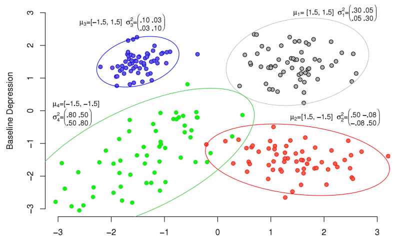
Nicholas， 2018

Dirichlet process mixture



Dirichlet distribution

$$f(x_1,\ldots,x_K;\alpha_1,\ldots,\alpha_K)=\frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{i=1}^{K}x_i^{\alpha_i-1}$$

- Normalizing constant

$$\mathrm{B}(\boldsymbol{\alpha})=\frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)},\qquad \boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_K).$$

Dirichlet process: distribution over distributions

$$G\sim DP(\alpha,G_0),X_n|G\sim G$$

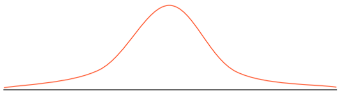
- Note

  - G_0 is a base distribution
  - α is a positive scaling parameter

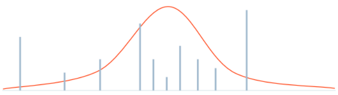
  G is a random probability measure that has the same support as G_0

Example

Consider Gaussian G_0



G ~ DP(α, G_0)



- Comparison

  G_0 is continuous, so the probability that any two samples are equal is precisely zero.

  However, G is a discrete distribution, made up of a countably infinite number of point masses [Blackwell]

  - Therefore, there is always a non-zero probability of two samples colliding

Sampling from DP

$$X_n|X_1,\ldots,X_{n-1}=\begin{cases}X_i & \text{with probability } \frac{1}{n-1+\alpha}\\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha}\end{cases}$$

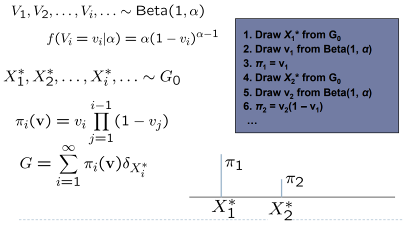
$$P(X_1,\ldots,X_N)=P(X_1)P(X_2|X_1)\ldots P(X_N|X_1,\ldots,X_{N-1})$$

Chain rule

$$=\frac{\alpha^K\prod_{k=1}^{K}(\text{num}(X_k^*)-1)!}{\alpha(1+\alpha)\ldots(N-1+\alpha)}\prod_{k=1}^{K}G_0(X_k^*)$$

P(partition) P(draws)

- Stick breaking viewpoint



$$V_1, V_2, \ldots, V_i, \ldots \sim \text{Beta}(1, \alpha)$$

$$f(V_i = v_i|\alpha) = \alpha(1 - v_i)^{\alpha-1}$$

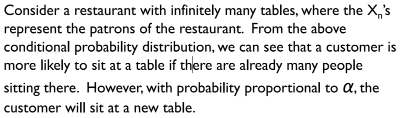
$$X_1^*, X_2^*, \ldots, X_i^*, \ldots \sim G_0$$

$$\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1} (1 - v_j)$$

$$G = \sum_{i=1}^{\infty} \pi_i(\mathbf{v})\delta_{X_i^*}$$

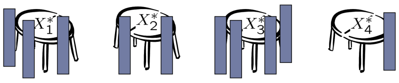
1. Draw X_1* from G_0
2. Draw v_1 from Beta(1, α)
3. $\pi_1 = v_1$
4. Draw X_2* from G_0
5. Draw v_2 from Beta(1, α)
6. $\pi_2 = v_2(1 - v_1)$
 …

- Chinese restaurant process: posterior of DP



Consider a restaurant with infinitely many tables, where the X_n's represent the patrons of the restaurant. From the above conditional probability distribution, we can see that a customer is more likely to sit at a table if there are already many people sitting there. However, with probability proportional to α, the customer will sit at a new table.

  - Clustering effect



- Dirichlet process mixture model
  - Background
    Given a data set, and are told that it was
    generated from a mixture of Gaussian distributions. But no one has any idea how many Gaussians produced the data.
  - Finite mixture models

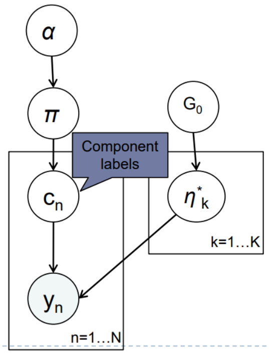
π ~ Dirichlet(α/K,…, α/K)

c_n ~ Multinomial(π)

η_k ~ G_0

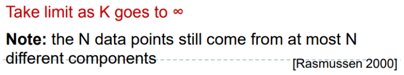
y_n | c_n, η_1,… η_K ~ F(· | η_{c_n})

    - Assumption
      A finite mixture model assumes that the data come from a mixture of a finite number of distributions
    - Illustration



Component labels

      - Infinite mixture models

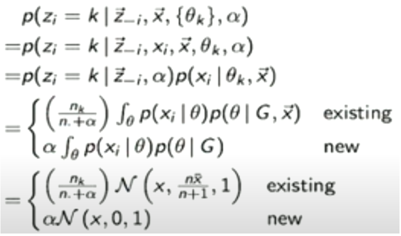


Take limit as K goes to ∞

Note: the N data points still come from at most N different components

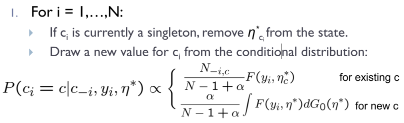
[Rasmussen 2000]

- Inference for DPMM
  - Procedure (infer)
    - Take a random guess
    - Calculate mean for each cluster
      - Let #clusters grow or shrink
    - Compute cluster assignment for each sample



$$
\begin{aligned}
&p(z_i = k \mid \vec{z}_{-i}, \vec{x}, \{\theta_k\}, \alpha)\\
=&p(z_i = k \mid \vec{z}_{-i}, x_i, \vec{x}, \theta_k, \alpha)\\
=&p(z_i = k \mid \vec{z}_{-i}, \alpha) p(x_i \mid \theta_k, \vec{x})\\
=&\begin{cases} \left(\frac{n_k}{n.+\alpha}\right) \int_\theta p(x_i \mid \theta) p(\theta \mid G, \vec{x}) & \text{existing} \\ \alpha \int_\theta p(x_i \mid \theta) p(\theta \mid G) & \text{new} \end{cases}\\
=&\begin{cases} \left(\frac{n_k}{n.+\alpha}\right) \mathcal{N}\left(x, \frac{n\bar{x}}{n+1}, 1\right) & \text{existing} \\ \alpha \mathcal{N}(x, 0, 1) & \text{new} \end{cases}
\end{aligned}
$$

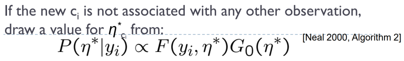
      - Update cluster by sampling from the distribution
  - Procedure (generate)
    - Generate a new label assignment



1. For i = 1,…,N:
   - If c_i is currently a singleton, remove $\eta^*_{c_i}$ from the state.
   - Draw a new value for c_i from the conditional distribution:

$$P(c_i = c | c_{-i}, y_i, \eta^*) \propto \begin{cases} \frac{N_{-i,c}}{N-1+\alpha} F(y_i, \eta^*_c) & \text{for existing c} \\ \frac{\alpha}{N-1+\alpha} \int F(y_i, \eta^*) dG_0(\eta^*) & \text{for new c} \end{cases}$$

      - For unique assignment

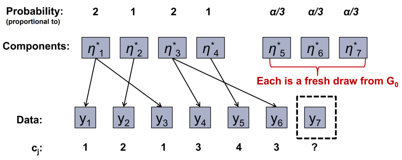


If the new c_i is not associated with any other observation, draw a value for $\eta^*_{c_i}$ from:

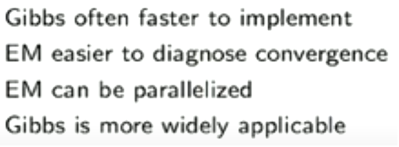
$$P(\eta^* | y_i) \propto F(y_i, \eta^*) G_0(\eta^*)$$

[Neal 2000, Algorithm 2]

    - Illustration



  - Comparison between EM and Gibbs



Gibbs often faster to implement
EM easier to diagnose convergence
EM can be parallelized
Gibbs is more widely applicable

Main model

- Model components
  - DPMM

$$
\begin{aligned}
&M \sim \text{Gamma}(\Psi_1, \Psi_2)\\
&G|M \sim \text{DPMM}(M, G_0), G_0 = N(0, \sigma_\tau^2)\\
&\tau_i|G \sim G
\end{aligned}
$$

- Prior specification

$$\Psi_1 = 2, \Psi_2 = 0.1$$

- BART

$$m \sim \mathrm{BART}(\alpha, \beta, k, J)$$

- Prior specification

$$\alpha = 0.95, \beta = 2, k = 2, J = 200$$

- Assumptions

$$y = \hat{\mu}_{AFT} + \epsilon, \epsilon \sim N(0, \hat{\sigma}^2_{AFT})$$

- Centered transformation

$$y_i^{tr} = y_i \exp\{-\hat{\mu}_{AFT}\}$$

- Prior of terminal nodes

$$\mu_{j,l} \sim N(0, \frac{4\hat{\sigma}^2_{AFT}}{Jk^2})$$

- Prior of BART

$$m(A, x) \sim N(0, \frac{4\hat{\sigma}^2_{AFT}}{k^2})$$

- Nonparametric AFT model

$$\log T_i = m(A_i, x_i) + W_i$$
$$\sigma^2 \sim \kappa\nu/\chi^2_\nu, \kappa = \sigma^2_\tau$$
$$W_i|\tau_i, \sigma^2 \sim N(\tau_i, \sigma^2)$$

- Prior specification

$$P(Var(W|G, \sigma) \leq \hat{\sigma}^2_W) = 0.5 \rightarrow \kappa = \hat{\sigma}^2_W/4, \sigma^2_\tau = \kappa$$

- Posterior inference

- Blocked Gibbs sampling

AFT model

- Accelerated failure time (AFT) model

$$\log T_i = x_i^T\beta + \sigma\epsilon_i, \epsilon_i \sim f \text{ (completely unspecified)}$$

- Expression

        - Hazard function

          $$\lambda(t/x) = \exp(\beta' x)\lambda_0(\exp(\beta' x)t)$$

        - Survival function

          $$S_i(t) = S_0(e^{x_i^T \beta} t)$$

    - Common distributions of AFT models
        - Weibull distribution and exponential AFT model

          $$S(t) = \exp(-\lambda t^{\alpha})$$

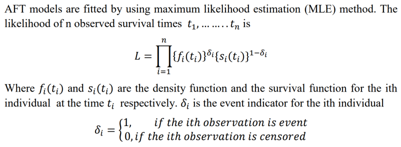
            - Weibull distribution

              $$f(t; \lambda, \alpha) = \alpha \lambda t^{\alpha-1} \exp(-\lambda t^{\alpha})$$

            - Hazard function

              $$h(t) = \lambda \alpha t^{\alpha-1}$$

                - Cumulated hazard function
            - 自由主题
        - Log-normal AFT model
        - Log-logistic distribution
    - Estimation of AFT models



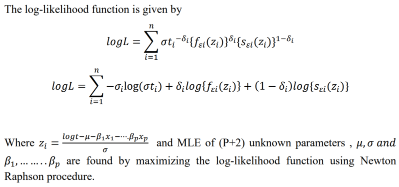
AFT models are fitted by using maximum likelihood estimation (MLE) method. The likelihood of n observed survival times $t_1, \dots \dots . . t_n$ is

$$L = \prod_{i=1}^{n} \{f_i(t_i)\}^{\delta_i} \{s_i(t_i)\}^{1-\delta_i}$$

Where $f_i(t_i)$ and $s_i(t_i)$ are the density function and the survival function for the ith individual at the time t_i respectively. δ_i is the event indicator for the ith individual

$$\delta_i = \begin{cases} 1, & if\ the\ ith\ observation\ is\ event \\ 0, if\ the\ ith\ observation\ is\ censored \end{cases}$$

        - MLE estimation



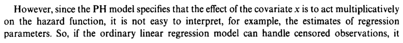
The log-likelihood function is given by

$$logL = \sum_{i=1}^{n} \sigma t_i{}^{-\delta_i} \{f_{\varepsilon i}(z_i)\}^{\delta_i} \{s_{\varepsilon i}(z_i)\}^{1-\delta_i}$$

$$logL = \sum_{i=1}^{n} -\sigma_i \log(\sigma t_i) + \delta_i log\{f_{\varepsilon i}(z_i)\} + (1-\delta_i) log\{s_{\varepsilon i}(z_i)\}$$

Where $z_i = \frac{logt - \mu - \beta_1 x_1 - \cdots . \beta_p x_p}{\sigma}$ and MLE of (P+2) unknown parameters , $\mu, \sigma\ and$ $\beta_1, \dots \dots . . \beta_p$ are found by maximizing the log-likelihood function using Newton Raphson procedure.

    - Discussion
        - Advantage
            -

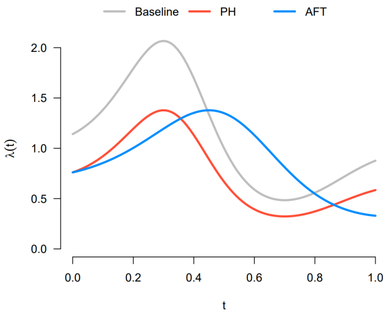


However, since the PH model specifies that the effect of the covariate *x* is to act multiplicatively on the hazard function, it is not easy to interpret, for example, the estimates of regression parameters. So, if the ordinary linear regression model can handle censored observations, it

        - Statistical issues

- Illustration

$$\text{PH}: \lambda_i(t) = \lambda_0(t)e^{\eta_i}$$
$$\text{AFT}: \lambda_i(t) = \lambda_0(e^{-\eta_i}t)e^{-\eta_i}$$



- Nonparametric AFT model

$$\log(T) = m(A, x) + W, A \in \{0, 1\}$$

Individual treatment effect

- Ratio in expected survival time

$$\xi(x) = \frac{E(T|A=1, x, m))}{E(T|A=0, x, m)} = \exp(\theta(x))$$

- Heterogeneity of treatment effect

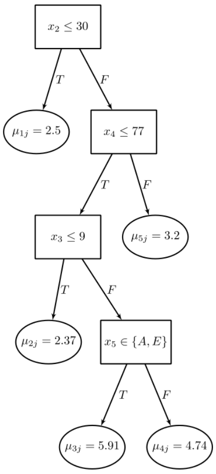
$$D_i = P(\theta(x_i) \geq \bar{\theta}|y, \delta), \bar{\theta} = \sum \theta(x_i)/n$$
$$D_i^* = \max\{1 - 2D_i, 2D_i - 1\}$$

Bayesian additive regression tree (BART)

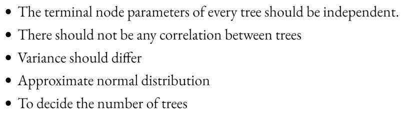
Building a sum of trees

$$y = \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon, \epsilon \sim N(0, \sigma^2)$$

- Binary regression tree



- regularization of prior

  - The terminal node parameters of every tree should be independent.
  - There should not be any correlation between trees
  - Variance should differ
  - Approximate normal distribution
  - To decide the number of trees

  - Prior of tree
    - Prior of tree at depth d nonterminal

      $$d_j = \frac{\alpha}{(1+d)^{\beta}}, \alpha \in (0,1), \beta \geq 0$$

    - The distribution on the splitting variable assignments at each interior node is the uniform prior on available variables
    - The distribution on the splitting rule assignment in each interior node, conditional on the splitting variable is the uniform prior on the discrete set of available splitting values
  - Prior of mean

    $$y \rightarrow [y_{\min} = -0.5, y_{\max} = 0.5] \Rightarrow \mu_{ij} \sim N(0, \sigma_{\mu}^2), \sigma_{\mu} = \frac{1}{2k\sqrt{m}}$$

  - Prior of σ

    $$\sigma^2 \sim \frac{\nu\lambda}{\chi_{\nu}^2}$$

    - Chipman (2010)

      $$(\nu, q) = (3, 0.9), P(\sigma < \hat{\sigma}) = q$$

  - Number of trees

    $$m = 200$$

- Posterior sampling
  - backfitting MCMC algorithm

    $$p((T_1, M_1), \ldots, (T_m, M_m), \sigma | y)$$

    - Partial residual

      $$R_j = y - \sum_{k \neq j} g(x; T_k, M_k)$$

    - Draw σ from full conditional

      $$\sigma | T_1, \ldots, T_m, M_1, \ldots, M_m, y$$

    - Gibbs sampling

      $$T_j | R_j, \sigma$$
      $$M_j | T_j, R_j, \sigma$$

- Particle Gibbs algorithms