Bayesian Method

Max Chen

Abstract

This is the summary notes of *Bayesian Method* by Yuling Max Chen, based on the lecture materials of Prof Geoff Nicholls. Most of the contents are directly from the the lecture notes and slides of *Bayesian Method*, although some reordering and rearrangements are made in sake of helping the readers to understand the materials. Some gaps (e.g. proofs and derivations) are filled to the original materials, based on the handwritten annotations during the lectures. Some personal ideas are also added, hence may not be 100% theoretically rigorous but should be helpful for the comprehension of the materials.
No person or party should use this notes for any purpose other than studying and understanding the notes itself.

Contents

1 Bayesian Inference Pipeline

1.1 Measure Theory and Bayes Rule

Frequentist : EDA $\implies$ data modeling $\implies$ parameter estimation (MLE) $\implies$ model selection (Likelihood Ratio Tests) $\implies$ goodness of fit checking $\implies$ reporting.
Bayesian : prior elicitation $\implies$ EDA $\implies$ data modeling $\implies$ parameter estimation (MLE) $\implies$ model selection (posterior mean and Bayes Factor) $\implies$ goodness of fit checking $\implies$ reporting.

Measure theory notation: Consider $\Omega \in \mathbb{R}^p$, $\mathcal{B}_\Omega$ (the Borel $\sigma-$algebra of subsets of Ω), $d\pi(\theta)$ (the general probability measure on Ω), $d\theta$ (the Lebesque volume measure in Ω), then:
(i) $d\pi(\theta) = \pi(d\theta) = \pi(\theta)d\theta$;
(ii) If $A \in \mathcal{B}_\Omega$, then $\pi(A) = \int_A \pi(d\theta)$ is a probability, i.e. **prior** $\pi : \mathcal{B}_\Omega \to [0,1]$.

Bayes Rule: $\pi(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{p(y)}$
where $p(y) = \int_\Omega p(y|\theta)\pi(\theta)d\theta$ is the **normalizing marginal likelihood (prior predictive distribution)** of the data.

Posterior distribution: $\pi(S|y) = Pr(\Theta \in S|Y = y) = \int_S \pi(\theta|y)d\theta$.

1.2 Prior Elicitation Checklist

1. Is the parameter θ generated by some process we can model? If so then the the distribution over θ determined by the process *is* the prior.
2. "Elements of reality": if the parameters correspond to real world quantities, it will be easier to identify prior knowledge. If introducing these parameters as latent variables, may make modelling easier.
3. Physically interpretable function $f(\theta)$ of the parameter: The distribution of $f(\theta)$ is determined by the prior so the prior is constrained to realise a priori plausible f -values.
4. Reliability: downweight unreliable priors, to ensure that carelessly imposed prior structure doesn't overwhelm data information for parameters which are poorly informed by the data.
5. Construct a prior which is non-informative with respect to the Scientific hypothesis/parameter.
• For example if we have a parameter $\theta \in [0,1]$ and we are interested in whether it is greater than 0.99 then the uniform prior $\theta \sim U(0,1)$ is strongly informative. If we are using the posterior as a summary then it will reflect this information. Non-informative does not in general equal uniform.
6 . The number of unknowns is unknown, put a prior on the number of thing we don't know.
7. The prior density models the prior knowledge. Once elicited, simulate the prior, and check the realised samples and physically meaningful functions of the samples are distributed as intended.
8. Check results are insensitive to a range of priors representing different states of knowledge. We are asking what conclusions another analyst would reach if they started with a different state of knowledge.

1.3 Bayes risk and Bayes Rule

For observation model $Y \sim p(\cdot,\theta), Y \in \mathcal{Y}$ and the $\Theta-$estimator $\delta : \mathcal{Y} \to \mathbb{R}^p$,
Risk: $\mathbb{R}(\theta,\delta) = E_{Y|\Theta=\theta}(L(\theta,\delta(Y))) = \int_\mathcal{Y} L(\theta,\delta(y))p(y \mid \theta)dy$
Expected Posterior Loss: $\rho(\pi,\delta \mid y) = E_{\Theta|Y=y}(L(\Theta,\delta(y))) = \int_\Omega L(\theta,\delta)\pi(\theta \mid y)d\theta$

Bayes risk: $\rho(\pi,\delta)=E_{\Theta,Y}(L(\Theta,\delta(Y)))=\int_\Omega\int_\mathcal{Y}L(\theta,\delta(y))p(y|\theta)\pi(\theta)dyd\theta=\int_\mathcal{Y}\rho(\pi,\delta(y)|y)p(y)dy$.
Bayes Rule: $\delta^\pi=\arg\min_\delta\rho(\pi,\delta)=\arg\min_\delta\rho(\pi,\delta|y)$

1.4 Admissibility

Inadmissible: δ_0 is inadmissible if $\exists\delta_1:\mathbb{R}(\theta,\delta_1)\leq\mathbb{R}(\theta,\delta_0)$ and $\exists$ at least one $\theta_0:\mathbb{R}(\theta_0,\delta_1)<\mathbb{R}(\theta_0,\delta_0)$.
• Admissible is the negation of inadmissible.
• Every admissible estimator is either a Bayes estimator or can be expressed as the limit of Bayes estimators.

Prop 1.1: If prior π is strictly positive on Ω with finite Bayes risk, and the risk $\mathbb{R}(\theta,\delta)$ is a continuous function of θ, then Bayes estimator δ^π is admissible.
Proof: By definition, Bayes estimator $\delta^\pi=\arg\min_\delta\rho(\pi,\delta)$.
$\implies\rho(\pi,\delta^\pi)\leq\rho(\pi,\delta),\forall\delta\overset{E[\cdot]}{\implies}\int_\Omega\mathbb{R}\left(\theta,\delta^\pi\right)\pi(\theta)d\theta>\int_\Omega\mathbb{R}\left(\theta,\delta\right)\pi(\theta)d\theta,\forall\delta$
$\implies\nexists\quad\delta':\mathbb{R}(\theta,\delta')<\mathbb{R}(\theta,\delta^\pi)$ [EOP]

1.5 Estimate the Posterior Expectation for f

$\hat{E}_{\Theta|Y=y}[f(\Theta)]=\hat{f}=\frac{1}{T}\sum_{t=1}^T f\left(\theta^{(t)}\right)$, where $\theta^{(t)}\sim\pi(\cdot|y)$
• If $S\in\mathcal{B}_\Omega$ and $f(\theta)=1_{\theta\in S}$,then $\hat{f}$ esimtiates $\pi(S|y)$.

Level α Highest Posterior Density (HPD) (C_α): $\int_{\Omega\cap C_\alpha}\pi(\theta\mid y)d\theta=1-\alpha$,
s.t $\theta\in C_\alpha$ and $\theta'\in\Omega\backslash C_\alpha\Rightarrow\pi(\theta\mid y)\geq\pi\left(\theta'\mid y\right)$.

Posterior Predictive Distribution: $p\left(y'\mid y\right)=\int_\Omega p\left(y'\mid\theta\right)\pi(\theta\mid y)d\theta$
• For model comparison and GOF: simulated data $y'\sim p(\cdot|y)$ should resemble the real data y, in a way that the summary computed on the real data lies in the tail of the posterior predictive distribution.

1.6 Model Selection

Introduce a new parameter $m\in\mathcal{M}$ as the model index. Then the parameter prior is $\Theta\sim\pi(\theta|m)$ and the observation model is $Y\sim p(y|\theta,m)$.
Hence the **posterior under model** m: $\pi(\theta\mid y,m)=\frac{p(y|\theta,m)\pi(\theta|m)}{p(y|m)}$, with the **marginal likelihood under the model** m: $p(y|m)=\int_{\Omega_m}p(y|\theta,m)\pi(\theta|m)d\theta$.
At the model level, the **posterior model probability** is: $\pi(m|y)=\frac{p(y|m)\pi_M(m)}{p(y)}$, where π_M is the prior probability that m is the correct model, i.e. the **priori preference** of the model; and $p(y)=\sum_{m\in\mathcal{M}}p(y|m)\pi_M(m)$ is the **marginal likelihood averaged over models**.
• Model selection and the 0-1 loss make sense when the number of models is small.

1.6.1 3 Odds

$$A_{m,m'}=\frac{\pi(m|y)}{\pi(m'|y)}=\frac{p(y|m)}{p(y|m')}\frac{\pi(m)}{\pi(m')}=B_{m,m'}C_{m,m'}$$

where $A_{m,m'}$ is the **posterior odds**, $B_{m,m'}$ is the **Bayes factor** and $C_{m,m'}$ is the **prior odds**.
• Favor m over m' if $A_{m,m'}>1$, i.e. model m is $A_{m,m'}$ times more likely a posterior than model

m'.
- The Bayes factor measures the relative support for the whole generative model coming from the data, i.e. how good the models are at predicting the data.
- higher model complexity $\implies$ probability mass put on the support of the likelihood decreases $\implies p(y|m) = E_{\Theta|M=m}[p(y|\Theta, m)] \downarrow$

1.6.2 Multiple Model Testing

Consider M_0 as the baseline and K alternatives M_k.
Bonferroni correction (Frequentist approach): Take $\alpha_k = \frac{\alpha}{K} \implies \Pr\left(P_{\min} < \tilde{\alpha}\right) = 1 - (1 - \tilde{\alpha})^K \simeq \alpha$.
Bayesian approach: consider there are K genes, hypothesis is that the disease is associated with either no gene (m=0) or the k-th gene (m=k). Set $\pi_M(0) = 1/2, \pi_M(k) = 1/2K$, then the posterior odds gives: $\frac{\pi(M=k|y)}{\pi(M=0|y)} = \frac{p(y|M=k)\pi_M(k)}{p(y|M=0)\pi_M(0)} = B_{k,0}/K$
- Simply looking for the largest $BF_{k,0}$ is problematic as:
(i) there might be equivalent BFs, and;
(ii) BF sets an uniform prior over the models meaning that the prior probability that one gene is associated to the disease is $K/K+1$ (which is supposed to be $1/2$).

1.7 Case Study: Radio Carbon Dating

Observation model: uncalibrated radiocarbon age y_i consists of the unknown true age $\theta_i \in [L, U]$ and the noise ϵ_i: $y_i = \mu(\theta_i) + \epsilon_i$, with $\epsilon_i \sim N(0, \sigma_c^2(\theta_i) + \sigma_i^2)$, where:
- σ_i^2 the measurement error;
- $\sigma_c^2(\theta_i)$ the standard deviation in the calibration map μ.
The **likelihood** is (under cond. indep.): $p(y|\theta) = \prod_{i=1}^n p\left(y_i \mid \theta_i\right) = \prod_{i=1}^n \frac{\exp\left(-(y_i-\mu(\theta_i))^2/2\left(\sigma_c(\theta_i)^2+\sigma_i^2\right)\right)}{\sqrt{2\pi\left(\sigma_c(\theta_i)^2+\sigma_i^2\right)}}$

Prior: consider 2 priors,
(i) A *uniform prior* over θ: $\pi_u(\theta) = (U-L)^{-n} \prod_{i=1}^n \mathbb{I}\left(L \leq \theta_i \leq U\right)$;
Hence the distribution of the span $s_u = \theta^+ - \theta^-$: $\pi_{S_u}\left(s_u\right) = \frac{n(n-1)}{(U-L)^n} s_u^{n-2}\left(U - L - s_u\right) \quad$ for $0 \leq s_u \leq U - L$
<u>proof</u>: The joint distribution of $\theta^- = \min(\theta)$ and $\theta^+ = \max(\theta)$ is $\pi_{u,\pm}\left(\theta^-, \theta^+\right) = \frac{n(n-1)}{(U-L)^n}\left(\theta^+ - \theta^-\right)^{n-2}$.
Because for $\theta = (\theta^-, \tilde{\theta}, \theta^+)$, $\underbrace{\pi_u(\theta)}_{\propto (U-L)^{-n}} = \underbrace{\pi_u(\tilde{\theta}|\theta^-, \theta^+)}_{\propto(\theta^+-\theta^-)^{-(n-2)}} \pi_u(\theta^-, \theta^+)$.
$\implies \pi_u(\theta^-, \theta^+) \propto \frac{(\theta^+-\theta^-)^{n-2}}{(U-L)^n}$ and there are $n(n-1)$ choices of (θ^-, θ^+).
Then, change of variable: $(\theta^-, \theta^+) \rightarrow (\theta^-, S_u) \implies \pi_u(s_u, \theta^-) = \pi(\theta^-, \theta^+) \propto s_u^{n-2}$
$\implies \pi_u(S_u) = \int_L^{u-s_u} \pi_u(s_u, \theta^-) d\theta^- \propto s_u^{n-2}(U - s_u - L)$.

(ii) A *Shrinkage prior*: Consider ψ_1, ψ_2 s.t. $L < \psi_1 < \psi_2 < U$ and $S_s = \psi_2 - \psi_1$. Assume dates θ are realisations of a Poisson process with rate λ over the interval $[\psi_1, \psi_2]$.
prior for θ: $\theta_{1:n} \sim Pois(\lambda|N = n) \implies \pi_s(\theta|\psi) \propto \frac{1}{(\psi_2-\psi_1)^n} 1_{(\psi_1<\theta_1,\ldots,\theta_n<\psi_2)}$
prior for ψ: $S_s = \psi_2 - \psi_1 \sim Unif(0, U - L) \implies \pi_s(\psi) \propto \frac{1}{(U-L)-(\psi_2-\psi_1)}$
Then, change of variable $(\psi_1, \psi_2) \rightarrow (\psi_1, S_s) \implies \pi_s(S_s) = \int_L^{U-S_s} \pi_s(\psi_1, S_s) d\psi_1 \propto 1_{S_s \in [0, U-L]}$
Joint prior: $\pi_s(\psi, \theta) = \pi_s(\psi)\pi_s(\theta|\psi) \propto \frac{1}{(\psi_2-\psi_1)^n} \frac{1}{(U-L-(\psi_2-\psi_1))}$

• The density of span in the shrinkage prior is uniform, hence desired.

Posterior:
$\pi_u(\theta \mid y) \propto p(y \mid \theta)\pi_u(\theta), \quad \pi_s(\theta, \psi \mid y) \propto p(y \mid \theta)\pi_s(\theta, \psi)$

Model comparison:
$B_{s,u} = \frac{p_s(y)}{p_u(y)} = \frac{\int_{\Omega_s} p(y|\theta)\pi_s(\theta,\psi)d\psi d\theta}{\int_{\Omega_u} p(y|\theta)\pi_u(\theta)d\theta}$, where:
- $\Omega_u = \{\theta \in [L,U]^n\}$
- $\Omega_s = \left\{(\theta, \psi) \in [L,U]^{n+2} : \psi_1 < \theta_i < \psi_2, i = 1, \ldots, n\right\}$

2 Markov Chain Monte Carlo Methods (MCMC)

2.1 MCMC

2.1.1 Irreducibility, Aperiodicity, Stationarity, Reversibility, Ergodicity

Markov Chain: $\{X_t\}_{t=0}^{\infty}$ a homogeneous Markov chain of random variables on Ω, with the *starting distribution* $X_0 \sim p^{(0)}$ and *n-step transition probability*: $P_{i,j}^{(n)} = \mathbb{P}\left(X_{t+n} = j \mid X_t = i\right)$.
• The Transition matrix P is:
(i) **irreducible** $\iff \forall i, j \in \Omega, \exists n : P_{i,j}^{(n)} > 0$;
(ii) **aperiodic** if $P_{i,j}^{(n)} \neq 0, \forall n$ sufficiently large.
• The target distribution of a MCMC in Bayesian inference is the posterior $p(\theta) = \pi(\theta|y)$.

Stationary Distribution: if $p^{(0)} = p$, then $p_j^{(1)} = \sum_{i \in \Omega} p_i^{(0)} P_{i,j} = p_j$, i.e. $pP = p$.

Detailed balance: $p_i P_{i,j} = p_j P_{j,i}, \forall i, j \in \Omega$.
• **Reversiblily** $\iff$ DB $\overset{\text{sufficient}}{\Longrightarrow}$ **Stationarity**.

Ergodic Theorem: If $\{X_t\}_{t=0}^{\infty}$ is an MCMC that is irreducible, aperiodic, and DB (wrt p). Then $\hat{f}_T = \frac{1}{T}\sum_t f(X_t) \overset{a.s.}{\to} E(f(X)), \forall f : \Omega \to \mathbb{R}$ (bounded).
• Then such MC is **ergodic** wrt target distribution p.
• CLT holds here, hence the CI: $\hat{f}_n \pm \sqrt{Var(\hat{f}_n)}$.

2.1.2 The Metropolis-Hastings (MH) Algorithm

Algorithm 1: Metropolis-Hastings (MH) Algorithm

```
Initialize proposal probability distribution $q(j|i) = Q_{i,j}$ s.t. $q(j|i) > 0 \iff q(i|j) > 0$.
Initialize the starting state $X_0 = i_0, p_{i0} > 0$.
for $t = 1, ..., T$ do
    Draw $j \sim q(\cdot|i)$ and $u \sim U[0,1]$
    if $u \leq \alpha(j|i) = \min\left\{1, \frac{p_j q(i|j)}{p_i q(j|i)}\right\}$ = acceptance probability, then
        $X_{t+1} = j$
    end
    else
        $X_{t+1} = X_t$
    end
end
Return $X_{1:T}$ as the sample from the targeting distribution $p$.
```

.

Lemma 2.3: If the MC from MH-algorithm is irreducible and aperiodic then it is ergodic with target p.
Proof: Need to show irreducibility, aperiodicity and DB.
To show irreducibility and aperiodicity, compute the transition probability: If $X_t = i$, then $P_{i,j}$ is the probability to propose j at step t times the probability to accept it at step $t+1$, i.e.

$$P_{i,j} = P\left(X_{t+1} = j \mid X_t = i\right) = q(j \mid i)\alpha(j \mid i) \geq q(j \mid i) > 0$$

To show DB:

$$\begin{aligned} p_i P_{i,j} &= p_i q(j \mid i)\alpha(j \mid i) \\ &= p_i q(j \mid i) \min\left\{1, \frac{p_j q(i \mid j)}{p_i q(j \mid i)}\right\} \\ &= \min\left\{p_i q(j \mid i), p_j q(i \mid j)\right\} \\ &= p_j q(i \mid j) \min\left\{\frac{p_i q(j \mid i)}{p_j q(i \mid j)}, 1\Big)\right\} \\ &= p_j q(i \mid j)\alpha(i \mid j) \\ &= p_j P_{j,i} \quad \text{[EOP]} \end{aligned}$$

Equal Mixture of Bivariate Normals: MH-MCMC targeting the density:

$$\pi(\theta) = (2\pi)^{-1}\left(0.5e^{-(\theta-\mu_1)\Sigma_1^{-1}(\theta-\mu_1)/2} + 0.5e^{-(\theta-\mu_2)\Sigma_2^{-1}(\theta-\mu_2)/2}\right), \quad \theta = (\theta_1, \theta_2)$$

Initialization: proposal distribution is $\theta_i' \sim U(\theta_i - a, \theta_i + a)$ for some constant *jump size* $a > 0$, hence $q(\theta'|\theta) = q(\theta|\theta') = 1/4a^2$ (as we jump uniformly within a box of side $2a$).
Sampling: θ_1' and θ_2' are sampled from proposal independently, then form θ'.
Acceptance: $\alpha(\theta'|\theta) = \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\right\}$, as the proposals are cancelled out.
• Small jump size a will cause the chain can't move easily between modes through the saddle, and lead to small acceptance probability (as a path between modes must include a pair of state $\theta^{(t)}, \theta^{(t+1)}$ with $\pi^{(t)} << \pi^{(t+1)}$). Whereas large jump size a will make the sampler move cross the modes easily, and will also lead to small proposals in the tails of the density due to large jumps.

Mixing Updates for Multivariate Targets: If $\theta = (\theta_1, \cdots, \theta_p)$, then can set/fix $\theta'_{-i} = \theta_{-i}$ and only update $\theta'_i \sim q_i(\cdot|\theta)$ at each step, where we randomly chose to update $q_i(\cdot|\theta)$ with probability ξ_i.
- The overall transition matrix: $P(\theta, \theta') = \sum_i \xi_i P_i(\theta, \theta') = \sum_i \xi_i q_i(\theta'_i|\theta)\alpha(\theta'|\theta)$.
- $q_i(\theta'|\theta)$ reversible wrt $\pi(\theta) \implies P_i(\theta, \theta')$ reversible $\implies P(\theta, \theta')$ reversible (as summation preserves reversibility).

2.2 Output Analysis

Initialization bias: MCMC sample is biased if the initialization is not sampled from the target.

2.2.1 Convergence and Mixing

Namely *bias* and *variance* of the MCMC-sample. To deal with:
(i) Bias: large number of MCMC-samples (i.e. large T and long chain) and **burn-in** (cut-off at the beginning);
(ii) Variance: large number of MCMC-samples and check ACF.

2.2.2 MCMC variance $var(\bar{f}_n)$ in equilibrium

Effective Sample Size (ESS): The number of independent samples which would give the same variance reduction as our n correlated samples, $ESS = \frac{var(f(X))}{var(\bar{f}_n)}$.
- If MCMC is indenpendent, then $ESS = n$; most commonly, $ESS << n$.

We consider 2 approaches to estimate $var(\bar{f}_n)$:
(i) The simple K-runs approach: make K MCMC-samples of the same length n, $\left(\theta^{(k,t)}\right)_{t=1,k=1}^{t=T,k=K}$. Then, (independent across K runs while dependent among T steps within each run)

$$var(\bar{f}_T) \simeq \hat{\sigma}^2_{\bar{f},T} = \frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{f}_{k,n} - K^{-1}\sum_{j=1}^{K}\bar{f}_{j,T}\right)^2, \quad \bar{f}_{k,T} = \frac{1}{T}\sum_t f\left(\theta^{(k,t)}\right)$$

- $ESS \simeq \hat{\sigma}^2_f / \hat{\sigma}^2_{\bar{f},T}$, a measure of the precision gain afforded by our n correlated samples.

(ii) **Binning**: use a single long run instead of K runs, assuming each block within the long-run is independent of each other as they are far apart.

$$\begin{aligned}
\operatorname{var}\left(\bar{f}_n\right) &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}\left(f\left(X_i\right), f\left(X_j\right)\right) \\
&= \sigma^2 n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_{|i-j|}, \quad \rho_s = \frac{\operatorname{cov}\left(f\left(X_t\right), f\left(X_{t+s}\right)\right)}{\operatorname{var}\left(f\left(X_t\right)\right)}, \text{ the correlation at lag } s \\
&= \sigma^2 n^{-1}\left[1 + 2\sum_{s=1}^{n-1}\left(1 - \frac{s}{n}\right)\rho_s\right] \\
&\simeq \sigma^2 n^{-1}\left[1 + 2\sum_{s=1}^{n-1}\rho_s\right], \quad \because n >> s \\
&= \sigma^2\frac{\tau_f}{n}, \text{where } \tau_f \text{ is the } \textbf{Integrated Autocorrelation Time (IACT)}
\end{aligned}$$

Where $\hat{\tau}_f = 1 + 2\sum_{s=1}^{M} \hat{\rho}_s = 1 + 2\sum_{s=1}^{M} \frac{\hat{\gamma}_s}{\hat{\gamma}_0} = 1 + 2\sum_{s=1}^{M} \frac{n^{-1}\sum_{i=1}^{n-s}\big(f(X_i)-\hat{f}\big)\big(f(X_{i+s})-\hat{f}\big)}{var(f(X))}$,
and M a cut-off as $\hat{\rho}_s \overset{s\uparrow\infty}{\to} 0$ and is dominated by estimation noise at large s. (Choose M to be the least t s.t. $\rho_t + \rho_{t+1} > 0$ and $\hat{\rho}_t + \hat{\rho_{t+1}} > \hat{\rho}_{t-1} + \hat{\rho}_t$)
•$ESS \simeq {}^{n}/_{\tau_f}$

2.2.3 MCMC Convergence

No sufficient conditions, but can check the necessary conditions:
(i) Make multiple runs of different initializations and check marginal distributions agree;
(ii) Plot ACF and check it falls off to vary around 0;
(iii) Compute ESS and check it reasonably large (¿ 100 good, ¿ 1000 very sound);
(iv) Plot MCMC traces of the variables and the key functions, and check they are stationary after burn-in.

2.3 Gibbs Samplers and Data Augmentation

2.3.1 Gibbs Samplers

Random scan Gibbs: a multi-component Metropolis Hastings sampler that,
(i) selects components $i = 1, \cdots, p$ with probability $\xi_i = {}^{1}/_{p}$, and;
(ii) takes as proposal the conditional density: $q_i(\theta_i|\theta) = \pi(\theta_i|\theta_{-i}) = \frac{\pi(\theta_i)}{\pi(\theta_{-i})}$, hence;
(iii) acceptance probability = 1.

Sequential scan Gibbs: update $q_i(\theta_i|\theta)$ from $i = 1, \cdots, p$ sequentially rather than a random selection.

Prop 2.6: For sequential-scan Gibbs with $\theta \in \mathbb{R}^p$, the process is stationary wrt π after p steps of updates.
Proof: WLOG, consider $p = 2$ and hence $\theta = (\theta_1, \theta_2)$.

$$\begin{aligned} p\left(\theta_1', \theta_2'\right) &= \int \pi\left(\theta_1, \theta_2\right) q_1\left(\theta_1' \mid \theta_2\right) q_2\left(\theta_2' \mid \theta_1'\right) d\theta_1 d\theta_2 \\ &= \int \pi\left(\theta_1, \theta_2\right) \frac{\pi\left(\theta_1', \theta_2\right)}{\pi\left(\theta_2\right)} \frac{\pi\left(\theta_1', \theta_2'\right)}{\pi\left(\theta_1'\right)} d\theta_1 d\theta_2 \\ &= \int \pi\left(\theta_1 \mid \theta_2\right) \pi\left(\theta_2 \mid \theta_1'\right) \pi\left(\theta_1', \theta_2'\right) d\theta_1 d\theta_2 \\ &= \pi\left(\theta_1', \theta_2'\right) \quad \text{[EOP]} \end{aligned}$$

2.3.2 Data Augmentation (DA)

Treat missing data z as another set of parameters like θ, with prior being the observation model $z \sim p(z|\theta)$, then work with the joint posterior density of both the missing data and the parameter: $p(\theta, z|y) \propto p(y|z,\theta)p(z|\theta)p(\theta)$.
• Avoids the integration in the single posterior of the parameters: $\pi(\theta|y) \propto \pi(\theta)\int p(y|z,\theta)p(z|\theta)dz$.

2.3.3 A Gibbs sampler (with DA) for Probit regression

Observation model: $y_i \sim Bern(\Phi(\eta_i(\theta))), \theta = (\theta_1, \cdots, \theta_p)$, with the *Inverse link function* Φ is the cdf of standard Gaussian.

Posterior: $\pi(\theta \mid y) \propto \pi(\theta) \prod_{i=1}^{n} \Phi\left(\eta_i(\theta)\right)^{y_i}\left(1-\Phi\left(\eta_i(\theta)\right)\right)^{1-y_i}$, with $\pi(\theta)$ the *prior.*

Gibbs sampler is prohibited as the conditionals $\pi(\theta_i|\theta_{-i}, y)$ is unavailable since θ is inside Φ. But we can introduce a latent parameter $z_i \sim \pi(z_i|\theta_i) = N(\eta_i(\theta_i), 1), i = 1, ..., n,$ and Gibbs sample targeting $\pi(\theta, z|y)$.

Algorithm 2: A Gibbs sampler (with DA) for Probit regression

```
Initialize $X_0 = (\theta^{(0)}, z^{(0)})$.
Initialize prior $\pi(\theta)$.
for $t = 1, \cdots, T$ do
    for $i = 1, \cdots, n$ do
        (i) $z_i^{(t+1)} \sim \pi(z_i^{(t)}|y, \theta_i) \propto N(\eta_i, 1)\mathbb{I}_{y_i=\mathbb{I}_{z_i^{(t+1)}>0}}$
    end
    for $j = 1, \cdots, p$ do
        (ii) $\theta_j^{(t+1)} \sim \pi(\theta^{(t)}|y, z^{(t+1)}) \propto \pi(z^{(t+1)}|\theta^{(t)})\pi(\theta^{(t)}) \propto \pi(\theta^{(t)}|z^{(t+1)})$
    end
end
$X_{(t+1)} = (\theta^{(t+1)}, z^{(t+1)})$
```

(i) For the z-update:
$z_i = \eta_i + \epsilon_i$ ($\epsilon_i \sim N(0, 1)$) $\implies P(z > 0) = P(\epsilon > -\eta) = \Phi(\eta)$ by symmetry.
Setting $y_i = \begin{cases} 1, z_i > 0 \\ 0, z_i \leq 0 \end{cases} \implies P(y_i = 1|\theta) = \Phi(\eta_i)$ and y_i is know for certain give z_i, i.e. $p(y_i|z_i) = \mathbb{I}_{y_i=\mathbb{I}_{z_i^{(t+1)}>0}}$.
The joint posterior augmented with z gives:
$\pi(\theta, z \mid y) \propto p(y \mid z)\pi(z \mid \theta)\pi(\theta) = \pi(z \mid \theta)\pi(\theta) \prod_i \mathbb{I}_{y_i=\mathbb{I}_{z_i}>0} \implies \pi(z|\theta, y) \propto p(y|z)\pi(z|\theta)$

(ii) For the $\theta-$update: $\pi(\theta|y, z) \propto \pi(z|\theta)\pi(\theta)$ because y is certain given z.

2.4 Estimation of Marginal Likelihoods

The posterior: $\pi(\theta|y, m) = \frac{\pi(\theta|m)p(y|\theta,m)}{p(y|m)}$. Want to estimate $\hat{p}(m) = \hat{p}(y|m)$.

(i) **The Naive Estimate**: Simulate $\theta^{(t)} \sim \pi(\theta|m)$ from prior, and average the likelihood

$$\hat{p}_m = \frac{1}{T} \sum_t p(y|\theta^{(t)}, m)$$

• The prior is diffuse over the parameter space while likelihood is small except on a small set of θ. Hence a naive simulation from the prior will only hit this set with a small proportion.

(ii) **The Harmonic Mean Estimate**: importance sampling targeting the posterior.

$$\hat{p}_m = \left[\frac{1}{T} \sum_t \frac{1}{p\left(y \mid \theta^{(t)}, m\right)}\right]^{-1}$$

Derivation: Simulate $\theta^{(t)} \sim \pi(\theta|y,m)$ from posterior.
Consider $w_t = \frac{\pi\left(\theta^{(t)}|m\right)}{\pi\left(\theta^{(t)}|y,m\right)} \implies \hat{p}'_m = \frac{1}{T}\sum_t w_t p\left(y|\theta^{(t)}, m\right)$, a consistent and unbiased estimator for $p(y|m)$ via importance sampling.
$\because\ E_{\theta^{(1:T)}|y,m}\left(\hat{p}'_m\right) = T^{-1}\sum_t \int_\Omega w_t p\left(y \mid \theta^{(t)}, m\right)\pi\left(\theta^{(t)} \mid y, m\right)d\theta^{(t)} = \int_\Omega p(y \mid \theta, m)\pi(\theta|m)d\theta = p(y|m)$
But $\pi\left(\theta^{(t)} \mid y, m\right)$ requires the marginal likelihood, which is unknown as this is what we want to estimate.
Consider instead $\tilde{w}_t = \frac{1}{p(y|\theta^{(t)},m)} \propto w_t$, and hence:
$\implies E_{\theta^{(t)}|y,m}\left(\tilde{w}_t\right) = \int_\Omega \frac{\pi\left(\theta^{(t)}|y,m\right)}{p\left(y|\theta^{(t)},m\right)}d\theta^{(t)} = \int_\Omega \frac{\pi(\theta|m)}{p(y|m)}d\theta = p(y \mid m)^{-1}$

Bridge Estimate: Simulate $\{\theta^{(1,t)}\}_{t=1}^T \sim \pi(\theta)$ from prior and $\{\theta^{(2,t)}\}_{t=1}^T \sim \pi(\theta|y)$ from posterior.

$$\hat{p}(y) = \frac{\sum_t \pi\left(\theta^{(1,t)}\right) p\left(y \mid \theta^{(1,t)}\right) h\left(\theta^{(1,t)}\right)}{\sum_t \pi\left(\theta^{(2,t)}\right) h\left(\theta^{(2,t)}\right)}$$

- Choice of h: $h \propto \frac{1}{\sqrt{\tilde{p}_1\tilde{p}_2}}$ is near optimal for bridging densities $\frac{\tilde{p}_1}{Z_1}$ and $\frac{\tilde{p}_2}{Z_2}$. In this case, $\tilde{p}_1 = \pi(\theta)$ and $\tilde{p}_2 = \pi(\theta)p(y|\theta)$, which gives: $h(\theta) = \pi(\theta)^{-1}p(y|\theta)^{-1/2} \implies \hat{p} = \frac{\sum_t p\left(y|\theta^{(1,t)}\right)^{1/2}}{\sum_t p\left(y|\theta^{(2,t)}\right)^{-1/2}}$
- This leads to lower **Relative Mean Square Error (RMSE)**: $E\left[\frac{(\hat{p}-p(y))^2}{p(y)}\right]$.
- Bridge estimator is inspired by Prop 2.7.

Prop 2.7: Let $h : \Omega \to \mathbb{R}$ be s.t. $E[h(\theta)] < \infty$ and $E[h(\theta)] < \infty$, then the identities hold:
(i)

$$p(y|m) = \frac{E_{\theta\sim\pi(\theta|m)}(\pi(\theta|m)p(y \mid \theta, m)h(\theta|m))}{E_{\theta\sim\pi(\theta|y,m)}(\pi(\theta|m)h(\theta|m))}$$

(ii)

$$B_{m',m} = \frac{p\left(y \mid m'\right)}{p(y \mid m)} = \frac{E_{\theta|y,m'}(\pi(\theta \mid m)p(y \mid \theta, m)h(\theta))}{E_{\theta|y,m}\left(\pi\left(\theta \mid m'\right)p\left(y \mid \theta, m'\right)h(\theta)\right)}$$

- The Bayes factor under the Bridge estimate with $\{\theta^{(1,t)}\}_{t=1}^T \sim \pi(\theta|y, m = 1)$, $\{\theta^{(2,t)}\}_{t=1}^T \sim \pi(\theta|y, m = 2)$, and $h(\theta) = \left(\pi(\theta \mid m)p(y \mid \theta, m)\pi\left(\theta \mid m'\right)p\left(y \mid \theta, m'\right)\right)^{-1/2}$ gives:

$$\hat{B}_{m',m} = \frac{\sum_t \left(\frac{\pi\left(\theta^{(2,t)}|m\right)p\left(y|\theta^{(2,t)}|m\right)}{\pi\left(\theta^{(2,t)}|m'\right)p\left(y|\theta^{(2,t)}|m'\right)}\right)^{1/2}}{\sum_t \left(\frac{\pi\left(\theta^{(1,t)}|m'\right)p\left(y|\theta^{(1,t)}|m'\right)}{\pi\left(\theta^{(1,t)}|m\right)p\left(y|\theta^{(1,t)}|m\right)}\right)^{1/2}}$$

2.5 Using Simulation to Check a Prior

Omitted, see notes Section 2.4.2 and 2.4.3.

3 Savage Axioms

- If we have a collection of prior preferences expressed as "A is more likely than B", and those preferences satisfy the Savage Axioms, then there is a prior probability distribution π s.t. $\pi(A) > \pi(B)$.

3.1 Utility Theory

Utility $U(r) \in \mathbb{R}$ with **reward** $r \in \{r_{\min}, r_{\min+1}, \cdots, r_{\max}\}$ is the opposite of loss: $L(\theta, \delta) = c - U(r(\theta, \delta))$, with $\theta \in \mathbb{R}^p$ the parameter, δ the predictor (action), and c the largest attainable utility.

Reward Distribution: $P_\delta(r) = \int_\Omega \mathbb{I}_{r(\theta,\delta(y))=r} \pi(\theta \mid y) d\theta$, with $\sum_{r_{\min}}^{r_{\max}} P_\delta(r) = 1$.

Expected Utility: $E_{P_\delta}[U(R)] = \sum_{r_{\min}}^{r_{\max}} P_\delta U(r)$.

Prop 3.1: Expected utility has the opposite sign to the expected posterior loss, i.e.

$$E_{P_\delta}(U(R)) = c - \int_\Omega L(\theta, \delta) \pi(\theta \mid y) d\theta$$

(Hence choosing δ to maximize utility is equivalent to loss minimization.)
Proof:

$$\begin{aligned} LHS = \sum_{r=r_{\min}}^{r_{\max}} U(r) P_\delta(r) &= \sum_{r=r_{\min}}^{r_{\max}} U(r) \int_\Omega \mathbb{I}_{r(\theta,\delta)=r} \pi(\theta \mid y) d\theta \\ &= \sum_{r=r_{\min}}^{r_{\max}} \int_\Omega \mathbb{I}_{r(\theta,\delta)=r} U(r(\theta,\delta)) \pi(\theta \mid y) d\theta \\ &= c - \sum_{r=r_{min}}^{r_{\max}} \int_\Omega \mathbb{I}_{r(\theta,\delta)=r} L(\theta,\delta) \pi(\theta \mid y) d\theta \\ &= c - \int_{\Omega'} L(\theta,\delta) \pi(\theta \mid y) d\theta \end{aligned}$$

where $\Omega' = \bigcup_{r=r_{\min}}^{r_{\max}} \{\theta \in \Omega : r(\theta, \delta) = r\} = \{\theta \in \Omega : r_{\min} \le r(\theta, \delta) \le r_{\max}\} = \Omega$ [EOP].

Example: draw and predict ball colors from a single urn
A ball with color $\theta = \{\text{black, red}\}$ is draw uniformly from the urn.
Suppose:
(i) there are n balls in total and $\phi = P(\theta = \text{black})$ the proportion of black balls, with $\phi \sim \pi(\cdot)$ prior, and;
(ii) prediction of the color is $\delta = \{\text{black, red}\}$, and;
(iii) the reward is $r(\theta, \delta) = \mathbb{I}_{\theta=\delta}$, and;
(iv) the utility is $u(r) = \begin{cases} u > 0, r = 1 \\ 0, r = 0 \end{cases}$.
Then,
(a) If we predict the next ball is black, i.e. $\delta =$black, then
- the probability of the reward: $P_\delta(r = 1) = E_\phi\left(E\left(\mathbb{I}_{\theta = \text{ black }} \mid \phi\right)\right) = E(\phi)$, and;
- the expected utility of choosing black: $E_{P_\delta}(U(r(\theta, \text{ black }))) = P_\delta(0)U(0) + P_\delta(1)U(1) \; uE(\phi)$
(b) If instead predict red, $\delta' =$red, then:
- the expected utility of choosing red: $E_{P'_\delta}(U(r(\theta, \text{red}))) = uE(1 - \phi)$

$\implies$ optimal action is $\delta^* = \begin{cases} \text{black, if } E(\phi) \ge 1/2 \\ \text{red, if } E(\phi) \le 1/2 \end{cases}$

3.2 Coherence

3.2.1 Expected utility hypothesis

Ordering on reward distribution:
$P_\delta > P_{\delta'} \quad \Leftrightarrow \quad E_{P_\delta}(U(r)) > E_{P_{\delta'}}(U(r))$
$P_\delta \sim P_{\delta'} \quad \Leftrightarrow \quad E_{P_\delta}(U(r)) = E_{P_{\delta'}}(U(r))$

Example: Choose urn and draw a black ball
Suppose:
(i) there are 2 urns $\delta = \{1, 2\}$ (each containing both black and red balls), and;
(ii) the proportion of black balls in urn δ: $\phi_\delta \in [0,1]$, with $\phi_\delta \sim \pi_\delta(\phi_\delta)$ prior and;
(iii) positive reward is given to the black ball draw $r(\theta,\delta) = \mathbb{I}_{\theta_\delta = \text{black}}$, and;
(iv) utility is $u(r) = \begin{cases} u > 0, r = 1 \\ 0, r = 0 \end{cases}$.
Then,
(a) The reward distribution: $P_\delta = (P_\delta(0), P_\delta(1)) = (1 - E_{\pi_\delta}(\phi_\delta), E_{\pi_\delta}(\phi_\delta))$, and;
(b) The expected utilities: $E_{P_\delta}(U(R)) = uE_{\pi_\delta}(\phi_\delta)$.
$\implies \delta^* = \arg\max_{\delta=1,2} E_{\pi_\delta}(\phi_\delta) = \arg\max_{\delta=1,2} P_\delta(1)$, i.e. the one with the highest prior preference of choosing black.

3.2.2 Coherent inference

Choose the action that maximises the expected utility (possible if the utility function and reward distributions exist).

Prior exists (and unique) $\implies$ **coherent belief** and reward distribution exists $\implies$ exists a preference order over reward distributions $\implies$ **coherent inference** if $\exists U(r)$ satisfying the expected utility hypothesis.

Suppose:
(i) 2 sets $A, B \in \mathcal{B}_\sigma$ (a $\sigma-$field containing all the sets of interest) and $A, B \subseteq \Omega$, and;
(ii) action space $\delta \in \{A, B\}$ and reward $r(\theta,\delta) = \mathbb{I}_{\theta \in \delta}$, and;
(iii) the probability space $(\Omega, \mathcal{B}_\sigma, \pi)$ with prior π, and;
(iv) utility is $u(r) = \begin{cases} u > 0, r = 1 \\ 0, r = 0 \end{cases}$.
Then,
(a) The reward distribution: $P_\delta = (P_\delta(0), P_\delta(1)) = (1 - E_{\pi_\delta}(\mathbb{I}_{\theta=\delta}), E_{\pi_\delta}(\mathbb{I}_{\theta=\delta})) = (1 - \pi_\delta, \pi_\delta)$,
(b) The expected utility: $E_\theta(U(r(\theta,\delta))) = P_\delta(0)U(0) + P_\delta(1)U(1) \propto u\pi_\delta$.
$\implies \delta^* = \arg\max_{\delta \in \{A,B\}} u\pi_\delta$, choice of action is coherent to the prior preference.

3.2.3 The Ellsberg paradox

- *Preferences are inconsistent with any prior.*

Suppose:
(i) 2 urns: A contains half black half red, B contains unknown black and/or red, and;

(ii) rewards $r \in \{-1000, 1000\}$, and utility is $u(r) = \begin{cases} u > 0, r = 1000 \\ 0, r = -1000 \end{cases}$, and;
(iii) $\phi = P(b_B)$ proportion of black balls in urn B, with $\phi \sim \pi(\phi)$ prior, and;
(iv) 4 bets, where each option chooses an urn, predict the color and draw from that urn:

	Option 1 (Color$_{\text{urn}}$)	Option 2 (Color$_{\text{urn}}$)
Bet 1	r_A	b_A
Bet 2	r_B	b_B
Bet 3	r_A	b_B
Bet 4	r_B	b_A

Then,
(a) Bet 1 and Bet 2 are neutral, hence:
- $E(U|r_A) = E(U|b_A) = \frac{u}{2}$, and;
- $E(U|b_B) = uE_\pi(\phi) = E(U|r_B) = u(1 - E_\pi(\phi)) \implies E_\pi(\phi) = \frac{1}{2}$.
(b) Choose the certainty:
- Choose r_A in Bet 3 $\implies E_\pi(\phi) < \frac{1}{2}$;
- Choose b_A in Bet 4 $\implies E_\pi(\phi) > \frac{1}{2}$.
$\implies$ Clear contradiction!

3.2.4 The Allais paradox

• *Preferences are inconsistent with any utility function.*

Suppose:
(i) Probability of winning in round i: $p^{(i)} = (p_1, p_2, p_3)$, and;
(ii) Known reward: $r = (\$0, \$500,000, \$750,000)$ and utility function is $U(r) = (0, u, 1)$;
(iii) 2 lotteries:

	Option 1	Option 2	Choice
L1	$P^{(A)} = (0, 1, 0)$	$P^{(B)} = (0.01, 0.89, 0.1)$	A
L2	$P^{(C)} = (0.89, 0.11, 0)$	$P^{(D)} = (0.9, 0, 0.1)$	D

Then,
(a) Expected utility for the 4 options are:
$\implies$ Clear contradiction!

	Option 1	Option 2	Choice implication
L1	$E(U|A) = u$	$E(U|B) = 0.1 + 0.89u$	$u > 0.1 + 0.89u \implies u > \frac{10}{11}$
L2	$E(U|C) = 0.11u$	$E(U|D) = 0.1$	$0.11u < 0.1 \implies u < \frac{10}{11}$

3.3 Savage Axioms

3.3.1 Probability Space

A **Probability Space** is $(S, \mathcal{S}, \pi)$, where:
(i) S is a sample space;

(ii) $\mathcal{S}$ is a $\sigma-$field of sets in S s.t.:
- $S \in \mathcal{S}$,
- $A \in \mathcal{S} \implies A^C \in \mathcal{S}$,
- $A_1, A_2, \cdots \in \mathcal{S} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{S}$;

(iii) $\pi : \mathcal{S} \to [0,1]$ a probability distribution satisfying the axioms of probability:
- $\pi(A) \geq 0, \forall A \in \mathcal{S}$,
- $\pi(\mathcal{S}) = 1$,
- Disjoint $A_1, A_2, \cdots \in \mathcal{S} \implies \pi(\cup_i A_i) = \sum_i \pi(A_i)$ (countably additive).

Theorem 3.4: Given a system of preferences over $A, B \in \mathcal{S}$, a probability distribution $\pi : \mathcal{S} \to [0,1]$ exists and is unique iff the preference relation satisfies the Savage Axioms of Probability, i.e. Axiom 1-5.

3.3.2 Axioms of Probability

Axiom 1: $\forall A, B \in \mathcal{S}$, exactly one of the relations must hold: $A > B, A < B, A \sim B$.
• $A \geq B$ represents a class of sets satisfying exactly one of $A > B, A \sim B$.

Axiom 2: $A_1 \cap A_2 = B_1 \cap B_2 =$ and $A_i \geq B_i, i = 1, 2 \implies A_1 \cup A_2 \geq B_1 \cup B_2$.
• If either $A_1 > B_1$ or $A_2 > B_2$, then $A_1 \cup A_2 > B_1 \cup B_2$

Axiom 3: $A \in \mathcal{S} \implies \leq A$

Following Axiom 1-3, we have:
(i) order transitivity, i.e. $A \leq B, B \leq C \implies A \leq C$;
(ii) $A \leq B \implies A^C \geq B^C$.

Axiom 4: $A_1 \supset A_2 \supset \cdots$ (a decreasing sequence of events) s.t. $A_i \geq B, \forall i$ with B a fixed event $\implies \bigcap_{i=1}^{\infty} A_i \geq B$.

Axiom 5: $\forall p \in [0,1], \exists A_p \in \mathcal{S} : \pi(A_p) = p$.

3.3.3 Axioms of Utility

Theorem 3.8: There exists a utility function U which expresses our preference relations over $P \in \mathcal{P}$ iff our preference satisfies Savage Axioms of Utility, i.e. Axioms 6-10.
• Where $\mathcal{P}$ is the set of distribution P over rewards $r \in \{r_{\min}, r_{\min+1}, \cdots, r_{\max}\}$, s.t. $P([r_{\min}, r_{\max}]) = \Pr(r_{\min} \leq R \leq r_{\max}) = 1, \forall R \sim P$

Axiom 6: $\forall P, P' \in \mathcal{P}$ exactly one of the relations must hold: $P > P', P < P', P \sim P'$

Axiom 7: $P \geq P', P' \geq P'' \implies P \geq P''$

Axiom 8: $P' \geq P'' \iff \alpha P' + (1-\alpha)P \geq \alpha P'' + (1-\alpha)P, \forall \alpha \in (0,1), P \in \mathcal{P}$

Axiom 9: (omitted, see notes section 3.4)

Axiom 10: (omitted, see notes section 3.4)

4 Exchangeability

• Both Exchangeability and Savage Axioms give sufficient conditions for existence of prior. Exchangeability says "if the data are exchangeable then a prior exists", while Savage Axioms say "if your preferences are coherent then a prior exists".

4.1 Exchangeability in Finite Sequences

Exchangeability: For $X \in \mathcal{X}^n$, joint distribution is unchanged by permutation of the indices, $(X_1, \cdots, X_n) \sim (X_{\sigma_1}, \cdots, X_{\sigma_n})$, i.e. $p_{1:n}(x_1, \cdots, x_n) = p_{1:n}(x_{\sigma_1}, \cdots, x_{\sigma_n})$, $\forall$ permutation $\sigma \in \mathcal{P}_n$ and $\forall (x_1, \cdots, x_n) \in \mathcal{X}^n$.
• iid $\implies$ exchangeability (but not conversely).

Example: iid and exchangeability:
Consider hypergrometric distribution: $p_{1:n}(x_1, \ldots, x_n) = \left(\begin{array}{c} K \\ k(x) \end{array}\right)\left(\begin{array}{c} N-K \\ n-k(x) \end{array}\right) / \left(\begin{array}{c} N \\ n \end{array}\right)$, where $X_i \in \{0,1\}$, N the population size, K the population number of 1's, n is the draw sample size (without replacement) and k(x) is the number of 1's in the draw.
(i) Since $k(x)$ is the same for any order of x, probability the last 3 are 1's is the same as the probability that the first 3 are 1's, by exchangeability:
$P(X_{n-2} = 1, X_{n-1} = 1, X_n = 1) = p_{n-2:n}(1,1,1) = \frac{K(K-1)(K-2)}{N(N-1)(N-2)} = P_{1:3}(1,1,1)$
(ii) $P(X_2 = 1 | X_1 = 1) = \frac{K-1}{N-1} \neq P(X_2 = 1 | X_1 = 0) = \frac{K}{N-1} \implies$ dependence between X_1, X_2.

4.2 Infinite Exchangeable Sequence

Infinite Exchangeable Sequence (IES): infinite sequence of random variables s.t. $X_1, \cdots, X_n$ are exchangeable $\forall n \geq 1$.
• Any subset of IES is exchangeable.

Exchangeability in Hierarchical Model:
WTS: $\exists X_1, X_2, \cdots$ (IES) with marginals:

$$p_{1:n}(x) = p_{1:n}(x_1, \cdots, p_n) = N\left(x; 0_n, \Sigma^{(n)}\right)$$

where $\Sigma^{(n)}$ is an $n \times n$ covariance matrix with on-diagonal = 1 (unit variance) and off-diagonal = ρ (equal covariance).
Proof: Fix any n,
(i) check exchangeability: for any $\sigma \in \mathcal{P}_n$,

$$\begin{aligned} p_{1:n}(x_{\sigma_1}, \ldots, x_{\sigma_n}) &= N\left(x_\sigma; 0_n, \Sigma^{(n)}\right) \\ &= N\left(x; 0_n, \Sigma^{(n)}\right) \\ &= p_{1:n}(x_1, \ldots, x_n) \end{aligned}$$

(ii) Simulate $\theta \sim N(0,\rho)$ and set $X_i = \theta + \epsilon_i, \epsilon \overset{iid}{\sim} N(0, 1-\rho)$
$\implies E(X_i) = 0, Var(X_i) = 1, Cov(X_i, X_j) = \rho$, and:

$$p_{1:n}\left(x_1, \ldots, x_n\right) = \int_{-\infty}^{\infty} \prod_{i=1}^{n} N\left(x_i; \theta, 1-\rho\right) N(\theta; 0, \rho) d\theta = N\left(x; 0_n, \Sigma^{(n)}\right) \quad [EOP]$$

.
• $\Sigma^{(n)}$ is positive definite if $\rho \geq 0$. proof omitted, see Example 4.7 at P56 on notes.

Marginal consistency: a probability distributions is marginally consistent if every marginal of every distribution in the set is also in the set.
• (discrete case) $p_{1:n}\left(x_1, \ldots, x_n\right) = p_{1:n+1}\left(x_1, \ldots, x_n, 0\right) + p_{1:n+1}\left(x_1, \ldots, x_n, 1\right)$
• If PMF is not marginally consistent, then IES does not exist. (an example of this is omitted, see Example 4.9 at P57 on notes.)

4.3 de Finetti's Theorem

Let $(X_i)_{i=1}^{\infty}$ be an infinite sequence of binary RVs with PMF:

$$p_{1:n}\left(x_1, \ldots, x_n\right) = \Pr\left(X_1 = x_1, \ldots, X_n = x_n\right), n \geq 1,$$

then $(X_i)_{i=1}^{\infty}$ is exchangeable iff $\exists F(\theta) = P(\Theta \leq \theta) = P\left(\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} X_i \leq \theta\right) \in [0,1]$ (distribution function) s.t.

$$p_{1:n}\left(x_1, \ldots, x_n\right) = \int_0^1 \underbrace{\left[\prod_{i=1}^{n} p\left(x_i \mid \theta\right)\right]}_{p_{1:n}(x_1,\ldots,x_n|\Theta=\theta)\sim Bern(n,\theta)} dF(\theta), \text{with } p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$
$$= \int_0^1 \prod_{i=1}^{n} p\left(x_i \mid \theta\right) \pi(\theta) d\theta, \text{ if } F(\theta) \text{ is CDF}$$

• IES is distributed as a mixture of iid random variable.

<u>Proof</u>: ($\Longleftarrow$) is trivial, hence just show ($\Longrightarrow$).
Let $S_n = \sum_{i=1}^{n} X_i$ for $n = 1, 2, \cdots$, $X_i \in \{0,1\}$, and let r, s be 2 integers satisfying $0 \leq r < s \leq N$, then,

$$\Pr\left(S_n = r\right) = \binom{n}{r} p_{1:n}\left(x_1, \ldots, x_n\right) (*), \quad \Pr\left(S_n = r \mid S_N = s\right) = \frac{\binom{s}{r}\binom{N-s}{n-r}}{\binom{N}{n}}$$

When $S_n = r = \sum_{i=1}^{n} x_i$, there are at least $(n-r)$ 0's, hence:

$$\Pr\left(S_n = r\right) = \sum_{s=r}^{N-(n-r)} \Pr\left(S_n = r \mid S_N = s\right) \Pr\left(S_N = s\right)$$
$$= \sum_{s=r}^{N-(n-r)} \Pr\left(S_n = r \mid S_N/N = \theta(s)\right) \Pr\left(S_N/N = \theta(s)\right), \text{ with} \theta(s) = \frac{s}{N}$$

Define a RV $\Theta_N \sim \frac{S_N}{N}$ taking values $\{0, \frac{1}{N}, \cdots, 1\}$, then the CDF and density are:

$$F_N(\theta) = \Pr\left(\Theta_N \leq \theta\right), \forall\theta \in [0,1] = \sum_{s=0}^{N} \Pr\left(S_N = N\theta(s)\right) \mathbb{I}_{\theta(s)\leq\theta}$$
$$f_N(\theta) = \sum_{s=0}^{N} \Pr\left(S_N = N\theta\right) \delta_{\theta(s)}(\theta)$$

where $\delta_{\theta(s)}(\theta)$ is the Dirac delta-function putting a single point mass at $\theta = \theta(s)$.
• The discontinuities at $\theta(s)$ of F_N are associated with point masses $P(S_N = N\theta)\delta_{\theta(s)}$ in f_N.
Then,

$$\begin{aligned}
\Pr\left(S_n = r\right) &= \sum_{s=0}^{N} \int_0^1 g_N(\theta) \Pr\left(S_N = N\theta\right) \delta_{\theta(s)}(\theta) d\theta, \\
&\quad \text{with } g_N(\theta) = \mathbb{I}_{r\leq N\theta\leq N-(n-r)} \Pr\left(S_n = r \mid S_N = N\theta\right) \\
&= \sum_{s=0}^{N} \int_0^1 g_N(\theta) \Pr\left(S_N = N\theta\right) \delta_{\theta(s)}(\theta) d\theta \\
&= \int_0^1 g_N(\theta) f_N(\theta) d\theta \\
&= \int_{r/N}^{1-(n-r)/N} \Pr\left(S_n = r \mid S_N = N\theta\right) dF_N(\theta), \quad \because dF_N(\theta) \equiv f(\theta)d\theta \\
&= \int_0^1 \mathbb{I}_{r\leq N\theta\leq N-(n-r)} \Pr\left(S_n = r \mid S_N = N\theta\right) f_N(\theta) d\theta \\
&\to \binom{n}{r} \int_0^1 \theta^r (1-\theta)^{n-r} dF(\theta) \text{ as } N \to \infty
\end{aligned}$$

$\overset{(*)}{\Longrightarrow} p_{1:n}\left(x_1, \ldots, x_n\right) = \int_0^1 \theta^r (1-\theta)^{n-r} dF(\theta)$ [EOP]

Bayesian Prior Elicitation:
For $\{X_1 = x_1, \cdots, X_n = x_n\}$ a realization of n samples in an IES, ∃ generative model:

$$\Theta \sim F$$
$$X_i \mid \Theta = \theta \overset{iid}{\sim} p(\cdot \mid \theta)$$

• F is the *natural prior.*
• Posterior predictive distribution: $p\left(x_{m+1:n} \mid x_{1:m}\right) = \frac{p(x_{1:n})}{p(x_{1:m})} = \int p\left(x_{m+1:n} \mid \theta\right) \frac{p(x_{1:m}|\theta)dF(\theta)}{p(x_{1:m})}$, with $dF\left(\theta \mid x_1, \ldots, x_m\right) \propto p\left(x_1, \ldots, x_m \mid \theta\right) dF(\theta)$.

5 Approximate Bayesian Computation (ABC)

• A likelihood-free Bayesian method.

5.1 Doubly Intractable

Doubly Intractable Problem: posterior is doubly intractable if either the likelihood odds $(\frac{p(y|\theta)}{p(y|\theta')})$ or the prior odds $(\frac{\pi(\theta)}{\pi(\theta')})$ is intractable.

- Arises when the observation model takes the form $p(y|\theta) = \frac{p(y,\theta)}{c(\theta)}$ with $c(\theta) = \int_{\mathcal{Y}} p(y,\theta)dy$ intractable.
- But nevertheless possible to simulate the generative model: $y \sim p(\cdot|\theta), \theta \sim \pi(\cdot)$.

5.1.1 Ising model

omitted, see Example 5.7 at P65 on notes.

5.2 ABC Posterior

5.2.1 ABC Posterior approximation of $\pi(\theta|y_{obs})$

:

$$\begin{aligned}\pi_{ABC}\left(\theta \mid y_{obs}\right) = \pi(\theta|Y \in \Delta_\delta(y_{obs})) &= \frac{p\left(\Delta_\delta\left(y_{obs}\right) \mid \theta\right)\pi(\theta)}{p\left(\Delta_\delta\left(y_{obs}\right)\right)} \\ &= \int_{\Delta_\delta(y_{obs})} \pi(\theta \mid y) p\left(y \mid Y \in \Delta_\delta\left(y_{obs}\right)\right) dy \ \text{(Prop 5.14)}\end{aligned}$$

where:
(i) $\Delta_\delta\left(y_{obs}\right) = \left\{y' \in \mathcal{Y} : D\left(S\left(y_{obs}\right), S\left(y'\right)\right) \leq \delta\right\}$, a ball of radius δ centered on y_{obs}, s.t. $p\left(\Delta_\delta\left(y_{obs}\right) \mid \theta\right) = \int_{\Delta_\delta(y_{obs})} p(y \mid \theta)dy$;
(ii) $S : \mathcal{Y} \to \mathbb{R}^p$ summary statistic, and $D : \mathbb{R}^p \times \mathbb{R}^p \to [0,\infty)$ distance measure, and;
(iii) $p\left(y \mid Y \in \Delta_\delta\left(y_{obs}\right)\right) = \frac{p(y)\mathbb{I}_{y \in \Delta_\delta(y_{obs})}}{p(\Delta_\delta(y_{obs}))}$.
Proof of Prop 5.14:

$$\begin{aligned}\int_{\Delta_\delta(y_{obs})} \pi(\theta \mid y) p\left(y \mid Y \in \Delta_\delta\left(y_{obs}\right)\right) dy &= \int_{\Delta_\delta(y_{obs})} \frac{\pi(\theta \mid y)p(y)}{p\left(\Delta_\delta\left(y_{obs}\right)\right)} dy = \frac{\int_{\Delta_\delta(y_{obs})} p(y \mid \theta)\pi(\theta)dy}{p\left(\Delta_\delta\left(y_{obs}\right)\right)} \\ &= \frac{p\left(\Delta_\delta\left(y_{obs}\right) \mid \theta\right)\pi(\theta)}{p\left(\Delta_\delta\left(y_{obs}\right)\right)} = \pi_{ABC}\left(\theta \mid y_{obs}\right) \ \text{[EOP]}\end{aligned}$$

5.2.2 Simulation of ABC Posterior

Rejection Sampling of ABC Posterior:

```
Algorithm 3: Rejection Sampling of ABC Posterior
Observe data $y_{obs}$, initialize $n = 0$.
while $y_n \notin \Delta_\delta(y_{obs})$ do
    Simulate $\theta_n \sim \pi(\cdot), y_n \sim p(\cdot|\theta_n)$;
    $n = n + 1$
end
Return $(\Theta_{ABC} = \theta_n, Y_{ABC} = y_n, N = n)$, with $\Theta_{ABC} \sim \pi(\cdot|Y \in \Delta_\delta(y_{obs}))$ and
 $Y_{ABC} \sim Y|Y \in \Delta_\delta(y_{obs})$.
Regression Adjustment:
for $t = 1, ..., n$ do
    $s^{(t)} = S(y^{(t)})$
end
Regress $\theta^{(1:n)}$ against $\alpha(s^{(1:n)} - s_{obs})\beta$ and get $\hat{\alpha}, \hat{\beta}$.
Return $\theta_{adj}^{(t)} = \theta^{(t)} - (s^{(t)} - s_{obs})\hat{\beta}$
```

.

Proof: WTS $Pr(\Theta_{ABC} \in A) = \pi_{ABC}(A|y_{obs}), A \subseteq \Omega$.
Note that we get the output $(\Theta_{ABC} = \theta_n, Y_{ABC} = y_n, N = n)$ iff $(y_i)_{i=1}^{n-1} \notin \Delta_\delta(y_{obs})$, with $P(Y \in \Delta_\delta(y_{obs})) = p(\Delta_\delta(y_{obs}))$ and $N \sim Geom(p(\Delta_\delta(y_{obs})))$.

$$\begin{aligned}
&\Pr\left(\Theta_{ABC} \in A, Y_{ABC} \in \Delta_\delta\left(y_{obs}\right), N = n\right) \\
&\quad = \Pr\left(\Theta_{ABC} \in A, Y_{ABC} \in \Delta_\delta\left(y_{obs}\right), N = n \mid N > n-1\right) \times \Pr(N > n-1)) \\
&\quad = \Pr\left(\left(\theta_n, y_n\right) \in A \times \Delta_\delta\left(y_{obs}\right) \mid N > n-1\right)\left(1 - p\left(\Delta_\delta\left(y_{obs}\right)\right)\right)^{n-1}, \quad \because N \sim Geom(p(\Delta_\delta(y_{obs}))) \\
&\quad = \Pr\left((\theta, y) \in A \times \Delta_\delta\left(y_{obs}\right)\right)\left(1 - p\left(\Delta_\delta\left(y_{obs}\right)\right)\right)^{n-1}, \quad \because (\theta_i, y_i) \text{ are independent of each other.}
\end{aligned}$$

$$\begin{aligned}
\Pr\left(\Theta_{ABC} \in A\right) &= \sum_{n=1}^{\infty} \int_{\Delta_\delta(y_{obs})} \Pr\left(\Theta_{ABC} \in A, Y_{ABC} \in \Delta_\delta(y_{obs}), N = n\right) dy \\
&= \sum_{n=1}^{\infty} \Pr\left(\Theta_{ABC} \in A, N = n\right), \because Y_{ABC} \in \Delta_\delta(y_{obs}) \text{ for certain} \\
&= \sum_{n=1}^{\infty} \Pr\left((\theta, y) \in A \times \Delta_\delta\left(y_{obs}\right)\right)\left(1 - p\left(\Delta_\delta\left(y_{obs}\right)\right)\right)^{n-1} \\
&= \frac{\int_A \int_{\Delta_\delta(y_{obs})} \pi(\theta) p(y \mid \theta) dy d\theta}{p\left(\Delta_\delta\left(y_{obs}\right)\right)} = \frac{\int_A p\left(\Delta_\delta\left(y_{obs}\right) \mid \theta\right) \pi(\theta) d\theta}{p\left(\Delta_\delta\left(y_{obs}\right)\right)} \\
&= \int_A \pi\left(\theta \mid Y \in \Delta_\delta\left(y_{obs}\right)\right) d\theta = \pi_{ABC}(A \mid y) \quad [EOP]
\end{aligned}$$

5.2.3 Regression Adjustment of Samples

Proposition 5.23: If $(\theta, y) \sim \pi(\theta)p(y|\theta)$ with sufficient statistics $s = S(y)$, then the adjusted sample:

$$\theta_{adj} = \theta(s - s_{obs})\beta \sim \pi(\cdot|y_{obs})$$

Proof:

$$\begin{aligned}
\theta_{adj} &= \theta(s - s_{obs})\beta = \mu(s_{obs}) + \theta - \mu(s), \quad \because \text{ mean is a linear function of } s \\
&= \mu(s_{obs}) + \epsilon, \text{ with } \epsilon|y \sim \pi(\mu(s) + \epsilon|y) = \pi(\mu(s_{obs}) + \epsilon|y_{obs}) \\
\implies \theta_{adj} &\sim \pi(\theta_{adj}|y_{obs}) \quad [EOP]
\end{aligned}$$

($\epsilon|y \sim \pi(\mu(s) + \epsilon|y)$ because $\mu(s)$ is certain given y.)

- Regression adjustment allows us to take δ large, hence more samples are accepted.

5.2.4 Ising Model

Omitted, see section 5.3 at P70-71 on notes.

6 Model Averaging

- The key idea is to consider the model index m as a parameter.

6.1 Model averaging distributions and Decisions

Extended parameter space: $\Omega^* = \bigcup_{m\in\mathcal{M}}\bigcup_{\theta\in\Omega_m}\{(\theta,m)\}$.

Joint Posterior Distribution for the model and parameter:

$$\begin{aligned}\pi(\theta,m|y) &= \pi(\theta|y,m)\pi(m|y), \quad (\theta,m)\in\Omega^*\\ &= \left(\frac{p(y|\theta,m)\pi(\theta|m)}{p(y|m)}\right)\left(\frac{p(y|m)\pi_M(m)}{p(y)}\right)\\ &\propto p(y\mid\theta,m)\pi(\theta\mid m)\pi(m)\end{aligned}$$

where $p(y|m)=\int_{\Omega_m}p(y|\theta,m)\pi(\theta|m)d\theta$ (marginal likelihood under model $M=m$).

Model-Averaged...
(i) **marginal likelihood**: $p(y)=\sum_{m\in\mathcal{M}}p(y|m)\pi_M(m)$.
(ii) **posterior**: $\pi(\theta|y)=\sum_{m\in\mathcal{M}}\pi(\theta,m|y)\propto\sum_{m\in\mathcal{M}}\pi(\theta|y,m)p(y|m)\pi_M(m), \quad \theta\in\Omega=\bigcup_{m\in\mathcal{M}}\Omega_m$.

Example 6.5: Averaging over link functions
Suppose:
(i) observation model: $y_i\sim Bern(\mu_m(\beta_1+\beta_2x_i)), i=1,\cdots,n$;
(ii) 2 models associated with different link functions: $m=1$ (logistic), $m=2$ (probit), with priors $\pi(m=1)=\pi(m=2)=\frac{1}{2}$;
Then,
Model averaged posterior:

$$\begin{aligned}\pi(\beta\mid y) &= \pi(\beta\mid m=1,y)\pi(m=1\mid y)+\pi(\beta\mid m=2,y)\pi(m=2\mid y)\\ &= \pi(\beta\mid m=1,y)\frac{B_{1,2}}{1+B_{1,2}}+\pi(\beta\mid m=2,y)\frac{1}{1+B_{1,2}}\end{aligned}$$

$\because B_{1,2}=\frac{p(y|m=1)}{p(y|m=2)}=\frac{\pi(m=1|y)p(y)/\pi(m=1)}{\pi(m=2|y)p(y)/\pi(m=2)}=\frac{\pi(m=1|y)}{\pi(m=2|y)}$ and $\pi(m=1|y)=1-\pi(m=2|y)$.

Prop 6.6: Under squared error loss $(h-\delta)^2$ with estimator δ for the truth h, modelling averaging always minimizes the Bayes Risk $\rho(\pi,\delta(y))$:

$$\rho\left(\pi,E_{\theta|y,m}(h)\right)\geq\rho\left(\pi,E_{\theta,m|y}(h)\right)$$

where, for any functional $h:\Omega\to\mathbb{R}$ defined on each parameter space Ω_m,
(i) The modelling averaging posterior mean: $E_{\theta,m|y}(h(\theta))=\sum_{m\in\mathcal{M}}\int_{\Omega_m}h(\theta)\pi(\theta,m\mid y)d\theta$;
(ii) The single model posterior mean: $E_{\theta|y,m^*}(h(\theta))=\int_{\Omega_{m^*}}h(\theta)\pi\left(\theta\mid y,m^*\right)d\theta$
Proof: The expected posterior loss is,

$$\begin{aligned}\rho(\pi,\delta\mid y) &= \sum_{m\in\mathcal{M}}\int_{\Omega_m}(\delta-h(\theta))^2\pi(\theta,m\mid y)d\theta\\ \Longrightarrow \frac{\partial\rho}{\partial\delta} &= \sum_{m\in\mathcal{M}}\int_{\Omega_m}(2\delta-2h(\theta))\pi(\theta,m\mid y)d\theta=0\\ \Longrightarrow \delta(y) &= E_{\theta,m|y}(h(\theta)) \quad [EOP]\end{aligned}$$

6.2 Spike-and-Slab Priors

Suppose:
(i) Regression model $Y \sim N\left(X\theta_z, \sigma^2\right)$, with $\theta_z = (z_1\theta_1, \ldots, z_p\theta_p)$ and $z_i \in \mathcal{M} = [0,1]^p$;
$\implies \pi(\theta, \sigma, z|y) \propto p(y|\theta, \sigma, z)\pi(\theta, \sigma)\pi(z)$ with $p(y|\theta, \sigma, z) = N(y; X\theta_z, \sigma^2)$ and $\pi(\theta, \sigma|z) = \pi(\theta, \sigma)$.
(ii) All independent priors: $\pi(\theta, \sigma, z) = \pi_\sigma(\sigma) \prod_i \pi\left(\theta_i\right) p\left(z_i\right)$, with the total parameter space: $\Omega^* = \mathbb{R}^p \times \mathbb{R}^+ \times \mathcal{M}$.
Then,
(a) The **prior CDF of** $\theta_{z,i} = \theta_i z_i$ is defined by summing over $z_i = 0, 1$:

$$\Pr\left(\Theta_{z,i} \leq c\right) = w\mathbb{I}_{c\geq 0} + (1-w)\int_{-\infty}^{c} \pi\left(\theta_i\right) d\theta_i$$

•$w = p(z_i = 0)$ and we know for certain that $\theta_{z,i} = 0 \iff z_i = 0$.
(b) The **Spike-and-Slab Prior** is:

$$\frac{\partial}{\partial c}\Pr\left(\Theta_{z,i} \leq c\right) = \pi_{\Theta_{z,i}}\left(\theta_{z,i}\right) = \underbrace{w\delta_0\left(\theta_{z,i}\right)}_{\text{spike}} + \underbrace{(1-w)\pi\left(\theta_{z,i}\right)}_{\text{slab}}$$

6.2.1 Example: Polynomial Regression

Suppose:
(i) Regression model: $Y_i = \sum_{j=1}^{p} z_j\theta_j x_{i,j}^{j-1} + \epsilon_i$, with $\epsilon_i \sim N(0, \sigma^2)$ and $z_i \in \{0, 1\}$;
(ii) Set $p = 6 \implies X = (1, x, , x^2, \cdots, x^5)$;
(iii) Set **Priors**:
- $\theta_i \sim N(0, 9)$;
- $\sigma \sim \frac{1}{\sigma}$ (Jefery's Prior);
- $z \sim \pi(z) = \prod_{i=1}^{p} \xi^{z_i}(1 - \xi^{z_i})$, with $\xi = \frac{c}{p} = \frac{3}{5}$ (prior of z gives the expected number of covariates c in the model.)

Then,
(a) **Posterior**: under the "all independent priors" assumption,

$$\pi(\theta, z, \sigma \mid y) \propto N\left(y; X\theta_z, \sigma^2\right) \times N\left(\theta; 0, 9I_p\right) \times \sigma^{-1} \times \xi^{|z|}(1-\xi)^{p-|z|}$$

(b) **MCMC targeting** $\pi(\theta, z, \sigma \mid y)$: see Algorithm 4
(c) **Posterior mean of the observation model**: let $v(x) = (1, x, \cdots, x^5)^T$

$$\mu(x) = v(x)^T E_{\theta,\sigma,z|y}(\theta_z) \implies \hat{\mu}(x) = v(x)^\top \frac{1}{T}\sum_{t=1}^{T} \theta_z^{(t)}$$

(d) **Posterior probability of the model index**:

$$\hat{\pi}(z|y) = \frac{1}{T}\sum_{t} \mathbb{I}_z^{(t)} = z$$

Algorithm 4: MCMC targeting $\pi(\theta, z, \sigma \mid y)$

```
Initialize parameters $(\theta^{(0)}, \sigma^0, z^{(0)})$;
Initialize $a > 0$.
for $t = 1, \cdots, T$ do
    The parameters are now $(\theta^{(t)}, \sigma^t, z^{(t)}))$ and hence $\theta_z = \left(z_1^{(t)}\theta_1^{(t)}, \ldots, z_p^{(t)}\theta_p^{(t)}\right)$
    $\theta-$update:
    Randomly choose $i \sim U(1, 2, \cdots, p)$ and simulate $\theta'_i \sim U(\theta_i^{(t)} - a, \theta_i^{(t)} + a)$;
    Replace $\theta_i^{(t)}$ with $\theta'_i$ and set $\theta'_z = \left(z_1^{(t)}\theta_1^{(t)}, \ldots, z_{i-1}^{(t)}\theta_{i-1}^{(t)}, z_i^{(t)}\theta'_i, z_{i+1}^{(t)}\theta_{i+1}^{(t)}, \ldots, z_p^{(t)}\theta_p^{(t)}\right)$;
    Simulate $u_\theta \sim U(0, 1)$;
    if $u_\theta < \alpha_{\theta_i}\left(\theta'_z \mid \theta_z^{(t)}\right) = \min\left\{1, \frac{N\left(y; X\theta'_z, \sigma^2\right)N\left(\theta'_i; 0,9\right)}{N(y; X\theta_z, \sigma^2)N(\theta_i; 0,9)}\right\}$ then
        $\theta_i^{(t+1)} = \theta'_i$ and $\theta_{-i}^{(t+1)} = \theta_{-i}^{(t)}$
    end
    else: $\theta^{(t+1)} = \theta^{(t)}$ unchanged.
    $z-$update:
    Randomly choose $j \sim U(1, 2, \cdots, p)$ and set $z'_j = 1 - z_j^{(t)}$;
    Replace $z_j^{(t)}$ with $z'_j$ and set $\theta_{z'} = \left(z_1^{(t)}\theta_1^{(t+1)}, \ldots, z'_j\theta_j^{(t+1)}, \ldots z_p^{(t)}\theta_p^{(t+1)}\right)$;
    Simulate $u_z \sim U(0, 1)$;
    if $u < \alpha_{z_i}\left(\theta_{z'} \mid \theta_z^{(t)}\right) = \min\left\{1, \frac{N\left(y; X\theta_{z'}, \sigma^2\right)\xi^{z'_i}(1-\xi)^{1-z'_i}}{N(y; X\theta_z, \sigma^2)\xi^{z_j^{(t)}}(1-\xi)^{1-z_j^{(t)}}}\right\}$ then
        $z_j^{(t+1)} = z'_j$ and $z_{-j}^{(t+1)} = z_{-j}^{(t)}$
    end
    else: $z^{(t+1)} = z^{(t)}$ unchanged.
    $\sigma-$update: random walk on a log scale, see Algorithm 6
end
Return the sample $(\theta^t, z^{(t)}, \sigma^{(t)})$.
```

7 Reversible-Jump MCMC

• Reversible-Jump MCMC deals with problems when the models $m \in \mathcal{M}$ we consider have parameter space of different dimensions ($\Omega_m = \mathbb{R}^{p_m}$), i.e. when the number of unknowns is one of the things unknown.

7.1 Transition Kernel and Detailed Balance

Suppose we have N kernels $K_{1:N}$, each step we randomly pick one K_j with probability ξ_j and use it to update the state. Hence: $K(\theta, d\theta') = \sum_{j=1}^{N} \xi_j K_j(\theta, d\theta')$, with $(\theta = \mathbb{R}, \mathcal{B}, \pi)$ the target probability space.

For any fixed $j = 1, ..., N$:
Transition Kernel for Metropolis-Hasting MCMC:

$$K_j\left(\theta, d\theta'\right) = \alpha_j\left(\theta' \mid \theta\right) q_j\left(d\theta' \mid \theta\right) + c_j(\theta)\delta_\theta\left(d\theta'\right)$$

where: (for simplicity, remove the subscripts)
(i) $\theta' \sim q(\cdot|\theta)$ proposal;
(ii) $\alpha(\theta'|\theta)$ acceptance probability of a proposed θ';
(iii) $c(\theta) = 1 - \int_\Omega \alpha\left(\theta' \mid \theta\right) q\left(d\theta' \mid \theta\right)$ rejection probability of a proposed θ'.
Proof: Transition from state X_t to X_{t+1} includes either acceptance (with probability of a proposal times the probability of accepting that proposal), or rejection (with probability of rejection).

$$\operatorname{Pr}\left(X_{t+1} \in A \mid X_t = \theta\right) = c(\theta)\mathbb{I}_{\theta \in A} + \int_A \alpha\left(\theta' \mid \theta\right) q\left(d\theta' \mid \theta\right) = \int_A K\left(\theta, d\theta'\right) \quad [EOP]$$

Detailed Balance of Transition Kernel:
Prop: 7.6: Transition Kernel satisfies Detailed Balance iff: for $\theta \in A, \theta' \in B$,

$$\int_B \int_A \pi\left(d\theta'\right) q\left(d\theta \mid \theta'\right) \alpha\left(\theta \mid \theta'\right) = \int_A \int_B \pi(d\theta) q\left(d\theta' \mid \theta\right) \alpha\left(\theta' \mid \theta\right)$$

Proof:

$$\int_B \int_A \pi\left(d\theta'\right) q\left(d\theta \mid \theta'\right) \alpha\left(\theta \mid \theta'\right) = \int_A \int_B \pi(d\theta) q\left(d\theta' \mid \theta\right) \alpha\left(\theta' \mid \theta\right)$$
$$\overset{(*)}{\Longrightarrow} \int_B \int_A \pi\left(d\theta'\right)\left[q\left(d\theta \mid \theta'\right) \alpha\left(\theta \mid \theta'\right) + c(\theta')\delta_\theta\left(d\theta\right)\right] = \int_A \int_B \pi(d\theta)\left[q\left(d\theta' \mid \theta\right) \alpha\left(\theta' \mid \theta\right) + c(\theta)\delta_\theta\left(d\theta'\right)\right]$$
$$\Longrightarrow \int_B \int_A \pi(d\theta')K(\theta', d\theta) = \int_A \int_B \pi(d\theta)K(\theta, d\theta')$$
$$\Longrightarrow \pi(d\theta')K(\theta', d\theta) = \pi(d\theta)K(\theta, d\theta'), \text{ by Definition, DB holds.}$$

where we can do $(*)$ because:

$$\int_B \int_A \pi(d\theta')c(\theta')\delta_\theta(d\theta) = \int_B \pi(d\theta')c(\theta')\mathbb{I}_{\theta \in A} = \int_A \pi(d\theta)c(\theta)\mathbb{I}_{\theta \in B} = \int_A \int_B \pi(d\theta)c(\theta)\delta_\theta(\theta') \quad [EOP]$$

7.2 Jacobian-Based MCMC

7.2.1 Proposal Transformations

Goal: Given the chosen proposal $q(\theta'|\theta)$, want to find a density $g(u)$ and a function $\theta' = \psi_1(\theta, u)$ to simulate it.

Proposal function: an invertible differentiable function $\psi_1(\theta, u) : \Omega \times \mathcal{U} \to \Omega$, where,
(i) $\mathcal{U} \subseteq \mathcal{R}$ s.t. **proposal variable** $u \in \mathcal{U}$ with a density $g(u)$
• We simulate proposal by first simulating $u \sim g(u)$ and then set $\theta' = \psi_1(\theta, u)$.
•$\forall\theta' \in \{\psi_1(\theta, u) : u \in \mathcal{U}\}, !\exists u : \theta' = \psi_1(\theta, u)$ and the mapping $u \to \theta'$ is 1-1 and invertible.

Conditional proposal distribution: $q(d\theta'|\theta) = g(u)du$
Conditional proposal density: $q(\theta'|\theta) = g(u)\left|\frac{\partial\theta'}{\partial u}\right|^{-1} = g(u)\left|\frac{\partial\psi_1(\theta,u)}{\partial u}\right|^{-1}$
•$u(\theta')$ solves $\theta' = \psi_1(\theta, u)$

Reversibility of the Proposal function: the **proposal variable for the reverse update** u' solves $\theta = \psi_1(\theta', u') = \psi_1(\psi_1(\theta, u), u')$
• The reverse proposal function $\psi_2 : \Omega \times \mathcal{U} \to \mathcal{U}$ s.t. $u' = \psi_2(\theta, u)$ is invertible and differentiable.

Prop 7.18: $\theta' = \psi_1(\theta, u) \& u' = \psi_2(\theta, u) \implies \psi_2(\theta', u') = u$.
<u>Proof</u>: Suppose $\psi_2(\theta', u') = x$. By definition of ψ_2, x solves $\theta' = \psi_1(\psi_1(\theta', u'), x) = \psi_1(\theta, x)$, as $\theta = \psi_1(\theta', u')$. But also note that u solves $\theta' = \psi_1(\theta, x)$ and the solution is unique due to invertibility of ψ_1. Therefore, $x = u$. [EOP]

Prop 7.19: Function $\psi = (\psi_1, \psi_2)$ mapping $\psi : \Omega \times \mathcal{U} \to \Omega \times \mathcal{U}$ is an invertible, differentiable involution, i.e. $(\theta, u) = \psi(\psi(\theta, u))$.
<u>Proof</u>: $\psi(\psi(\theta, u)) = \psi(\psi_1(\theta, u), \psi_2(\theta, u)) = \psi(\theta', u') = (\psi_1(\theta', u'), \psi_2(\theta', u')) = (\theta, u)$ [EOP]

Example 7.9 & 7.17 & 7.20: For $a > 0, u \sim U(0,1)$, set $\theta' = \theta + a(2u-1)$ as the standard *random-walk* proposal. Then, $g(u) = \mathbb{I}_{0<u<1}, \quad \psi_1(\theta, u) = \theta + a(2u-1)$.
- Invertible at $\theta \implies u = \frac{a+\theta'-\theta}{2a}$.
- $u' = \psi_2(\theta, u) = 1-u$ because $\psi_1(\theta', u') = \psi_1(\theta + a(2u-1), 1-u) = \theta + a(2u-1) + a(2(1-u)-1) = \theta$.
- $\psi(\psi(\theta, u)) = \psi(\theta + a(2u-1), (1-u)) = (\theta + a(2u-1) + a(2(1-u)-1), 1-(1-u)) = (\theta, u)$

7.2.2 MCMC with Transformation

Algorithm 5: MCMC with Transformation

```
Initialize $(\theta^{(0)}, u^{(0)})$ and $t = 0$.
Initialize proposal $u \sim g(u)$ and the transformation function $\psi = (\psi_1, \psi_2)$
for $t = 1, \cdots, T$ do
    Simulate $u' \sim g(u)$;
    Compute $\theta' = \psi_1(\theta^{(t)}, u^{(t)})$ and $u' = \psi_2(\theta^{(t)}, u^{(t)})$;
    Simulate $k \sim Unif(0,1)$ and compute Jacobian $J_\psi(\theta^{(t)}, u^{(t)}) = \left|\frac{\partial(\theta', u')}{\partial(\theta^{(t)}, u^{(t)})}\right|$
    if $k < \alpha(\theta'|\theta^{(t)}) = \min\left\{1, \frac{\pi(\theta')g(u')}{\pi(\theta^{(t)})g(u^{(t)})} J_\psi(\theta^{(t)}, u^{(t)})\right\}$ then
        Set $(\theta^{(t+1),u^{(t+1)}}) = (\theta', u')$
    end
    else: $(\theta^{(t+1)}, u^{(t+1)}) = (\theta^{(t)}, u^{(t)})$
end
Return $(\theta^{(t)}, u^{(t)})$ for $t = 1, \cdots, T$.
```

.

Theorem 7.21: The acceptance probability above satisfies DB.
<u>Proof</u>: Let $r(\theta', u'|\theta, u) = \frac{\pi(\theta')g(u')}{\pi(\theta)g(u)} J_\psi(\theta, u)$ and assume $r(\theta', u'|\theta, u) = (r(\theta, u|\theta', u'))^{-1}$ $(*)$ (show later).
<u>WTS</u>: $\pi(d\theta) q(d\theta' \mid \theta) \alpha(\theta' \mid \theta) = \pi(d\theta') q(d\theta \mid \theta') \alpha(\theta \mid \theta')$.
WLOG, suppose $r(\theta', u'|\theta, u) \leq 1$ and denote $(\theta'(\theta, u), u'(\theta, u)) = (\psi_1(\theta, u), \psi_2(\theta, u))$, then

$$\begin{aligned}
LHS &= \pi(\theta)g(u)\alpha\left(\theta'(\theta, u) \mid \theta\right) dud\theta = \pi(\theta)g(u)\frac{\pi\left(\theta'(\theta, u)\right) g\left(u'(\theta, u)\right)}{\pi(\theta)g(u)} J_\psi(\theta, u) dud\theta \\
&= \pi\left(\theta'(\theta, u)\right) g\left(u'(\theta, u)\right) \left|\frac{\partial\left(\theta', u'\right)}{\partial(\theta, u)}\right| dud\theta = \pi\left(\theta'\right) g\left(u'\right) du'd\theta' \\
RHS &= \pi(\theta')g(u')du'd\theta', \quad \because \alpha(\theta|\theta') = 1 \\
\implies LHS &= RHS
\end{aligned}$$

To show assumption $(*)$, recall that *Jacobian of the inverse transformation is the inverse of the Jacobian of the transformation*, so: $J_{\psi^{-1}}(\theta', u') = (J_\psi(\theta, u))^{-1}$

$\implies J_\psi(\theta, u) = \left(J_{\psi^{-1}}(\theta', u')\right)^{-1} = (J_\psi(\theta', u'))^{-1}$, as ψ is an involution.
$\implies (r(\theta, u \mid \theta', u'))^{-1} = \frac{\pi(\theta')g(u')}{\pi(\theta)g(u)} J_\psi(\theta', u')^{-1} = \frac{\pi(\theta')g(u')}{\pi(\theta)g(u)} J_\psi(\theta', u') = r(\theta', u' | \theta, u)$. [EOP]

• *Dimension matching*: $dim(\theta, u) = dim(\theta', u')$ and hence J_ψ is squared matrix (non-singular).

Random-walk on a Log Scale: Targeting $\pi(\theta) \sim \exp(-\theta)$

Algorithm 6: Random-walk on a Log Scale

```
Initialize $(\theta^{(0)}, u^{(0)})$ and $t = 0$.
Initialize proposal $u \sim U(\frac{1}{2}, 2)$ and the transformation function
  $\theta' = \psi_1(\theta, u) = u\theta, u' = \psi_2(\theta, u) = \frac{1}{u}$ $(\because \psi_1(\theta', u') = u\theta\frac{1}{u} = \theta)$.
for $t = 1, \cdots, T$ do
    Simulate $u' \sim g(u) = \frac{1}{2-1/2}\mathbb{I}_{0.5<u<1}$;
    Set $\theta' = u\theta^{(t)}$;
    Compute $\theta' = \psi_1(\theta^{(t)}, u^{(t)})$ and $u' = \psi_2(\theta^{(t)}, u^{(t)})$;
    Simulate $k \sim Unif(0, 1)$;
    Compute Jacobian $J_\psi(\theta^{(t)}, u^{(t)}) = \left|\frac{\partial(\theta', u')}{\partial(\theta^{(t)}, u^{(t)})}\right| = \begin{vmatrix} u^{(t)} & 0 \\ \theta^{(t)} & -1/\left(u^{(t)}\right)^2 \end{vmatrix} = \frac{1}{u^{(t)}}$
    if $k < \alpha(\theta'|\theta^{(t)}) = \min\left\{1, \frac{\pi(\theta')g(u')}{\pi(\theta^{(t)})g(u^{(t)})} J_\psi\left(\theta^{(t)}, u^{(t)}\right)\right\} = \min\left\{1, \frac{\exp\left(-\theta'+\theta^{(t)}\right)}{u}\right\}$ then
        Set $(\theta^{(t+1), u^{(t+1)}}) = (\theta', u')$
    end
    else: $(\theta^{(t+1)}, u^{(t+1)}) = (\theta^{(t)}, u^{(t)})$
end
Return $(\theta^{(t)}, u^{(t)})$ for $t = 1, \cdots, T$.
```

7.2.3 Matched Proposals

Matched kernels:

$$K_i\left(\theta, d\theta'\right) = \alpha_i\left(\theta' \mid \theta\right) q_i\left(d\theta' \mid \theta\right) + c_i(\theta)\delta_\theta\left(d\theta'\right)$$

where,
(i) acceptance rate: $\alpha_i(\theta' \mid \theta) = \min\left\{1, \frac{\pi(\theta')\xi_{\sigma_i} q_{\sigma_i}(\theta|\theta')}{\pi(\theta)\xi_i q_i(\theta'|\theta)}\right\}$;
(ii) rejection rate: $c_i(\theta) = 1 - \int_\Omega \alpha_i(\theta' \mid \theta) q_i(d\theta' \mid \theta)$;
(iii) $\sigma \in \mathcal{P}_N$ a permutation of $\{1, \cdots, N\}$.

Overall kernel: a weighted sum of the matched kernels, $K(\theta, d\theta') = \sum_{i=1}^N \xi_i K_i(\theta, d\theta')$.

Prop 7.28: DB holds if: for (i, σ_i) pairs, $i = 1, \cdots, N$,

$$\pi(d\theta)\xi_i\alpha_i\left(\theta' \mid \theta\right) q_i\left(d\theta' \mid \theta\right) = \pi\left(d\theta'\right)\xi_{\sigma_i}\alpha_{\sigma_i}\left(\theta \mid \theta'\right) q_{\sigma_i}\left(d\theta \mid \theta'\right)$$

Proof: WTS $\sum_i \pi(d\theta) K_i(\theta, d\theta') = \sum_i \pi(d\theta') K_i(\theta', d\theta)$, as this is the definition of DB.
Suppose WLOG $\alpha_i(\theta'|\theta) \leq 1$, then: (note we require $\sigma_{\sigma_i} = i$)

$$\alpha_{\sigma_i}\left(\theta|\theta'\right) = \min\left\{1, \frac{\pi(\theta)\xi_{\sigma_{\sigma_i}} q_{\sigma_{\sigma_i}}\left(\theta' \mid \theta\right)}{\pi\left(\theta'\right)\xi_{\sigma_i} q_{\sigma_i}\left(\theta \mid \theta'\right)}\right\} = \min\left\{1, \underbrace{\frac{\pi(\theta)\xi_i q_i\left(\theta' \mid \theta\right)}{\pi\left(\theta'\right)\xi_{\sigma_i} q_{\sigma_i}\left(\theta \mid \theta'\right)}}_{\geq 1}\right\} = 1$$

$$\begin{aligned}
\pi(\theta)\xi_i\alpha_i\left(\theta' \mid \theta\right) q_i\left(\theta' \mid \theta\right) &= \pi(\theta)\xi_i\frac{\pi\left(\theta'\right)\xi_{\sigma_i}q_{\sigma_i}\left(\theta \mid \theta'\right)}{\pi(\theta)\xi_i q_i\left(\theta' \mid \theta\right)}q_i\left(\theta' \mid \theta\right)\\
&= \pi\left(\theta'\right)\xi_{\sigma_i}q_{\sigma_i}\left(\theta \mid \theta'\right)\\
&= \pi\left(\theta'\right)\xi_{\sigma_i}\alpha_{\sigma_i}\left(\theta \mid \theta'\right)q_{\sigma_i}\left(\theta \mid \theta'\right) \qquad \because \alpha_{\sigma_i}(\theta|\theta') = 1
\end{aligned}$$

Summing over i on both sides concludes the proof, $c_i(\theta)$ terms cancelled as in Prop 7.6. [EOP]

Prop 7.30 omitted here, come back later.

7.3 Reversible-Jump MCMC (RJMCMC)

Goal: Targeting $\pi(\theta, m|y) \propto p(y|\theta, m)\pi(\theta|m)\pi(m)$, with $(\theta, m) \in \Omega^* = \bigcup_{m\in\mathcal{M}} \bigcup_{\theta\in\Omega_m} \{(\theta, m)\}$.

7.3.1 Reversible Jump Proposals

(i) **Add dimension to the state**: $(\theta, u) \overset{\psi_1,\rho_{m,m'}}{\longrightarrow} (\theta', u')$, where,
- $\theta = (\theta_1, \cdots, \theta_m)$, with $m = card(\theta)$;
- $\theta' = (\theta_1, \cdots, \theta_m, \theta'_{m+1})$, with $m' = card(\theta') = m + 1$;
- proposal transformation $\psi_1 : \Omega_m \times \mathcal{U}_{m,m'} \to \Omega_{m+1}$ s.t. $\psi_1(\theta, u) = (\theta_1, \cdots, \theta_m, \theta'_{m+1}(\theta, u))$;
- $\rho_{m,m'}$ probability to propose a move from m to m';
- $u \sim g_{m,m'}(\cdot)$ (generating density for the update) makes up the missing dimension.

(ii) **Removing dimension from the state**: $(\theta', u') \overset{\psi_1,\rho_{m',m}}{\Longrightarrow} (\theta, u)$, where,
- $\mathcal{U}_{m',m} = \emptyset, g_{m',m}(\emptyset) = 1 \implies$ directly delete the last entry of θ';
- proposal transformation $\psi_1 : \Omega_{m'} \times \mathcal{U}_{m',m} \to \Omega_m$ s.t. $\psi_1(\theta', \emptyset) = (\theta_1, \cdots, \theta_m)$;

• assume $\rho_{1,0} = 0$, as we do not want the update to propose a none-model.

(iii) **Transformation**: For $(\theta, u) \in \Omega_m \times \mathcal{U}_{m,m'}$ and $(\theta', u') \in \Omega_{m'} \times \mathcal{U}_{m',m}$ with $(m' = m + 1)$,
- *forward*: $(\theta', u') = \psi(\theta, u) = (\psi_1(\theta, u), \psi_2(\theta, u))$;
- *reverse*: $(\theta, u) = \psi(\theta', \emptyset) = (\psi_1(\theta', \emptyset), \psi_2(\theta', \emptyset))$.

•$\psi(\cdot)$ is an involution mapping $\left(\Omega_m \times \mathcal{U}_{m,m'}\right) \bigcup \left(\Omega_{m'} \times \mathcal{U}_{m',m}\right)$.

7.3.2 RJMCMC Algorithm

Algorithm 7: Reversible-Jump MCMC

```
Initialize $(\theta^{(0)}, m^{(0)})$ and transformation function $\psi = (\psi_1, \psi_2)$.
for $t = 1, \cdots, T$ do
    Set $m' = \begin{cases} m^{(t)} + 1, & \text{with probability} \rho_{m,m+1} \text{ and simulate } u \sim g_{m^{(t)},m^{(t)}+1}(\cdot) \\ m^{(t)} - 1, & \text{with probability} \rho_{m^{(t)},m^{(t)}-1} = 1 - \rho_{m^{(t)},m^{(t)}+1} \text{ and set } u = \emptyset \end{cases}$
    Set $(\theta', u') = (\theta, u)$.
    if $m' = m^{(t)} + 1$ then
        increase dimension acceptance:
        $\alpha\left(\theta', m' \mid \theta^{(t)}, m^{(t)}\right) = \min\left\{1, \frac{\pi(\theta',m'|y)\rho_{m',m^{(t)}}}{\pi(\theta^{(t)},m^{(t)}|y)\rho_{m^{(t)},m'} g_{m^{(t)},m'}(u)} J_\psi(\theta^{(t)}, u^{(t)})\right\}$
        with $J_\psi(\theta, u) = \left|\frac{\partial\theta'(\theta,u)}{\partial(\theta,u)}\right|$
    end
    if $m' = m^{(t)} - 1$ then
        decrease dimension acceptance:
        $\alpha\left(\theta', m' \mid \theta^{(t)}, m^{(t)}\right) = \min\left\{1, \frac{\pi(\theta',m'|y)\rho_{m',m^{(t)}} g_{m',m^{(t)}}(u')}{\pi(\theta^{(t)},m^{(t)}|y)\rho_{m^{(t)},m'}} J_\psi(\theta^{(t)}, \emptyset)\right\}$
        with $J_\psi(\theta, \emptyset) = \left|\frac{\partial(\theta',u')}{\partial\theta}\right|$
    end
    Simulate $k \sim U(0,1)$.
    if $k < \alpha\left(\theta', m' \mid \theta^{(t)}, m^{(t)}\right)$ then
        $(\theta^{(t+1)}, m^{(t+1)}) = (\theta', u')$
    end
    else: $(\theta^{(t+1)}, m^{(t+1)}) = (\theta^{(t)}, m^{(t)})$
end
```

.

Prop 7.32: DB holds for the update between pairs of transition kernels associated with the addition and deletion proposal kernels in Algorithm 7 targetting $\pi(\theta, m|y)$.

• A transition kernel of RJMCMC is $K_{m,m'}(\theta, d\theta') = \alpha(\theta', m'|\theta, m)q_{m,m'}(d\theta'|\theta) + c_{m,m'}(\theta)\delta_\theta(d\theta')$: where, - the *probability of proposing the move* from m to m', i.e. $\rho_{m,m'}$ and;

- the *conditional probability of proposing the parameter* θ' (given θ) under the proposed move $m \to m'$, i.e. $q_{m,m'}(\theta'|\theta)$, and;

- the *probability of accepting the proposal*, i.e. $\alpha(\theta', m'|\theta, m)$.

- the *probability of rejecting the proposal*, i.e. $c_{m,m'} = 1 - \int_{\Omega_{m'}} \alpha(\theta', m'|\theta, m)q_{m,m'}(\theta'|\theta)d\theta'$.

Proof: To prove DB, sufficient to show for $A \subset \Omega_m$ and $B \subset \Omega_{m'}$,

$$\int_B \int_A \pi\left(d\theta', m' \mid y\right) \rho_{m',m} q_{m',m}\left(d\theta \mid \theta'\right) \alpha\left(\theta, m \mid \theta', m'\right) \overset{(*)}{=} \int_A \int_B \pi(d\theta, m \mid y)\rho_{m,m'} q_{m,m'}\left(d\theta' \mid \theta\right) \alpha\left(\theta', m' \mid \theta, m\right)$$

Suppose WLOG $m' = m - 1$ hence $u' = \emptyset$, and $\alpha(\theta', m'|\theta, m) \leq 1$, then the proposal densities: (by definition of conditional proposal)

$$\textcircled{1} \quad q_{m,m'}\left(d\theta' \mid \theta\right) = \delta_{\theta_{1:m}}\left(d\theta'_{1:m}\right) g_{m,m'}(u)du$$

$$\textcircled{2} \quad q_{m',m}\left(d\theta \mid \theta'\right) = \delta_{\theta'_{1:m}}\left(d\theta_{1:m}\right)$$

$$
\begin{aligned}
RHS &\overset{\textcircled{1}}{=} \int_A \int_B \pi(\theta, m \mid y) \rho_{m,m'} g_{m,m'}(u) \delta_{\theta_{1:m}} \left(d\theta'_{1:m}\right) \frac{\pi\left(\theta', m' \mid y\right) \rho_{m',m}}{\pi(\theta, m \mid y) \rho_{m,m'} g_{m,m'}(u)} \left|\frac{\partial \theta'(\theta, u)}{\partial(\theta, u)}\right| du d\theta \\
&= \int_A \int_B \delta_{\theta_{1:m}} \left(d\theta'_{1:m}\right) \pi\left(\theta', m' \mid y\right) \rho_{m',m} \underbrace{\left|\frac{\partial \theta'(\theta, u)}{\partial(\theta, u)}\right|}_{J_\psi(\theta,u)} du d\theta \\
&= \int_A \int_B \pi\left(\theta', m' \mid y\right) \rho_{m',m} \delta_{\theta_{1:m}} \left(d\theta'_{1:m}\right) \left|\frac{\partial \theta'_{m+1}}{\partial u}\right| du d\theta, \quad \because J_\psi(\theta, u) = \begin{vmatrix} \frac{\partial \theta'_{1:m}}{\partial \theta} & \frac{\partial \theta'_{1:m}}{\partial u} \\ \frac{\partial \theta'_{m+1}}{\partial \theta} & \frac{\partial \theta'_{m+1}}{\partial u} \end{vmatrix} = \begin{vmatrix} I_{m\times m} & 0_{m1} \\ \frac{\partial \theta'_{m+1}}{\partial \theta} & \frac{\partial \theta'_{m+1}}{\partial u} \end{vmatrix} \\
&= \int_A \int_B \pi\left(\theta', m' \mid y\right) \rho_{m',m} \delta_{\theta'_{1:m}} \left(d\theta_{1:m}\right) \left|\frac{\partial \theta'_{m+1}}{\partial u}\right| \begin{vmatrix} \frac{\partial u}{\partial \theta'_{1:m}} & \frac{\partial u}{\partial \theta'_{m+1}} \\ \frac{\partial \theta}{\partial \theta'_{1:m}} & \frac{\partial \theta}{\partial \theta'_{m+1}} \end{vmatrix} d\theta' \overset{CoV}{\Longleftarrow} (u, \theta) \to \theta' = (\theta'_{1:m}, \theta'_{m+1}) \\
&= \int_A \int_B \pi\left(\theta', m' \mid y\right) \rho_{m',m} \delta_{\theta'_{1:m}} \left(d\theta_{1:m}\right) d\theta', \quad \because \frac{\partial u}{\partial \theta'_{1:m}} = \frac{\partial \theta}{\partial \theta'_{m+1}} = 0, \frac{\partial \theta}{\partial \theta'_{1:m}} = 1 \\
&\overset{\textcircled{2}}{=} \int_A \int_B \pi\left(d\theta', m' \mid y\right) \rho_{m',m} q_{m',m} \left(d\theta \mid \theta'\right) = RHS, \quad \because \alpha(\theta, m | \theta', m') = 1, \quad \text{[EOP]}
\end{aligned}
$$

- $\delta_{\theta_{1:m}}(d\theta'_{1:m})$ is to ensure $\theta_{1:m} = \theta'_{1:m}$.
- $m' = m - 1$ can be shown similarly.

7.3.3 Sampling a Semi-Random Variable via RJMCMC

Consider $X = \begin{cases} \frac{1}{2}, \text{ with prob} = \frac{1}{3} \\ V \sim F_V(v) = v^2, \text{ with prob} = \frac{2}{3} \end{cases} \implies \begin{cases} CDF : F_X(x) = \frac{2}{3}x^2 + \frac{1}{3}\mathbb{I}_{x \geq 1/2} \\ PDF : f_X(x) = \frac{4}{3}x + \frac{1}{3}\delta_{1/2}(x) \end{cases}$

Hence the X is generated from the process:

$$
\pi(x, m) = \pi(x|m)\pi(m) = \begin{cases} \mathbb{I}_{x=1/2} \times \frac{1}{3}, \text{ if } m = 1, x = \frac{1}{2}; \\ 2x \times \frac{2}{3}, \text{ if } m = 2, x \in (0, 1) \\ 0, o/w \end{cases}, \quad \Omega^* = \{(1/2, 1)\} \cup \bigcup_{x \in (0,1)} \{(x, 2)\}
$$

Algorithm 8: Sampling a Semi-random variable via RJMCMC

```
Initialize $(x^{(0)}, m^{(0)})$, with $x \begin{cases} \in \Omega_1 = \{\frac{1}{2}\}, \text{ if } m = 1 \\ \in \Omega_2 = (0,1), \text{ if } m = 2 \end{cases}$.
for $t = 1, \cdots, T$ do
    if $m^{(t)} = 1$ then
        Increase dimension:
        Propose $m' = 2$ with probability $\rho_{1,2} = 1$;
        Simulate $u \sim g_{1,2}(u) = \text{Beta}\,(u; \alpha = 1/2, \beta = 1/2)$;
        Set $x' = \psi_1(x^{(t)}, u) = u$ and $u' = \psi_2(x^{(t)}, u) = \emptyset$;
        Compute Jacobian:$J_\psi(x^{(t)}, u) = \left|\frac{\partial\psi(x^{(t)},u)}{\partial(x^{(t)},u)}\right| = \left|\frac{\partial x'}{\partial u}\right| = 1$
        Compute acceptance: $\alpha\left(x', m' \mid x^{(t)}, m^{(t)}\right) =$
         $\min\left\{1, \frac{\pi(x',m')\rho_{m',m^{(t)}} g_{2,1}(u')}{\pi(x^{(t)},m^{(t)})\rho_{m^{(t)},m'} g_{1,2}(u)} J_\psi(x^{(t)}, u)\right\} = \min\left\{1, \frac{4x'/3}{\text{Beta}(x';\alpha,\beta)/3}\right\}$
        - with $g_{2,1}(u') = \rho_{m',m^{(t)}} = \rho_{m^{(t)},m'} = 1$
    end
    if $m^{(t)} = 2$ then
        Decrease dimension:
        Propose $m' = 1$ with probability $\rho_{2,1} = 1$;
        Simulate $u \sim g_{2,1}(\emptyset) = 1 \implies u = \emptyset$;
        Set $x' = \psi_1(x^{(t)}, \emptyset) = \frac{1}{2}$ and $u' = \psi_2(x^{(t)}, \emptyset) = x^{(t)}$;
        Compute Jacobian:$J_\psi(x^{(t)}, u) = \left|\frac{\partial\psi(x^{(t)},u)}{\partial(x^{(t)},u^{(t)})}\right| = \left|\frac{\partial u'}{\partial x^{(t)}}\right| = 1$
        Compute acceptance: $\alpha\left(x', m' \mid x^{(t)}, m^{(t)}\right) =$
         $\min\left\{1, \frac{\pi(x',m')\rho_{m',m^{(t)}} g_{1,2}(u')}{\pi(x^{(t)},m^{(t)})\rho_{m^{(t)},m'} g_{2,1}(u)} J_\psi(x^{(t)}, u)\right\} = \min\left\{1, \frac{\text{Beta}\left(x^{(t)};\alpha,\beta\right)/3}{4x^{(t)}/3}\right\}$
        - with $g_{1,2}(u') = \text{Beta}\left(x^{(t)}; \alpha, \beta\right); \rho_{m',m^{(t)}} = \rho_{m^{(t)},m'} = 1$
    end
    Simulate $k \sim U(0,1)$.
    if $k < \alpha\left(x', m' \mid x^{(t)}, m^{(t)}\right)$ then
        $(x^{(t+1)}, m^{(t+1)}) = (x', u')$
    end
    else: $(x^{(t+1)}, m^{(t+1)}) = (x^{(t)}, m^{(t)})$
end
```

- proposal matches the change in dimension $\dim(\Omega_1 \times \mathcal{U}_{1,2}) = \dim(\Omega_2 \times \mathcal{U}_{2,1}) = 1$.

7.3.4 Sampling from Mixture Models via RJMCMC

Consider a Mixture Gaussian problem with the **observation model**:

$$p(y \mid \mu, \sigma, w, m) = \prod_{i=1}^{n}\left[\sum_{j=1}^{m} w_j N\left(y_i; \mu_j, \sigma_j^2\right)\right]$$

with: $\theta_i = (\mu_i, \sigma_i, w_i)$ and
- mixture means $\mu = (\mu_1, \ldots, \mu_m)$,
- mixture standard deviations $\sigma = (\sigma_1, \ldots, \sigma_m)$,
- mixture weights $w = (w_1, \ldots, w_m)$ s.t. $\sum_i w_i = 1$.

Priors: $\pi(\mu, \sigma, w|m)$: $\begin{cases} \mu \sim N(\mu_0 1_m, v_0 1_m), \text{ with } \mu_0 = 20, v_0 = 10^2 \\ \sigma_j \overset{iid}{\sim} \text{Gamma}(1.5, 0.5) \\ w \sim \text{Dirichlet}(\alpha 1_m) \\ m \sim \text{Possion}(\lambda) \end{cases}$

Targeting Posterior:

$$\begin{aligned} &\pi(\theta, m \mid y) \propto p(y \mid \theta, m)\pi(\theta \mid m)\pi(m) \\ &\quad \propto p(y \mid \mu, \sigma, w, m) \times \text{Dirichlet}\left(w; \alpha 1_m\right) \times \prod_{j=1}^{m} N\left(\mu_j; \mu_0, v_0\right) \times \text{Gamma}\left(\sigma_j; 1.5, 0.5\right) \times \text{Poisson}(m; \lambda) \times m! \end{aligned}$$

Predictive posterior: with the RJMCMC sample $(\theta^{(t)}, m^{(t)})_{t=1}^{T}$,

$$\widehat{p\left(y' \mid y\right)} = \frac{1}{T}\sum_{t=1}^{T} p\left(y' \mid \theta^{(t)}, m^{(t)}\right)$$

Key notes of Algorithm 9:
(i) $\rho_{m,m'} = \frac{1}{5}$;
(ii) For reverse proposal of move = 1 (increase dimension):
- $\frac{1}{m+1}$ = probability of simulating the i-th entry of μ, σ to remove;
- $\frac{1}{m}$ = probability of simulating w_j to add with w_i.
(iii) For reverse proposal of move = 2 (decrease dimension):
- $\frac{1}{m-1}$ = probability of simulating w_j to split;
- $\frac{1}{w_i+w_j}$ = probability of simulating the additional weight $w'_{m'}$.

.

Algorithm 9: Sampling Mixture Models via RJMCMC

```
Initialize $(\mu^{(0)}, \sigma^{(0)}, w^{(0)}, m^{(0)})$ from prior.
for $t = 1, \cdots, T$ do
    Randomly choose move $\sim U\{1, 2, 3, 4, 5\}$.
    if $move = 1$ then
        Increase dimension by 3: set $m' = m + 1$
        Simulate $\mu'_{m+1}, \sigma'_{m+1} \sim q_{\mu\sigma}\left(\mu'_{m+1}, \sigma'_{m+1}\right)$ (Normal-Gamma prior);
        Split weight $j \sim U(1, 2, \cdots, m)$ by: $w'_{m+1} \sim U(0, w_j)$ and set
        $w'_k = \begin{cases} w_k & k = 1, .., m, k \neq j \\ w_k - w'_{m+1} & k = j \qquad \text{(s.t. } \sum_{k=1}^{m} w_k = 1) \\ w'_{m+1} & k = m + 1 \end{cases}$
        Set $\theta'_{m+1} = \left(\mu'_{m+1}, \sigma'_{m+1}, w'_{m+1}\right)$ and $\theta' = \theta \cup \theta'_{m+1}$;
        Compute forward proposal:
        $Q(\mu', \sigma', w', m' | \mu, \sigma, w, m) = \rho_{m,m'} q_{\mu\sigma}\left(\mu'_{m+1}, \sigma'_{m+1}\right) \frac{1}{m} \times \frac{1}{w_j}$;
        Compute reverse proposal: $Q(\mu, \sigma, w, m | \mu', \sigma', w', m') = \rho_{m',m} \frac{1}{m} \times \frac{1}{m+1}$;
        Compute acceptance: $\alpha(\theta', m' | \theta, m) = \min\left\{1, \frac{\pi(\theta', m' | y) \times \text{reverse}}{\pi(\theta, m | y) \times \text{forward}}\right\} =$
        $\min\left\{1, \frac{\pi(\mu', \sigma', w', m' | y) Q(\mu, \sigma, w, m | \mu', \sigma', w', m')}{\pi(\mu, \sigma, w, m | y) Q(\mu', \sigma', w', m' | \mu, \sigma, w, m)}\right\}$
    end
    if $move = 2$ then
        Decrease dimension by 3: set $m' = m - 1$
        if $m' = 0$ then
            Set $\theta^{(t+1)} = \theta^{(t)}$
        end
        else: Simulate $i \sim U\{1, 2, \cdots, m\}$ and set $\mu' = \mu_{-i}, \sigma' = \sigma_{-i}$;
        Simulate $j \sim U\{1, \cdots, i - 1, i + 1, \cdots, m\}$, replace $w_j \leftarrow w_j + w_i$ and set $w' = w_{-i}$;
        Compute acceptance:
        $\alpha(\theta', m | \theta, m) = \min\left\{1, \frac{\pi(\mu', \sigma', w', m' | y) \rho_{m',m} q_{\mu\sigma}(\mu_i, \sigma_i) \times \frac{1}{m-1} \times \frac{1}{w_i + w_j}}{\pi(\mu, \sigma, w, m | y) \rho_{m,m'} \times \frac{1}{m(m-1)}}\right\}$
    end
    if $move = \{3, 4, 5\}$ then
        Fixed dimension update of $\mu, \sigma, w$ respectively
    end
    Simulate $k \sim U(0, 1)$.
    if $k < \alpha\left(x', m' \mid x^{(t)}, m^{(t)}\right)$ then
        $(x^{(t+1)}, m^{(t+1)}) = (x', u')$
    end
    else: $(x^{(t+1)}, m^{(t+1)}) = (x^{(t)}, m^{(t)})$
end
```

8 Dirichlet Process (DP)

Motivation: Making less model assumptions as more data is available, hence going from parametric to non-parametric Bayesian approach.
Setup: By De Finitte's, $y_{1:n}$ is IES $\implies$ $p(y_{1:n}) = \int_{\Omega} p(y_{1:n}|\theta) dG(\theta)$, this distribution $dP(y_{1:n}))$

and the posterior $d\pi(\theta|y_{1:n})dG(\theta)$ all exist and are unique. Let:
- $\mathcal{G}$: the **space of probability distributions**, $G \in \mathcal{G}$;
- G: the **unknown true generative distribution** for the parameter $\theta \sim G$;
- Π: a **probability distribution over** $\mathcal{G}$, i.e. $G \sim \Pi$, hence $d(\Pi(G))$ is the prior for the prior $\pi(\theta|G)$;

Then,
- the **joint distribution (likelihood)** is: $d\Pi(G,\theta) = dG(\theta)d\Pi(G)$,
- the **posterior** is: $d\Pi(G,\theta|y) \propto p(y|\theta)dG(\theta)d\Pi(\theta)$,
- the **marginal distribution** is $d\pi(\theta) \propto \int_{\mathcal{G}} dG(\theta)d\Pi(G)$.

8.1 Dirichlet Process (DP) and the Chinese Restaurant Process (CRP)

8.1.1 Dirichlet Distribution

$w \sim Dir(\alpha_{1:M})$ with density:

$$\pi(w_{1:M}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} w_1^{\alpha_1 - 1} \cdots w_M^{\alpha_M - 1}$$

where $w \in \{(0,1)^M : \sum_{k=1}^M w_k = 1\}$.
Property of Dirichlet Distribution:
(i) **Agglomeration**: $w_1 + w_2, w_3, \cdots, w_M \sim Dir(\alpha_1 + \alpha_2, \alpha_3, \cdots, \alpha_M)$;
(ii) **Conjugate Prior for the multinomial distribution:** if $n_{1:M} \sim Multinomial(n, w)$, then $w|n_{1:M} \sim Dir(\alpha_1 + n_1, \cdots, \alpha_M + n_M)$.

8.1.2 Multinomial Dirichlet Process (MDP)

Algorithm 10: Multinomial Dirichlet Process (MDP)

```
Initialize base distribution: $H$.
Initialize dimension $M \geq 1$ and $\alpha \geq 0$.
for $k = 1, \cdots, M$ do
    Sample $\theta_k^* \sim H$
end
Sample $w_1, \ldots, w_M \sim$ Dirichlet($\alpha/M$).
Set $dG_M(\theta) = \sum_{k=1}^M w_k \delta_{\theta_k^*}(d\theta)$ or equivalently $G_M(\theta) = \sum_{k=1}^M w_k \delta_{\theta_k^*}$.
```

.

Prop 8.3: The random distribution G_M is "centered" in the base distribution, i.e. $E(G_M) = H(A) = \int_\Omega \mathbb{I}_{\theta \in A} h(\theta) d\theta$.
Proof:

$$\begin{aligned}
E\left(G_M(A)\right) = E\left[\int_\Omega \mathbb{I}_{\theta \in A} dG_M(\theta)\right] &= \int_A \sum_{k=1}^M w_k \delta_{\theta_k^*}(d\theta) = \sum_{k=1}^M w_k \mathbb{I}_{\theta_k^* \in A} \\
&= \sum_{k=1}^M E\left(w_k \mathbb{I}_{\theta_k^* \in A}\right) = \sum_{k=1}^M E\left(w_k\right) E\left(\mathbb{I}_{\theta_k^* \in A}\right), \quad E(w_k) = \frac{\alpha_k/M}{\sum_j (\alpha_j/M)} \\
&= \sum_k \frac{\alpha_k}{\sum_j \alpha_j} H(A) = H(A) \quad \text{[EOP]}
\end{aligned}$$

• Both MDP and DP puts atoms of probability mass w_k at points θ_k^* in Ω. (Hence, $\forall \theta_1, \theta_2 \sim G_M \sim \Pi_M(\alpha, H), P(\theta_1 = \theta_2) > 0$)

8.1.3 Dirichlet Process (DP)

$G \sim \Pi(\alpha, H)$ is a DP iff $\forall$ partition $A_1, \cdots, A_r$ of Ω (with $A_k \in \mathcal{B}$), $G\left(A_1\right), \ldots, G\left(A_r\right) \sim$ Dirichlet $\left(\alpha H\left(A_1\right), \ldots, \alpha H\left(A_r\right)\right)$.

• For each partition (A_k), the DP G is unique and $\sum_k G(A_k) = 1$.

• $G \sim \Pi(\alpha, H)$ (DP exists) $\implies \theta_1, \cdots, \theta_n \overset{iid}{\sim} G$ is IES. Conversely, Given $\theta_1, \cdots, \theta_n$ an IES and given α, H, the DP G with distribution $d\Pi(G)$ exist.

8.1.4 Properties of DP

Prop 8.11: $G \sim \Pi(\alpha, H) \implies \forall A \in \mathcal{B}, E(G(A)) = H(A)$.
Proof: Since $H(A^C) = 1 - H(A)$, we have,

$$\begin{aligned} G(A), G\left(A^c\right) &\sim \operatorname{Dirichlet}(\alpha H(A), \alpha(1-H(A))) \\ &\sim \operatorname{Beta}(\alpha H(A), \alpha(1-H(A))) \\ \implies E(G(A)) &= \frac{\alpha H(A)}{\alpha H(A) + \alpha(1-H(A))} \end{aligned}$$

• Dirichlet Distribution with two components is a Beta distribution.

Prop 8.12:
(i) $G \sim \Pi(\alpha, H)$ and $\theta \sim G \implies \theta \sim \Pi$ marginally.
Proof: $\Pr(\theta \in A) = E_G[P(\theta \in A|G)] = E_G\left(E_{\theta|G}\left(\mathbb{I}_{\theta \in A} \mid G\right)\right) = E(G(A)) = H(A)$ [EOP]
(ii) $G \sim \Pi(\alpha, H)$ and $\theta \sim G \implies \forall B \subseteq A$ (measurable), $\Pr(\theta \in B \mid \theta \in A) = \frac{\Pr(\theta \in B)}{\Pr(\theta \in A)} = \frac{H(B)}{H(A)}$

8.1.5 DP Generative Models and Predictive Distributions

Generative model:

$$\begin{aligned} &G \sim \Pi(\alpha, H) \\ &\theta_i \sim G, i = 1, 2, \ldots, n \end{aligned}$$

Marginal distribution: $d\pi(\theta) = d\pi\left(\theta_n \mid \theta_{1:n-1}\right) d\pi\left(\theta_{n-1} \mid \theta_{1:n-2}\right) \ldots d\pi\left(\theta_1\right)$,
with,
- $d\pi\left(\theta_1\right) = H(d\theta_1)$, the marginal distribution as in **Prop 8.12**.
- $d\pi\left(\theta_{i+1} \mid \theta_{1:i}\right)$ are predictive distributions.

Predictive distributions: $d\pi\left(\theta_{i+1} \mid \theta_{1:i}\right) = \tilde{H}_j\left(d\theta_{j+1}\right) = \frac{\alpha H(d\theta_{j+1}) + \sum_{i=1}^j \delta_{\theta_i}(d\theta_{j+1})}{\alpha + j}$

Marginal density: $\pi(\theta) = \pi\left(\theta_n \mid \theta_{1:n-1}\right) \pi\left(\theta_{n-1} \mid \theta_{1:n-2}\right) \ldots \pi\left(\theta_1\right) = \prod_{j=0}^{n-1} \frac{\alpha h(\theta_{j+1}) + \sum_{i=1}^j \delta_{\theta_i}(\theta_j)}{\alpha + j}$
• $\sum_{i=1}^j \delta_{\theta_i}\left(\theta_j\right) = 0$ if $j = 0$.

Prop 8.19: If $\theta_{1:n} \sim G \sim \Pi(\alpha, H)$, then,

$$G \mid \theta_{1:n} \sim \Pi\left(\tilde{\alpha}_n, \tilde{H}_n\right) \text{ with,} \tilde{\alpha}_n = \alpha + n, \quad \tilde{H}_n = \frac{\alpha H + \sum_{i=1}^n \delta_{\theta_i}}{\alpha + n}$$

and marginal density defined as above.
Proof: (by induction, here we prove Prop 8.21 on the notes, which is the $n=1$ case of Prop 8.19.)
WTS: conditional $G \mid \theta_1 \sim DP\left(\alpha+1, \frac{\alpha H+\delta_{\theta_1}}{\alpha+1}\right)$ and the "jump density" $d\pi\left(\theta_2 \mid \theta_1\right)=\frac{\alpha H\left(d\theta_2\right)+\delta_{\theta_1}\left(d\theta_2\right)}{\alpha+1}$.
Suppose:
(i) $A_{1:r}$ a partition of Ω, with $\theta_1 \in A_j$ for fixed $j \in\{1, \cdots, r\}$;
(ii) Let $g_i=G(A_i) \geq 0, i=1, \cdots, r$ s.t. $\sum_{i=1}^r g_i=1$, denote $g=(g_{1:r})$;
(iii) Let $H_i \equiv H(A_i)$;
(iv) Let $f(g)=$ Dirichlet $(g ; \alpha H_1, \ldots \alpha H_r)$ be the Dirichlet density of g for fixed partition $A_{1:r}$.
Goal: want $f(g|\theta_1) \propto \pi(\theta_1|g) f(g)$.

$$
\begin{aligned}
\pi\left(\theta_1 \mid g\right) & =\pi\left(\theta_1, \theta_1 \in A_j \mid g\right), \quad \because \theta_1 \in A_j \text{ is given} \\
& =\pi\left(\theta_1 \mid \theta_1 \in A_j, g\right) \pi\left(\theta_1 \in A_j \mid g\right) \overset{(*)}{=} \pi\left(\theta_1 \mid \theta_1 \in A_j\right) \pi\left(\theta_1 \in A_j \mid g\right. \\
& =h\left(\theta_1 \mid \theta_1 \in A_j\right) g_j, \text{ by Prop 8.12 (ii)} \\
\Longrightarrow & f\left(g \mid \theta_1\right) \propto h\left(\theta_1 \mid \theta_1 \in A_j\right) \times g_j \times g_1^{\alpha H_1-1} \times \ldots \times g_r^{\alpha H_r-1} \\
& \propto g_1^{\alpha H_1-1+\mathbb{I}_{\theta_1 \in A_1}} \times \ldots \times g_r^{\alpha H_r-1+\mathbb{I}_{\theta_1 \in A_r}} \\
\Longrightarrow & G\left(A_1\right), \ldots, G\left(A_r\right) \mid \theta_1 \sim \operatorname{Dirichlet}\left(\alpha H_1+\mathbb{I}_{\theta_1 \in A_1}, \ldots, \alpha H_r+\mathbb{I}_{\theta_1 \in A_r}\right)
\end{aligned}
$$

•$(*)$ because $\theta_1 \perp g \mid \theta_1 \in A_j$, as A_j contains all information about θ_1.
Hence taking $\tilde{\alpha}_n=\alpha+n, \quad \tilde{H}_n=\frac{\alpha H+\sum_{i=1}^n \delta_{\theta_i}}{\alpha+n}$, we ensure $\tilde{\alpha}_1 \tilde{H}_1\left(A_j\right)=\alpha H\left(A_j\right)+\mathbb{I}_{\theta_1 \in A_j}, \forall j$ and the conditional is proved.

To show the "jump density", observe that it is the predictive posterior density $\theta_2|\theta_1 \sim \tilde{H}_1$, and with $\theta_2 \sim G|\theta_1$ and $G|\theta_1 \sim \Pi(\tilde{\alpha}_1, \tilde{H}_1)$, $\theta_2 \sim \tilde{H}_1$ marginally by Prop 8.12. [EOP]

Updated base distribution for $G|\theta_1$: $\tilde{H}_1\left(d\theta_2\right)=\frac{\alpha}{\alpha+1} h\left(\theta_2\right) d\theta_2+\frac{1}{\alpha+1} \delta_{\theta_1}\left(d\theta_2\right)$
• To simulate the marginal (θ_1, θ_2), we have 2 ways:
(i) simulate $G \sim \Pi(\alpha, H)$, then simulate $\theta_1, \theta_2 \sim G$; OR
(ii) simulate $\theta_1 \sim H$ and simulate $\theta_2|\theta_1 \sim \tilde{H}_1$:
- by simulating $\theta_2 \sim h$ with $\frac{\alpha}{\alpha+1}$ or just set $\theta_2=\theta_1$ otherwise.
• This leads to the repeated value problem (having too many $\theta_2=\theta_1$), hence we update the simulation process as in Algorithm 11.

8.1.6 Sequential Simulation and the Repeated Value Problem

```
Algorithm 11: Sequential Simulation
  Initialize $\alpha, H$(base distribution).
  Initialize the number of clusters: $K = 1$.
  Initialize the set of indices within clusters: $S_1 = \{1\}$.
  Initialize the set of index-sets for each cluster: $S = \{S_1\}$.
  Simulate $\theta_1^* \sim H$.
  for $j = 1, \cdots, n-1$ do
      simulate $u \sim U(0, 1)$
      if $u \leq \frac{\alpha}{\alpha+j}$ then
          Simulate $\theta_{K+1}^* \sim H$, set $S_{K+1} = \{j+1\}$ and $S = S \cup S_{K+1}$.
          $K \leftarrow K + 1$
          (Generate a new $\theta_{j+1} = \theta_{K+1}^*$ and start a new cluster $K + 1$.)
      end
      else
          for $k = 1, \cdots, K$ do
              Set $n_k = |S_k|$ and simulate $k^* \sim U(\frac{n_1}{j}, \cdots, \frac{n_K}{j})$ and Set $S_{k^*} \leftarrow S_{k^*} \cup \{j + 1\}$
              (Generate $\theta_{j+1}$ by equating it to an old $\theta$, with the probability weighted by the size of the index-sets)
          end
      end
  end
```

- Here, the base distribution is reexpressed as $\tilde{H}_n = \frac{\alpha}{\alpha+n} H + \frac{1}{\alpha+n} \sum_{k=1}^{K} n_k \delta_{\theta_k^*}$

Joint distribution of (θ^*, S):

$$d\pi\left(\theta^*, S\right) = \pi_S(S) d\pi\left(\theta^* \mid S\right) = \pi_S(S) \prod_{k=1}^{K} H(d\theta_k^*) = \pi_S(S) \prod_{k=1}^{K} h(\theta_k^*) d\theta_k^*$$

with
- $(\theta^*, S) \in \Omega^* = \bigcup_{S \in \Xi_{[n]}} \Omega^{K(S)} \times \{S\}$;
- $\Xi_{[n]}$ set of all partitions of $[n] = \{1, \cdots, n\}$;
- $K(S)$ the number of clusters in the partition $S = (S_1, \cdots, S_{K(S)})$ of $[n]$.

•

$$\Pr\left((\theta^*, S) \in A\right) = \sum_{S \in \Xi_{[n]}} \int_{\Omega^{K(S)}} \mathbb{I}_{(\theta^*, S) \in A} \left[P_{\alpha,[n]}(S) \prod_{k=1}^{K(S)} h\left(\theta_k^*\right) \right] d\theta_1^*, \ldots, d\theta_{K(S)}^*$$

where,
- any subset $A \in \Omega^*$;
- $P_{\alpha,[n]}(S = s) = \pi_S(S = s)$ is defined as the proportion of times $S = s$ is realized by the process in an infinite sequence of trials.

8.1.7 Chinese Restaurant Process

- To compute $P_{\alpha,[n]}(S = s) = \pi_S(S = s)$ from above.

Algorithm 12: Chinese Restaurant Process (CRP)

```
Initialization: After the first customer arrives, there are $j = 1$ customers seated at table
 $k = 1$, and $K_1 = 1$ tables are occupied.
for $j = 1, \cdots, n-1$ do
    simulate $u \sim U(0,1)$. if $u \leq \frac{\alpha}{\alpha+j}$ then
        The $(j+1)-$th customer arrives and chooses a new table $K_{j+1} = K_j + 1$
    end
    else
        The $(j+1)-$th customer arrives and chooses table $k$ with probability $\frac{n_k^{(j)}}{\alpha+j}$
    end
    After the $j+1$-th customer is seated, there are $n_k^{(j+1)}$ people at table $k$ and $K_{j+1}$
     tables are occupied.
end
After all $n$ customers are seated, $K_n := K$ tables are occupied, each table seats $n_k^{(n)} := n_k$
 customers.
Return: the set of customer-lists at each table: $S = (S_k)_{k=1}^K$, with $|S_k| = n_k$.
```

.

Expected number of clusters: $E(K) = \sum_{i=1}^{n} \frac{\alpha}{\alpha+i-1}$

- proof omitted, to be shown in PS4.

Probability of the Partition: S is a partition of the n customers in K clusters, with probability:

$$P_{\alpha,[n]}(S) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \alpha^K \prod_{k=1}^{K} \Gamma(n_k)$$

Proof: Suppose ultimately n customers occupied K tables, then there are $K-1$ customers that chose a new table after the first customer seated. For each table k, we have $S_k = \{i_1, i_2, \cdots, i_{n_k}\}$ the list of customers seated at table k. Customer i_1 chose table k with probability $\frac{\alpha}{\alpha+i_1-1}$, whereas the rest "followers" seated at table k with probability $\left(\frac{(j-1)}{\alpha+i_j-1}\right)_{j=2}^{n_k}$.

$$\begin{aligned} P_{\alpha,[n]}(S) &= \alpha^{K-1} \prod_{k=1}^{K} (n_k - 1)! \prod_{i=2}^{n} \frac{1}{\alpha+i-1} \\ &= \alpha^{K} \prod_{k=1}^{K} (n_k - 1)! \prod_{i=1}^{n} \frac{1}{\alpha+i-1} \\ &= \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \alpha^K \prod_{k=1}^{K} \Gamma(n_k) \quad \text{[EOP]} \end{aligned}$$

- $\sum_{S \in \Xi_{[n]}} P_{\alpha,[n]}(S) = 1$.
- Example omitted, see Remark 8.40 at P108 on the notes.

8.2 Inference for a DP Mixture

8.2.1 DP Mixture in General

Mixture observation model: $f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta_i) = \prod_{k=1}^{K} f(y_{S_k}|\theta_k^*) = f(y|\theta^*, S)$, where:
- θ's are equal within each cluster k (θ_k^*);
- $f(y_{S_k}|\theta_k^*) = \prod_{i\in S_k} f(y_i|\theta_i = \theta_k^*)$.

DP Prior for the Mixture: $\theta \sim G$ with $G \sim \Pi(\alpha, H)$.

DP Posterior for the Mixture: With $\theta = \theta(\theta^*, S)$ and $(\theta^*, S) \in \Omega^* = \bigcup_{S\in\Xi_{[n]}} \Omega^{K(S)} \times \{S\}$,

$$\pi\left(\theta^*, S \mid y\right) \propto f\left(y \mid \theta^*, S\right) \pi\left(\theta^* \mid S\right) P_{\alpha,[n]}(S)$$

$$\text{under base dist'n: } \propto \prod_{k=1}^{K} \left[\alpha\Gamma\left(n_k\right) h\left(\theta_k^*\right) f\left(y_{S_k} \mid \theta_k^*\right)\right], \text{ by joint dist'n and partition prob}$$

8.2.2 DP Gaussian Mixture

In the Gaussian setting, $\theta_k^* = (\mu_k^*, \sigma_k^{*2})$.
DP Prior for Gaussian Mixture under base distribution: $h\left(\theta_k^*\right) = h_\mu\left(\mu_k^*\right) h_\sigma\left(\sigma_k^{*2}\right)$, with,
- $h_\mu\left(\mu_k^*\right) = N\left(\mu_k^*; \mu_0, \sigma_0^2\right)$ and $h_\sigma\left(\sigma_k^{*2}\right) = \text{I}\Gamma\left(\sigma_k^{*2}; \alpha_0, \beta_0\right)$ (Inverse Gamma)
• Normal-InverseGamma prior is a conjugate prior, allowing us to integrate out $\theta^* = (\mu^*, \sigma^*)$ completely and sample the discrete $\pi(S|y)$. (so-called *collapsed Gibbs sampler*).

DP Observation model for Gaussian Mixture: $y_i|S, \mu^*, \sigma^* \sim N(\mu_{k_i}^*, \sigma_{k_i}^{*\;2})$

DP Posterior for Gaussian Mixture:

$$\pi\left(S, \mu^*, \sigma^* \mid y\right) \propto f\left(y; \mu^*, \sigma^*, S\right) \pi\left(\mu^*, \sigma^* \mid S\right) P_{\alpha,[n]}(S), \text{ by DP posterior for mixture}$$

$$\propto \prod_{k=1}^{K} \prod_{i\in S_k} N\left(y_i; \mu_k^*, \sigma_k^{*2}\right) \times \prod_{k=1}^{K} N\left(\mu_k^*; \mu_0, \sigma_0^2\right) \text{I}\Gamma\left(\sigma_k^{*2}; \alpha_0, \beta_0\right) \times \alpha^K \prod_{k=1}^{K} \Gamma\left(n_k\right)$$

8.2.3 Gibbs Sampler for DP Gaussian Mixture

Algorithm 13: Gibbs Sampler for Gaussian Mixture targeting $\pi(\theta^*, S|y)$

```
Initialize partition $S = (S_k)_{k=1}^K$, the mixture parameters $\theta^* = (\mu^*, \sigma^*)$, and $[n] = \{1, \cdots, n\}$
Initialize Priors as in DP Prior for Gaussian Mixture above and the base distribution $H$.
Sample Mixture Parameters: Fix the partition $S$,
for $t = 1, \cdots, T$ do
    Sample $(\mu^*, \sigma^*)$ iteratively:
    (i) Simulate $\mu_k^*|\sigma_k^*, y \sim N(a, b)$, with $a = \left(\frac{n_k \bar{y}_k}{\sigma_k^{*2}} + \frac{\mu_0}{\sigma_0^2}\right)$,   $b = \left(\frac{n_k}{\sigma_k^{*2}} + \frac{1}{\sigma_0^2}\right)^{-1}$, $n_k = |S_k|$
     and $\bar{y}_k = \frac{1}{n_k} \sum_{i \in S_k} y_i$
    (ii) Simulate $\sigma_k^*|\mu_k^*, y \sim I\Gamma(c, d)$, with $c = \alpha_0 + n_k/2$,   $d = \beta_0 + \frac{1}{2} \sum_{i \in S_k} (y_i - \mu_k^*)^2$
end
Sample Partition: Fix the Mixture Parameters $(\mu^*, \sigma^*)$,
for $j \in [n]$ do
    Remove the $j$-th and form $(\theta_{-j}^*, S^{-j})$, where $S^{-j} = \left(S_k^{-j}\right)_{k=1}^{K^{-j}}$ and $\theta_{-j}^* = \left(\theta_{-j,k}^*\right)_{k=1}^{K^{-j}}$.
    • $K^{-j} = \begin{cases} K-1, \text{ if a cluster is empty (hence dropped) after removing } j \\ K, o/w \end{cases}$
    Re-simulate the $j$-th: $\theta_{-j,K^{-j}+1}^* \sim H$, and simulate $u \sim U(0,1)$.
    Set $\theta^* = (\theta_{-j}^*, \theta_{-j,K^{-j}+1}^*)$
    if $u \leq \alpha/\alpha+n-1$ then
        Set $K \leftarrow K^{-j} + 1$, $S_K = \{j\}$, and $S = S^{-j} \cup S_K$. Simulate $y_j \sim f(\cdot|\theta_{-j,K^{-j}+1}^*)$.
        (Put $j$ in a new cluster $S_{K^{-j}+1}$ with probability $\alpha f\left(y_j \mid \theta_{-j,K^{-j}+1}^*\right)$).
    end
    else
        Set $K \leftarrow K^{-j}$, and set $n_k^{-j} = |S_k^{-j}|, k = 1, \cdots, K$.
        Simulate $k^* \sim U(\frac{n_1^{-j}}{n-1}, \cdots, \frac{n_k^{-j}}{n-1})$ ($\because$ by removing $j$-th, there are $n-1$ remaining.)
        Set $S_{k^*}^{-j} \leftarrow S_{k^*}^{-j} \cup \{j\}$ and set $S \leftarrow S^{-j}$. Simulate $y_j \sim f(\cdot|\theta_{-j,k^*}^*)$.
        (Put $j$ in an old cluster $S_k^{-j}$ with probability $n_k^{-j} f(y_j|\theta_k^*)$ for $k = 1, \cdots, K^{-j}$.)
    end
    (For each given $j$, set a permutation of $[n]$ with $j$ being the last $\{j_1, \cdots, j_{n-1}, j\}$. Can
    also update the $j$th-removed-partition $S^{-j}$ wrt $j_1, \cdots, j_{n-1}$, and simultaneously sample
    $y_{-j}$. See details Prop 8.52 at P112 in the notes.)
end
Return: $(\theta^*, S, y)$
```

.

Conditional Probability for $j \in S_k$ given everything else:

$$\Pr\left(j \in S_k \mid S^{-j}, \theta_{-j}^*, \theta_{-j,K^{-j}+1}^*, y\right) \propto \begin{cases} n_k^{-j} f\left(y_j \mid \theta_{-j,k}^*\right), & \text{if } k \in \left\{1, \ldots, K^{-j}\right\} \\ \alpha f\left(y_j \mid \theta^* - j, k\right), & \text{if } k = K^{-j} + 1 \end{cases}$$

with $\theta_{-j,K^{-j}+1}^* \sim H$ an independent draw from the prior under base distribution H.

8.2.4 DP Mixture Predictive Posterior

$$
\begin{aligned}
& p\left(y' \mid y\right) = \sum_{S \in \Xi_{[n]}} \int_{\Omega^{K(S)}} p\left(y' \mid \theta^*, S\right) \pi\left(\theta^*, S \mid y\right) d\theta^*_1, \ldots, d\theta^*_{K(S)} \\
\Longrightarrow\ & p\widehat{\left(y' \mid y\right)} = \frac{1}{T}\sum_{t=1}^{T} f\left(y' \mid \theta^{*,(t)}, S^{(t)}\right) = \frac{1}{T}\sum_{t=1}^{T}\int_{\Omega} p(y', \theta' | \theta^{*,(t)}, S) d\theta' = \frac{1}{T}\sum_{t=1}^{T}\int_{\Omega} f\left(y' \mid \theta'\right) p\left(\theta' \mid \theta^{*,(t)}, S\right) d\theta' \\
& = \frac{1}{T}\sum_{t=1}^{T}\int_{\Omega} f\left(y' \mid \theta'\right)\left[\frac{\alpha h\left(\theta'\right) + \sum_{k=1}^{K^{(t)}} \delta_{\theta_k^{*,(t)}}\left(\theta'\right)}{\alpha + n}\right] d\theta', \quad \because \theta' \sim H \text{ and } \theta^{*,(t)} = (\theta_k^{*,(t)})_{k=1}^{K^{(t)}} \\
& = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\alpha}{\alpha+n} p(y') + \sum_{k=1}^{K^{(t)}} f\left(y' \mid \theta_k^{*,(t)}\right) \frac{n_k^{(t)}}{\alpha+n}\right], \quad \because \int_{\Omega} f(y'|\theta')\delta_{\theta_k^{*,(t)}}(\delta')d\theta' = f(y'|\theta_k^{*,(t)}), \forall k \\
& \simeq \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\alpha}{\alpha+n} f\left(y' \mid \theta^{*,(t)}_{K^{(t)}+1}\right) + \sum_{k=1}^{K^{(t)}} f\left(y' \mid \theta_k^{*,(t)}\right) \frac{\left|S_k^{(t)}\right|}{\alpha+n}\right], \quad \because \hat{p}(y') \leftarrow f\left(y' \mid \theta^{*,(t)}_{K^{(t)}+1}\right) \text{ unbiased}
\end{aligned}
$$

with $K^{(t)} = K(S^{(t)})$ and $\theta^*_{K+1} \overset{iid}{\sim} H$.