Clustering Impossibility and Stability

B. Clarke[1]

[1]Dept of Medicine, CCS, DEPH
University of Miami
Joint with H. Koepke,
Stat. Dept., U Washington

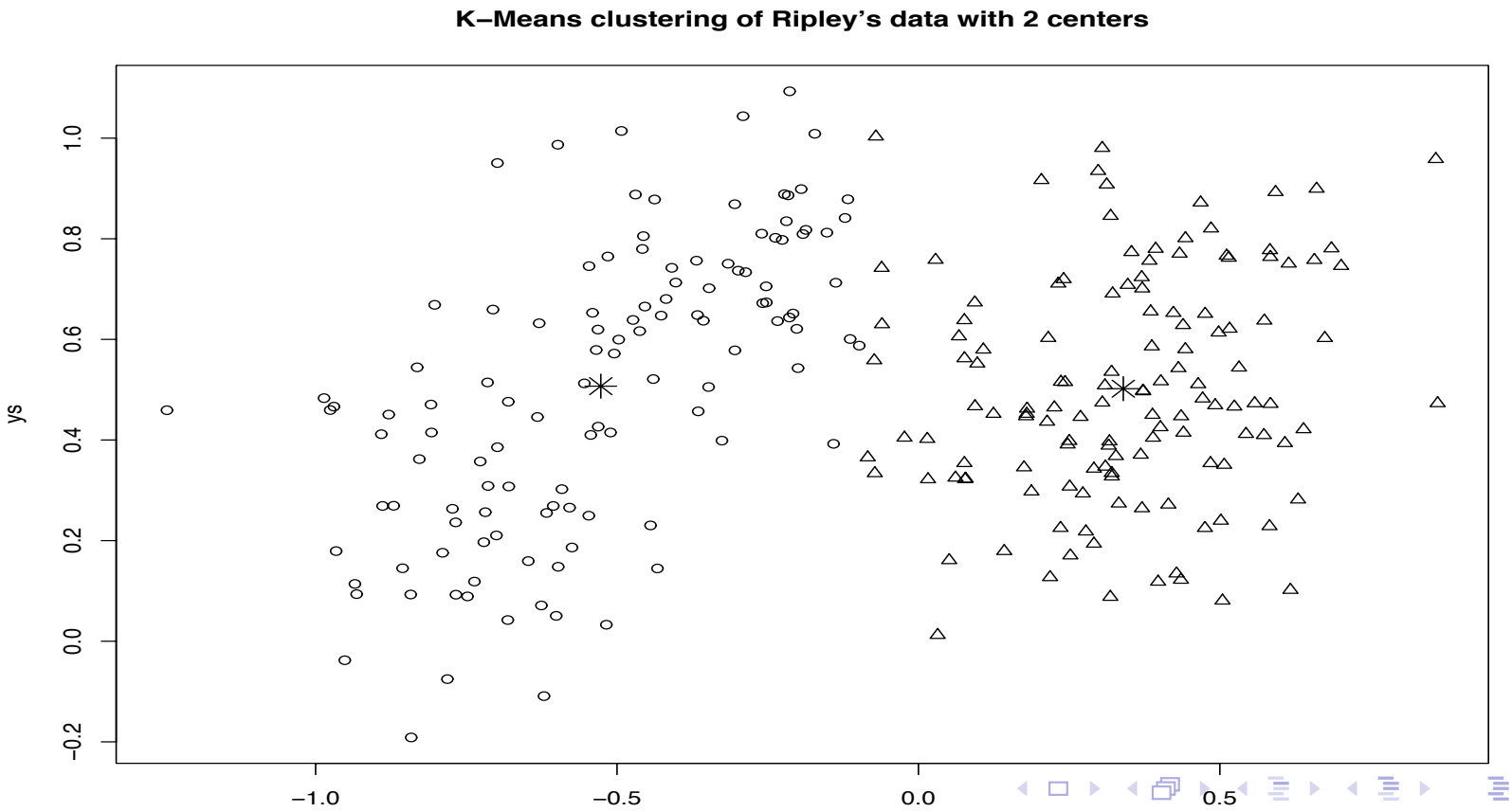
8 November 2011
Stat. Dept., FSU

Outline

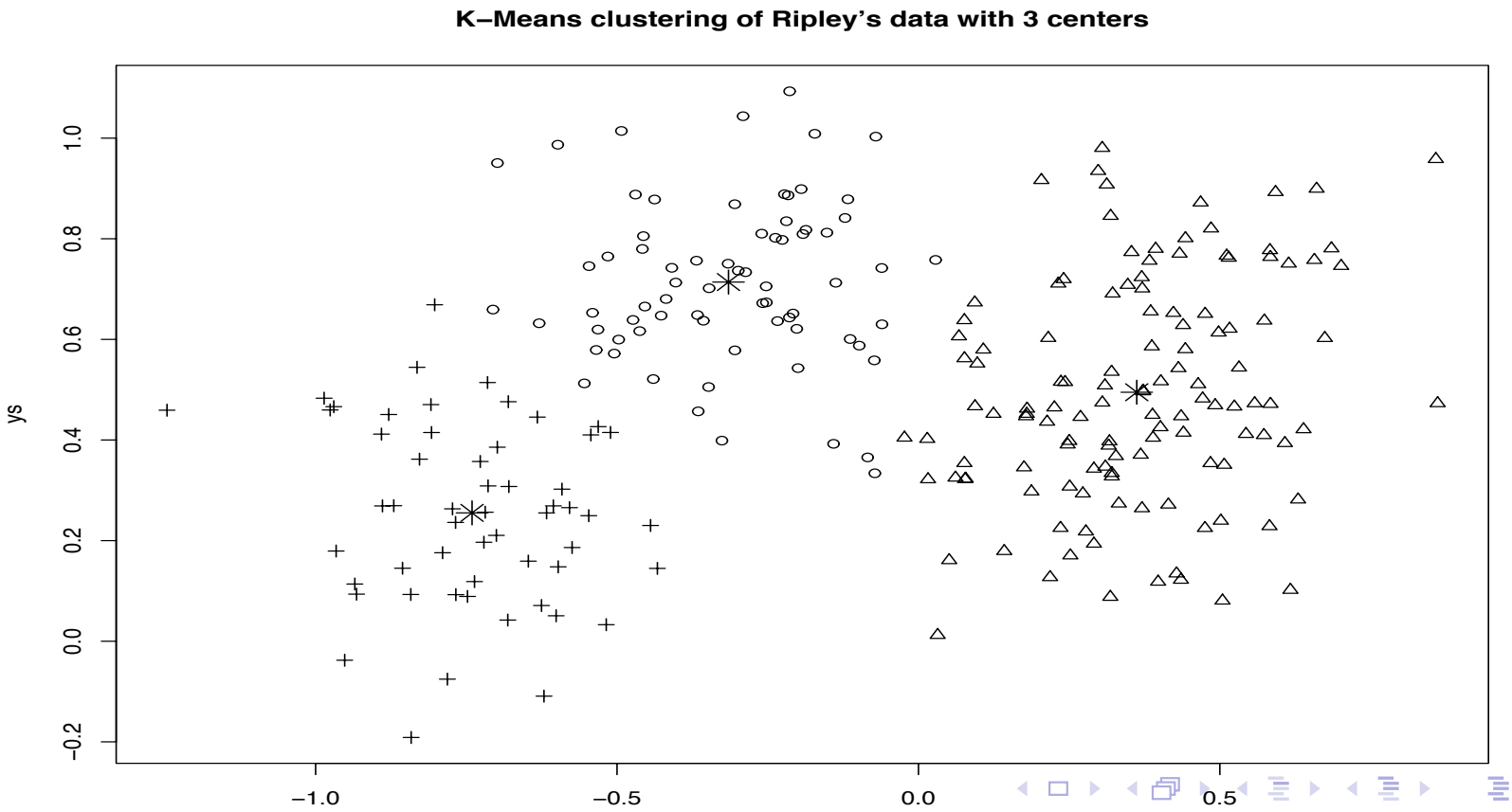
Basic Setting

- Imagine n points in D-dimensional space, say $x_i = (x_{1,i}, \ldots, x_{D,i})$ for $i = 1, \ldots, n$. They often group together with some points closer to each other and some points farther apart.
- Our goal is to put the points that 'belong together' in the same set and define different sets for the points that don't belong together.
- Such a set is called a cluster; a set of clusters is called a clustering (of the points).
- Thus we have $\mathcal{P} = \{P_1, \ldots, P_K\}$ where the P_k's are disjoint and $\cup_k P_k = S = \{x_i, \ldots, x_n\}$.
- Consider an example with $D = 2$, the Ripley data set. For this case $n = 250$.

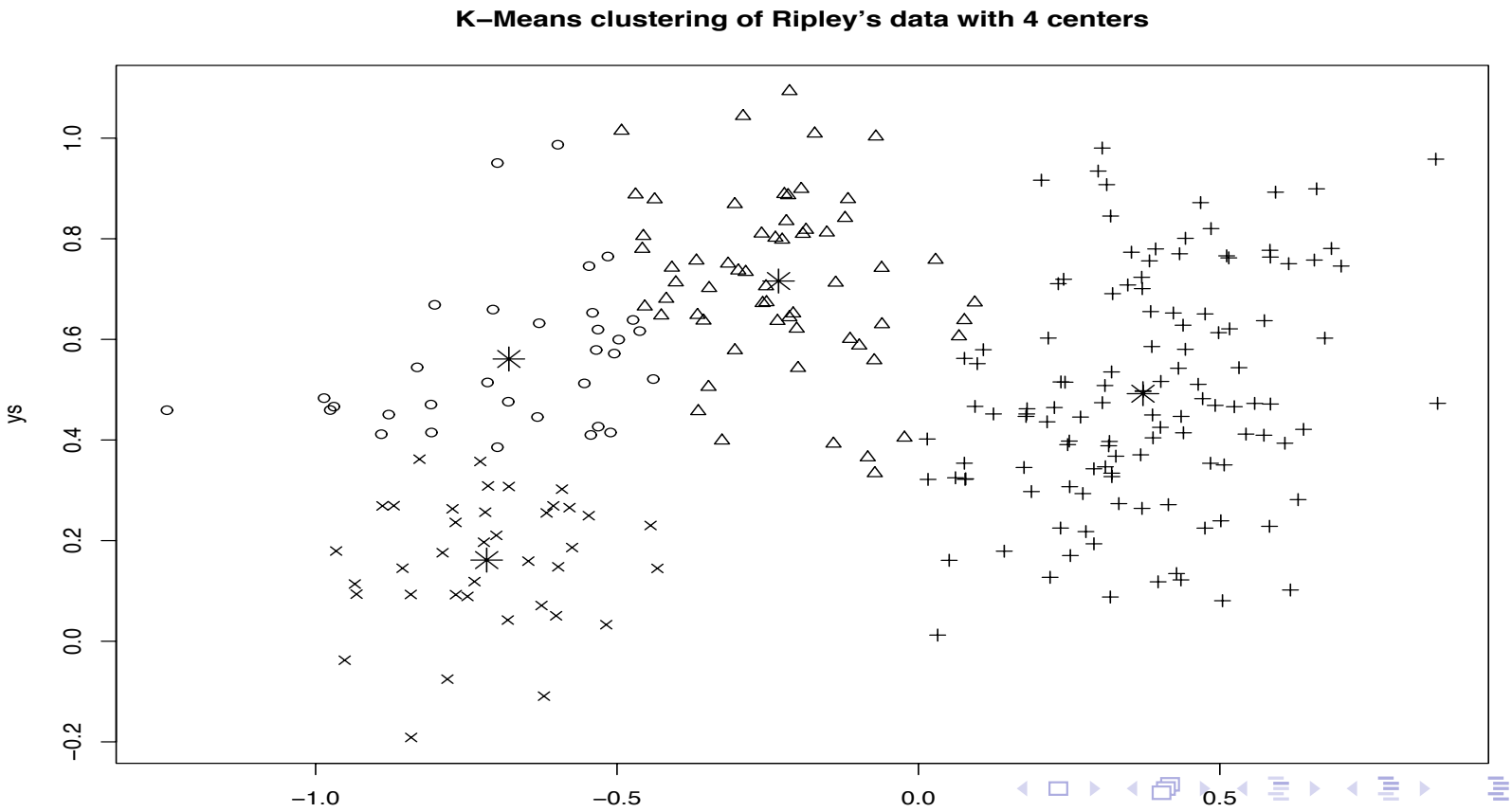
Ripley, 2 Centers



Ripley, 3 Centers



Ripley, 4 Centers



Interpretation

- This is synethetic data with two classes; the data for each class were generated by a mixture of two normal distributions. 125 data points from each of the two classes.
- Roughly, there is a left cluster and a right cluster. However, over many regions it is impossible to cluster perfectly because at many points there is a nonzero probability of each class occurring.
- As we allow more clusters, the 'P_1' on the left from $K = 2$ is split in $K = 3$ and split again for $K = 4$.
- Can regard the clusters as coming from (y_i, x_i) but the y_i's are missing (and we don't know the definition of Y).

How Were these Clusterings Found?

- K-means is the most common clustering technique partially because it scales up to high dimensions.
- Centroid based, invented by MacQueen (1967). The analyst picks the number of clusters K and makes initial guesses about the cluster means. The procedure starts with those K means, assigning points to whichever mean is closest to them, recomputing the mean, and repeating the procedure, hoping for convergence.
- The mean is like the pure type the cluster represents; regard each cluster as a modal region in $\mathbb{R}^D$.
- There are many, many, many clustering techniques – and many, many, many ways to assess stablity.

Without Y, Stability Is Like Fit

- We don't really have anything besides stability because we have no Y so we can't assess fit, predictive success, etc.
- On the basis of stability, we'd probably pick 2 or 3 clusters. 4 clusters looks OK too (and in some sense is right!) but we start wondering whether we've got too many.
- Problem: There are cases where no clustering is stable!
- OTOH, if we use a stability method to choose the number of clusters we want to be sure we got the right number.
- Think in terms of squared error and K-means but our results are much more general.

Statistical Model

- Think in terms of a signal plus noise model

$$\mathbf{Y} = \mathbf{x} + \varepsilon,$$

  where $\mathbf{Y}$, $\mathbf{x}$, and ε are $D \times n$ dimensional matrices.
- The D-dimensional data points in the columns of $\mathbf{Y}$ come from n non-random but unknown D-dimensional columns $\mathbf{x}_i$ of $\mathbf{x}$ plus a column from the random noise matrix ε.
- The entries in $\mathbf{Y}$ are the only values that are available to the experimenter.
- The $\mathbf{x}_i$'s are non-stochastic, represent 'centroids' and include multiplicity.
- Think of high dimensional, low sample size, i.e. large D and small n.

Cluster over Samples

- Two ways: Cluster over samples, i.e., over n vectors of length D, to find relationships among subjects.
- Or: Cluster over variables, i.e., over D vectors of length n to find relationships among explanatory variables.
- We focus on the first since that is often the primary goal.
- The problem: Evaluating different clusterings by a squared error cost function is only possible when the sum of squared distances between the $\mathbf{x}_i$'s, determined by the clusterings, has a rate at least $\sqrt{D}$ as D increases.
- Otherwise, meaningful clustering is not possible: Any ordering over clusterings is indistinguishable from random.
- Implication: Must do variable selection before clustering.

Not a Surprise...

- For finite D and n, Kleinberg (2003) defines 3 properties:
- Scale invariance gives insensitivity to changes in the unit of measurement.
- Richness means that the range of clusterings a procedure gives is large.
- Consistency encapsulates the idea that shrinking the distance between points in the same cluster or expanding the distance between points in different clusters should not affect the clustering itself.
- Kleinberg (2003) proves: No clustering procedure satisfies all three properties.

Possibility and Impossibility

- Beyer et al. (1999), Hinneburg et al. (2000) argue that as D increases min and max distances between points go to 0. Steinbach et al. (2003) observes this destroys clustering in the extreme case.
- Murtagh (2008 + earlier), Hall et al. (2005), look at high dimensional geometry in clustering argue impossibility.
- Devroye ey al. (2007) uses level sets to show consistency for number of clusters; this is density estimation which converges slowly.
- Pollard (1981, 1982) gives consistency of K-means clustering but under technical and restrictive conditions.
- Analogous results for classification, Jin (2009), Fan and Fan (2007), Devroye et al. (1996).

Cost Function

- Given n points and a number of clusters $K \leq n$, a partitioning $\mathcal{P} = \{P_1, P_2, ..., P_K\}$ is a set of K non-empty, disjoint exhaustive subsets of $\{1, 2, ..., n\}$.
- Given a partitioning $\mathcal{P} = \{P_1, P_2, ..., P_K\}$ on a set of data points $\mathbf{Y} \in \mathbb{R}^{D \times n}$, the squared error cost function is

$$\text{cost}(\mathbf{Y}, \mathcal{P}) = \sum_k \sum_{i \in P_k} \|\mathbf{Y}_{:i} - \overline{\mathbf{Y}}_k\|_2^2$$

where $\mathbf{Y}_{:i} = (Y_{1i}, Y_{2i}, ..., Y_{Di})$, $\overline{\mathbf{Y}}_\mathbf{k} = \text{mean}\{\mathbf{Y}_{:i} i \in P_k\}$ is the k-th cluster mean.

Differences of Cost Functions

- Let $\mathbf{Y}_d = (Y_{d1}, \ldots, Y_{dn})$, $\mathbf{x}_d = (x_{d1}, \ldots, x_{dn})$, and $\varepsilon_d = (\varepsilon_{d1}, \ldots, \varepsilon_{dn})$ for each $d = 1, \ldots, D$.
- Rewrite cost into dimensional components to see there is an $n \times n$ matrix $\mathbf{A} = \mathbf{A}(\mathcal{P})$ so that

$$\text{cost}(\mathbf{Y}, \mathcal{P}) = \sum_{d=1}^{D} \mathbf{Y}_d^T \mathbf{A} \mathbf{Y}_d = \text{trace}[\mathbf{Y}^T \mathbf{A} \mathbf{Y}].$$

- Given two partitions $\mathcal{P}$ and $\mathcal{Q}$, each has it's matrix $\mathbf{A}$ so there exists a matrix $\mathbf{B} = \mathbf{B}(\mathcal{P}, \mathcal{Q})$

$$\text{cost}(\mathbf{Y}, \mathcal{P}) - \text{cost}(\mathbf{Y}, \mathcal{Q}) = \text{trace}[\mathbf{Y}^T \mathbf{B} \mathbf{Y}].$$

Properties of $\mathbf{B} = \mathbf{B}(\mathcal{P}, \mathcal{Q})$

- Write $Z_d = \mathbf{Y}_d^T \mathbf{B} Y_d$ where $Y_d = x_d + \varepsilon_d$. Not hard to show:

$$\begin{aligned} E\varepsilon_d^T \mathbf{B} \varepsilon_d &= 0 \\ EZ_d &= \mathbf{x}_d^T \mathbf{B} \mathbf{x}_d \\ Z_d &= \text{cost}(\mathbf{Y}_d, \mathcal{P}) - \text{cost}(\mathbf{Y}_d, \mathcal{Q}) \\ &= (\mathbf{x}_d + \varepsilon_d)^T \mathbf{B} (\mathbf{x}_d + \varepsilon_d) \end{aligned}$$

- As events, $\left\{ \sum_{d=1}^D Z_d \geq 0 \right\} = \{\text{cost}(\mathbf{Y}, \mathcal{P}) \geq \text{cost}(\mathbf{Y}, \mathcal{Q})\}$.
- So, if $P(\sum_{d=1}^D Z_d \geq 0) \to 1/2$ means $\mathcal{P}$ is as good as $\mathcal{Q}$.

Impossibility as $D \to \infty$

- Let $\mathbf{Y}_d$, $\mathbf{x}_d$, and ε_d as before and suppose $\mathcal{P}$ and $\mathcal{Q}$ are any two distinct partitions of the n data points into K clusters, with cost difference matrix $\mathbf{B}$. If Condition F holds and if

$$\frac{1}{\sqrt{D}} \sum_{d=1}^{D} \mathbf{x}_d^T \mathbf{B} \mathbf{x}_d \ \to \ 0$$

  then

$$P\left(\text{cost}(\mathbf{Y}, \mathcal{P}) \leq \text{cost}(\mathbf{Y}, \mathcal{Q})\right) \ \to \ \frac{1}{2}$$

  as $D \to \infty$.
- This rests on a CLT for the Z_d's.
- Condition F holds whenever the ε's are continuous with IID components.

Standard Cases

- Note that $\sum_d \mathbf{x}_d^T \mathbf{B} \mathbf{x}_d = o_P(\sqrt{D})$ is trivially satisfied if $\sum_d \|\mathbf{x}_d\|_2^2 = o_P(\sqrt{D})$.
- The condition on the $\mathbf{x}_d$'s is tight. If

$$\sum_{d=1}^{D} \mathbf{x}_d^T \mathbf{B} \mathbf{x}_d = \mathcal{O}(\sqrt{D})$$

  then $\sum_d Z_d/\sqrt{D}$ may converge to a normal distribution shifted by a non-zero constant having a non-zero mean.
- More, a higher rate of growth would mean that the informative components eventually win out over the noise.

Corollary for Finite Dimensional Subspaces

- It is often assumed that the true data is 'sparse' in the sense that a small number of features contain almost all the information.
- However, we do not know which those are.
- The Corollary considers this case to emphasize that considering all the components of the dataset can make matters worse.
- Corollary: Suppose $\mathbf{Y} = \mathbf{x} + \varepsilon$, and suppose the columns of $\mathbf{x}$ vary over a fixed finite-dimensional subspace $S \subset \mathbb{R}^D$ as D increases. If the components of ε are IID then

$$\xi_D = P\left(\text{cost}(\mathbf{Y}, \mathcal{P}) \leq \text{cost}(\mathbf{Y}, \mathcal{Q})\right) \to \frac{1}{2} \quad \text{as} \quad D \to \infty.$$

Berry-Esseen Bounds on ξ_D

- In the sparse case we can bound ξ_D as a function of D.
- Berry-Esseen Theorem: Let $V_1, \ldots, V_D$ be IID with $EV_d = 0$, $EV_d^2 = \sigma^2$, and $E|V_d|^3 = \rho < \infty$. Let $\overline{V_D} = \frac{1}{D}\sum_{d=1}^{D} V_d$, and let F_D be the cumulative distribution function of $\sqrt{D}\overline{V_D}/\sigma$.
- Then there exists a constant δ such that

$$|F_n(t) - \Phi(t)| \leq \frac{\delta\rho}{\sigma^3\sqrt{D}}$$

  $\Phi(t)$ is the DF of $N(0,1)$ and $\delta \leq 0.7655$.
- Assume the ε_{id}'s have finite sixth moment and be IID along the dimension component d.

Decomposition: Signal vs. Noise:

- Suppose the first c dimension components are the only ones with non-zero signals.
- We have

$$\sum_{d=1}^{c} Z_d = \left[\sum_{d=1}^{c} \mathbf{x}_d^T \mathbf{B} \mathbf{x}_d\right] + \left[\sum_{d=1}^{c} \varepsilon_d^T \mathbf{B} \varepsilon_d + \sum_{d=1}^{c} \varepsilon_d^T \mathbf{B} \mathbf{x}_d + \sum_{d=1}^{c} \mathbf{x}_d^T \mathbf{B} \varepsilon_d\right].$$

$$= C + V_c$$

- This defines C as a constant and V_c as a sum of normal and Chi-square random variables.

$\sqrt{D}$ bounds on ξ_D

- Suppose the later $D - c$ components are drawn from an IID noise distribution with finite sixth moment. Then for $\alpha = \alpha(D)$ satisfying

$$\frac{e^{-\alpha(D)/8}}{\sqrt{D}} \to 0$$

we have that

$$\xi_D \in [\Phi^*(-a_D) - b_D, \Phi^*(-a_D) + b_D]$$

where and Φ^* indicates the result of integrating out α' from a normal distribution conditioned on α' where $V_c = \alpha'$ for $\alpha' < \alpha$ and multiplied by $1/P(\{V_c \leq \alpha\})$; $-a_D$ is the argument over which the integration is done.

More notation...

- In the theorem,

$$a_D = \frac{C + \alpha'}{\sigma\sqrt{D-c}}, \quad b_D = \frac{\delta\rho}{\sigma^3\sqrt{D-c}}$$

$$\sigma^2 = E(\text{cost}(\mathbf{Y}_d, \mathcal{P}) - \text{cost}(\mathbf{Y}_d, \mathcal{Q}))^2 = E(\varepsilon_d^T \mathbf{B} \varepsilon_d)^2,$$

$$\rho = E|\text{cost}(\mathbf{Y}_d, \mathcal{P}) - \text{cost}(\mathbf{Y}_d, \mathcal{Q}^3) = E|\varepsilon_d^T \mathbf{B} \varepsilon_d|^3$$

- The confidence intervals are distorted by the integration, however, the rate is preserved for each $\alpha' > \alpha$ giving an overall $\sqrt{D}$ convergence.
- We require $\alpha = o(\ln D)$ to control a probability conditioned on $V_c \geq \alpha$ to apply a Berry-Esseen Theorem pointwise in $\alpha' < \alpha$.

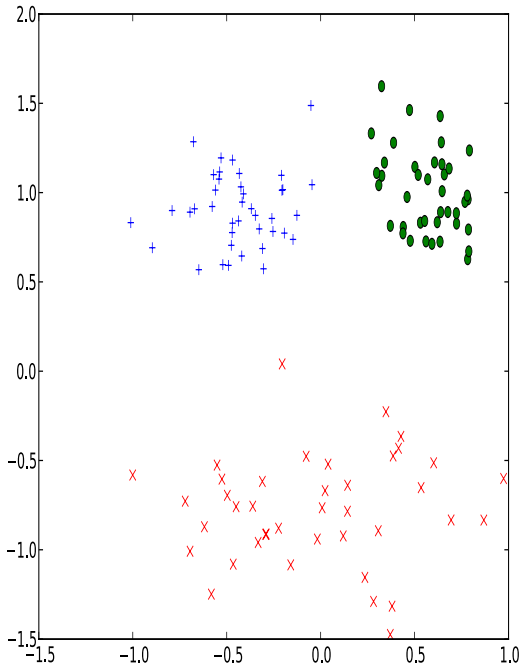
Corollary

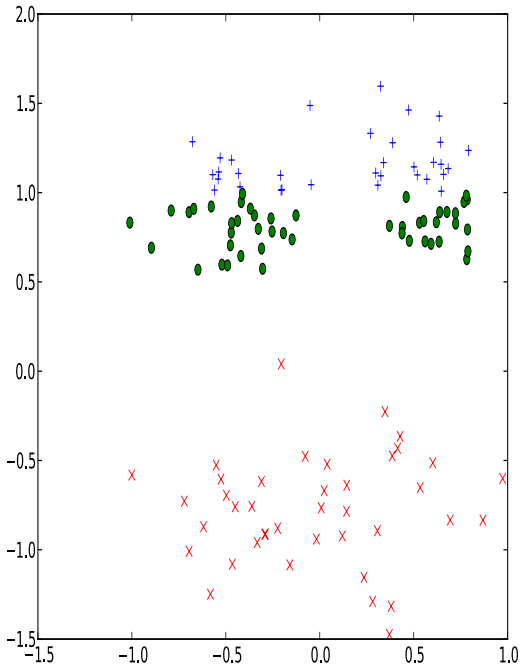
- In principle $\alpha = o(\ln D)$, can swamp the effect of C. However, in calculating these bounds on the cost curves we used $\alpha = 0$ and obtained reasonable results. This may mean the $o(\ln D)$ only takes effect for very large D or that the bound using α is loose.
- Corollary: The asymptotic convergence of $\xi_D - 1/2$ to 0 has rate at most $\mathcal{O}(1/\sqrt{D})$.
- Can generalize: Other cost functions, weaker hypotheses...

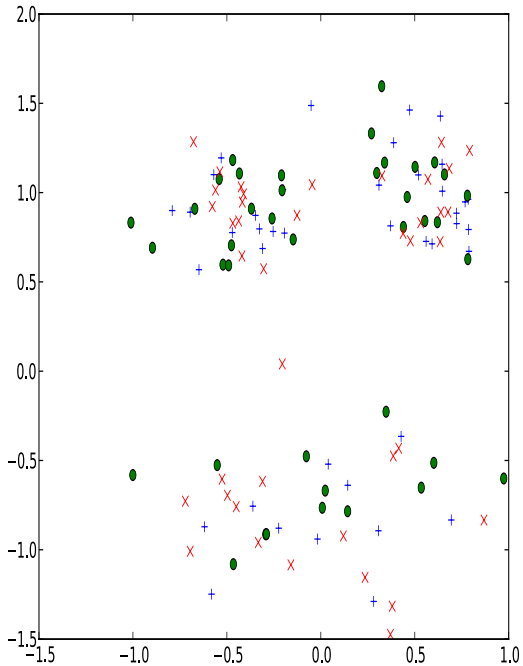
Increasing Noise Dimensions

- If D for a set of n vectors grows and the difference in costs of one clustering over another is calculated repeatedly then a curve $\xi = \xi_D$ can be given.
- We assume that the number of informative dimensions is much smaller than the apparent D, a sort of sparsity.
- Suppose a 2-dimensional data set of size $n = 120$ is generated by taking 40 IID data points from $N((-0.5, 1), \text{diag}(.2^2, .25^2))$, $N((0.5, 1), \text{diag}(.15^2, .25^2))$ and $N((0, -0.75), \text{diag}(.45^2, .35^2))$.
- The next panel shows the correct clustering, $\mathcal{P}_{best}$, a bad clustering $\mathcal{P}_{bad}$, and a terrible clustering $\mathcal{P}_{random}$.

Good, Bad, and Random Clusterings



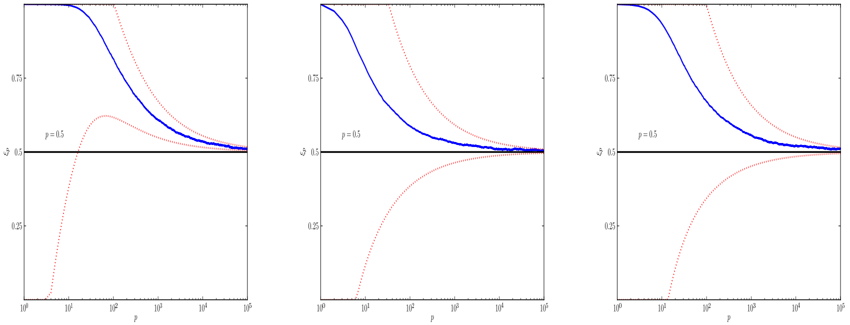




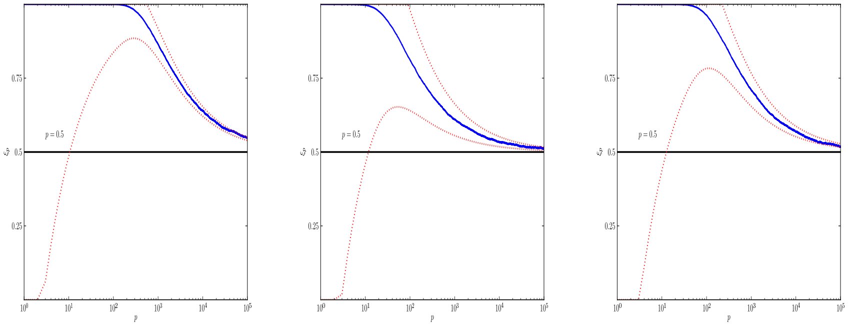
Adding Noise Dimensions

- We extend the data to data of dimension $D = 3, 4, \ldots$ by adding $D - 2$ pure noise coordinates.
- Then we computed ξ_D for 6 scenarios: Two choices of partitions $\mathcal{P}_{best}$ vs $\mathcal{P}_{bad}$ and $\mathcal{P}_{best}$ vs $\mathcal{P}_{rand}$ with three choices of noise, $Normal(0, 1)$, $\chi^2_2 - 2$, and a Student-t_4.
- The blue curves are the actual curves of ξ_D.
- The red curves are from the Berry-Esseen bounds. The vertical distance between the two curves for fixed D is a sort of 'confidence interval' for ξ_D.

Bad vs Good for Normal, χ^2_2, t_4



Random vs Good for Normal, χ^2_2, t_4



Computational Details

- We calculated ξ_D for each value of D using a Monte Carlo simulation. For speed of computation, we treated the difference in costs for the noisy components, Z_d, as a random variable and estimated its distribution using a pool of 10^7 samples, each one from an IID draw of ε.
- Given this EDF we sampled N values of Z_d, $Z_{d,1}, \ldots, Z_{d,N}$ until $\frac{1}{N}\sum_{j=1}^{N} \mathbb{I}_{\sum_{d=1}^{D} Z_{d,j} \geq 0}$ converged (we chose $N = 50000$); this gave our estimate of $\xi_D = P(\sum_{d=1}^{D} Z_d \geq 0)$ for D between 1 and 10^5.
- For the bounds on ξ_D from the theorem we took $\alpha' = 0$, found $\hat{\sigma}$ and $\hat{\rho}$ empirically, and used them to estimate a_D and b_D.

First observations

- Our simulations show clustering starts to have very appreciable probability of spuriousness in relatively benign settings.
- With $\mathcal{P}_{bad}$ and $\mathcal{P}_{good}$ we see that for $n = 120$ and 2 informative dimensions, by the time there are 20 to 30 variables the probability of distinguishing a good clustering from a bad one can fall to .7 or less in squared error.
- In all 3 cases with $\mathcal{P}_{bad}$, by the time around $D = 50$-ish, it becomes unreasonable to declare $\mathcal{P}_{bad}$ worse than $\mathcal{P}_{best}$.
- While it is easier to distinguish between $\mathcal{P}_{random}$ and $\mathcal{P}_{best}$, ξ_D still gets close enough to $1/2$ once D is much over 100 to cause problems.

Interpretations

- Reliability drops fastest for asymmetric noise ($\chi^2_2 - 2$), slowest for normal. The t_4 is in between.
- This because (1) the normal is very tight and so more noise has to accumulate to throw off the inferences, and
- (2) the asymmetry of the χ^2_2 with exponential tails provides more distortion than the symmetric t- distribution does even though it has heavier tails.
- The bounds reflect ξ_D well for large D.
- a_D gives the midpoint and tracks the blue line well.
- b_D controls the interval width and this narrows with D.
- The bounds therefore give a pretty good indication of the reliability of a clustering.

More generally....

- This example is highly favorable to cost function based evaluation and is pretty good... responding well to two informative components out of 20 or so with $n = 120$ is pretty impressive.
- In a genomic data set from Gordon et al (2005) when there are 5 informative variables among $D = 100$ variables, the probability that a good clustering is distinguishable from a poor clustering is around .6 for n in the low 30's.
- We suggest that clustering results from sparse data, i.e., $D >> n$ and few important variables, should be regarded as unreliable in the absence of further analysis.

Proposed Stability Assessment

- Fix D-dimensional data $x_1, \ldots, x_n$ and assume that for each K we have a clustering of size K $\hat{\mathcal{P}}_K = \{\hat{P}_{K1}, \ldots, \hat{P}_{KK}\}$.
- Assume it's centroid based with the property that

$$\forall j \; x \in \hat{P}_{Kj} \Leftrightarrow d(x, \hat{\mu}_{Kj}) \leq d(x, \hat{\mu}_{Kj'}) \quad j \neq j'$$

where

$$\hat{\mu}_{Kj} = \frac{\sum_{i=1}^{n} x_i \chi_{x_i \in \hat{P}_{Kj}}}{\sum_{i=1}^{n} \chi_{x_i \in \hat{P}_{Kj}}}$$

and d is a metric on $\mathbb{R}^D$.

Assumptions

- Each $\hat{\mathcal{P}}_K$ has a limit: $\exists \mathcal{P}_K = \{P_{K1}, \ldots, C_{KK}\}$ with

$$\mu(P_{Kj} \triangle \hat{P}_{Kj}) \to 0.$$

- This means that there are μ_{Kj}'s so that $\hat{\mu}_{Kj} \to \mu_{Kj}$.
- Let $\lambda_1, \ldots, \lambda_K \geq 0$ IID have continuous prior DF F.
- Bayesian empirical stability objective function is $nQ_n(K)$ is

$$\sum_{j=1}^{K} \sum_{i=1}^{n} \int \mathbb{I}\left(\forall \ell \neq j \; \lambda_j d(x_i, \hat{\mu}_{Kj}) \leq \lambda_\ell d(x_i, \hat{\mu}_{K\ell})\right) \mathbb{I}\left(x_i \in \hat{P}_{Kj}\right) dF(\lambda_1^K)$$

Population Version

- Bayesian empirical stability objective function is $Q_\infty(K)$ is

$$\sum_{j=1}^{K} E \int \mathbb{I}\left(\forall \ell \neq j \; \lambda_j d(x_i, \mu_{Kj}) \leq \lambda_\ell d(x_i, \mu_{K\ell})\right) \mathbb{I}\left(x_i \in P_{Kj}\right) dF(\Lambda_1^K)$$

- Let $\hat{K} = \arg\max Q_n(K)$ and $K_T = \arg\max Q_\infty(K)$ assuming K varies over a compact set.
- Consistency theorem seems possible!
- Newey-McFadden Theorem: If $\hat{Q}_n(K) \to Q_\infty(K)$ then $\hat{K} \to K_T$.

Does the Convergence Hold?

- Write

$$\hat{\phi}_j(x) = \int \mathbb{I}\left(\{\forall \ell \neq j \ \lambda_j d(x, \hat{\mu}_{Kj}) \leq \lambda_\ell d(x, \hat{\mu}_{K\ell})\}\right) dF(\lambda_1^K)$$

and

$$\phi_j(x) = \int \mathbb{I}\left(\{\forall \ell \neq j \ \lambda_j d(x, \mu_{Kj}) \leq \lambda_\ell d(x, \mu_{K\ell})\}\right) dF(\lambda_1^K)$$

- Then, it's enough to show that for $j = 1, \ldots, K$,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\phi}(X_i)\mathbb{I}\left(x_i \in \hat{P}_{Kj}\right) \to E\phi_j(X)\mathbb{I}\left(X \in P_{Kj}\right).$$

Why Should This Hold?

- We have assumed that for each K the clusterings have limits, so $\hat{P}_{Kj} \to P_K$.
- It's not hard to see that $\hat{\mu}_{Kj} \to \mu_K$ because

$$\hat{\mu}_{Kj} \approx \frac{\sum_{i=1}^{n} x_i \chi_{x_i \in \hat{P}_{Kj}}}{nP(X \in P_{Kj})} \to EX\mathbb{I}(X \in P_{Kj}) = \mu_{Kj}$$

- Moreover, we think $\sqrt{n}(\hat{\mu}_{Kj} - \mu_{Kj})$ is $AN(0, c)$ for some $c > 0$.

Upper and Lower Bounds within $\hat{\phi}_j$

- Using the usual properties of metrics we have

$$d(x, \mu_{Kj}) - d(\mu_{Kj}, \hat{\mu}_{Kj}) \leq d(x, \hat{\mu}_{Kj}) \leq d(x, \mu_{Kj}) + d(\mu_{Kj}, \hat{\mu}_{Kj})$$

  for the lower bound in $\hat{\phi}_j$.
- Also, we have

$$d(x, \mu_{\ell}) - d(\mu_{Kj}, \hat{\mu}_{K\ell}) \leq d(x, \hat{\mu}_{K\ell}) \leq d(x, \mu_{\ell}) + d(\mu_{K\ell}, \hat{\mu}_{K\ell})$$

  for the upper bound.
- Problem: There are n errors of order $\mathcal{O}(1/n)$ so if this doesn't work we may have to use a stronger technique like Hoeffding's ineqaulity or other large deviation property for the $\hat{\phi}_j(X_i)$'s.

Reasonable to Interpret High $Q_\infty(K)$ as Stability

- For the two cluster case, let μ_1, μ_2 be the centroids and $D_1 = d(X, \mu_1)$, $D_2 = d(X, \mu_2)$. Let $\Lambda_1 = \lambda_2/\lambda_1$, $\Lambda_2 = \lambda_1/\lambda_2$ and let G_{Λ_u} be the survival function for Λ_u.
- Can show:

$$Q_\infty(2) = E\mathbb{I}_{D_1/D_2 \leq 1} G_{\Lambda_1}(D_1/D_2) + E\mathbb{I}_{D_2/D_1 \leq 1} G_{\Lambda_2}(D_2/D_1).$$

- So, if D_1/D_2 small on P_1 then the first term is near $P(P_1)$ and P_1 is stable. If D_2/D_1) small on P_2 then the second term is near $P(P_2)$ This means $Q_\infty(2)$ is near 1 and so should $\hat{Q}_n(2)$ be. Generalizes to K clusters.
- Reverse the argument for D_1/D_2 large on P_1, D_2/D_1 large on P_2, i.e., many points near the boundary.

Relation to Silhouette Distance:

- Bayesian version of slihouette distance. Roughly: for each x_i let $a(i)$ be the within cluster dissimilarity of x_i. Compare $a(x_i)$ with the data of another cluster of which x_i is not a member.
- Let $b(x_i)$ be the lowest average dissimilarity of x_i with another cluster. The silhouette distance of a data point is

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

- Clearly, like ϕ_j, $-1 \leq s(i) \leq 1$. $s(x_i)$ is close to 1 if $a(x_i) << b(x_i)$: Small $a(x_i)$ means x_i is in the right cluster and large $b(x_i)$ means that not other cluster is close.

Relation to Silhouette distance, ctd.

- If $s(x_i)$ is near -1, then x_i really belongs in its nearest neighboring cluster.
- The average of $s(x_i)$ over a cluster is a measure of how tightly grouped all the data in the cluster are and the average of $s(x_i)$ over all the data is a measure of how good the clustering is.
- If K is too large or too small, some clusters will typically display much smaller silhouettes than the rest. Thus silhouette plots and averages may be used to determine the natural number of clusters within a dataset.

Other Assessments of Stabiity

- Gap statistic: Tibshirani et al. (2001). Idea: Look at $D_k = \sum_{i,j \in C_k} d(x_i, x_j)$ and find $W = \sum_k \alpha_k D_k$.
- Define $Gap(k) = E \log W - \log W$ and choose the smallest k so that $Gap(k) \geq Gap(k+1) - s_{k+1}$ where s_{k+1} is calculated from a simulation procedure.
- Looks for where validity measure levels off with K.
- Data-perturbation methods: These have largely used in a 'standard before best' heuristic: Choose the K which is within one SE of the best K under some criterion.
- Some of these techniques work in examples, but there aren't many theorems, interpretations for what sort of stability they mean, or demonstrations that they routinely give intuitively reasonable answers.

Does our Method Give Reasonable Answers?

- If there are K modes then $Q_\infty(K) \to 1$ as the modes separate. So, large values suggest stability.
- Look at mixtures of normals? OK, but must do this formally.
- Note that our method does not use data perturbation (it's Bayesian) and does not introduce subjectivity apart from the prior. It focusses on the how big/small factors must be to reverse the ordering of distances between points and their cluster centers.
- Ends up focussing on the boundary.

What next?

- Finish the proof the the consistency for choosing K.
- Finish giving an interpretation for the sense of stability the method is evaluating...how proximity to cluster boundaries affect $Q_\infty(K)$.
- Must verify more extensively that the optimization gives an intuitively reasonable number of clusters in standard cases.

What does this mean?

- The impossibility theorem and rates applies to clusters – doesn't matter how they were generated.
- Result not dependent on loss function or strong hypotheses; just how separated cluster centers are.
- For typical n, say 30-50, and typical clusterings, you really want 10% or more non-noise variables for reliable clustering. For n large, say 100-200, must have 5%.
- Stability looks like it can be used to get a consistent selection of the number of clusters – if a convergent collection of clusterings $\mathcal{P}_K$ is used.
- Seems to respond to boundary regions...promising....