1.1– Outline

- Mixture and composition of kernels.
- “Hybrid” algorithms.
- Examples
2.1– Mixture of proposals

• If $K_1$ and $K_2$ are $\pi$-invariant then the mixture kernel

$$K(\theta, \theta') = \lambda K_1(\theta, \theta') + (1 - \lambda) K_2(\theta, \theta')$$

is also $\pi$-invariant.

• If $K_1$ and $K_2$ are $\pi$-invariant then the composition

$$K_1 K_2(\theta, \theta') = \int K_1(\theta, z) K_2(z, \theta') \, dz$$

is also $\pi$-invariant.
2.1– Mixture of proposals

- **Important**: It is not necessary for either $K_1$ or $K_2$ to be irreducible and aperiodic to ensure that the mixture/composition is irreducible and aperiodic.

- For example, to sample from $\pi(\theta_1, \theta_2)$ we can have
  
  - the kernel $K_1$ updates $\theta_1$ and keeps $\theta_2$ fixed whereas
  - the kernel $K_2$ updates $\theta_2$ and keeps $\theta_1$ fixed.
2.2– Applications of Mixture and Composition of MH algorithms

- For $K_1$, we have $q_1 (\theta, \theta') = q_1 ((\theta_1, \theta_2), \theta'_1) \delta_{\theta_2} (\theta'_2)$ and

$$r_1 (\theta, \theta') = \frac{\pi (\theta'_1, \theta_2) q_1 ((\theta'_1, \theta_2), \theta_1)}{\pi (\theta_1, \theta_2) q_1 ((\theta_1, \theta_2), \theta'_1)} = \frac{\pi (\theta'_1 | \theta_2) q_1 ((\theta'_1, \theta_2), \theta_1)}{\pi (\theta_1 | \theta_2) q_1 ((\theta_1, \theta_2), \theta'_1)}$$

- For $K_2$, we have $q_2 (\theta, \theta') = \delta_{\theta_1} (\theta'_1) q_2 ((\theta_1, \theta_2), \theta'_2)$ and

$$r_2 (\theta, \theta') = \frac{\pi (\theta_1, \theta'_2) q_2 ((\theta_1, \theta'_2), \theta_2)}{\pi (\theta_1, \theta_2) q_2 ((\theta_1, \theta_2), \theta'_2)} = \frac{\pi (\theta'_2 | \theta_1) q_2 ((\theta_1, \theta'_2), \theta_2)}{\pi (\theta_2 | \theta_1) q_2 ((\theta_1, \theta_2), \theta'_2)}$$

- We then combine these kernels through mixture or composition.
2.3– Composition of MH algorithms

Assume we use a composition of these kernels, then the resulting algorithm proceeds as follows at iteration \( i \).

**MH step to update component 1**

- Sample \( \theta^*_1 \sim q_1 \left( \left( \theta_1^{(i-1)}, \theta_2^{(i-1)} \right), \cdot \right) \) and compute

\[
\min \left( 1, \frac{\pi \left( \theta^*_1 | \theta_2^{(i-1)} \right) q_1 \left( \left( \theta^*_1, \theta_2^{(i-1)} \right), \theta_1^{(i-1)} \right)}{\pi \left( \theta_1^{(i-1)} | \theta_2^{(i-1)} \right) q_1 \left( \left( \theta_1^{(i-1)}, \theta_2^{(i-1)} \right), \theta^*_1 \right)} \right)
\]

- With probability \( \alpha_1 \left( \left( \theta_1^{(i-1)}, \theta_2^{(i-1)} \right), \left( \theta^*_1, \theta_2^{(i-1)} \right) \right) \), set \( \theta_1^{(i)} = \theta_1^* \) and otherwise \( \theta_1^{(i)} = \theta_1^{(i-1)} \).
2.3– Composition of MH algorithms

MH step to update component 2

- Sample $\theta^*_2 \sim q_2 \left( \left( \theta^{(i)}_1, \theta^{(i-1)}_2 \right), \cdot \right)$ and compute

$$\alpha_2 \left( \left( \theta^{(i)}_1, \theta^{(i-1)}_2 \right), \left( \theta^{(i)}_1, \theta^*_2 \right) \right) = \min \left( 1, \frac{\pi \left( \theta^*_2 \mid \theta^{(i)}_1 \right) q_2 \left( \left( \theta^{(i)}_1, \theta^*_2 \right), \theta^{(i-1)}_2 \right)}{\pi \left( \theta^{(i-1)}_2 \mid \theta^{(i)}_1 \right) q_2 \left( \left( \theta^{(i)}_1, \theta^{(i-1)}_2 \right), \theta^*_2 \right)} \right)$$

- With probability $\alpha_2 \left( \left( \theta^{(i)}_1, \theta^{(i-1)}_2 \right), \left( \theta^{(i)}_1, \theta^*_2 \right) \right)$, set $\theta^{(i)}_2 = \theta^*_2$ otherwise $\theta^{(i)}_2 = \theta^{(i-1)}_2$. 
2.4– Properties

- It is clear that in such cases both $K_1$ and $K_2$ are NOT irreducible and aperiodic.

$\Rightarrow$ Each of them only update one component!!!!

- However, the composition and mixture of these kernels can be irreducible and aperiodic because then all the components are updated.
2.5– Back to the Gibbs sampler

• Consider now the case where

\[ q_1 ( ( \theta_1, \theta_2 ), \theta'_1 ) = \pi ( \theta'_1 | \theta_2 ). \]

then

\[ r_1 ( \theta, \theta' ) = \frac{\pi ( \theta'_1 | \theta_2 ) q_1 ( ( \theta'_1, \theta_2 ), \theta_1 )}{\pi ( \theta_1 | \theta_2 ) q_1 ( ( \theta_1, \theta_2 ), \theta'_1 )} = \frac{\pi ( \theta'_1 | \theta_2 ) \pi ( \theta_1 | \theta_2 )}{\pi ( \theta_1 | \theta_2 ) \pi ( \theta'_1 | \theta_2 )} = 1 \]

• Similarly if \( q_2 ( ( \theta_1, \theta_2 ), \theta'_2 ) = \pi ( \theta'_2 | \theta_1 ) \) then \( r_2 ( \theta, \theta' ) = 1. \)

• If you take for proposal distributions in the MH kernels the full conditional distributions then you have the Gibbs sampler!
• Generally speaking, to sample from $\pi(\theta)$ where $\theta = (\theta_1, ..., \theta_p)$, we can use the following algorithm at iteration $i$.

• Iteration $i; \ i \geq 1$:

  For $k = 1 : p$

  • Sample $\theta_k^{(i)}$ using an MH step of proposal distribution $q_k \left((\theta_{-k}^{(i)}, \theta_k^{(i-1)}), \theta_k'\right)$ and target $\pi \left(\theta_k | \theta_{-k}^{(i)}\right)$.

  where $\theta_{-k}^{(i)} = \left(\theta_1^{(i)}, ..., \theta_{k-1}^{(i)}, \theta_{k+1}^{(i-1)}, ..., \theta_p^{(i-1)}\right)$.
2.6– General hybrid algorithm

- If we have $q_k (\theta_1:p, \theta'_k) = \pi (\theta'_k | \theta_{-k})$ then we are back to the Gibbs sampler.

- We can update some parameters according to $\pi (\theta'_k | \theta_{-k})$ (and the move is automatically accepted) and others according to different proposals.

- **Example:** Assume we have $\pi (\theta_1, \theta_2)$ where it is easy to sample from $\pi (\theta_1 | \theta_2)$ and then use an MH step of invariant distribution $\pi (\theta_2 | \theta_1)$. 
At iteration $i$. 

- Sample $\theta_1^{(i)} \sim \pi \left( \theta_1 | \theta_2^{(i-1)} \right)$.

- Sample $\theta_2^{(i)}$ using one MH step of proposal distribution $q_2 \left( \left( \theta_1^{(i)}, \theta_2^{(i-1)} \right), \theta_2 \right)$ and target $\pi \left( \theta_2 | \theta_1^{(i)} \right)$.

**Remark:** There is NO NEED to run the MH algorithm multiple steps to ensure that $\theta_2^{(i)} \sim \pi \left( \theta_2 | \theta_2^{(i-1)} \right)$.
3.1– Alternative acceptance probabilities

• The standard MH algorithm uses the acceptance probability

\[ \alpha (\theta, \theta') = \min \left( 1, \frac{\pi (\theta') q (\theta', \theta)}{\pi (\theta) q (\theta, \theta')} \right). \]

• This is not necessary and one can also use any function

\[ \alpha (\theta, \theta') = \frac{\delta (\theta, \theta')}{\pi (\theta) q (\theta, \theta')} \]

which is such that

\[ \delta (\theta, \theta') = \delta (\theta', \theta) \text{ and } 0 \leq \alpha (\theta, \theta') \leq 1 \]

• Example (Baker, 1965):

\[ \alpha (\theta, \theta') = \frac{\pi (\theta') q (\theta', \theta)}{\pi (\theta') q (\theta', \theta) + \pi (\theta) q (\theta, \theta')} . \]
3.1– Alternative acceptance probabilities

• Indeed one can check that

\[ K(\theta, \theta') = \alpha(\theta, \theta') q(\theta, \theta') + \left(1 - \int \alpha(\theta, u) q(\theta, u) \, du\right) \delta_{\theta}(\theta') \]

is \( \pi \)-reversible.

• We have

\[ \pi(\theta) \alpha(\theta, \theta') q(\theta, \theta') = \pi(\theta) \frac{\delta(\theta, \theta')}{\pi(\theta) q(\theta, \theta')} q(\theta, \theta') \]

\[ = \delta(\theta, \theta') \]

\[ = \delta(\theta', \theta) \]

\[ = \pi(\theta') \alpha(\theta', \theta) q(\theta', \theta) . \]

• The MH acceptance is favoured as it increases the acceptance probability.
4.1– Logistic Regression Example

- In 1986, Challenger exploded; the explosion being the result of an O-ring failure. It was believed to be a result of a cold weather at the departure time: 31°F.

- We have access to the data of 23 previous flights which give for flight $i$: Temperature at flight time $x_i$ and $y_i = 1$ failure and zero otherwise (Robert & Casella, p. 15).

- We want to have a model relating $Y$ to $x$. Obviously this cannot be a linear model $Y = \alpha + x\beta$ as we want $Y \in \{0, 1\}$. 
4.1– Logistic Regression Example

• We select a simple logistic regression model

\[
\Pr (Y = 1 \mid x) = 1 - \Pr (Y = 0 \mid x) = \frac{\exp (\alpha + x\beta)}{1 + \exp (\alpha + x\beta)}.
\]

• Equivalently we have

\[
\text{logit} = \log \left( \frac{\Pr (Y = 1 \mid x)}{\Pr (Y = 0 \mid x)} \right) = \alpha + x\beta.
\]

• This ensures that the response is binary.
We follow a Bayesian approach and select

\[ \pi(\alpha, \beta) = \pi(\alpha|b) \pi(\beta) = b^{-1} \exp(\alpha) \exp(-b^{-1} \exp(\alpha)) ; \]

i.e. exponential prior on \( \exp(\alpha) \) and flat prior on \( \beta \).

- \( b \) is selected as the data-dependent prior such that \( E(\alpha) = \hat{\alpha} \) where \( \hat{\alpha} \) is the MLE of \( \alpha \) (Robert & Casella).

- As a simple proposal distribution, we use

\[ q((\alpha, \beta), (\alpha', \beta')) = \pi(\alpha'|b) \mathcal{N}(\beta'; \beta^{(i-1)}, \hat{\sigma}_\beta^2) \]

where \( \hat{\sigma}_\beta^2 \) is the associated variance at thr MLE \( \hat{\beta} \).
The algorithm proceeds as follows at iteration $i$

- Sample $(\alpha^*, \beta^*) \sim \pi(\alpha|b) \mathcal{N}(\beta; \beta^{(i-1)}, \hat{\sigma}_\beta^2)$ and compute

$$\zeta \left( \left( \alpha^{(i-1)}, \beta^{(i-1)} \right), (\alpha^*, \beta^*) \right) = \min \left( 1, \frac{\pi(\alpha^*, \beta^*| \text{data}) \pi(\alpha^{(i-1)}|b)}{\pi(\alpha^{(i-1)}, \beta^{(i-1)}| \text{data}) \pi(\alpha^*|b)} \right)$$

- Set $(\alpha^{(i)}, \beta^{(i)}) = (\alpha^*, \beta^*)$ with probability $\zeta \left( \left( \alpha^{(i-1)}, \beta^{(i-1)} \right), (\alpha^*, \beta^*) \right)$, otherwise set $(\alpha^{(i)}, \beta^{(i)}) = (\alpha^{(i-1)}, \beta^{(i-1)})$. 
4.1– Logistic Regression Example

Plots of $\frac{1}{k} \sum_{i=1}^{k} \alpha^{(k)}$ (left) and $\frac{1}{k} \sum_{i=1}^{k} \beta^{(i)}$ (right).
Histogram estimates of $p(\alpha|\text{data})$ (left) and $p(\beta|\text{data})$ (right).
Predictive \( \Pr(Y = 1|x) = \int \Pr(Y = 1|x, \alpha, \beta) \pi(\alpha, \beta|\text{data}) \), predictions of failure probability at 65°F and 45°F.
4.2– Probit Regression Example

- We consider the following example: we take 4 measurements from 100 genuine Swiss banknotes and 100 counterfeit ones.

- The response variable $y$ is 0 for genuine and 1 for counterfeit and the explanatory variables are

  - $x_1$: the length,

  - $x_2$: the width of the left edge

  - $x_3$: the width of the right edge

  - $x_4$: the bottom margin width

All measurements are in millimeters.
Left: Plot of the status indicator versus the bottom margin width.
Right: Boxplots of the bottom margin width for both counterfeit status.
4.2– Probit Regression Example

- Instead of selecting a logistic link, we select a probit one here

\[
\Pr(Y = 1|x) = \Phi(x^1\beta_1 + \ldots + x^4\beta_4)
\]

where

\[
\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(-\frac{v^2}{2}\right) dv
\]

- For \( n \) data, the likelihood is then given by

\[
f(y_{1:n} | \beta, x_{1:n}) = \prod_{i=1}^{n} \Phi(x_i^T\beta)^{y_i} (1 - \Phi(x_i^T\beta))^{1-y_i}.
\]
4.2– Probit Regression Example

• We assume a vague prior where \( \beta \sim \mathcal{N}(0, 100I_4) \) and we use a simple random walk sampler with \( \hat{\Sigma} \) the covariance matrix associated to the MLE (estimated using simple deterministic method).

• The algorithm is thus simply given at iteration \( i \) by
  
  - Sample \( \beta^* \sim \mathcal{N}(\beta^{(i-1)}, \tau^2\hat{\Sigma}) \) and compute
    
    \[
    \alpha\left(\beta^{(i-1)}, \beta^*\right) = \min\left(1, \frac{\pi(\beta^* | y_{1:n}, x_{1:n})}{\pi(\beta^{(i-1)} | y_{1:n}, x_{1:n})}\right).
    \]

  - Set \( \beta^{(i)} = \beta^* \) with probability \( \alpha(\beta^{(i-1)}, \beta^*) \) and \( \beta^{(i)} = \beta^{(i-1)} \) otherwise.

• Best results obtained with \( \tau^2 = 1 \).
4.2– Probit Regression Example

Traces (left), Histograms (middle) and Autocorrelations (right) for \( (\beta_1^{(i)}, \ldots, \beta_4^{(i)}) \).
4.2– Probit Regression Example

• One way to monitor the performance of the algorithm of the chain \( \{X^{(i)}\} \) consists of displaying \( \rho_k = \text{cov} \left[ X^{(i)}, X^{(i+k)} \right] / \text{var} \left( X^{(i)} \right) \) which can be estimated from the chain, at least for small values of \( k \).

• Sometimes one uses an effective sample size measure

\[
N^{\text{ess}} = N \left( 1 + 2 \sum_{k=1}^{N_0} \hat{\rho}_k \right)^{-1/2}.
\]

This represents approximately the sample size of an equivalent i.i.d. samples.

• One should be very careful with such measures which can be very misleading.
• We found for $E(\beta|y_{1:n},x_{1:n}) = (-1.22, 0.95, 0.96, 1.15)$ so a simple plug-in estimate of the predictive probability of a counterfeit bill is

$$\hat{p} = \Phi (-1.22x^1 + 0.95x^2 + 0.96x^3 + 1.15x^4)$$

For $x = (214.9, 130.1, 129.9, 9.5)$, we obtain $\hat{p} = 0.59$.

• A better estimate is obtained by

$$\int \Phi (\beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4) \pi (\beta|y_{1:n},x_{1:n}) d\beta$$
4.3– Gibbs sampling for Probit Regression

- It is impossible to use Gibbs to sample directly from $\pi(\beta | y_{1:n}, x_{1:n})$.

- Introduce the following unobserved latent variables

$$Z_i \sim \mathcal{N}(x_i^T \beta, 1),$$

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

- We have now define a joint distribution

$$f(y_i, z_i | \beta, x_i) = f(y_i | z_i) f(z_i | \beta, x_i).$$
4.3– Gibbs sampling for Probit Regression

• Now we can check that

\[ f(y_i = 1| x_i, \beta) = \int f(y_i, z_i| \beta, x_i) \, dz_i = \int_0^\infty f(z_i| \beta, x_i) \, dz_i = \Phi(x_i^T \beta). \]

⇒ We haven’t changed the model!

• We are now going to sample from \( \pi(\beta, z_{1:n}| x_{1:n}, y_{1:n}) \) instead of \( \pi(\beta| x_{1:n}, y_{1:n}) \) because the full conditional distributions are simple

\[
\pi(\beta| y_{1:n}, x_{1:n}, z_{1:n}) = \pi(\beta| x_{1:n}, z_{1:n}) \text{ (standard Gaussian!)},
\]

\[
\pi(z_{1:n}| y_{1:n}, x_{1:n}, \beta) = \prod_{i=1}^n \pi(z_k| y_k, x_k, \beta)
\]

where

\[
z_k| y_k, x_k, \beta \sim \begin{cases} 
\mathcal{N}_+(x_k^T \beta, 1) & \text{if } y_k = 1 \\
\mathcal{N}_-(x_k^T \beta, 1) & \text{if } y_k = 0.
\end{cases}
\]
4.3– Gibbs sampling for Probit Regression

Traces (left), Histograms (middle) and Autocorrelations (right) for $(\beta_1^{(i)}, \ldots, \beta_4^{(i)})$.
4.3– Gibbs sampling for Probit Regression

- The results obtained through Gibbs are very similar to MH.

- We can also adopt an Zellner’s type prior and obtain very similar results.

- Very similar were also obtained using a logistic fonction using the MH (Gibbs is feasible but more difficult).