Clustering Stability: Impossibility and possibility

B. Clarke

1Dept of Medicine, CCS, DEPH
University of Miami
Joint with H. Koepke,

15 December 2011
Outline

1. The Problem
2. Clustering Impossibility
3. Rate of Impossibility
4. Simulations
5. Bayesian Stability
6. Conclusions
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 $\bigcup_k P_k = S = \{x_i, \ldots, x_n\}$. 
Think in terms of a signal plus noise model

\[ Y = x + \varepsilon, \]

where \( Y, x, \) and \( \varepsilon \) are \( D \times n \) dimensional matrices.

The \( D \)-dimensional data points in the columns of \( Y \) come from \( n \) non-random but unknown \( D \)-dimensional columns \( x_i \) of \( x \) plus a column from the random noise matrix \( \varepsilon \).

The entries in \( Y \) are the only values that are available to the experimenter.

The \( x_i \)'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 $x_i$’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.
Given $n$ points and a number of clusters $K \leq n$, a partitioning $\mathcal{P} = \{P_1, P_2, \ldots, P_K\}$ is a set of $K$ non-empty, disjoint exhaustive subsets of $\{1, 2, \ldots, n\}$.

Given a partitioning $\mathcal{P} = \{P_1, P_2, \ldots, P_K\}$ 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 $\mathbf{Y}_{:i} = (Y_{1i}, Y_{2i}, \ldots, Y_{Di})$, $\overline{\mathbf{Y}}_k = \text{mean}\{\mathbf{Y}_{:i} | i \in P_k\}$ is the $k$-th cluster mean.
Differences of Cost Functions

- Let $Y_d = (Y_{d1}, \ldots, Y_{dn})$, $x_d = (x_{d1}, \ldots, x_{dn})$, 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 $A = A(\mathcal{P})$ so that

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

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

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

- Write $Z_d = Y_d^T B Y_d$ where $Y_d = x_d + \epsilon_d$. Not hard to show:

$$
E\epsilon_d^T B \epsilon_d = 0
$$

$$
EZ_d = x_d^T B x_d
$$

$$
Z_d = \text{cost}(Y_d, \mathcal{P}) - \text{cost}(Y_d, Q)
$$

$$
= (x_d + \epsilon_d)^T B (x_d + \epsilon_d)
$$

- As events, $\left\{ \sum_{d=1}^{D} Z_d \geq 0 \right\} = \{ \text{cost}(Y, \mathcal{P}) \geq \text{cost}(Y, Q) \}$.

- So, if $P(\sum_{d=1}^{D} Z_d \geq 0) \to 1/2$ means $\mathcal{P}$ is as good as $Q$. 

Let $Y_d, x_d, \text{ and } \varepsilon_d$ 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 $B$. If Condition F holds and if

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

then

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

as $D \to \infty$.

This rests on a CLT for the $Z_d$'s.

Condition F holds whenever the $\varepsilon$'s are continuous with IID components.
Standard Cases

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

\[
\sum_{d=1}^{D} x_d^T B x_d = \mathcal{O}(\sqrt{D})
\]

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 $Y = x + \varepsilon$, and suppose the columns of $x$ vary over a fixed finite-dimensional subspace $S \subset \mathbb{R}^D$ as $D$ increases. If the components of $\varepsilon$ are IID then

$$\xi_D = P(\text{cost}(Y, \mathcal{P}) \leq \text{cost}(Y, \mathcal{Q})) \rightarrow \frac{1}{2} \quad \text{as} \quad D \rightarrow \infty.$$
In the sparse case we can bound $\xi_D$ 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 $\delta$ 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 $\varepsilon_{id}$’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

$$\sum_{d=1}^{c} Z_d = \left[ \sum_{d=1}^{c} x_d^T B x_d \right] + \left[ \sum_{d=1}^{c} \varepsilon_d^T B \varepsilon_d + \sum_{d=1}^{c} \varepsilon_d^T B x_d \right] + \sum_{d=1}^{c} x_d^T B \varepsilon_d$$

$$= C + V_c$$

- This defines $C$ as a constant and $V_c$ as a sum of normal and Chi-square random variables.
Suppose the later $D - c$ components are drawn from an IID noise distribution with finite sixth moment. Then for $\alpha = \alpha(D)$ satisfying

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

we have that

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

where and $\Phi^*$ indicates the result of integrating out $\alpha'$ from a normal distribution conditioned on $\alpha'$ 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.
In the theorem,

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

\[ \sigma^2 = E(\text{cost}(Y_d, P) - \text{cost}(Y_d, Q))^2 = E(\varepsilon_d^T B \varepsilon_d)^2, \]

\[ \rho = E|\text{cost}(Y_d, P) - \text{cost}(Y_d, Q^3)| = E|\varepsilon_d^T B \varepsilon_d|^3 \]

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 $\alpha$ is loose.

- Corollary: The asymptotic convergence of $\xi_D - 1/2$ to 0 has rate at most $O(1/\sqrt{D})$.

- Can generalize: Other cost functions, weaker hypotheses...
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 $N((-0.5, 1), \text{diag}(0.2^2, 0.25^2))$, $N((0.5, 1), \text{diag}(0.15^2, 0.25^2))$ and $N((0, -0.75), \text{diag}(0.45^2, 0.35^2))$.

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 $\xi_D$ for 6 scenarios: Two choices of partitions $\mathcal{P}_{\text{best}}$ vs $\mathcal{P}_{\text{bad}}$ and $\mathcal{P}_{\text{best}}$ vs $\mathcal{P}_{\text{rand}}$ with three choices of noise, $\text{Normal}(0, 1)$, $\chi^2_2 - 2$, and a Student-$t_4$.
- The blue curves are the actual curves of $\xi_D$.
- 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 $\xi_D$. 
Bad vs Good for Normal, $\chi^2_2$, $t_4$
Random vs Good for Normal, $\chi^2_2$, $t_4$
Problems even in 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 $\mathcal{P}_{best}$, $\xi_D$ still gets close enough to $1/2$ once $D$ is much over 100 to cause problems.

- Reliability drops fastest for asymmetric noise ($\chi_2^2 - 2$), slowest for normal. The $t_4$ is in between.
Proposed Stability Assessment

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

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

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 $\mathbb{R}^D$. 

B Clarke  Cluster Impos. + Stab.
Assumptions

- Each $\hat{P}_K$ has a limit: $\exists \mathcal{P}_K = \{P_{K1}, \ldots, P_{KK}\}$ with
  \[\mu(P_{Kj} \triangle \hat{P}_{Kj}) \to 0\].

- Assume that in the limit
  \[\forall j \ x \in P_{Kj} \iff d(x, \mu_{Kj}) \leq d(x, \mu_{Kj'}) \quad j \neq j'\]
  where
  \[\mu_{Kj} = EX_1 | X_1 \in P_{Kj}\].

- This means $\hat{\mu}_{Kj} \to \mu_{Kj}$.
- Let $\lambda_1, \ldots, \lambda_K \geq 0$ IID have continuous prior DF $F$.
- Consider the set
  \[\hat{S}_{ij}(\lambda_1, \ldots, \lambda_K) = \{\forall l \neq j \ \lambda_j d(x_i, \hat{\mu}_{Kj}) \leq \lambda_l d(x_i, \hat{\mu}_{Kl})\}\]
The further apart the $d(x_i, \hat{\mu}_{Kj})$’s are, the bigger the set of $\lambda_j$’s for which the inequality holds.

Integrating over $\lambda^K = (\lambda_1, \ldots, \lambda_K)$, restricting to $\hat{P}_{Kj}$, summing over $j$, and averaging over $i = 1, \ldots, n$ gives a Bayesian empirical stability objective function by setting

$$Q_n(K) = \sum_{j=1}^{K} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{x_i \in \hat{P}_{Kj}\}} \int \mathbb{I}_{\hat{S}_{ij}(\lambda^K)}(X_i) dF(\lambda^K_1)$$
Population Version

- Consider the set

\[ S_{ij}(\lambda_1, \ldots, \lambda_K) = \{ \forall \ell \neq j \, \lambda_j d(x_i, \mu_{Kj}) \leq \lambda_\ell d(x_i, \mu_{K\ell}) \} \]

- Integrating over \( \lambda^K = (\lambda_1, \ldots, \lambda_K) \), restricting to \( P_{Kj} \), summing over \( j \), and averaging over \( i = 1, \ldots, n \) gives a Bayesian empirical stability objective function by setting

\[ Q_\infty(K) = \sum_{j=1}^{K} E_{\{X_1 \in P_{Kj}\}} \int I_{S_{1j}(\lambda^K)}(X_1) dF(\lambda^K) \]

- We want \( Q_n(K) \to Q_\infty(K) \).
Does $Q_n(K) \to Q_\infty(K)$?

- Write

$$\hat{\phi}_j(x) = \int \mathbb{I}(\{\forall \ell \neq j \lambda_j d(x, \mu_{K_j}) \leq \lambda_{\ell} d(x, \mu_{K_\ell})\}) \ dF(\lambda^K_1)$$

and

$$\phi_j(x) = \int \mathbb{I}(\{\forall \ell \neq j \lambda_j d(x, \mu_{K_j}) \leq \lambda_{\ell} d(x, \mu_{K_\ell})\}) \ dF(\lambda^K_1)$$

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

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_j(X_i) \mathbb{I}(x_i \in \hat{P}_{K_j}) \to E\phi_j(X) \mathbb{I}(X \in P_{K_j}).$$
Convergence result for $Q_n(K)$

- When $\hat{\mu}_j \to \mu_j$ for $j = 1, \ldots, K$ it can be shown that
  $$Q_n(K) \to Q_\infty(K).$$

- For any finite range of $K$ we also have
  $$\sup_{K \in [K_1, K_2]} |Q_n(K) - Q_\infty(K)| \to 0$$
as $n \to \infty$.

- Now, for each $K$ choose a single clustering, perhaps by $K$-means (optimal for that $K$) or by different choices of cutoff on a dendrogram for hierarchical clustering.
Consistency for $K$

- Let
  \[ \hat{K} = \arg\max_{K \in [K_1, K_2]} Q_n(K) \]
  and let
  \[ K_T = \arg\max_{K \in [K_1, K_2]} Q_{\infty}(K). \]

- So, if we have that on $[K_1, K_2]$ all $\hat{\mu}_{K_j} \to \mu_{K_j}$, then we have that for all $K$, $Q_n(K) \to Q_{\infty}(K)$ uniformly.

- Since $[K_1, K_2]$ is compact and $Q_{\infty}(K)$ is (trivially) continuous on $[K_1, K_2]$ we can invoke the Newey-McFadden Theorem.

- Conclusion: $\hat{K} \to K_T$, i.e., we have consistency for the choice of $K$ subject to $Q_{\infty}$ being an intuitively reasonable encapsulation of how many clusters there should be.
Properties of $Q_{\infty}(K)$

- For $K = 2$, let $\mu_j = E(X|C_j)$ and $D_j = d(X, \mu_j)$. Let $\Lambda_1 = \lambda_2/\lambda_1$, $\Lambda_2 = \lambda_1/\lambda_2$ and let $G_{\Lambda_u}$ be the survival function for $\Lambda_u$.

- 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 $P(P_{21}$ and $P_{21}$ is stable. If $D_2/D_1$ small on $P_{22}$ then the second term is near $P(P_{22})$. This means $Q_{\infty}(2)$ is near 1 and so should $\hat{Q}_n(2)$ be. Generalizes to $K$ clusters.

- That is, if the distribution of $X$ concentrates at $\mu_1$ and $\mu_2$ then $Q_{\infty}(2)$ goes to 1.
For $D_1/D_2$ large on $P_{21}$, i.e., $D_1/D_2 \to 1$ we expect many points in $P_{21}$ to be close to the boundary between $P_{21}$ and $P_{22}$. Similarly if $D_2/D_1$ close to 1.

In these cases,

$$Q_\infty(2) \to P(P_{21}) G_{\Lambda_1}(1) + P(P_{22}) G_{\Lambda_2}(1) = 1/2.$$ 

Since $1/2 \leq Q_\infty(2) \leq 1$, it seems reasonable to regard $Q_\infty(K)$ as indicating stability.

In general, $1/K \leq \phi_\infty(K) \leq 1$.

If there are $K$ modes then $Q_\infty(K) \to 1$ as the modes separate. If the $K$ modes get closer together, $Q_\infty(K) \to 1/K$.

Again, $\phi_\infty(K)$ seems to assess stability.
What next?

- 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. Maybe look at mixtures of normals?
The impossibility theorem and rates applies to clusterings – 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 reasonable collection of clusterings $\mathcal{P}_K$ is used.

Stability criterion seems to respond to boundary regions.