Clustering for Gritches

Shuhei Mano

The Institite of Statistical Mathematics, Japan

Deta characterization meeting

November 25, 2013

Harmonic Analysis

Schuster, A. (1897) On Lunar and Solar Periodicities of Earthquakes. Proc Roy Soc Lond 61: 455-465.

Number of events at periods $z_1, z_2, ..., z_N$, where $N = 2q + 1$.

$$z_t = a_0 + \sum_{i=1}^{q} (a_i \cos 2\pi f_i t + b_i \sin 2\pi f_i t) + \epsilon_t, \qquad f_i = \frac{i}{N}.$$

Least Squares Estimates:

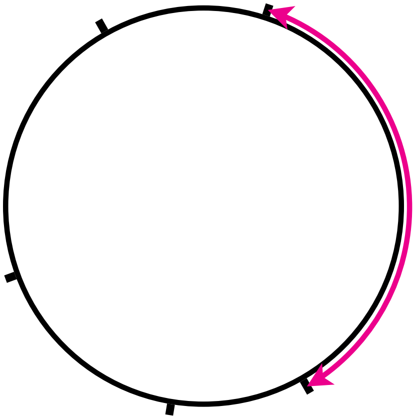
$$\hat{a}_i = \frac{2}{N} \sum_{t=1}^{N} z_t \cos 2\pi f_i t, \qquad \hat{b}_i = \frac{2}{N} \sum_{t=1}^{N} z_t \sin 2\pi f_i t.$$

Periodgram:

$$I_i = \frac{N}{2} (\hat{a}_i{}^2 + \hat{b}_i{}^2), \qquad i = 1, 2, ..., q.$$

Extreme Sizes in Random Partition

Fisher, R.A. (1929) Tests of Significance in Harmonic Analysis. Proc Roy Soc Lond 125: 54-59.



The test based on the distribution of $X_{(1)}$ in $(X_1, X_2, ..., X_n) \sim Dir(1)$.

Dirichlet Process

Definition (Ferguson,1973)

Let μ be a finite measure on $(\mathcal{X}, \mathcal{B})$. A random measure D on $\mathcal{X}$ is called a Dirichlet process if for every finite measureable partition $\{B_1, ..., B_k\}$ of $\mathcal{X}$, $(D(B_1), ..., D(B_k)) \sim Dir(\mu(B_1), ..., \mu(B_k))$.

Theorem (Ferguson, 1973)

1. Y_i, *i.i.d. with a probability measure* $\mu(\cdot)/\theta$ *with* $\theta = \mu(\mathcal{X})$.
2. *Let* ρ_t *be the gamma process with* $\rho_0 = 0$ *whose increment follows* $Gamma(\theta t, 1)$ *and* $(Z_1, Z_2, ...)$ *be the jump sizes up to* $t = 1$.
3. $D(\cdot) = \sum_{i=1}^{\infty} \frac{Z_i}{\rho_1} \delta_{Y_i}(\cdot)$ *is a Dirichlet process with parameter* μ.

Poisson-Dirichlet Distribution

Definition (Kingman,1978; Pitman,1995)

Let

$$P_1 = W_1, \qquad P_i = W_i \prod_{j=1}^{i-1}(1 - W_j),\ i = 2, 3, ...,$$

where $W_i \sim Beta(1 - \alpha, \theta + i\alpha)$, i.i.d. with $0 \leq \alpha < 1$ and $\theta > -\alpha$, or $\alpha < 0$ and $\theta = -\alpha m$, $m \in \mathbb{N}$. The distribution of the ranked sequence of P is called the 2-parameter Poisson-Dirichlet distribution $PD(\alpha, \theta)$.

Remark

- For $\alpha = 0$, $(P_i) \overset{d}{=} ({\rho_1}^{-1} Z_i)$ and $D(\cdot) = \sum_{i=1}^{\infty} P_i \delta_{Y_i}(\cdot)$ is the 2-parameter generalization of the Dirichlet process.
- For $\alpha < 0$, $PD(\alpha, -\alpha m)$ is the m-dimensional symmetric $Dir(-\alpha)$.
- Most general such that P is invariant under size-biased permutation.

Ewens-Pitman Random Partition

A partition of $\boldsymbol{n} \in \mathbb{N}$ by integers is identified by multiplicity $\boldsymbol{c}$ (size index) such that

$$||\boldsymbol{s}|| := \sum_{i=1}^{n} s_i = k_n, \qquad |\boldsymbol{s}| := \sum_{i=1}^{n} i s_i = n.$$

Example

$10 = 5 + 2 + 2 + 1$ gives $k_{10} = 4$, $s_5 = s_1 = 1,\ s_2 = 2$.

Definition (Ewens,1972; Pitman,1992)

An exchangeable random partition

$$P(S = s, K_n = k) = \frac{\left(\frac{\theta}{\alpha}\right)_k}{(\theta)_n}(-1)^{n-k} n! \prod_{j=1}^{n} \left(\begin{array}{c} \alpha \\ j \end{array} \right)^{s_j} \frac{1}{s_j!},$$

where $0 \leq \alpha < 1$ and $\theta > -\alpha$, or $\alpha < 0$ and $\theta = -\alpha m$, $m = 1, 2, \ldots$

Chinese restaurant process

The Ewens-Pitman random partition is the sampling distribution from $PD(\alpha, \theta)$. Deonote the ranked sizes $\{i : C_i > 0\}$ by $L^{(n)}_{(1)}, L^{(n)}_{(2)}, \ldots$

$$n^{-1}(L^{(n)}_{(1)}, L^{(n)}_{(2)}, \ldots) \xrightarrow{d} (P_{(1)}, P_{(2)}, \ldots), \qquad n \to \infty.$$

Suppose n person occupy k tables (cluster) and n_i people sit at table i. The next person

- sits at an empty table with probability $\dfrac{\theta + k\alpha}{\theta + n}$,
- sits at the table i with probability $\dfrac{n_i - \alpha}{\theta + n}$.

The random partition is numbers of people at each table.

Dirichlet Process Mixture Distribution

Assume σ^2 and σ_0 are known for simplicity. A gaussian mixture distribution is

$$f(x|P,\mu) = \sum_{i=1}^{m} P_i \phi(x|\mu_i, \sigma^2)$$

with $\mu_i \sim N(0, \sigma_0^2)$, $P_i \sim Dir(-\alpha)$.

How to choose the number of clusters?

- Try each and choose optimal one by some criteria (information criterion, cross validation,...)
- Dirichlet process with $\alpha \geq 0$: $D \sim DP(\alpha, \theta; N(0, \sigma_0^2))$.

Assignment

By Bayes' rule,

$$
\begin{aligned}
P(C_i|C_{-i}, X) &\propto P(X_i|C, X_{-i})P(C_i|C_{-i}, X_{-i}) \\
&\hat{=} \phi(X_i|\hat{\mu}_i(C_{-i}, X_{-i}), \sigma^2)P(C_i|C_{-i})
\end{aligned}
$$

The second factor is given by CRP. Sampling from $P(C|X)$ is possible by using the Gibbs sampler.
For a gritch clustering, X is not a position but a wave form. Constructing the likelihood is the key issue.