2. Properties of the finite-dimensional Dirichlet distribution.

The Dirichlet distribution is one of the possible multivariate extension of the Beta.
We start considering:

$$U_1, \dots, U_k \quad \text{s.t.} \quad U_j \sim \text{Gamma}(\alpha_j, \beta) \qquad j = 1, \dots, k$$

We define: $X_j = \dfrac{U_j}{U_1 + \dots + U_k}$

We obtain: $(X_1, \dots, X_k) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)$.

We impose $\sum_{j=1}^k U_j = 1$ and so $(X_1, \dots, X_k) \in \Delta_{k-1} = \left\{ \underline{x} \in \mathbb{R}^k : \sum_{j=1}^k x_j = 1, \; x_j \in [0,1] \; \forall j \right\}$

The p.d.f. of the Dirichlet distribution is:

$$f(x_1, \dots, x_k; \alpha_1, \dots, \alpha_k) = \frac{1}{B(\underline{\alpha})} \prod_{j=1}^k x_j^{\alpha_j - 1} \; \mathbb{1}_{\Delta_{k-1}}(\underline{x})$$

where $B(\underline{\alpha}) = \dfrac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)}{\Gamma(\alpha_1 + \dots + \alpha_k)}$.

Notice that, because of Δ_{k-1}: $f(\underline{x}; \underline{\alpha}) = \dfrac{1}{B(\underline{\alpha})} \left[\prod_{j=1}^{k-1} x_j^{\alpha_j - 1} \right] \left(1 - \sum_{j=1}^{k-1} x_j \right)^{\alpha_k - 1} \mathbb{1}_{\Delta_{k-1}}(\underline{x})$

For simplicity: $\alpha_0 := \sum_{j=1}^k \alpha_j$.

Properties:

1. $\mathbb{E}[X_j] = \dfrac{\alpha_j}{\alpha_0} \qquad j = 1, \dots, k$
2. $\text{Var}(X_j) = \dfrac{\alpha_j (\alpha_0 - \alpha_j)}{\alpha_0^2 (\alpha_0 + 1)} \qquad j = 1, \dots, k$
3. $\text{Cov}(X_i, X_j) = -\dfrac{\alpha_j \alpha_i}{\alpha_0^2 (\alpha_0 + 1)} \qquad i \neq j$
4. Marginal distribution: $X_j \sim \text{Beta}(\alpha_j, \alpha_0 - \alpha_j) \qquad j = 1, \dots, k$
5. Aggregation property:

   $(X_1, \dots, X_K) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)$

   $(Z_1, \dots, Z_m) := \left(\sum_{j \in C_1} X_j, \dots, \sum_{j \in C_m} X_j \right)$ s.t. $\begin{cases} C_j \subseteq \{1, \dots, n\} & \forall j \\ C_i \cap C_j = \emptyset & \forall i \neq j \\ \bigcup_{j=1}^m C_j = \{1, \dots, n\} \end{cases}$

   $\longrightarrow (Z_1, \dots, Z_m) \sim \text{Dirichlet}(\alpha_1^*, \dots, \alpha_m^*)$

   $\alpha_j^* = \sum_{j \in C_j} \alpha_j$

6. If $(X_1, \dots, X_k) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)$ then:

$$\begin{cases} X_1 \sim \text{Beta}(\alpha_1, \alpha_0 - \alpha_1) \\ X_1 \perp\!\!\!\perp \left(\dfrac{X_2}{X_1}, \dots, \dfrac{X_k}{X_1} \right) \\ \left(\dfrac{X_2}{X_1}, \dots, \dfrac{X_k}{X_1} \right) \sim \text{Dirichlet}(\alpha_2, \dots, \alpha_k) \end{cases} \leftarrow \text{we can take out one dimension}$$

3. Bayes' Theorem for dominated models

Consider a random sample $\underline{X}_n|\theta = (X_1,\dots,X_n)|\theta \sim P_\theta$, $\theta \in \Theta \subset \mathbb{R}^k$. Suppose that the model is dominated ($\exists$ a density (the R-N density) : $f(\underline{x}|\theta)$ = density of the joint distribution of $\underline{X}_n$ w.r.t. some σ-finite measure on $\mathbb{R}^n$ ($:= \lambda^{(n)}$)).
Suppose $\theta \sim \pi$ (prior for θ).

$\Longrightarrow$ The posterior distribution for θ given $\underline{X}_n = \underline{x}$ can be expressed as:

$$\mathbb{P}(\theta \in B | \underline{X}_n = \underline{x}) = \frac{\int_B f(\underline{x}|\theta)\,\pi(d\theta)}{\int_\Theta f(\underline{x}|\theta)\,\pi(d\theta)} \qquad \forall B \in \mathcal{B}(\Theta)$$

proof.

1. We define $\gamma :=$ joint distribution of $\underline{X}$ and θ :

$$\gamma(A\times B) = \int_B P_\theta(A)\,\pi(d\theta) \overset{*}{=} \int_B \int_A f(\underline{x}|\theta)\,\lambda^{(n)}(d\underline{x})\,\pi(d\theta) \qquad A \in \mathcal{B}(\mathbb{R}^n),\ B \in \mathcal{B}(\Theta)$$

FT

$$= \int_A \left(\int_B f(\underline{x}|\theta)\,\pi(d\theta)\right)\lambda^{(n)}(d\underline{x})$$

* is valid because P_θ is dominated $\Rightarrow$ we can express it as integral of the density.

2. We want now to define the marginal :

$$\mu_n(A) = \gamma(A\times\Theta) = \int_A \int_\Theta f(\underline{x}|\theta)\,\pi(d\theta)\,\lambda^{(n)}(d\underline{x}) := \int_A m_n(\underline{x})\,\lambda^{(n)}(d\underline{x}).$$

We needed this marginal because the theorem is obtained from:

$$\chi(\underline{x},\theta) = \chi(\underline{x}|\theta)\,\chi(\theta) = \chi(\theta|\underline{x})\,\chi(\underline{x}) \longrightarrow \text{we needed } \chi(\underline{x}).$$

3. We now fix $B \in \mathcal{B}(\Theta)$ and we define $\gamma(A\times B)$ varying A.
$\gamma(\cdot \times B)$ is a measure on $\mathcal{B}(\mathbb{R}^n)$, moreover: $\gamma(\cdot \times B) \ll \mu_n$
(i.e. $\gamma(\cdot \times B)$ is absolutely continuous w.r.t. μ_n :
consider A : $\mu_n(A) = 0 \Longrightarrow \gamma(A\times B) \le \gamma(A\times\Theta) \overset{\text{def.}}{=} \mu_n(A) \overset{\text{Hp.}}{=} 0$).
We can apply the R-N theorem:

$\forall B$ fixed $\exists$ function $\pi(\underline{x},B)$: $\gamma(A\times B) = \int_A \pi(\underline{x},B)\,\mu_n(d\underline{x})$

$$\longrightarrow \gamma(A\times B) = \int_A \pi(\underline{x},B)\,m_n(\underline{x})\,\lambda^{(n)}(d\underline{x}) = \int_A \pi(\underline{x},B)\int_\Theta f(\underline{x}|\theta)\,\pi(d\theta)\,\lambda^{(n)}(d\underline{x})$$

4. Considering that : $\gamma(A\times B) = \mathbb{P}(\underline{X}\in A, \theta \in B) = \int_A \mathbb{P}(\theta \in B|\underline{X}=\underline{x})\,\mu_n(d\underline{x})$
we have that : $\pi(\underline{x},B) = \mathbb{P}(\theta \in B|\underline{X}=\underline{x})$:

$$\gamma(A\times B) = \int_A \mathbb{P}(\theta\in B|\underline{X}=\underline{x}) \int_\Theta f(\underline{x}|\theta)\pi(d\theta)\,\lambda^{(n)}(d\underline{x})$$

5. comparing :

$$\begin{cases} \gamma(A\times B) = \int_A \int_B f(\underline{x}|\theta)\,\pi(d\theta)\,\lambda^{(n)}(d\underline{x}) \\ \gamma(A\times B) = \int_A \mathbb{P}(\theta\in B|\underline{X}=\underline{x}) \int_\Theta f(\underline{x}|\theta)\pi(d\theta)\,\lambda^{(n)}(d\underline{x}) \end{cases}$$

$$\Longrightarrow \int_B f(\underline{x}|\theta)\pi(d\theta) = \mathbb{P}(\theta\in B|\underline{X}=\underline{x}) \int_\Theta f(\underline{x}|\theta)\,\pi(d\theta) \qquad \text{a.s. w.r.t. } \mu_n$$

$$\Longrightarrow \mathbb{P}(\theta\in B|\underline{X}=\underline{x}) = \frac{\int_B f(\underline{x}|\theta)\,\pi(d\theta)}{\int_\Theta f(\underline{x}|\theta)\pi(d\theta)} \qquad \text{a.s. w.r.t. } \mu_n$$

Moreover : $\mu_n\left(\{\underline{x} : \int_\Theta f(\underline{x}|\theta)\,\pi(d\theta) = m_n(\underline{x}) = 0\}\right) = 0$ ■

If $\theta \sim \pi(\theta)$:

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\,\pi(\theta)}{\int_\Theta f(\underline{x}|\theta)\,\pi(\theta)\,d\theta} \Longrightarrow$$

The posterior distribution is proportional to the likelihood times the prior

1. Joint of $\underline{X},\theta$ (I)
2. Marginal of $\underline{X}$
3. Radon-Nikodym : $\gamma(\cdot\times B)$, $\mu_n(\cdot)$ (B fixed)
4. Joint of $\underline{X},\theta$ (II)
5. (I) = (II)

4. Posterior mean as the functional minimizing the posterior quadratic loss function (multivariate case), posterior median as the functional minimizing the posterior absolute value loss function.

We suppose: $X_1, \dots, X_n | \theta \sim f(x, \theta)$, $\theta \sim \pi(\theta)$ s.t. $\underline{\theta} \in \Theta \subseteq \mathbb{R}^k$.
We have a set of possible choices (actions) $\mathcal{A}$: $a \in \mathcal{A}$ is a single action (in terms of point estimation). We define a loss function: $l : (\theta, a) \mapsto l(\theta, a)$. The loss function tells us how much we lose choosing the action a when the (true) state is θ.

We have: prior expectation loss : $\mathbb{E}_\pi[l(\theta, a)]$
posterior expectation loss : $\mathbb{E}_\pi[l(\theta, a) | \underline{X}]$

The optimal choices (by Bayes) are all the a^* s.t.:

$$a^* = \arg\min_{a \in \mathcal{A}} \mathbb{E}_\pi[l(\theta, a) | \underline{X}]$$

1. Quadratic loss function (multivariate): ($\Theta \subseteq \mathbb{R}^p$)

$\mathcal{A} \equiv \Theta$, $l(\underline{\theta}, \underline{a}) = (\underline{\theta} - \underline{a})^T Q (\underline{\theta} - \underline{a})$ Q simmetric, positive semidef. matrix.

$$\Rightarrow \tilde{\underline{\theta}} = \mathbb{E}_\pi[\underline{\theta} | \underline{X}] = \arg\min_{\underline{a} \in \mathcal{A}} \mathbb{E}_\pi[l(\underline{\theta}, \underline{a}) | \underline{X}]$$

proof.

$$\frac{d}{d\underline{a}} \mathbb{E}_\pi[l(\underline{\theta}, \underline{a}) | \underline{X}] = \frac{d}{d\underline{a}} \int_\Theta (\underline{\theta} - \underline{a})^T Q (\underline{\theta} - \underline{a}) \pi(\underline{\theta} | \underline{X}) d\underline{\theta} := \frac{d}{d\underline{a}} \int_\Theta F(\underline{a}) \pi(\underline{\theta} | \underline{X}) d\underline{\theta}$$

where $F(\underline{a}) = \sum_{i,j=1}^p q_{ij} (\theta_i - a_i)(\theta_j - a_j)$, $q_{ij} = [Q]_{ij}$

$$\Rightarrow \frac{d}{da_\ell} F(\underline{a}) = -\sum_{j=1}^p q_{\ell j} (\theta_j - a_j) - \sum_{i=1}^p q_{i\ell} (\theta_i - a_i) = -2 \sum_{j=1}^p q_{\ell j} (\theta_j - a_j) = [-2Q(\underline{\theta} - \underline{a})]_\ell$$

$$\Rightarrow \frac{d}{d\underline{a}} \mathbb{E}_\pi[l(\underline{\theta}, \underline{a}) | \underline{X}] = \int_\Theta -2Q(\underline{\theta} - \underline{a}) \pi(\underline{\theta} | \underline{X}) d\underline{\theta} = 0$$

$$\Rightarrow 2Q \int_\Theta \underline{\theta} \pi(\underline{\theta} | \underline{X}) d\underline{\theta} = 2Q \int_\Theta \underline{a} \pi(\underline{\theta} | \underline{X}) d\underline{\theta} \Rightarrow \underline{a}$$

$$\Rightarrow \underline{a} = \tilde{\underline{\theta}} = \int_\Theta \underline{\theta} \pi(\underline{\theta} | \underline{X}) d\underline{\theta} = \mathbb{E}_\pi[\underline{\theta} | \underline{X}] \quad \blacksquare$$

2. Absolute value loss function:

$$l(\theta, a) = |\theta - a| = \begin{cases} \theta - a & \theta > a \\ a - \theta & \theta \le a \end{cases}$$

$\Rightarrow \tilde{\theta}$ = quantile of $\pi(\theta | \underline{X})$ of order $\frac{1}{2}$ = median of $\pi(\theta | \underline{X})$

proof.

$$\begin{aligned} \mathbb{E}_\pi[l(\theta, a) | \underline{X}] &= \int_a^\infty (\theta - a) \pi(\theta | \underline{X}) d\theta + \int_{-\infty}^a (a - \theta) \pi(\theta | \underline{X}) d\theta \\ &= \int_a^\infty \int_a^\theta dy \, \pi(\theta | \underline{X}) d\theta + \int_{-\infty}^a \int_\theta^a dy \, \pi(\theta | \underline{X}) d\theta \\ &= \int_a^\infty \int_y^\infty \pi(\theta | \underline{X}) d\theta \, dy + \int_{-\infty}^a \int_{-\infty}^y \pi(\theta | \underline{X}) d\theta \, dy \\ &= \int_a^\infty P(\theta > y | \underline{X}) dy + \int_{-\infty}^a P(\theta < y | \underline{X}) dy \end{aligned}$$

$$\begin{aligned} \frac{d}{da} \mathbb{E}_\pi[l(\theta, a) | \underline{X}] &= -P(\theta > a | \underline{X}) + P(\theta < a | \underline{X}) \\ &= -(1 - P(\theta \le a | \underline{X})) + P(\theta < a | \underline{X}) \\ &= 0 \end{aligned}$$

$$\Rightarrow P(\theta \le \tilde{\theta} | \underline{X}) = \frac{1}{2} \Rightarrow \tilde{\theta} = \text{median of } \pi(\theta | \underline{X}) \quad \blacksquare$$

5. Use the MC std error to bound (in probability) the Monte Carlo error.

All we want to know about a distribution can be achieved by simulating many draws from it. Suppose $\vartheta \sim \pi$ distribution (which we consider as posterior of the Bayesian model): $\vartheta \in \Theta \subseteq \mathbb{R}^p$. Let's consider a measurable function $h: \Theta \to \mathbb{R}$ s.t. $\mathbb{E}_\pi[|h(\theta)|] < \infty$. We define $\bar{h} := \int_\Theta h(\theta)\,\pi(d\theta)$. We want to evaluate $\bar{h}$.

Strong law of large Numbers (SLLN):

If $\vartheta^{(1)}, \vartheta^{(2)}, \vartheta^{(3)}, \ldots \overset{iid}{\sim} \pi$ and $\bar{h}^{(T)} := \frac{1}{T}\sum_{j=1}^{T} h(\vartheta^{(j)})$ $\quad \left(h: \Theta \to \mathbb{R} : \mathbb{E}_\pi[|h(\theta)|] < \infty\right)$

$$\Rightarrow \bar{h}^{(T)} \xrightarrow{T\to\infty} \bar{h} \quad a.s.$$

$$\left(\frac{1}{T}\sum_{j=1}^{T} h(\vartheta^{(j)}) \xrightarrow{T} \mathbb{E}_\pi[h(\theta)] = \int_\Theta h(\theta)\,\pi(d\theta)\right)$$

Therefore: $err^{(T)} := \bar{h}^{(T)} - \bar{h} \xrightarrow{T\to\infty} 0 \quad a.s.$

The point is: how fast the error goes to zero?

Central Limit Theorem (CLT):

Let $\theta^{(1)}, \theta^{(2)}, \ldots$ be an infinite sequence iid from π.
Consider $h: \Theta \to \mathbb{R}$, $\bar{h} := \mathbb{E}_\pi[h(\theta)]$ s.t. $0 < Var(h(\theta)) < \infty$
Let's define $\sigma^2 := Var(h(\theta))$ and $\sigma^{2(T)} := \frac{1}{T}\sum_{j=1}^{T}\left(h(\vartheta^{(j)}) - \bar{h}\right)^2$

$$\Rightarrow \begin{cases} \sqrt{T}\left(\bar{h}^{(T)} - \bar{h}\right) \xrightarrow{d} N(0, \sigma^2) \\ \sigma^{2(T)} \xrightarrow{T\to\infty} \sigma^2 \quad a.s. \end{cases}$$

The first conclusion is saying: $\bar{h}^{(T)} - \bar{h}$ goes to zero at $(\sqrt{T})^{-1}$

$$\Rightarrow err^{(T)} = \bar{h}^{(T)} - \bar{h} \approx N\left(0, \frac{\sigma^2}{T}\right) \simeq N\left(0, \frac{\sigma^{2(T)}}{T}\right)$$

And so we know how to bound the error:

$$\mathbb{P}(|err^{(T)}| > c) = \mathbb{P}\left(\frac{|\bar{h}^{(T)} - \bar{h}|}{\sqrt{\frac{\sigma^{2(T)}}{T}}} > \frac{c}{\sqrt{\frac{\sigma^{2(T)}}{T}}}\right) \overset{CLT}{\approx} 2\left(1 - \phi\left(\frac{c}{\sqrt{\frac{\sigma^{2(T)}}{T}}}\right)\right)$$

$\Rightarrow$ The Monte Carlo standard error $\sqrt{\frac{\sigma^{2(T)}}{T}}$ controls in probability how large the Monte Carlo error is: the smaller the std error, the smaller the error.

6. Definition of the method of composition and its use in Bayesian Statistic.

Suppose that we want to sample from $\mathcal{L}(X_1)$ but we're not able to.
Suppose that we can sample from $\mathcal{L}(X_1|X_2)$ and $\mathcal{L}(X_2)$ (where X_2 is just an auxiliary random variable).
Then we recall: $\mathcal{L}(X_1) = \int \mathcal{L}(X_1, dX_2) = \int \mathcal{L}(X_1|X_2)\,\mathcal{L}(dX_2)$.
And so, we can now sample from $\mathcal{L}(X_1)$:

$\rightarrow$ it's enough to sample from $\mathcal{L}(X_1, X_2)$: $(X_1^{(j)}, X_2^{(j)})$ MC sample from $\mathcal{L}(X_1, X_2)$.
How can we sample from $\mathcal{L}(X_1, X_2)$?

1. We sample $X_2^{(j)} \sim \mathcal{L}(X_2)$
2. We sample $X_1^{(j)} \sim \mathcal{L}(X_1 | X_2 = X_2^{(j)})$

$\rightarrow$ We obtain samples of the joint distribution $\mathcal{L}(X_1, X_2)$.
How can we obtain samples of $\mathcal{L}(X_1)$?
We consider only the first component of each couple.

In Bayesian Statistic we use this when we need to sample from the posterior distribution.

7. Generalized inverse distribution function sampling method:

(i) sampling from a truncated distribution function on a real interval $(X|_{(a,b)})$

(ii) sampling from a Gamma(m,b) m integer.

Inverse transform method:

We want to sample $X \sim F_X(\cdot)$ with $F_X(\cdot)$ c.d.f.: $F_X: \mathbb{R} \to [0,1]$ monotonically increasing, right continuous, $F(-\infty)=0$, $F(+\infty)=1$.

Prop. If $U \sim \mathcal{U}([0,1]) \Rightarrow X = F^{-1}(U) \sim F_X$ $\quad (F^{-1}(u) = \inf\{x : F(x) \geq u\}$ = generalized inverse of $F)$

proof.

$x_0 \in \mathbb{R}$: $A_1 = \{u \in [0,1] : F^{-1}(u) \leq x_0\}$

$A_2 = \{u \in [0,1] : F(x_0) \geq u\}$

Consider $u \in A_1$: $F^{-1}(u) \leq x_0$. For right continuity: $u \leq F(x_0) \Rightarrow u \in A_2$.

Consider $u \in A_2$: $u \leq F(x_0)$. By the def. of $F(\cdot)$: $F^{-1}(u) \leq x_0 \Rightarrow u \in A_1$.

$\Rightarrow A_1 = A_2$

$\Rightarrow \mathbb{P}(X \leq x_0) = \mathbb{P}(F^{-1}(U) \leq x_0) = \mathbb{P}(U \leq F(x_0)) = F_X(x_0)$

(construction; $A_1 = A_2$; $U \sim \mathcal{U}([0,1])$) ■

$\Rightarrow$ **Algorithm:** (to sample $X \sim F(\cdot)$) for $i = 1, \dots, n_{rep}$:

1. Sample $U_i \sim \mathcal{U}([0,1])$
2. Set $X_i = F^{-1}(U_i)$

(i) Truncated distribution

Let $X \sim F_X(\cdot)$ absolutely continuous. Let $A := (a,b) \subseteq \mathbb{R}$. Let $Y = X|_{X \in A}$.

We want to sample from Y.

$$F_Y(y) = F_{X|X\in A}(y) = \mathbb{P}(X \leq y \mid X \in A) = \frac{\mathbb{P}(X \leq y, X \in A)}{\mathbb{P}(X \in A)} = \begin{cases} 0 & y \leq a \\ \frac{F_X(y) - F_X(a)}{F_X(b) - F_X(a)} & a < y \leq b \\ 1 & y > b \end{cases}$$

$$f_Y(y) = \frac{dF_Y}{dy} = \frac{1}{\mathbb{P}(y \in A)} f_X(y) \mathbb{1}_A(y)$$

$\Rightarrow U = F_Y(y) = \frac{F_X(y) - F_X(a)}{F_X(b) - F_X(a)} \Rightarrow F_X(y) = [F_X(b) - F_X(a)]U + F_X(a) := U^*$

$\Rightarrow$ We sample $U^* \sim \mathcal{U}([F_X(a), F_X(b)])$ and then $Y = F^{-1}(U^*)$.

(ii) Gamma distribution (Gamma(m,b))

1. Exponential:

$X \sim \mathcal{E}(\lambda)$: $F(x) = 1 - e^{-\lambda x}$

$U = F(X) = 1 - e^{-\lambda X} \Rightarrow 1 - U = e^{-\lambda X} \Rightarrow \log(1-U) = -\lambda X$

$\Rightarrow X = -\frac{1}{\lambda}\log(1-U) \overset{d}{=} -\frac{1}{\lambda}\log(U)$

We sample $U_i \sim \mathcal{U}([0,1])$ and then $X_i = -\frac{1}{\lambda}\log(U_i)$

2. Gamma:

Remark 1. $X_1, \dots, X_m$: $X_i \sim \text{Gamma}(a_i, b) \Rightarrow \sum_{i=1}^m X_i \sim \text{Gamma}(\sum a_i, b)$

Remark 2. Gamma$(1,b) = \mathcal{E}(b)$

$X \sim$ Gamma(m,b):

for $i = 1, \dots, n_{rep}$:

1. Sample $Y_1, \dots, Y_m \overset{iid}{\sim} \text{Gamma}(1,b) = \mathcal{E}(b)$
2. Set $X_i = \sum_{j=1}^m Y_j$

8. Simulation from a univariate (Box-Müller) and multivariate Gaussian distribution.

Box-Muller method.

We want to sample from $X \sim N(0,1)$.
We start from $(X,Y) \overset{iid}{\sim} N(0,1)$: $f(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}$

We move to polar coordinates:

$$\begin{cases} \rho = x^2 + y^2 & \in [0,\infty) \\ \varphi = \tan^{-1}\left(\frac{y}{x}\right) & \in [0, 2\pi) \end{cases} \Longrightarrow \begin{cases} x = \sqrt{\rho}\cos(\varphi) \\ y = \sqrt{\rho}\sin(\varphi) \end{cases}$$

$$J = \begin{bmatrix} \frac{\partial x}{\partial \rho} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial \rho} & \frac{\partial y}{\partial \varphi} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2\sqrt{\rho}}\cos(\varphi) & -\sqrt{\rho}\sin(\varphi) \\ -\frac{1}{2\sqrt{\rho}}\sin(\varphi) & -\sqrt{\rho}\cos(\varphi) \end{bmatrix}, \quad |J| = \frac{1}{2}$$

$$\Longrightarrow f(\rho,\varphi) = f_{\underline{X}}\left(\sqrt{\rho}\cos(\varphi), \sqrt{\rho}\sin(\varphi)\right) \cdot |J| \cdot \mathbb{1}_{(0,\infty)}(\rho)\, \mathbb{1}_{[0,2\pi)}(\varphi)$$

$$= \underbrace{\frac{1}{2\pi}}_{\sim \mathcal{U}([0,2\pi])} \underbrace{e^{-\frac{1}{2}\rho} \cdot \frac{1}{2}}_{\sim \mathcal{E}(\frac{1}{2})}$$

$\Longrightarrow$ Algorithm: for $i = 1, \dots, n_{rep}/2$:

1. Sample $U_1, U_2 \sim \mathcal{U}([0,1])$
2. Set $\rho = -2\log(U_1)$, $\varphi = (2\pi)U_2$
3. Set $X = \sqrt{\rho}\cos(\varphi)$, $Y = \sqrt{\rho}\sin(\varphi)$

We obtain two independent gaussian random variables X and Y.
If we want :

- $Y \sim N(\mu, \sigma^2) \Rightarrow$ sample $X \sim N(0,1)$, set $Y = \sigma X + \mu$
- $\underline{Y} \sim N_k(\underline{\mu}, \Sigma)$, $\Sigma = A^T A \Rightarrow$ sample $\underline{X} \sim N_k(\underline{0}, I)$, set $\underline{Y} = A\underline{X} + \underline{\mu}$

9. Acceptance-rejection method to sample from a (univariate) target density f.

Application: sampling from $\text{Gamma}(\alpha,\beta)$ with $\alpha > 1$ when the proposal is $\text{Gamma}(m,b)$ with m integer. Parameters choice to maximize the efficiency.

Acceptance-rejection method.

Suppose we want to sample from $X \sim f(x)$ but we're not able to do it. We introduce an instrumental random variable $Y \sim g(y)$ s.t. :

1. $S_X \subseteq S_Y \subseteq \mathbb{R}^k$
2. $f(x) \le M\, g(x) \quad \forall x \qquad (\exists M > 0)$

Prop. Assume $f(x)$ and $g(x)$ densities s.t. 1. and 2. holds. Then, to simulate from $f(x)$ we can simulate from $Y \sim g(y)$ and $U|Y=y \sim \mathcal{U}([0, M\cdot g(y)])$ until $0 < U < f(y)$.

The proposition is saying : $\mathbb{P}\left(Y \le x \,\middle|\, U \le \frac{f(Y)}{Mg(Y)}\right) = \mathbb{P}(X \le x)$

proof.

$$\mathbb{P}\left(Y \le x \,\middle|\, U \le \frac{f(Y)}{Mg(Y)}\right) = \frac{\mathbb{P}\left(Y \le x, U \le \frac{f(Y)}{Mg(Y)}\right)}{\mathbb{P}\left(U \le \frac{f(Y)}{Mg(Y)}\right)} = \frac{\int \mathbb{P}\left(Y \le x, U \le \frac{f(Y)}{Mg(Y)} \,\middle|\, Y=y\right) g(y)\, dy}{\int \mathbb{P}\left(U \le \frac{f(Y)}{Mg(Y)} \,\middle|\, Y=y\right) g(y)\, dy}$$

$$= \frac{\int \mathbb{1}_{(-\infty,x)}(y)\, \mathbb{P}\left(U \le \frac{f(y)}{Mg(y)}\right) g(y)\, dy}{\int \mathbb{P}\left(U \le \frac{f(y)}{Mg(y)}\right) g(y)\, dy} \quad \text{unif c.d.f.}$$

$$= \frac{\int \mathbb{1}_{(-\infty,x)}(y)\, \frac{f(y)}{Mg(y)}\, g(y)\, dy}{\int \frac{f(y)}{Mg(y)}\, g(y)\, dy}$$

$$= \frac{\int \mathbb{1}_{(-\infty,x)}(y)\, f(y)\, dy}{\int f(y)\, dy} = \mathbb{P}(X \le x) \qquad \blacksquare$$

sampling from $f(\cdot)$ given $g(\cdot)$.

- when does it work?
  $\exists M > 0$: 1., 2.
- find M / $M(\theta)$ given $g(\cdot)$:
  $\frac{f(x)}{g(x)} \le h(x,\theta)$
  $\frac{d}{dx} h(x,\theta) = 0 \Rightarrow x^* = \arg\max_x h(\cdot,\cdot)$
  $M(\theta) = h(\theta, x^*)$
- acceptance probability?
  $\int_{\mathbb{R}} \frac{f(x)}{Mg(x)} g(x)\, dx = \frac{1}{M(\theta)}$
  $\Rightarrow \theta^* = \arg\max \frac{1}{M(\theta)}$

(9.) The algorithm is: for $i = 1, \dots, n_{rep}$:

1. Sample $Y \sim g(y)$
2. Sample $U \sim \mathcal{U}([0,1])$
3. If $U \le \frac{f(y)}{Mg(y)}$ set $X_i = Y$, otherwise go to 1.

Sampling from Gamma$(\alpha, 1)$, $\alpha > 1$.

We want to sample from $f(x) \overset{d}{=} \text{Gamma}(\alpha, 1)$. The proposal is $g(x) \overset{d}{=} \text{Gamma}(m, b)$ with $b = \lfloor \alpha \rfloor$. We need to find M s.t. $f \le Mg$:

$$\frac{f}{g} = \frac{\frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}}{\frac{1}{\Gamma(m)} b^m x^{m-1} e^{-bx}} = \frac{\Gamma(m)}{\Gamma(\alpha)} b^{-m} \underbrace{x^{\alpha-m} e^{-(1-b)x}}_{\propto \text{Gamma}(\alpha-m+1,\, 1-b)}$$

conditions: $\begin{cases} 1-b > 0 & (b < 1) \\ \alpha - m + 1 > 0 & (\checkmark) \end{cases}$

For $b \in (0,1)$ the maximum of the Gamma is in its mode: $\frac{\alpha - \lfloor \alpha \rfloor}{1-b}$

$$\Rightarrow \frac{f}{g} \le \frac{\Gamma(\lfloor \alpha \rfloor)}{\Gamma(\alpha)} b^{-\lfloor \alpha \rfloor} \left(\frac{\alpha - \lfloor \alpha \rfloor}{1-b}\right)^{\alpha - \lfloor \alpha \rfloor} e^{-(1-b)\left(\frac{\alpha - \lfloor \alpha \rfloor}{1-b}\right)} := M(b)$$

Optimize in terms of b = minimize $M(b)$ = maximize $M^{-1}(b)$. (since the probability of accepting is $1/M$ and we want to maximize it)

$$\frac{1}{M(b)} \propto b^{\lfloor \alpha \rfloor} (1-b)^{\alpha - \lfloor \alpha \rfloor} \propto \text{Beta}(\lfloor \alpha \rfloor + 1, \lfloor \alpha \rfloor - \alpha + 1)$$

To maximize $M^{-1}(b)$ we consider the mode of Beta$(\cdot,\cdot)$, which is in $\frac{\lfloor \alpha \rfloor}{\alpha}$

$$\Rightarrow b^{\text{optimal}} = \frac{\lfloor \alpha \rfloor}{\alpha} \Rightarrow M^{\text{optimal}} = M(b^{\text{optimal}})$$

10. Optimal choice for the proposal distribution of the importance sampler, as the one minimizing the variance of the estimator.

Importance sampling.

Let f_X be a density and $h(\cdot)$ a measurable function. We're interested in evaluating $\mathbb{E}_f[h(X)] = \int_{\mathbb{R}} h(x) f(dx)$. If we're not able to sample from f_X (and so use a Monte Carlo method) we can introduce an instrumental random variable with density $g(\cdot)$.

We have:
- Generate: $X_1, \dots, X_m \sim g(x)$
- $\mathbb{E}_f[h(X)] \approx \frac{1}{m} \sum_{j=1}^{m} \frac{f(x_j)}{g(x_j)} h(x_j)$

In fact: $\frac{1}{m} \sum_{j=1}^{m} \frac{f(x_j)}{g(x_j)} h(x_j) \xrightarrow{a.s.} \mathbb{E}_g\left[\frac{f(x)}{g(x)} h(x)\right] = \int \frac{f(x)}{g(x)} h(x) g(x)\, dx = \mathbb{E}_f[h(x)]$

The optimal choice of $g(\cdot)$ is the one that minimizes the variance of the importance sampler estimator, i.e.:

$$g^*(x) = \frac{|h(x)| f(x)}{\int |h(t)| f(t)\, dt}$$

proof.

$$\text{Var}_g\left(h(x) \frac{f(x)}{g(x)}\right) = \mathbb{E}_g\left[\left(h(x) \frac{f(x)}{g(x)}\right)^2\right] - \mathbb{E}_g^2\left[h(x) \frac{f(x)}{g(x)}\right]$$

$$\mathbb{E}_g\left[h(x) \frac{f(x)}{g(x)}\right] = \int h(x) \frac{f(x)}{g(x)} g(x)\, dx = \int h(x) f(x)\, dx \quad \perp\!\!\!\perp \; g(x)$$

$\Rightarrow$ $g(x)$ that minimizes $\text{Var}_g(\cdot\cdot)$ = $g(x)$ that minimizes $\mathbb{E}_g[(\cdot\cdot)^2]$

$$\mathbb{E}_g\left[\left(h(x) \frac{f(x)}{g(x)}\right)^2\right] \ge \left(\mathbb{E}_g\left[|h(x)| \frac{f(x)}{g(x)}\right]\right)^2 = \left(\int |h(x)| f(x)\, dx\right)^2$$

If $g(x) = \frac{|h(x)| f(x)}{\int |h(x)| f(x)\, dx}$ we obtain the equality ($\uparrow$)

and so such $g(x)$ is optimal. ■

(*) $\int_E P(x,A)\pi(x)\,dx = \int_E \left(\int_A p(x,y)\,dy + r(x)\mathbb{1}_A(x)\right)\pi(x)\,dx$

$= \int_A \left(\int_E p(x,y)\pi(x)\,dx\right)dy + \int_A r(x)\pi(x)\,dx$

R $= \int_A \left(\int_E p(y,x)\,dx\right)\pi(y)\,dy + \int_A r(x)\pi(x)\,dx$

$= \int_A (1-r(y))\pi(y)\,dy + \int_A r(x)\pi(x)\,dx = \int_A \pi(y)\,dy = \pi(A)$ ■

11. Reversability of the Metropolis-Hasting algorithm w.r.t. the target distribution π.

Metropolis-Hastings:

$\pi(x)$ = target distribution w.r.t. a measure μ

$Q(x,dy) := q(x,y)\mu(dy)$ = ~~transition probability~~ proposal density

$$\alpha(x,y) := \begin{cases} \min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right) & \text{if } \pi(x)q(x,y) > 0 \\ 1 & \text{otherwise} \end{cases}$$

= probability of accepting y (that comes from $Y \sim Q(x,\cdot)$) starting from x

$$p(x,y) := \begin{cases} q(x,y)\alpha(x,y) & x \neq y \\ 0 & x = y \end{cases}$$

$P(x,A) := \int_A p(x,y)\mu(dy) + r(x)\mathbb{1}_A(x)$ = transition probability

prob. of remaining in x

Reversability condition: $p(x,y)\pi(x) = p(y,x)\pi(y)$

proof.

since $\alpha(\cdot,\cdot)$ is a minimum we have 2 cases:

1. $\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} < 1$

$p(x,y)\pi(x) = q(x,y)\alpha(x,y)\pi(x) = q(x,y)\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\pi(x) = \pi(y)q(y,x)\underbrace{1}_{:=\alpha(y,x)}$

Is it correct to consider $\alpha(y,x) = 1$? Yes.

If $\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} < 1 \Rightarrow \frac{\pi(x)q(x,y)}{\pi(y)q(y,x)} > 1 \Rightarrow \alpha(y,x) = \min(\cdot,1) = 1$

2. $\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} > 1$

$p(y,x)\pi(y) = q(y,x)\alpha(y,x)\pi(y) = q(y,x)\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}\pi(y) = \pi(x)\cdot q(x,y)\underbrace{1}_{:=\alpha(x,y)}$

This implies that $\pi(\cdot)$ is invariant:

$\int_E P(x,A)\pi(x)\,dx = \pi(A) \quad \forall A$ (*) ■

12. Gaussian linear model with homoschedastic (iid) errors: conjugate prior when the variance is known/unknown, posterior and parameters updates; Jeffrey's prior (proportional to $1/\sigma^2$, where σ^2 is the error variance) and its posterior.

Suppose that we have the data $(y_i, \underline{x}_i)$ $i = 1,\dots,n$, where y_i are the responses and $\underline{x}_i$ are the covariates (p-dimensional). Suppose y_i to be continuous. We're interested in finding correlation between $\mathbb{E}[Y_i]$ and $\underline{X}_i$.

We assume:

$$\underline{Y} \mid X, \underline{\beta}, \sigma^2 \sim N_n(X\underline{\beta}, \sigma^2 I_n)$$

X = $n \times p$ matrix of covariates

due to the modelization: $Y_i = \underline{X}_i^T\underline{\beta} + \varepsilon_i$ $\left(\varepsilon_i \sim N(0,\sigma^2)\right)$.

The parameters of interest are $(\underline{\beta}, \sigma^2)$ or equivalently $(\underline{\beta}, \tau)$.

(We consider σ^2 the variance and $\tau = 1/\sigma^2$ the precision).

The likelihood of the model is:

$$L(\underline{\beta}, \sigma^2; \underline{y}) \propto \frac{1}{(\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta})}$$

Considering $\hat{\underline{\beta}}_{MLE} = (X^TX)^{-1}X^T\underline{y}$, $S^2 = (\underline{y} - X\hat{\underline{\beta}}_{MLE})^T(\underline{y} - X\hat{\underline{\beta}}_{MLE})$

$$(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta}) = (\underline{y} - X\hat{\underline{\beta}}_{MLE})^T(\underline{y} - X\hat{\underline{\beta}}_{MLE}) + (\underline{\beta} - \hat{\underline{\beta}}_{MLE})^TX^TX(\underline{\beta} - \hat{\underline{\beta}}_{MLE})$$

$$= S^2 + (\underline{\beta} - \hat{\underline{\beta}}_{MLE})^TX^TX(\underline{\beta} - \hat{\underline{\beta}}_{MLE})$$

$$\Rightarrow \boxed{L(\underline{\beta}, \sigma^2; \underline{y}) \propto \frac{1}{(\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\left[S^2 - (\underline{\beta} - \hat{\underline{\beta}}_{MLE})^TX^TX(\underline{\beta} - \hat{\underline{\beta}}_{MLE})\right]}}$$

(12.) **Conjugate prior when σ^2 is known.**

Prior: $\underline{\beta} \sim N_p(\underline{b}_0, B_0)$, B_0 invertible $p \times p$ matrix

Posterior: $\underline{\beta}|\underline{y} \propto \exp\left\{-\frac{\tau}{2}(\underline{\beta}-\hat{\underline{\beta}}_{MLE})^T X^T X(\underline{\beta}-\hat{\underline{\beta}}_{MLE}) - \frac{1}{2}(\underline{\beta}-\hat{\underline{b}}_0)^T B_0^{-1}(\underline{\beta}-\hat{\underline{b}}_0)\right\}$

$$\underline{\beta}|\underline{y} \sim N_p\left((\tau X^T X + B_0^{-1})^{-1}(\tau X^T X \hat{\underline{\beta}}_{MLE} + B_0^{-1}\underline{b}_0),\ (\tau X^T X + B_0^{-1})^{-1}\right)$$

The posterior mean is a weighted average of the prior mean $\underline{b}_0$ and $\hat{\underline{\beta}}_{MLE}$.

Conjugate prior when σ^2 is unknown.

We consider: $\pi(\underline{\beta}, \sigma^2) = \pi(\underline{\beta}|\sigma^2)\,\pi(\sigma^2)$

Prior:
$$\begin{cases} \underline{\beta}|\sigma^2 \sim N_p(\underline{b}_0, \sigma^2 B_0) \\ \sigma^2 \sim \text{inv-gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right) \end{cases}$$

$\underline{b}_0 \in \mathbb{R}^p$, B_0 invertible $p \times p$, $\nu_0, \sigma_0^2 > 0$

Again we consider: $\pi(\underline{\beta}, \sigma^2|\underline{y}) = \pi(\underline{\beta}|\sigma^2, \underline{y})\,\pi(\sigma^2|\underline{y})$

Posterior:
$$\begin{cases} \underline{\beta}|\sigma^2, \underline{y}, X \sim N_p(\underline{b}_n, \sigma^2 B_n) \\ \sigma^2|\underline{y}, X \sim \text{inv-gamma}\left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}\right) \end{cases}$$

$$B_n = (X^T X + B_0^{-1})^{-1}$$
$$\underline{b}_n = (X^T X + B_0^{-1})^{-1}(X^T X \hat{\underline{\beta}}_{MLE} + B_0^{-1}\underline{b}_0)$$
$$\nu_n = \nu_0 + n$$
$$\sigma_n^2 = \frac{1}{\nu_n}\left(\nu_0 \sigma_0^2 + \underline{b}_0^T B_0^{-1}\underline{b}_0 + \underline{y}^T \underline{y} - \underline{b}_n^T B_n^{-1}\underline{b}_n\right)$$

Jeffrey's: prior: $\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}\mathbb{1}_{(0,\infty)}(\sigma^2)$

posterior:
$$\begin{cases} \underline{\beta}|\sigma^2, \underline{y}, X \sim N_p(\hat{\underline{\beta}}_{MLE}, \sigma^2(X^T X)^{-1}) \\ \sigma^2|\underline{y}, X \sim \text{inv-gamma}\left(\frac{n-p}{2}, \frac{s^2}{2}\right) \end{cases}$$

proper if $X^T X$ is invertible and $n > p$

13. Formal approach to the model choice: calculation of the posterior probability of the model j, $j = 1, \dots, K$.

In the formal approach of model selection we fix prior probability masses for any model. Then we use the Bayes theorem to compute the posterior distribution. The best models are those with highest posterior probability.
Let m be the index describing K models $M_1, \dots, M_K$.

- $\mathbb{P}(m=j)$ = prior prob. to choose model M_j (typically $= 1/K$)
- model M_j: likelihood $f(\underline{y}|\underline{\theta}_j, M_j)$, prior $\pi(\underline{\theta}_j|M_j) = \pi(\underline{\theta}_j|m=j)$ where $\underline{\theta}_j$ is the vector of parameters

How does it work?

1: for any model M_j compute the posterior for the parameters:

$$\pi(\underline{\theta}_j|\underline{y}, M_j) = \frac{f(\underline{y}|\underline{\theta}_j, M_j)\,\pi(\underline{\theta}_j|M_j)}{m(\underline{y}|M_j)}, \qquad m(\underline{y}|M_j) = \int f(\underline{y}|\underline{\theta}_j, M_j)\,\pi(\underline{\theta}_j|M_j)\,d\underline{\theta}_j$$

($\pi(\underline{\theta}_j|\underline{y}, M_j)$: posterior under model M_j; $f(\underline{y}|\underline{\theta}_j, M_j)$: likelihood under model M_j; $\pi(\underline{\theta}_j|M_j)$: prior under model M_j; $m(\underline{y}|M_j)$: marginal distr. of data under model M_j)

2: compute the posterior probability masses of $M_1, \dots, M_K$:

$$\mathbb{P}(m=j|\underline{y}) = \frac{m(\underline{y}|M_j)\,\mathbb{P}(m=j)}{m(\underline{y})}, \qquad m(\underline{y}) = \sum_{j=1}^K m(\underline{y}|M_j)\,\mathbb{P}(m=j)$$

3: choose the model with the highest $\mathbb{P}(m=j|\underline{y})$. We can select 2/3 models and then compare these models through some goodness-of-fit criteria.

likelihood · prior / marginal

14. Computation of LPML (log-pseudo marginal likelihood) from a MCMC sample from the posterior, given all the datapoints.

Cross-validation.

We're going to split the dataset in training set and test set. The training we'll be used to compute the estimates, the test we'll be used for a comparison with the prediction computed through the training set.

Suppose to have n data. We split in $n-1$ and 1.

$\underline{y} = (y_1, \dots, y_n) \Longrightarrow \underline{y}_{-i} = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$ training, y_i test

For all i $(i=1,\dots,n)$ we find $\mathcal{L}(Y_i \mid \underline{Y}_{-i})$ and compute

$$CPO_i := f_i(y_i \mid \underline{y}_{-i}) = \text{CONDITIONAL PREDICTIVE ORDINATE } i$$

We suppose $Y_1, \dots, Y_n$ s.t. $Y_i \mid \theta \sim f_i$;

$$f_i(y_i \mid \underline{y}_{-i}) = \int f_i(y_i \mid \theta)\, \pi(\theta \mid \underline{y}_{-i})\, d\theta$$

$$= \int f_i(y_i \mid \theta)\, \frac{\pi(\theta) \prod_{j \neq i} f_j(y_j \mid \theta)}{\int_\Theta \prod_{j \neq i} f_j(y_j \mid \theta)\, \pi(\theta)\, d\theta}\, d\theta$$

= likelihood integrated w.r.t. the posterior distribution using all the data but the i-th

$$\frac{1}{f_i(y_i \mid \underline{y}_{-i})} = \frac{\int_\Theta \prod_{j \neq i} f_j(y_j \mid \theta)\, \pi(\theta)\, d\theta}{\int_\Theta \prod_{j=1}^n f_j(y_j \mid \theta)\, \pi(\theta)\, d\theta} = \frac{\int_\Theta \frac{1}{f_i(y_i \mid \theta)} \prod_{j=1}^n f_j(y_j \mid \theta)\, \pi(\theta)\, d\theta}{\int_\Theta \prod_{j=1}^n f_j(y_j \mid \theta)\, \pi(\theta)\, d\theta}$$

$$= \int_\Theta \frac{1}{f_i(y_i \mid \theta)}\, \pi(\theta \mid y_1, \dots, y_n)\, d\theta$$

posterior distribution using all the data

$$\approx \frac{1}{M} \sum_{m=1}^M \frac{1}{f_i(y_i \mid \theta^{(m)})} \qquad (\theta^{(m)} = \text{MCMC sample})$$

$$\Longrightarrow \hat{CPO}_i = \left(\frac{1}{M} \sum_{m=1}^M \frac{1}{f_i(y_i \mid \theta^{(m)})} \right)^{-1}$$

The highest the CPO_i, the better the model explains the i-th observation.

For model comparison we calculate the LOG-PSEUDO-MARGINAL-LIKELIHOOD

$$LPML := \sum_{i=1}^n \log(CPO_i) \qquad \text{(the higher the better)}$$

15. Probit regression model with latent variables: comp. of the full-conditionals

We consider the data $(y_i, \underline{x}_i)$ $i = 1, \dots, n$, $y_i \in \{0, 1\}$.

We assume:

$$Y_i \mid \underline{x}_i, \underline{\beta} \sim Be(\pi_i), \quad \pi_i = \Phi(\underline{x}_i^T \underline{\beta})$$

$$\underline{\beta} \sim \pi(\underline{\beta}) \quad (\text{commonly: } \underline{\beta} \sim N_p(\underline{b}_0, B_0))$$

This means that: $\mathbb{P}(Y_i = 1) = \pi_i = \Phi(\underline{x}_i^T \underline{\beta})$.

We introduce latent variables $Z_1, \dots, Z_n$:

$$Z_i = \underline{x}_i^T \underline{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0,1) \Longrightarrow Z_i \overset{iid}{\sim} N(\underline{x}_i^T \underline{\beta}, 1)$$

and we consider:

$$Y_i = \begin{cases} 1 & \text{if } Z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

$1 - \Phi(-k) =$

The models are equivalent:

$$\mathbb{P}(Y_i = 1) = \mathbb{P}(Z_i > 0) = 1 - \mathbb{P}(Z_i \leq 0) = 1 - \mathbb{P}\left(\frac{Z_i - \underline{x}_i^T \underline{\beta}}{1} \leq \frac{-\underline{x}_i^T \underline{\beta}}{1} \right) = \Phi(\underline{x}_i^T \underline{\beta})$$

(15.) The parameters now are : $(\underline{\beta}, \underline{z})$.
A Gibbs sampler can simply sample from $\pi(\underline{\beta}, \underline{z}|\underline{y})$, then we'll focus on the marginal of $\underline{\beta}$. ($\pi(\underline{\beta}|\underline{y})$)

$$\pi(\underline{\beta}, \underline{z}|\underline{y}) \propto \mathcal{L}(\underline{\beta}, \underline{z}, \underline{Y}) = \mathcal{L}(\underline{Y}|\underline{\beta}, \underline{z})\, \mathcal{L}(\underline{z}|\underline{\beta})\, \mathcal{L}(\underline{\beta})$$

$$\propto \prod_{i=1}^{n} \Big[\underbrace{\left(\mathbb{1}_{\{Y_i=1\}} \mathbb{1}_{\{z_i>0\}} + \mathbb{1}_{\{Y_i=0\}} \mathbb{1}_{\{z_i \le 0\}} \right)}_{\mathcal{L}(Y_i|\underline{z})} \cdot \underbrace{N\left(z_i ; \underline{x}_i^T \underline{\beta}, 1\right)}_{\mathcal{L}(\underline{z}|\underline{\beta})} \Big] \underbrace{\pi(\underline{\beta})}_{\mathcal{L}(\underline{\beta})}$$

The full-conditionals are:

- $[\underline{\beta}|\underline{z}, \underline{y}] \propto \mathcal{L}(\underline{Y}, \underline{z}, \underline{\beta}) = \mathcal{L}(\underline{Y}|\underline{\beta}, \underline{z})\, \mathcal{L}(\underline{z}|\underline{\beta})\, \mathcal{L}(\underline{\beta})$
  $\propto \mathcal{L}(\underline{z}|\underline{\beta})\pi(\underline{\beta})$. This last distribution is the posterior of a linear model with Gaussian likelihood where the data are $\underline{z}$:

  ~~if $\pi(\underline{\beta}) = \text{constant} \Rightarrow [\underline{\beta}|\underline{z}, \underline{y}] \sim N_p(\hat{\underline{\beta}}, (X^TX)^{-1})$~~

  if $\pi(\underline{\beta}) = N_p(\underline{b}_0, B_0) \Rightarrow [\underline{\beta}|\underline{z}, \underline{y}] \sim N_p(\tilde{\underline{\beta}}, \tilde{B})$

  where :
  $$\begin{cases} \tilde{B} = (X^TX + B_0^{-1})^{-1} \\ \tilde{\underline{\beta}} = \tilde{B}(X^TX\hat{\underline{\beta}} + B_0^{-1}\underline{b}_0) \\ \hat{\underline{\beta}} = (X^TX)^{-1}X\underline{z} \end{cases}$$

- $[\underline{z}|\underline{\beta}, \underline{y}] \propto \mathcal{L}(\underline{Y}, \underline{z}, \underline{\beta}) = \mathcal{L}(\underline{Y}|\underline{z}, \underline{\beta})\, \mathcal{L}(\underline{z}|\underline{\beta})\, \mathcal{L}(\underline{\beta})$
  $$\propto \prod_{i=1}^{n} \left[\left(\mathbb{1}_{\{Y_i=1\}} \mathbb{1}_{\{z_i>0\}} + \mathbb{1}_{\{Y_i=0\}} \mathbb{1}_{\{z_i \le 0\}} \right) \cdot N(z_i ; \underline{x}_i^T \underline{\beta}, 1) \right]$$

  because of the product :
  $$z_i|\underline{\beta}, y_i = \begin{cases} \mathbb{1}_{\{z_i>0\}} \cdot N(z_i ; \underline{x}_i^T\underline{\beta}, 1) & y_i = 1 \\ \mathbb{1}_{\{z_i \le 0\}} \cdot N(z_i ; \underline{x}_i^T\underline{\beta}, 1) & y_i = 0 \end{cases}$$

  $$\Rightarrow \begin{cases} z_i|\underline{\beta}, y_i=1 & \text{is a } N(\underline{x}_i^T\underline{\beta}, 1) \text{ truncated to } z_i \in (0, \infty) \\ z_i|\underline{\beta}, y_i=0 & \text{is a } N(\underline{x}_i^T\underline{\beta}, 1) \text{ truncated to } z_i \in (-\infty, 0] \end{cases}$$

  "truncated to $(0,\infty)$" :

16. Likelihood for right-censored (conditionally independent) data

T_i = lifetime / failure time $\quad i = 1,..,n$
C_i = time of censoring $\quad i = 1,..,n$
$\Big\} \perp\!\!\!\perp$

We have the data (y_i, δ_i) $i=1,..,n$: $y_i = \min(T_i, C_i)$, $\delta_i = \mathbb{1}\{T_i \le C_i\}$

We assume :

$T_i|\theta \sim f_i(t|\theta)$, S_i = survival function , h_i = hazard function

Then, the likelihood given the data is:

$$L(\theta|D) \propto \prod_{i=1}^{n} \left[(f_i(y_i|\theta))^{\delta_i} \, (S_i(y_i|\theta))^{1-\delta_i} \right]$$

which means:
the contribution to the likelihood is:

- $f_i(y_i|\theta)$ if $\delta_i = 1$
- $S_i(y_i|\theta)$ if $\delta_i = 0$

proof.

(Instead of $\mathbb{P}(T_i \in dy|\theta)$ we use $\mathbb{P}(T_i = y|\theta)$)
The contribution of the i-th observation is:

$$L(\theta|y_i,\delta_i) = \begin{cases} \mathbb{P}(T_i = y_i,\ \delta_i = 1|\theta) & \delta_i = 1 \\ \mathbb{P}(C_i = y_i,\ \delta_i = 0|\theta) & \delta_i = 0 \end{cases}$$

$$= \begin{cases} \mathbb{P}(T_i = y_i,\ T_i \le C_i|\theta) & \delta_i = 1 \\ \mathbb{P}(C_i = y_i,\ T_i > C_i|\theta) & \delta_i = 0 \end{cases}$$

$$= \begin{cases} \mathbb{P}(T_i = y_i,\ C_i \ge y_i|\theta) & \delta_i = 1 \\ \mathbb{P}(C_i = y_i,\ T_i > y_i|\theta) & \delta_i = 0 \end{cases}$$

$T_i \perp\!\!\!\perp C_i$

$$= \begin{cases} \mathbb{P}(T_i = y_i|\theta)\ \mathbb{P}(C_i \ge y_i|\theta) & \delta_i = 1 \\ \mathbb{P}(C_i = y_i|\theta)\ \mathbb{P}(T_i > y_i|\theta) & \delta_i = 0 \end{cases}$$

$C_i \perp\!\!\!\perp \theta$

$$= \begin{cases} f_i(y_i|\theta)\ \text{const}_1 & \delta_i = 1 \\ S_i(y_i|\theta)\ \text{const}_2 & \delta_i = 0 \end{cases}$$

■

The following assumptions must hold:

1. T_i and C_i are independent (independent censoring)
2. The distribution of C_i does not depend on parameters contained in the distribution of T_i (non-informative censoring). In our case this means $C_i \perp\!\!\!\perp \theta$
3. All the subjects in the sample are independent conditionally to all the parameters.

we proved what is the contribution to the likelihood in case of $\delta_i = 0$ or $\delta_i = 1$. The (total) likelihood is the joint distribution of all the data.

17. First and second moments of the Dirichlet Process P (included the covariance).

Finite-dimensional Dirichlet distribution:

$(D_1, \dots, D_{k-1}, D_k) \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_{k-1}, \alpha_k)$ $\quad \alpha_j > 0$, $D_j \in [0,1]$ $\forall j$, $\sum_j D_j = 1$

if the distribution of $(D_1, \dots, D_k)$ cannot be absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^k$, but the density of $(D_1, \dots, D_{k-1})$ is:

$$f(x_1, \dots, x_{k-1}) = \frac{\Gamma(\alpha_1 + \dots + \alpha_k)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \left[\prod_{j=1}^{k-1} x_j^{\alpha_j - 1} \right] (1 - x_1 - \dots - x_{k-1})^{\alpha_k - 1} \mathbb{1}_{S_{k-1}}(\underline{x})$$

where $S_{k-1} = \{ (x_1, \dots, x_{k-1}) : \in \mathbb{R}^{k-1} : 0 \le x_1 + \dots + x_{k-1} \le 1, \ 0 \le x_j \le 1 \ \forall j = 1, \dots, k-1 \}$

Dirichlet Process:

Let α be a finite measure on $\mathbb{R}$: $0 < \alpha(\mathbb{R}) < \infty$ $\quad (a := \alpha(\mathbb{R}),\ \alpha_0(A) = \frac{\alpha(A)}{a})$

we say that P is a Dirichlet Process if for any finite and measurable partition of $\mathbb{R}$, say $A_1, \dots, A_k$, we have:

$$(P(A_1), \dots, P(A_k)) \sim \mathrm{Dir}(\alpha(A_1), \dots, \alpha(A_k)) \quad : \quad P \sim DP(a, \alpha_0)$$

properties:

1. $P(A) \sim \mathrm{Beta}(\alpha(A), \alpha(A^c)) \quad \forall A \in \mathcal{B}(\mathbb{R})$
2. $\mathbb{E}[P(A)] = \alpha_0(A)$, $\quad \mathrm{Var}(P(A)) = \frac{\alpha_0(A)\,\alpha_0(A^c)}{a+1} \quad \forall A \in \mathcal{B}(\mathbb{R})$

proof.

$P(A) \sim \mathrm{Beta}(\cdot,\cdot) \Longrightarrow \mathbb{E}[P(A)] = \frac{\alpha(A)}{\alpha(A) + \alpha(A^c)} = \frac{\alpha(A)}{\alpha(\mathbb{R})} = \alpha_0(A)$

$$\mathrm{Var}(P(A)) = \frac{\alpha(A)\,\alpha(A^c)}{(\alpha(A)+\alpha(A^c))^2\,(\alpha(A)+\alpha(A^c)+1)} = \frac{\alpha(A)\,\alpha(A^c)}{a^2(a+1)} = \frac{\alpha_0(A)\,\alpha_0(A^c)}{a+1}$$ ■

3. $\mathbb{E}[P(A)P(B)] = \frac{\alpha(A\cap B) + \alpha(A)\,\alpha(B)}{a(a+1)}$

$\mathrm{Cov}(P(A), P(B)) = \frac{\alpha_0(A\cap B) - \alpha_0(A)\,\alpha_0(B)}{a+1}$ $\quad \forall A, B \in \mathcal{B}(\mathbb{R})$

proof.

1. Assume $A \cap B = \phi$. We generate the partition $\{A, B, (A\cup B)^c\}$ of $\mathbb{R}$.

$(P(A), P(B), P(A\cup B)^c) \sim \mathrm{Dir}(\alpha(A), \alpha(B), \alpha((A\cup B)^c))$

and do we know the density of $(P(A), P(B))$: (finite-dim. Dir)

$$\mathbb{E}[P(A)P(B)] = \int_{S_2} x_1 x_2 \frac{\Gamma(a)}{\Gamma(\cdot)\Gamma(\cdot)\Gamma(\cdot)} x_1^{\alpha(A)-1} x_2^{\alpha(B)-1} (1-x_1-x_2)^{\alpha((A\cup B)^c)-1}\, dx_1\, dx_2$$

$$= \frac{\Gamma(a)}{\Gamma(\cdot)\Gamma(\cdot)\Gamma(\cdot)} \cdot \frac{\Gamma(\alpha(A)+1)\,\Gamma(\alpha(B)+1)\,\Gamma(\alpha((A\cup B)^c))}{\Gamma(a+2)}$$

$$= \frac{\alpha(A)\,\alpha(B)}{(a+1)\,a}$$

$$\mathrm{Cov}(P(A), P(B)) = \mathbb{E}[P(A)P(B)] - \mathbb{E}[P(A)]\,\mathbb{E}[P(B)]$$

$$= \frac{\alpha(A)\alpha(B)}{(a+1)a} - \alpha_0(A)\alpha_0(B) = -\frac{\alpha_0(A)\,\alpha_0(B)}{a+1}$$

2. If $A \cap B \neq \phi$:

$A = (A\cap B) \cup (A \setminus B)$

$$\mathrm{Cov}(P(A), P(B)) = \mathrm{Cov}(P(A\cap B) + P(A\setminus B),\ P(A\cap B) + P(B\setminus A))$$

$$= \mathrm{Var}(P(A\cap B)) + \mathrm{Cov}(P(A\setminus B), P(A\cap B)) + \dots$$

(disjoint we are 1.)

■

18. Joint marginal distr. of a sample from a Dirichlet Process; its interpretation as the generalized Pólya urn.

$X_1, .., X_n | P \sim P$, $P \sim \mathcal{D}_\alpha$ — $P|\alpha, \alpha_0 \sim DP(\alpha, \alpha_0)$

We want to describe $\chi(X_1, .., X_n)$: $\chi(X_1, .., X_n) = \chi(X_1)\chi(X_2|X_1) .. \chi(X_n|X_1, .., X_{n-1})$

$$\mathbb{P}(X_1 \in A) = \int \mathbb{P}(X_1 \in A | P)\, \mathcal{D}_\alpha(dP) = \int \chi(X_1, dP) = \int \chi(X_1|P)\, \chi(dP)$$
$$= \int_{\mathcal{P}} P(A)\, \mathcal{D}_\alpha(dP) = \mathbb{E}[P(A)] = \alpha_0(A)$$

$\longrightarrow \chi(X_1) = \alpha_0$

$$\chi(X_2|X_1) = \int_{\mathcal{P}} \chi(X_2, dP|X_1) = \int_{\mathcal{P}} \chi(X_2|P, \cancel{X_1})\, \chi(dP|X_1)$$
$$= \int_{\mathcal{P}} P \cdot \mathcal{D}_{\alpha + \delta_{X_1}}(dP) = \mathbb{E}[P|X_1] = \frac{\alpha + \delta_{X_1}}{a+1} = \frac{a}{a+1}\alpha_0 + \frac{1}{a+1}\delta_{X_1}$$

Written differently: $X_2|X_1 = \begin{cases} \sim \alpha_0 & \text{with prob. } a/a+1 \\ = X_1 & \text{with prob. } 1/a+1 \end{cases}$

$$\chi(X_i|X_{i-1}, .., X_1) = \int_{\mathcal{P}} \chi(X_i, dP|X_{i-1}, .., X_1) = \int_{\mathcal{P}} \chi(X_i|P, \cancel{X_{i-1}, .., X_1})\, \chi(dP|X_1, .., X_{i-1})$$
$$= \int_{\mathcal{P}} P \cdot \mathcal{D}_{\alpha + \sum_{j=1}^{i-1}\delta_{X_j}}(dP) = \mathbb{E}[P|X_1, .., X_{i-1}]$$
$$= \frac{\alpha + \sum_{j=1}^{i-1}\delta_{X_j}}{a+i-1} = \frac{a}{a+i-1}\alpha_0 + \frac{1}{a+i-1}\sum_{j=1}^{i-1}\delta_{X_j}$$

Written differently:

$$X_i|X_{i-1}, .., X_1 = \begin{cases} \sim \alpha_0 & \text{with prob. } a/a+i-1 \\ = X_1 & \text{with prob. } 1/a+i-1 \\ = X_2 & \text{with prob. } 1/a+i-1 \\ \vdots & \vdots \\ = X_{i-1} & \text{with prob. } 1/a+i-1 \end{cases}$$

$$\longrightarrow \chi(X_1, .., X_n) = \alpha_0 \prod_{i=2}^{n} \left(\frac{a\alpha_0 + \sum_{j=1}^{i-1}\delta_{X_j}}{a+i-1} \right) := \text{GENERALIZED PÓLYA URN (//Chinese Restaurant Process)}$$

The generalized Pólya urn is a sampling scheme. We consider a urn. At each step we can sample a ball of the j-th color with probability proportional to the current number of balls of color j. We can also sample a ball of a new color from an other urn with an infinite number of balls $\alpha(\cdot)$, where each color is unique.

By sampling from a generalized Pólya urn we obtain a trajectory from a Dirichlet process.

Predictive for the Pólya urn: / for the Dirichlet Process

$$\chi(X_{n+1}|X_1, .., X_n) = \frac{a}{a+n}\alpha_0 + \frac{1}{a+n}\sum_{j=1}^{n}\delta_{X_j}$$

The resulting distribution over labels with this sampling scheme is the same as the distribution over values in a Dirichlet process.

Stick-breaking scheme:

- α = finite measure on $\mathbb{R}$
- $\theta_1, \theta_2, ..$ $\overset{iid}{\sim}$ α_0 $(= \frac{\alpha(\cdot)}{a})$
- $Y_1, Y_2, ..$ $\overset{iid}{\sim}$ Beta$(1, a)$
- $V_1 = Y_1$, $V_j = Y_j \prod_{i=1}^{j-1}(1 - Y_i)$ $j = 2, 3, ..$

$$\longrightarrow P := \sum_{j=1}^{+\infty} V_j \delta_{\theta_j} \sim \mathcal{D}_\alpha$$

Note: $\sum_{j=1}^{+\infty} V_j = 1$

19. Pólya urn (with 2 colors): the sequence of random v.s. representing the color of the sampled balls is exchangeable with Beta distr. as Finetti measure.

Characterization of exchangeability:

$X_1, \dots, X_n$ is ex. iff $\exists\, Q$ prob. meas. : $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \int \prod_{i=1}^n \mathbb{P}(X_i = x_i \mid \theta)\, Q(d\theta)$

We consider a **Pólya urn:**

we have a urn with balls of 2 colors, black and white. At each step we sample a ball with probability proportional to the number of balls of the same color currently in the urn. Once we sampled a ball, we return the ball in the urn and we add a new ball of the same color.

Let X_i = i-th Bernoulli sampling : $X_i = \begin{cases} 1 & \text{if black} \\ 0 & \text{if white} \end{cases}$,

W = # initial white balls, B = # initial black balls, S_n = # black balls at n trial

A sequence is exchangeable:

$$\mathbb{P}((1,0)) = \frac{B}{B+W} \cdot \frac{W}{B+W+1} = \mathbb{P}((0,1))$$

more generally: $\mathbb{P}(X_1, \dots, X_n) = \dfrac{\Gamma(B+S_n)\,\Gamma(W+(n-S_n))\,\Gamma(B+W)}{\Gamma(B)\,\Gamma(W)\,\Gamma(B+W+n)}$

From de Finetti theorem: $\mathbb{P}(X_1, \dots, X_n) = \int \prod_{i=1}^n \theta^{X_i}(1-\theta)^{1-X_i}\, Q(d\theta) = \int \theta^{S_n}(1-\theta)^{n-S_n}\, Q(d\theta)$

We call Q the de Finetti distr. $\lim_{n\to\infty} \overline{X}_n \sim \text{Beta}(B, W)$.

We have:

$$\mathbb{P}(X_1, \dots, X_n) = \int \theta^{S_n}(1-\theta)^{n-S_n} Q(d\theta) = \int \theta^{S_n}(1-\theta)^{n-S_n} \frac{\theta^{B-1}(1-\theta)^{W-1}}{B(B,W)}\, d\theta$$

$$= \frac{B(B+S_n, W+(n-S_n))}{B(B,W)} \int \frac{\theta^{B+S_n-1}(1-\theta)^{W+n-S_n-1}}{B(B+S_n, W+(n-S_n))}\, d\theta$$

$$= \frac{\Gamma(B+S_n)\,\Gamma(W+(n-S_n))\,\Gamma(B+W)}{\Gamma(B)\,\Gamma(W)\,\Gamma(B+W+n)} = \text{probability associated to an exchangeable sequence produced from a Pólya urn}$$

20. Predictive distribution under the DPM model, marginalizing (i.e. integrating out) the r.p.m. P (Dirichlet Process)

Dirichlet Process Mixtures

We assume:

$$\begin{cases} X_1, \dots, X_n \mid P \overset{iid}{\sim} f(x)(w) = \int_\Theta k(x;\theta)\, P(d\theta) \\ P \sim \mathcal{D}_\alpha, \quad \alpha = a \cdot \alpha_0 \end{cases} \iff \begin{cases} X_i \mid \theta_i \overset{ind}{\sim} k(\cdot;\theta_i) \quad i = 1, \dots, n \\ \theta_1, \dots, \theta_n \mid P \overset{iid}{\sim} P \\ P \sim \mathcal{D}_\alpha \\ X_1, \dots, X_n \mid \underline{\theta} \perp\!\!\!\perp P \mid \underline{\theta} \end{cases}$$

where $k(x;\theta)$ is a parametric kernel : a density or function of x and a measurable function as function of θ.

Predictive distribution:

$$\chi(X_{n+1} \mid X_1, \dots, X_n) = \int_\rho \chi(X_{n+1}, P \mid X_1, \dots, X_n)\, dP = \int_\rho \chi(X_{n+1} \mid P, \cancel{X_1, \dots, X_n})\, \chi(P \mid X_1, \dots, X_n)\, dP$$

$$= \int_\rho \chi(X_{n+1} \mid P)\, \chi(P \mid X_1, \dots, X_n)\, dP = \int_\rho \int_\Theta k(x;\theta)\, P(d\theta)\, \chi(P \mid X_1, \dots, X_n)\, dP$$

$$= \int_\Theta k(x;\theta) \int_\rho \chi(P \mid X_1, \dots, X_n)\, P(d\theta)\, dP = \int_\Theta k(x;\theta)\, \mathbb{E}[P(d\theta) \mid X_1, \dots, X_n]$$

$$= \int_\Theta k(x;\theta) \left[\frac{a}{a+n}\, \alpha_0(d\theta) + \sum_{j=1}^n \frac{1}{a+n}\, \delta_{X_j} \right]$$

$$= \frac{a}{a+n} \int_\Theta k(x;\theta)\, \alpha_0(d\theta) + \frac{1}{a+n} \sum_{j=1}^n k(x;\theta_j)$$

21. Computation of the full conditionals from the Pólya-urn scheme in DPM.

$$\begin{cases} X_1,\dots,X_n \mid P \overset{iid}{\sim} F(x)(\omega) = \int k(x;\theta)P(d\theta) \\ P \sim \mathcal{D}_\alpha, \quad \alpha = a\alpha_0 \end{cases} \iff \begin{cases} X_i \mid \theta_i \overset{ind}{\sim} k(\cdot;\theta_i) \\ \theta_1,\dots,\theta_n \mid P \overset{iid}{\sim} P \\ P \sim \mathcal{D}_\alpha \\ X_1,\dots,X_n \mid \underline{\theta} \perp\!\!\!\perp P \mid \underline{\theta} \end{cases}$$

$$P \mid X_1,\dots,X_n \sim \int \mathcal{D}_{\alpha + \sum_{i=1}^n \delta_{\theta_i}} H(d\theta_1,\dots,d\theta_n \mid X_1,\dots,X_n)$$

where:

$$H(d\theta_1,\dots,d\theta_n \mid X_1 = x_1,\dots,X_n = x_n) \propto \prod_{i=1}^n k(x_i;\theta_i)\, \alpha_0(d\theta_1) \prod_{i=2}^n \left[\alpha(d\theta_i) + \sum_{j=1}^{i-1} \delta_{\theta_j}(d\theta_i) \right]$$

22. Cluster estimates (definitions)

We're in a mixture model frameworks:

$$\begin{cases} X_i \mid \theta_i \overset{iid}{\sim} k(\cdot;\theta_i) \\ \theta_1,\dots,\theta_n \mid P \overset{iid}{\sim} P \\ P \sim \mathcal{D}_\alpha, \quad \alpha = a \cdot \alpha_0 \end{cases}$$

$\theta_1,\dots,\theta_n$ may show ties $\Longrightarrow$ we obtain a latent partition ρ_n in $\theta_1,\dots,\theta_n$. $\rho_n = \{A_1,\dots,A_k\}$ is a latent partition of $\{1,\dots,n\}$, where each A_j contains all the data indexes of the group j. The data are grouped as:

$$X_\ell \sim X_m \text{ (same group)} \iff \theta_\ell = \theta_m$$

Thanks to the ties we have a prior induced by the model. As clustering of the data we choose a summary of the posterior distribution of ρ_n, for instance the value of the random partition minimizing a posterior the loss function $\ell(\cdot,\cdot)$:

$$\rho_n^* = \arg\min_{\hat{\rho}_n \in \text{ space of partitions}} \mathbb{E}\left[\ell(\rho, \hat{\rho}_n) \mid X_1,\dots,X_n\right],$$

ρ_n^*: optimal partition when ρ = true partition

$\ell(\cdot,\cdot)$ = how much we're losing describing one part. in terms of the other (distance btw. partitions)

(23.) Binder's loss function.

It penalizes in different ways the partition different from the optimal (or true) one

Let's introduce $S_1,\dots,S_n$; $S_i = j$ if $X_i \in$ j-th group. (in this way $A_j = \{i : S_i = j\}$).

The Binder loss function penalizes with c_1 if two elements belong to the same group in the true partition but not in the considered one, with c_2 the opposite:

$$\ell_B(\rho, \hat{\rho}_n) = \sum_{i,j} \left[c_1 \mathbb{1}\{S_i = S_j\} \mathbb{1}\{\hat{S}_i \neq \hat{S}_j\} + c_2 \mathbb{1}\{S_i \neq S_j\} \mathbb{1}\{\hat{S}_i = \hat{S}_j\} \right].$$

It can be shown that the solution (opt.) is given by the partition which, in binary representation, minimizes the distance from the posterior similarity matrix M (where $M_{ij} = P(S_i = S_j \mid X_1,\dots,X_n)$) and S^*:

$$S^* = \arg\min_{\hat{S}} \left[\sum_{i,j} \left| \mathbb{1}\{\hat{S}_i = \hat{S}_j\} - M_{ij} \right| \right]$$

Ex. $\mathcal{B} = \{\{1,2,3\}, \{1,2\},3\}, \{1,2,3\}\}, \{\{1,3\},2\}, \{\{1\},2,3\}\}$

Binary rep. $A_1 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $A_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $A_3 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, $A_4 = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$, $A_5 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

and $M = \begin{bmatrix} 1 & 0.8 & 1 \\ 0.8 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{bmatrix}$ → posterior prob. that 1 and 2 are grouped together

$d(A_1, M) = 0.8 + 0.2 + 0.2 = 1.2$
$d(A_2, M) = 0.6$
$d(A_3, M) = 1.8$
$d(A_4, M) = 1.8$
$d(A_5, M) = 2.4$

$\Longrightarrow$ optimal = A_2