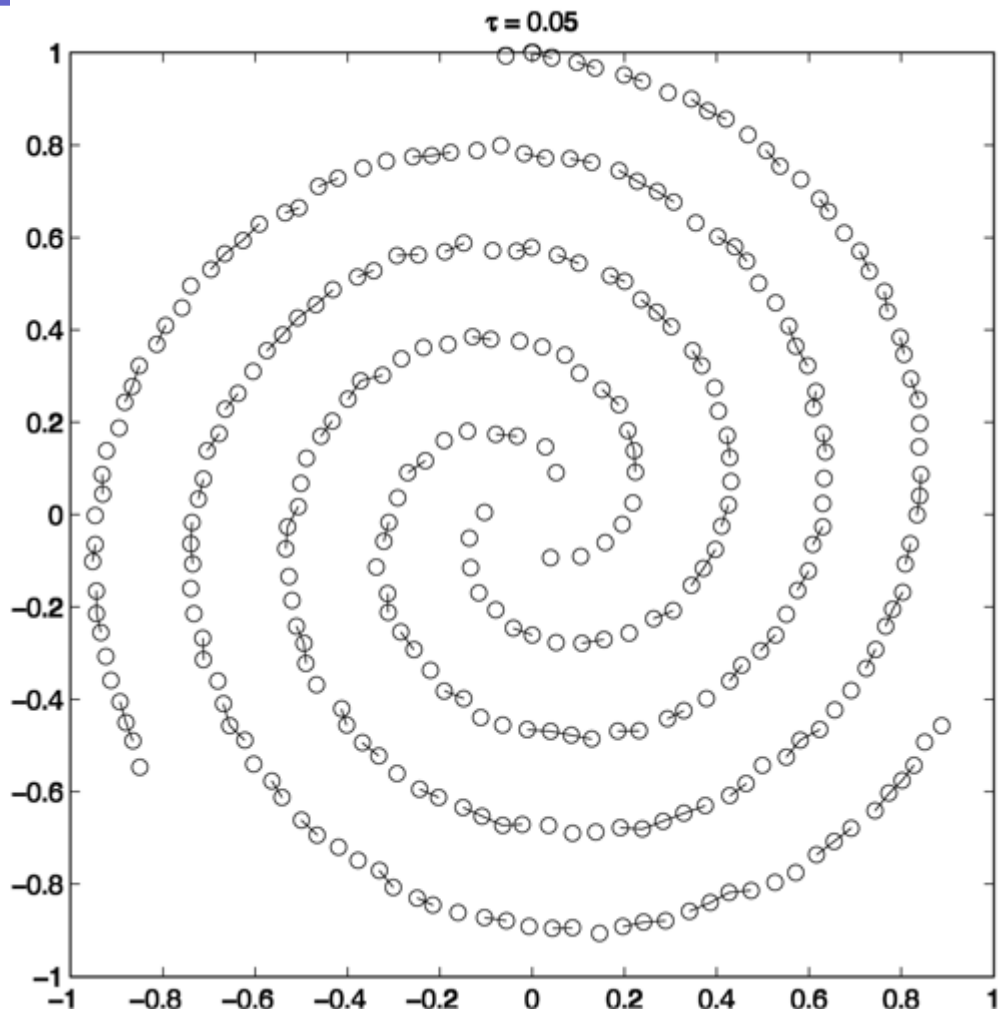
Single Linkage clustering of Swiss-roll – varying the cutoff distance



Single Linkage clustering of Swiss-roll – varying the cutoff distance

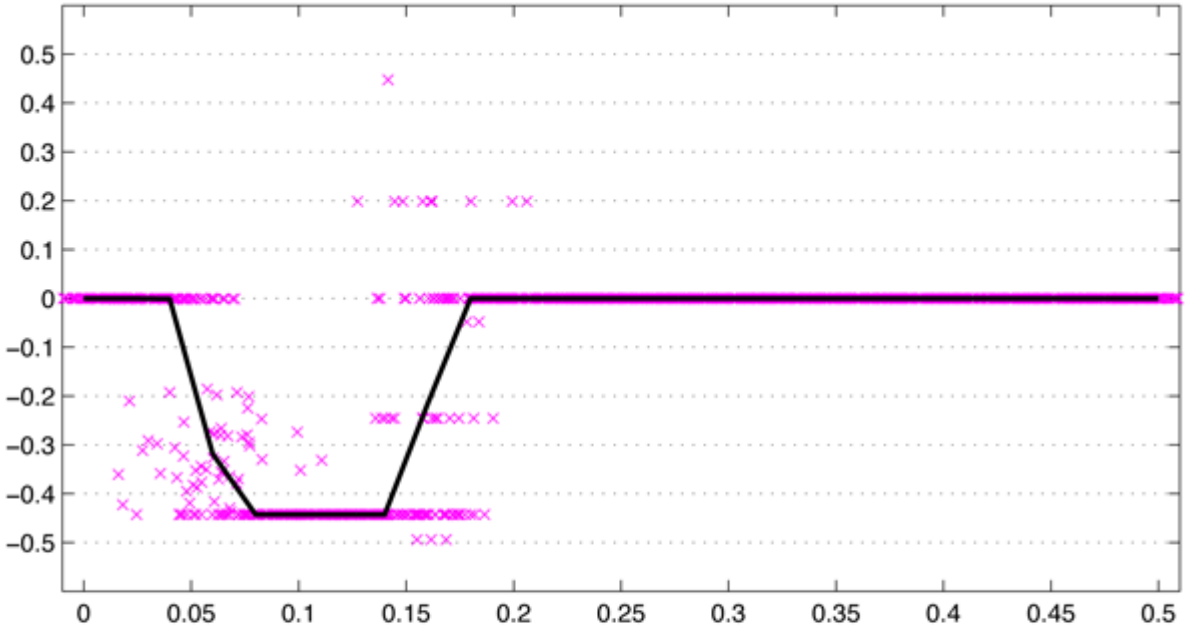


Single Linkage clustering of Swiss-roll – varying the cutoff distance

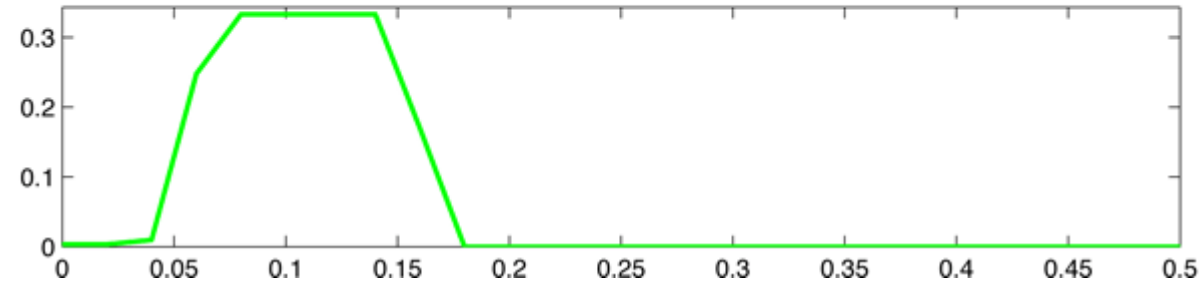


Instability detects the ‘correct’ clustering parameter

InStab



Size of 3rd largest component



Conclusions (as of Dec. 2005)

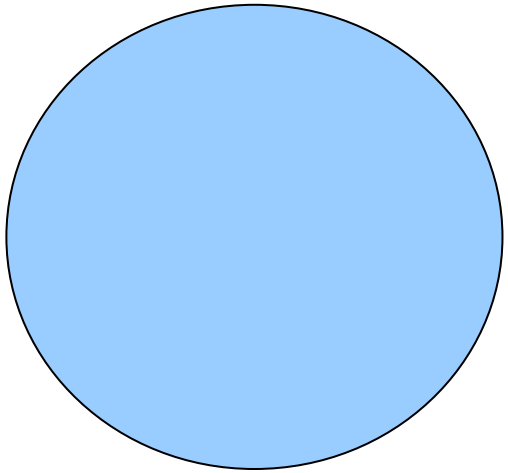
- ✓ We formally define a measure of statistical generalization for sampling-based clustering –***stability***.
- ✓ Stability is a necessary property for any clustering method to be considered 'meaningful'.
- ✓ Stability is viewed as a measure of the fit between a clustering function and an input data.
- ✓ We show that this measure can be reliably estimated from finite samples.

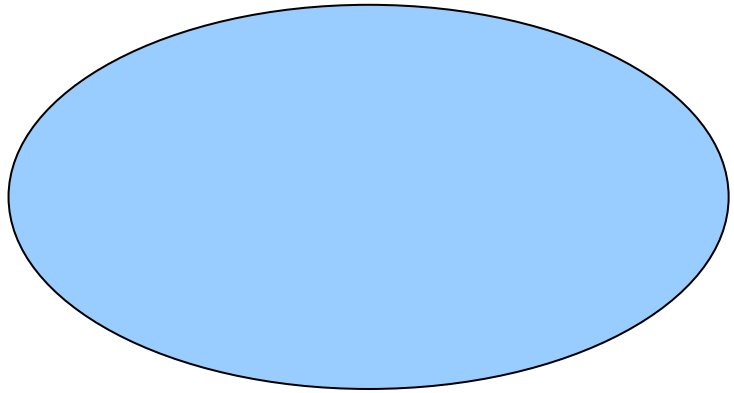
Have we found a good answer?

This is what we thought in January 2006,
when we (with Ule and David Pal) set
to prove that
"stability is a reliable model-selection tool".

(*Its another interesting question how can one provide a mathematical formulation of such a statement).*

Some bothersome examples





- A perfect ‘circular’ data is unstable for every k>1,
- Once the symmetry is broken, it becomes stable for every choice of k.

The bottom line of a formal analysis

- **Stability does a nice model-selection job on *simple synthetic distributions*, (as well as in many practical applications).**
- **We *characterize* it for the k-means optimization algorithms (BD-Luxburg-Pal, COLT06, BD-Pal-Simon, COLT07).**
- **We conclude that that success should be considered a lucky coincidence rather than a reliable rule.**

The formal results

We consider cost-minimizing clustering algorithms.

We say that an algorithm ***A is stable on a data set D*** if $\mathbf{Lim_{m->\infty} InStab_m(A,D)=0}$

Theorem (BD-Pal-Luxburg 06, DB-Pal-Simon 07):

A cost minimizing algorithm, *A*, is **stable** on data set *D* *if and only if*

there is a **unique** clustering solution to the cost minimization problem for *D*.

Proof Idea 1: Uniqueness implies stability

- [BD (COLT'04)] proves that, for any data set, the cost of (optimally) clustering samples uniformly converges to the optimal cost, as sample sizes go to infinity.
- If there is unique cost-minimizing solution, large enough samples are therefore bound to produce clusterings that are close to that optimum.

Proof idea (2): Multiple solutions imply instability

- Support of P is finite $\Longrightarrow$ finitely many different clusterings. (Note that $d_P(\mathcal{C}, \mathcal{D}) > 0$ for any two distinct clusterings $\mathcal{C}, \mathcal{D}$.)
- Some of them are P-optimal: **OPT** $= \{\mathcal{C}, \mathcal{D}, \mathcal{E}, \ldots\}$.
- Let m be large enough, so that, with high probability a sample $S \sim P^m$ is such that $A_k(S) \in$ **OPT**. This follows from the uniform convergence theorem. (We can ignore S's not having this property.)
- Pick two clusterings $\mathcal{C}, \mathcal{D} \in$ **OPT**, $d_P(\mathcal{C}, \mathcal{D}) > 0$. We will show that

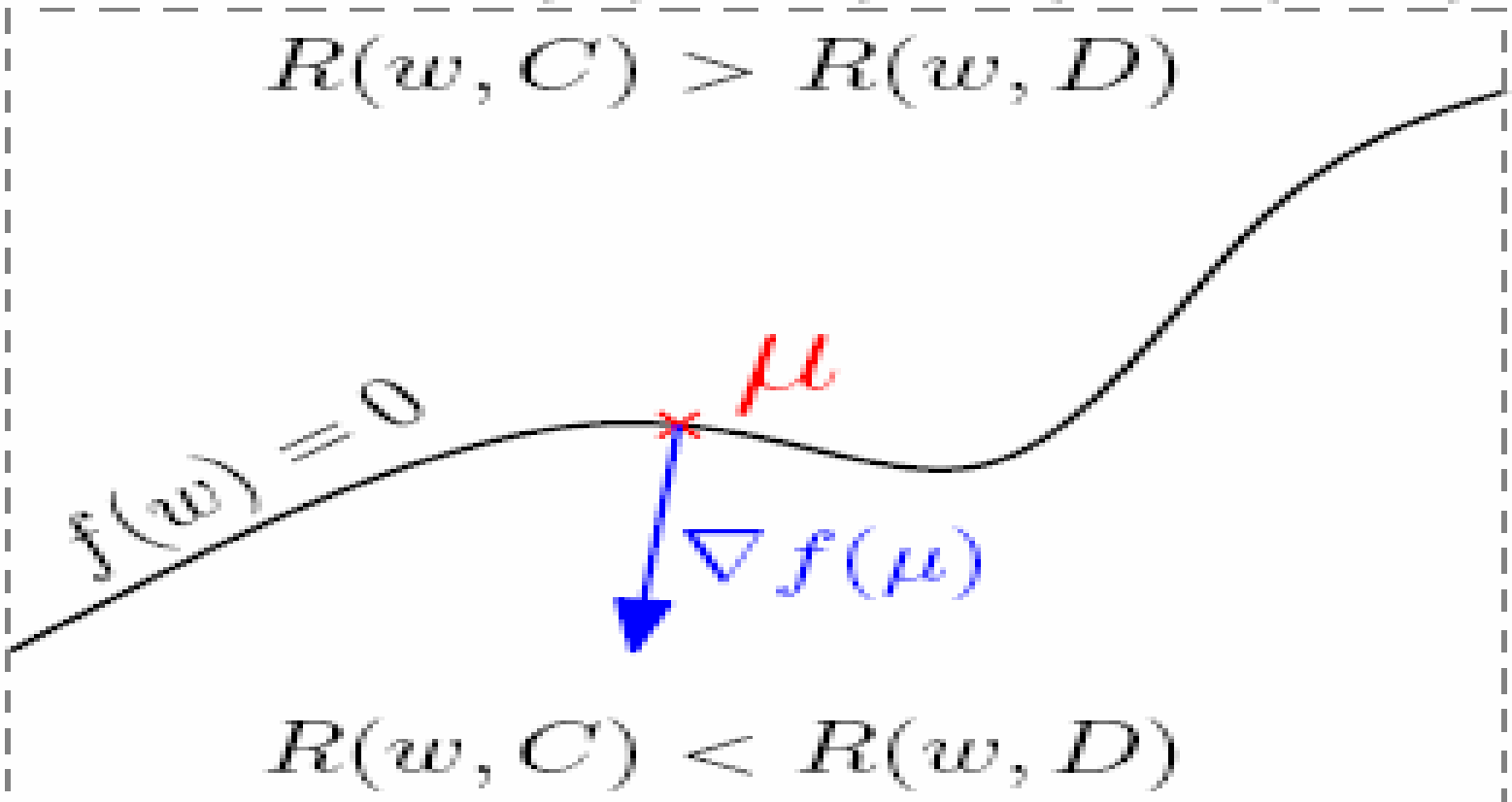
$$\lim_{m \to \infty} \Pr[R(S, \mathcal{C}) < R(S, \mathcal{D})] \quad \notin \quad \{0, 1\} .$$

- In other words, in the limit, there will be at least two P-optimal clusterings which are S-optimal with non-zero probability.

Proof idea (continued)

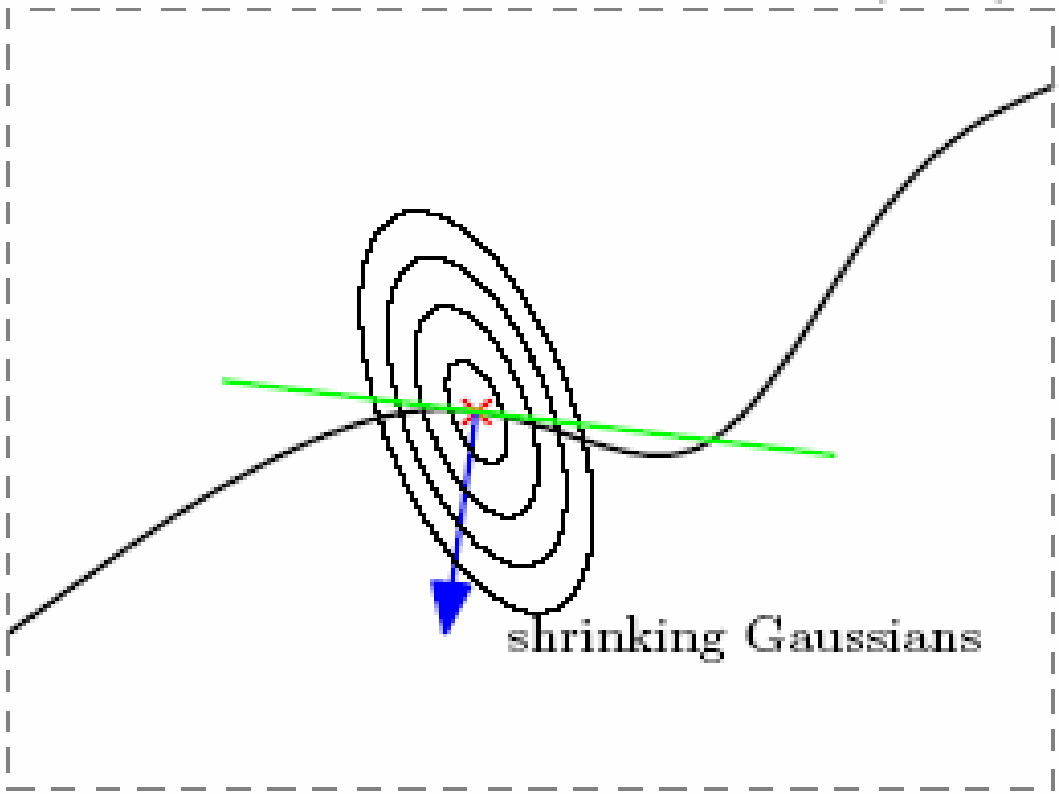
Consider the decision function:

$f : \mathbb{R}^n \longrightarrow \mathbb{R},\ f(w) = R(w, \mathfrak{D}) - R(w, \mathfrak{C})$



Proof Idea (continued)

By central limit theorem, as $m \to \infty$, $w \sim N(\mu, \Sigma)$ with $\Sigma \to 0$:



In a nut shell

Common Belief

A is stable on P iff

A picks correct number of clusters.

Our result

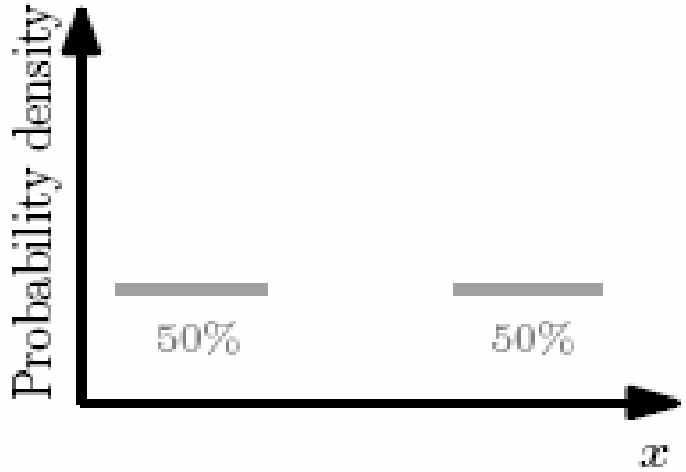
A is stable on P iff

the cost function that *A* minimizes has unique optimum over *P*.

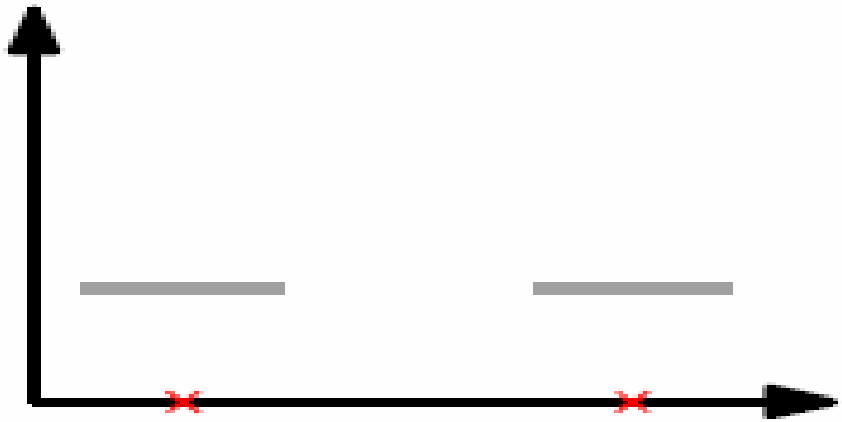
- Are the two statements the same?

Some Examples

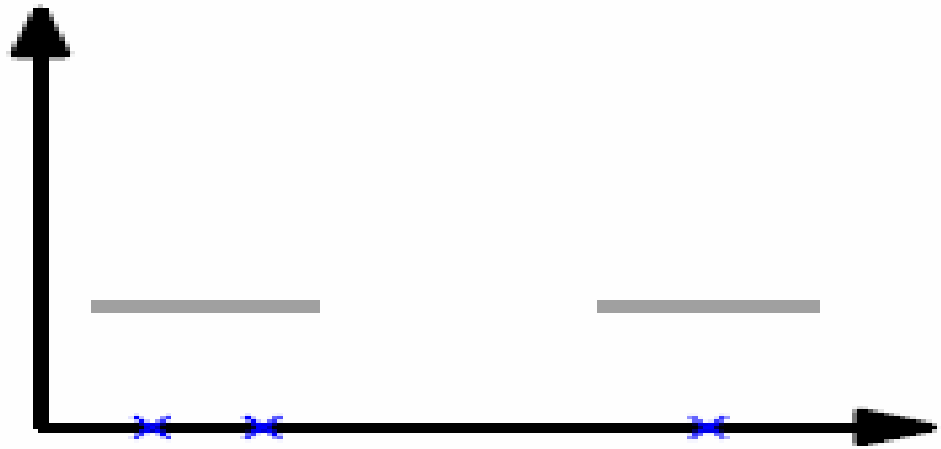
1D probability distribution



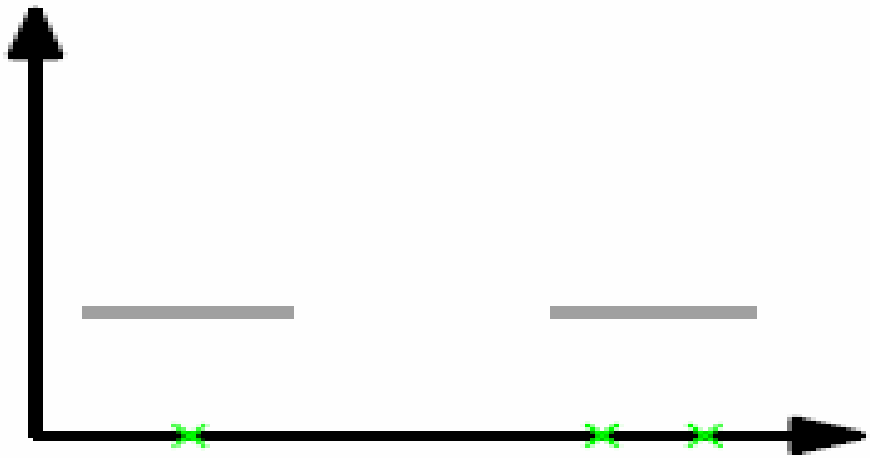
2 centers — stable



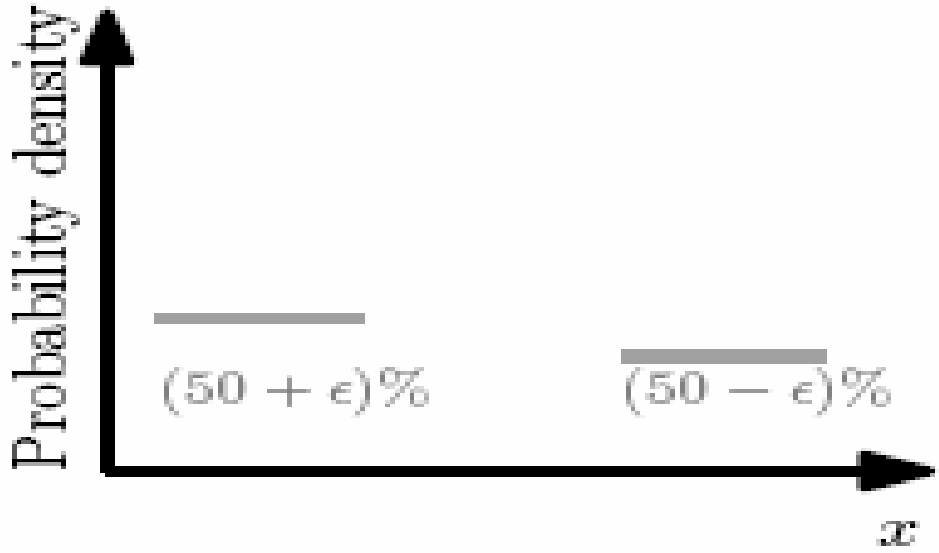
3 centers — solution #1



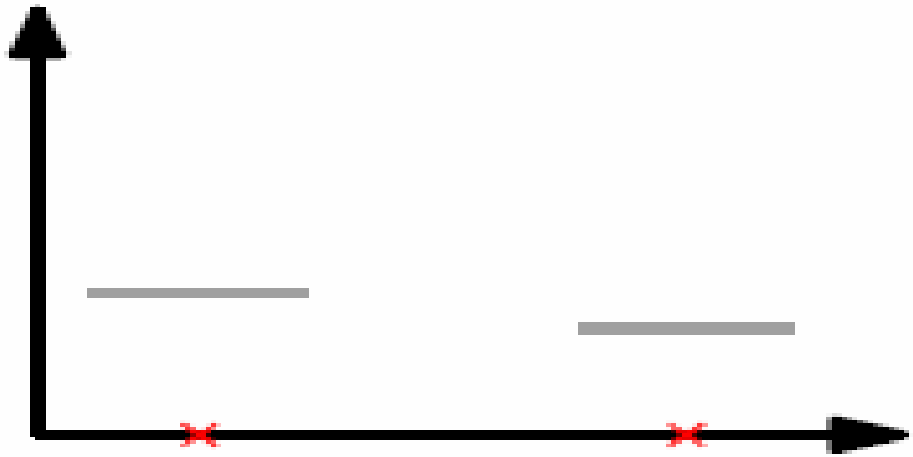
3 centers – solution #2 $\Longrightarrow$ unstable



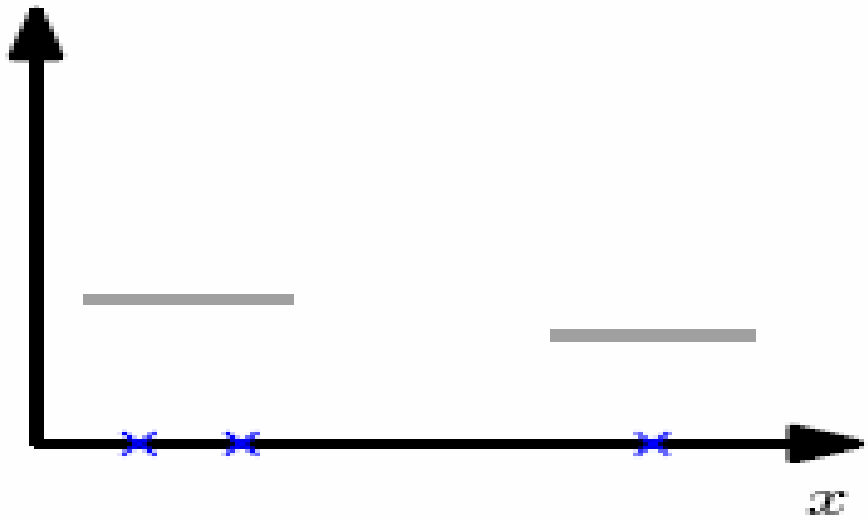
slightly asymmetric distribution



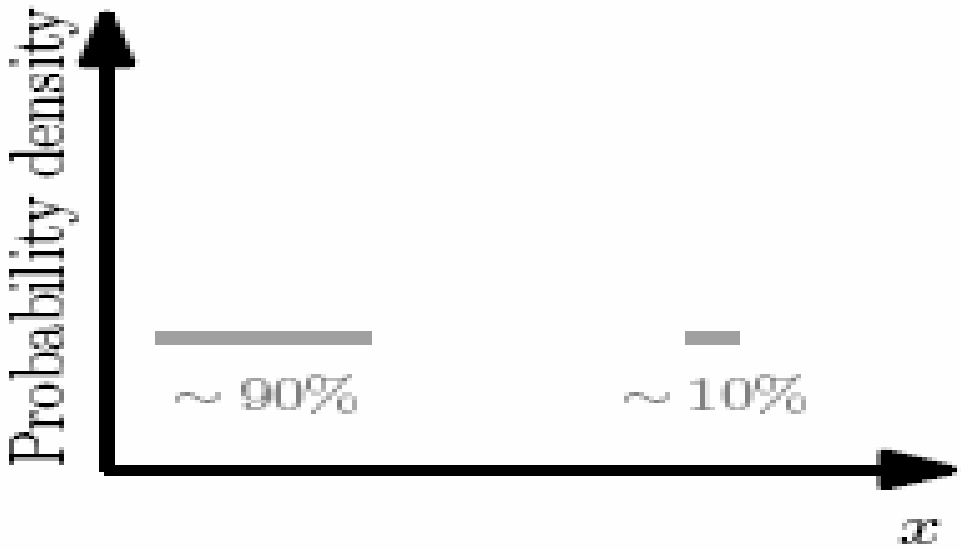
2 centers — stable



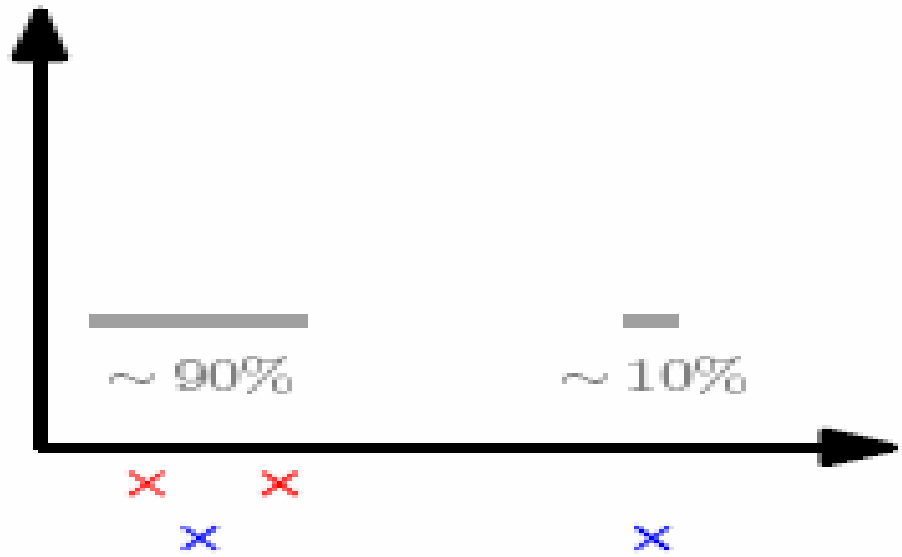
3 centers – stable



1D probability distribution



2 centers — unstable



The bottom line

- In practice, since no real data set is nicely symmetric, there will always be a unique cost-minimizing solution.

- Consequently, *Any choice of clustering parameters, on any real data set,*

  will always end up being stable.

Stability does not do the job we thought it did!

Take home message

Synthetic data sets can be misleading!

There is some inherent regularity in such data.

In some cases, such distinction between synthetic and real data, may crucially effect the behavior of a tested technique

(both in simulations and under mathematical analysis)

A Promising Direction

- In recent work, we consider the *relative* instability of two parameter settings

$$\frac{\mathrm{InStab}_m(A_k, P)}{\mathrm{InStab}_m(A_{k'}, P)}$$

[BD-Luxburg-Pal 07]
If A_k outputs boundaries that pass through considerably denser regions than those around the boundaries of $A_{k'}$'s output, then this ratio bounded below 1 (for all sufficiently large m).