Clustering - How to choose $K$?
- Ad hoc approaches: data-driven, but little theory
- Methods based on maximizing objective functions (M-estimators)
  e.g. clustering via K-means, spectral clustering
  has frequentist guarantees but hard to be data-driven
- Parametric methods such as AIC, BIC requires hidden assumptions

Model-based approach:
\[ P(X|\phi) = \sum_{k=1}^{K} \pi_k P(X|\phi_k) \]

Let \[ G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k} \]

Mixing measure
Model: \[ \phi | G \sim G \]
\[ X | \phi \sim P(X|\phi) \]

$G$ Random. Need a prior distribution on $G$.
want $K$ unbounded ($K = \infty$) \Rightarrow Stick-breaking prior.
This is a distribution over $G$ of form:
\[ G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \]
\[ \sum_{k=1}^{\infty} \pi_k = 1 \]

Chinese Restaurant Process (CRP)
A random process in which customers sit down in a Chinese Restaurant
with infinite number of tables.
CRP: first customer sits at the first table
- m-th subsequent customer sits at a table drawn from the following dist:
\[ \Pr(\text{sits at previously occupied table } i) = n_i \]
\[ \Pr(\text{sits in a new (unoccupied) table}) = \alpha \delta_0 \]
\[ n_i = \# \text{customers sitting at table } i \] (away previous m-1 customers)

CRP and Clustering:
CRP induces a partition of customers into clusters (subsets)
distinct and # of clusters (tables)

Let \( \pi_k \) be the proportion of customers sitting at table \( k \),
\( \forall k \in \mathbb{N} \setminus \{0\} \) (below \( n_k \)'s, \( \pi_k \)'s are).
Then we obtained a random measure:
\[ \mathcal{G} = \sum_{k=1}^{\infty} n_k \delta_{\pi_k} \]

What can we say about \( \mathcal{G} \)?

Polya urn model: CRP is a particular example of a general class of prob. models known as Polya urn.

Let \( \theta_1, \ldots, \theta_n \) iid \( \mathcal{G} \) there are implicitly among the \( \theta_i \)'s. The Polya urn model is as follows:
\[ \theta_1 \sim \mathcal{G}_0 \]
\[ \theta_i | \theta_1, \ldots, \theta_{i-1} \sim \alpha \mathcal{G}_0 + \sum_{j=1}^{i-1} \delta_{\theta_j} \]
\[ P(\theta_1, \theta_2) = P(\theta_1) P(\theta_2 | \theta_1) \]
\[ \leq P(\theta_1) \left( \int_0^1 G_0(d\theta_2) + \sum \delta_{\theta_i}(\theta_2) \right) \]
\[ \leq \alpha \int_0^1 G_0(d\theta_2) G_0(d\theta_1) + G_0(d\theta_1) \mathbb{1}(\theta_1 = \theta_2). \]

So, \[ P(\theta_1, \theta_2) = P(\theta_2, \theta_1). \]

\[ P(\theta_1, \theta_2, \theta_3) = P(\theta_1, \theta_2) P(\theta_3 | \theta_1, \theta_2) \]
\[ = P(\theta_2, \theta_1) P(\theta_3 | \theta_2, \theta_1) = P(\theta_2, \theta_1, \theta_3) \]

in general \( \theta_1, \theta_2, \ldots, \theta_n \) are infinitely exchangeable!

By De Finetti's theorem, \( \exists \) a random mixing dist \( G \) s.t.

\[ \theta_1, \ldots, \theta_n \mid G \overset{iid}{\sim} G. \]

So that \[ P(\theta_1, \ldots, \theta_n) = \int \prod P(\theta_i | G) d\Pi(G) \]

What is \( G \)?

Stick-breaking process:

Let \( \beta_k \overset{iid}{\sim} \text{Beta}(1, \alpha_0) \quad k = 1, 2, \ldots. \)

Define

\[ \pi_k = \beta_k \quad k = 1 \]
\[ \pi_k = \beta_k \prod_{k'=1}^{k-1} (1 - \beta_{k'}) \quad k = 2, 3, \ldots. \]

Easy to check that \( \sum_{k=1}^{\infty} \pi_k = 1. \)

To check:

\[ 1 - \sum_{k=1}^{\infty} \pi_k = 1 - \beta_1 - \beta_2 (1 - \beta_1) - \beta_3 (1 - \beta_2) \]
\[ = (1 - \beta_1) (1 - \beta_2) - \beta_3 (1 - \beta_2) (1 - \beta_1) \]
\[ = \cdots \]
\[ = (1 - \beta_1) \cdots (1 - \beta_k) \rightarrow 0 \quad \text{as} \quad k \rightarrow \infty. \]

Then: let \( G = \sum_{k=1}^{\infty} \delta_{\phi_k} \) where \( \phi_k \overset{iid}{\sim} G_0 \) then this is the dist that bests Polya's urn.
In fact, the random measure \( G \) is called a Dirichlet Process. Where does Dirichlet come in? Hint: Beta dist was used.

Recall Dirichlet Dist (finite) is conjugacy (Dir-Multinomial).

\[
P(t | \alpha) = \frac{P(t) \prod \alpha_i^{t_i} \alpha^{-1}}{\prod \Gamma(\alpha_i)}
\]

\( \prod \alpha \sim \text{Dir}(\alpha) \)

Let \( \gamma_1, ..., \gamma_n \mid \prod \sim \text{Malt} (\prod) \)

Then

\[
P(\gamma_1, ..., \gamma_n) = \int P(\prod) \prod P(\gamma_i | \prod) dP(\prod | \alpha)
\]

\[
= \int \prod_{i=1}^{n} \frac{\prod \gamma_i \prod_{j \neq i} \alpha_j}{\prod_{j} \Gamma(\alpha_j)} \prod_{j=1}^{n} \Gamma(\alpha_j) d\prod
\]

\[
= \frac{\prod(\alpha_1 + ... + \alpha_n)}{\prod_{k} \Gamma(\alpha_k)} \prod \frac{\Gamma(\alpha_1 + ... + \alpha_n + n)}{\prod(\alpha_1 + ... + \alpha_n + n)}
\]

\[
P(\gamma_1, ..., \gamma_n) = \frac{\prod(\alpha_1 + \alpha_n)}{\prod(\alpha_1 + ... + \alpha_n + n)}
\]

Suppose we can somehow let \( K \to \infty \) yet \( \alpha_k = \frac{\alpha}{K} \)

so that \( \sum \alpha_k = \alpha \) fixed, then

\[
P(n_1, n_2, ..., n_K) \sim \text{Mult} \quad \text{if } \gamma_n = k
\]

there is a random probability that \( \gamma_n \neq \gamma_1, ..., \gamma_{n-1} \)

the probability \( \to \frac{\alpha}{\alpha + n} \)

How is this related to DP?

\[\begin{align*}
\text{As } k \to \infty, & \quad \prod \to \text{Dirichlet priors} \\
\text{Order } (n_1, ..., n_K) & \text{ in decreasing order}
\end{align*}\]
• Now we're ready to state the original def of a Dirichlet process.
  - **Stochastic process vs. RV's.** Real-valued.
  - Elementary prob. is: RV's are function \( X : \Omega \rightarrow \mathbb{R} \).
  - \( \Omega \): set of events
  - \( P(X \leq t) = \mathbb{E}(\xi X(w) < t) \).
  - A stochastic process can be viewed as a dist/mixture in space of more complex objects such as functions or measures.
  - Therefore, random function or random measure.
  - **Gaussian process** as log dist in \( \mathbb{R}^T \)
    - is defined as a collection of RV's \( \xi X(t) : \mathbb{R} \rightarrow \mathbb{R} \) \( t \in T \).
    - s.t. every finite collection of \( \xi X(t) \) is a multivariate Gaussian
    - and that they are consistent with each other through marginalization.

• Kolmogorov's thin says that there is a meaningful dist in the space of functions \( \xi X : T \rightarrow \mathbb{R} \)
  - which is called a Gaussian process.
  - For Gaussian, consistency is obviously satisfied: \((X_1, X_2) \sim \mathcal{N} \) then \( X_1 \sim \mathcal{N} \).

• A probability measure (dist) \( G \) is a function that maps \( \Omega \) to \([0,1] \)
  - \( \xi \in \Omega \) \( \xi(A) = \text{probability } \xi \in A \) under \( G \).
  - i.e., if \( X \sim G \) then \( P(X \in A) = G(A) \).

• To define a Random measure \( G \):
  - \( G(A) \) is a collection of RV's.
**Def.:** \( \Theta \subset (\mathbb{R},+,\cdot) \) a prob. space.

\( G \) is a random measure \( \Theta \)-st. \((\mathbb{R},+,\cdot)\) s.t.

- for any partition \((A_1, A_2, \ldots, A_k)\) of \( \Omega \), \( k \in \mathbb{N} \)
- \( G(A_1), G(A_2), \ldots, G(A_k) \sim \text{Dir}(\alpha_1 G_0(A), \ldots, \alpha_k G_0(A)) \)

Then \( G \) is said to be distributed according to a Dirichlet process

\( G_0 \) is called a base measure \( \text{(mean measure)} \)

\( \alpha \) concentration parameter.

Now such DP exists due to Courant's properties of their dists.

**Dir dist.:** if \( (Y_1, \ldots, Y_k) \sim \text{Dir}(\alpha_1, \ldots, \alpha_k) \), then \( (Y_1+\delta, \ldots, Y_k+\delta) \sim \text{Dir}(\alpha_1+\delta, \ldots, \alpha_k+\delta) \)

**Fact:** if \( G \sim \text{DP}(\alpha_1, G_0) \)
- \( G(A) \sim \text{Beta}(\alpha_1 G_0(A), \alpha_1(1-G_0(A))) \)
- \( \mathbb{E} G(A) = \frac{G_0(A)}{\alpha_1} \)
- \( \text{Var} G(A) = \frac{G_0(A)(1-G_0(A))}{\alpha_1^2 (\alpha_1+1)} \)
- \( \frac{G_0(A)}{\alpha_1} = \frac{G_0(A)}{\alpha_1} \frac{1-G_0(A)}{\alpha_1} \)

**Question:** if \( G \sim \text{DP}(\alpha_1, G_0) \)
- \( \Theta | G \sim G \)

What is the posterior of \( [G | \Theta] ? \)

- \( a \in \Theta \)
  - \( \mathbb{P}(\text{VEA} | G(A)) = G(A) \) a.s.
  - \( \mathbb{P}(X \in E) = \mathbb{E}G(A) = G_0(A) \)

- Take a partition \((G(B_1), \ldots, G(B_k)) \sim \text{Dir}(\alpha G_0(B_1), \ldots, \alpha G_0(B_k)) \)

Thus \( [G(B_1), \ldots, G(B_k) | X] \sim \text{Dir}(\alpha G_0(B_1) + \delta_{B_1}, \ldots) \)
\[
\mathbb{P}(x \in A_i \land G(B_i) \leq \gamma_i \land \ldots \land G(B_k) \leq \gamma_k) = \sum_{i=1}^{K} \mathbb{P}(x \in B'_i) \mathbb{P}(G(B_i) \leq \gamma_i \land \ldots \land G(B_k) \leq \gamma_k \mid x \in B'_i).
\]

Note that \(G(B'_i), G(B''_i), \ldots, G(B'_k), G(B''_k) \sim \text{Dir}(\alpha G(B_i), \ldots)\).

\[
\mathbb{P}(G(B'_i) \leq \gamma_i \land \ldots \land G(B_k) \leq \gamma_k \mid x \in A_i) = \sum_{i=1}^{K} G_0(B'_i) \mathbb{P}(G(B'_i) \leq \gamma_i \land \ldots \land G(B_k) \leq \gamma_k \mid x \in B'_i).
\]

If \(A \subset B'_i\) then
\[
\mathbb{P}(G(B'_i) \leq \gamma_i \land \ldots \land G(B_k) \leq \gamma_k \mid x \in A_i) = \mathbb{P}(G(B'_i) \leq \gamma_i \land \ldots \land G(B_k) \leq \gamma_k \mid x \in A_i) = \mathbb{P}(G(B'_i) \leq \gamma_i \land \ldots \land G(B_k) \leq \gamma_k \mid x \in B'_i) \times G_0(B'_i) + \delta_{B'_i} \times \ldots
\]

Due to (A) \(G(B'_i), \ldots, G(B'_k) \sim \text{Dir}(\alpha G(B_i), \ldots)\).
Let \( B_{(\phi_{B})} \) and \( C \) be a partition.

Then

\[
G = \beta \epsilon_{e_{1}} + (1-\beta_{1}) \left[ \beta_{2} \epsilon_{e_{2}} + (1-\beta_{2}) \beta_{3} \epsilon_{e_{3}} + \ldots \right]
\]

\[
= \beta_{1} \epsilon_{e_{1}} + (1-\beta_{1}) \epsilon_{g}'
\]

So

\[
G \overset{d_{s}}{=} \beta_{1} \epsilon_{e_{1}} + (1-\beta_{1}) \epsilon_{g}'.
\]

Can easily check that

\[
\text{Dir}(\alpha \epsilon_{0}(B_{1}), \ldots, \alpha \epsilon_{0}(B_{x})) \overset{d_{s}}{=} \beta \left( \delta_{e_{1}}(B_{1}), \ldots, \delta_{e_{1}}(B_{x}) \right) + (1-\beta_{1}) \left( \theta_{0} \epsilon_{g}(B_{1}), \ldots, \theta_{0} \epsilon_{g}(B_{x}) \right)
\]

if \( \epsilon_{1} \in B_{i} \), then

\[
\text{Dir}(\alpha \epsilon_{0}(B_{1}), \ldots, \alpha \epsilon_{0}(B_{x})) \overset{d_{s}}{=} \text{Dir}(\alpha \epsilon_{0}(B_{1}), \ldots, \alpha \epsilon_{0}(B_{x})) + (1-\beta_{1}) \text{Dir}(\alpha \epsilon_{0}(B_{1}), \ldots, \alpha \epsilon_{0}(B_{x}))
\]

\[
\sum_{j \in \mathcal{B}_{2}} \text{Dir}(\alpha \epsilon_{0}(B_{1}), \ldots, \alpha \epsilon_{0}(B_{x})) = \text{Dir}(\alpha \epsilon_{0}(B_{1}), \ldots, \alpha \epsilon_{0}(B_{x}))
\]
Lemma: \( V, U, W \) RV's, \( W \in [-1,1] \)

- \( U, V \) take values in some vector space
- \( \mathbb{P}(|W| = 1) > 0 \) then \( V \) is unique.

If \( V \neq V' \)

- \( V_{n+1} = U_n + W_n V_n \)
- \( V_{n+1} = U_n + W_n V_n' \)

\[ |V_{n+1} - V_{n+1}'| = |W_n| |V_n - V_n'| \to 0 \text{ wp 1.} \]

This proves automatically that \( \mathbb{P} \) distribution Random measures are discrete wp 1.

---

Alternative proxy of discreteness

\( G \sim \mathbb{D}P(\alpha, G_0) \)

\( X \mid G \sim G \)

Then, \( G(X) \sim \mathbb{D}P(\alpha G_0 + \delta_x) \),

\( G(X, \ldots, X_n) \sim \mathbb{D}P(\alpha G_0 + \sum_{i=1}^{n} \delta_{x_i}) \)

\( \text{DP}\left( \alpha G_0 + \sum_{i=1}^{\infty} \frac{1}{\alpha x_n} \delta_{x_i} \right) \)

As \( n \to \infty \):

- \( \sum \delta_{x_n} \) converges to \( \infty \)
- \( \text{base measure} \quad \frac{W}{\alpha} \to \sum_{i=1}^{\infty} \delta_{x_i} \)