STATS 305C Notes

Kenneth Tay

1 Basic Bayes

- **Exchangeability:** An exchange sequence of random variables is a finite or infinite sequence $X_1, X_2, \dots$ such that for any finite permutation σ of the indices $1, 2, \dots$ (i.e. permutation only acts on finitely many indices, rest fixed), the joint probability distribution of $X_{\sigma(1)}, X_{\sigma(2)}, \dots$ is the same as that of $X_1, X_2, \dots$.
- **de Finetti's Theorem:** Suppose $X_1, X_2, \dots$ is an infinite exchangeable sequence of Bernoulli random variables. Then there is some probability distribution m on $[0, 1]$ and some random variable Y such that (i) the probability distribution of Y is m, (ii) $X_1, X_2, \dots$ are conditionally independent given Y, and (iii) for any i, $\mathbb{P}(X_i = 1 \mid Y) = Y$.
- Let $\tilde{y}$ denote future data, y current data, and assume that $\tilde{y}$ and y are conditionally independent given θ. Then $p(\tilde{y} \mid y) = \displaystyle\int p(\tilde{y} \mid \theta) p(\theta \mid y) d\theta$.
- **Marginalization:** Say we have 2 parameters θ_1 and θ_2. We have $p(\theta_1, \theta_2 \mid y) \propto p(y \mid \theta_1, \theta_2) \cdot p(\theta_1, \theta_2)$, and
$$p(\theta_1 \mid y) = \int p(\theta_1, \theta_2 \mid y) d\theta_2 = \int p(\theta_1 \mid \theta_2, y) p(\theta_2 \mid y) dy.$$
- **Odds ratios:**
$$\underbrace{\frac{p(\theta_1 \mid y)}{p(\theta_2 \mid y)}}_{\text{posterior odds}} = \underbrace{\frac{p(\theta_1)}{p(\theta_2)}}_{\text{prior odds}} \cdot \underbrace{\frac{p(y \mid \theta_1)}{p(y \mid \theta_2)}}_{\text{likelihood ratio}}.$$
- **Bayesian CLT:** Let $\ell(\theta) = \log p(y_i \mid \theta)$ be the log-likehood. For large n, $p(\theta \mid y)$ is approximately Gaussian: $p(\theta \mid y) \dot{\propto} \exp\left[\frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2\right]$, where $\hat{\theta}$ is the posterior mode. In the multivariate case, we have $\theta \mid y \dot{\sim} \mathcal{N}\left(\hat{\theta}, (\ell'')^{-1}_{\theta = \hat{\theta}}\right)$. Note that $-\ell''$ is the observed Fisher information.

  Possible exceptions:
  - Prior $p(\theta)$ has finite support.
  - Model is unidentifiable.
  - Label switching. (E.g. model $p(y \mid \theta) = \lambda_1 q(y \mid \beta_1) + \lambda_2 q(y \mid \beta_2)$, where $\lambda_1 = 1 - \lambda_2$. Then $\theta = (\lambda_1, \beta_1, \beta_2)$ gives the same model as $\theta = (\lambda_2, \beta_2, \beta_1)$. Will end up with 2 posterior modes. Possible fix: Restrict support of prior.)
  - Unbounded likelihood. This can give an unbounded posterior distribution.
  - Bounded likelihood with a mode at ∞. (This could happen in logistic regression with separation in the data.) If we use a flat prior with this, we will get an improper posterior.
- **Hierarchical model:** Likelihood $p(y_j \mid \theta_j)$, prior $p(\theta_j \mid \phi)$, hyperprior $p(\phi)$.

  - $p(\theta, \phi \mid y) \propto p(\theta, \phi) p(y \mid \theta, \phi) = p(\theta) p(\theta \mid \phi) p(y \mid \theta)$. To simplify computations, we can make $p(y \mid \theta)$ conjugate to $p(\theta \mid \phi)$.
  - Posterior $p(\phi \mid y) = \dfrac{p(\theta, \phi \mid y)}{p(\theta \mid \phi, y)}$.
  - For new samples: simulate $\phi \sim p(\phi \mid y)$, then $\theta \mid \phi \sim p(\theta \mid \phi)$, then $y \mid \theta \sim p(y \mid \theta)$.
  - Example: $y_j \sim \text{Binom}(n_j, \theta_j)$, $\theta_j \sim \text{Beta}(\alpha, \beta)$. BDA chooses hyperprior $\dfrac{\alpha}{\alpha + \beta} \sim \text{Unif}(0, 1)$, $(\alpha + \beta)^{-1/2} \sim \text{Unif}(0, 1)$. (This ends up giving $p(\alpha + \beta) \propto (\alpha + \beta)^{-5/2}$.)
  - Example: $Y_{ij} \mid \theta_j \sim \mathcal{N}(\theta_j, \sigma^2)$, $\theta_j \mid \mu, \tau \sim \mathcal{N}(\mu, \tau^2)$. We can take $p(\mu \mid \tau) \propto 1$. For $p(\tau)$, we could take $p(\tau) \propto 1$, or $p(\log \tau) \propto 1$, which is equivalent to $p(\tau) \propto \dfrac{1}{\tau}$.

2 Picking priors

- Picking **conjugate priors** can be a good idea; it'll simplify the posterior calculation. (We can always do this for exponential families.)

- Can try to match the moments of the prior to the empirical moments from the data.

- Can try to pick **non-informative/flat priors**. (In some cases, these priors might be improper.) Problem with uniform priors: this uniformity might not be preserved under transformations.

- **Jeffreys prior:** $p(\theta) \propto \sqrt{\text{Fisher information of } \theta}$. For any transformation ϕ of θ, we will still have $p(\phi) \propto \sqrt{\text{Fisher information of } \phi(\theta)}$.

  - For $y \sim \text{Binom}(n, \theta)$, $\sqrt{\mathcal{I}(\theta)} \propto \sqrt{\theta(1 - \theta)}$, so the Jeffreys prior is $\text{Beta}(1/2, 1/2)$.
  - For $y \sim \mathcal{N}(0, \sigma^2)$ (with $\theta = \sigma$), we get $\sqrt{\mathcal{I}(\theta)} \propto \sigma^{-1}$. This is a popular prior for scale parameters. (It is an improper prior.)
  - Multivariate version: $p(\theta) \propto \sqrt{\det(\mathcal{I}(\theta))}$. This is not really used anymore. (E.g. for $\mathcal{N}(\mu, \sigma^2)$ both unknown, we get $p(\sigma) \propto \sigma^{-2}$ and $p(\mu) \propto 1$.)

- **Reference priors:** For high dimensional θ. Group the components θ_j into sets $\theta_{(1)}, \ldots, \theta_{(m)}$ from most to least important, then work with $p(\theta_{(k)} \mid \theta_{(1)}, \ldots, \theta_{(k)})$. (BDA3 believes this is too much work for too little gain.)

- **Weakly informative priors:** Idea: A proper prior will always give a proper posterior. Limit the range to what is plausible, then try to be flat on that range.

  - For logistic regression Y on $X_1, \ldots, X_d$, could take a flat prior with $|\beta_j| \leq 10$.

3 Conjugate distributions

- Likelihood $p(y \mid \theta) \sim \text{Binom}(n, \theta)$. If prior $p(\theta) \sim \text{Beta}(\alpha, \beta)$, then posterior $p(\theta \mid y) \sim \text{Beta}(\alpha + y, \beta + n - y)$.

  (Can imagine α and β to be the number of "prior" successes and failures respectively.

- Likelihood $p(y \mid \theta) \sim \text{Multinom}(n, \theta)$ (say θ has k components). If prior $p(\theta) \sim \text{Dirichlet}(\alpha)$, then posterior $p(\theta \mid y) \sim \text{Dirichlet}(\alpha_1 + y_1, \ldots, \alpha_k + y_k)$.

- Likelihood $p(y_i \mid \theta) \overset{iid}{\sim} \text{Pois}(\theta)$. If prior $p(\theta) \sim \text{Gam}(\alpha, \beta)$, then posterior $p(\sigma^2 \mid y) \sim \text{Gam}\left(\alpha + \sum_{i=1}^n y_i, \beta + n\right)$. (Can think of the prior as observing a total of α counts in β samples.)

- Likelihood $p(y_i \mid \mu) \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where σ^2 is known. If prior $p(\mu) \sim \mathcal{N}(\mu_0, \tau_0^2)$, then posterior $p(\mu \mid y_1, \dots, y_n) \sim \mathcal{N}(\mu_n, \tau_n^2)$, where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \qquad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}.$$

- Likelihood $p(y_i \mid \mu) \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where μ is known. If prior $p(\sigma^2) \sim \text{InvGam}(\alpha, \beta)$ (i.e. $\dfrac{1}{\sigma^2} \sim \dfrac{\text{Gam}(\alpha)}{\beta}$), then posterior $p(\sigma^2 \mid y) \sim \text{InvGam}\left(\alpha + \dfrac{n}{2}, \beta + \dfrac{1}{2}\displaystyle\sum_{i=1}^n (y_i - \mu)^2\right)$.

- Likelihood $p(y_i \mid \mu) \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where μ is known. If prior $p(\sigma^2) \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$, (i.e. $\dfrac{1}{\sigma^2} \sim \dfrac{\chi^2_{\nu_0}/\nu_0}{\sigma_0^2}$), then posterior $p(\sigma^2 \mid y) \sim \text{Inv-}\chi^2\left(\nu_0 + n, \dfrac{\nu_0\sigma_0^2 + \sum_{i=1}^n (y_i - \mu)^2}{\nu_0 + n}\right)$.

  (Can think of σ_0^2 as the initial guess for the variance. Also, $\text{Inv-}\chi^2(\nu_0, \sigma_0^2) \overset{d}{=} \text{InvGam}\left(\dfrac{\nu_0}{2}, \dfrac{\nu_0\sigma_0^2}{2}\right)$.)

- Likelihood $p(y_i \mid \mu) \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. If the prior is $p(\sigma^2) \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$ and $p(\mu \mid \sigma^2) \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$, i.e. the $\mathcal{N}\text{-Inv-}\chi^2\left(\mu_0, \dfrac{\sigma_0^2}{\kappa_0}; \nu_0, \sigma_0^2\right)$ distribution, then the posterior has $\mathcal{N}\text{-Inv-}\chi^2\left(\mu_n, \dfrac{\sigma_n^2}{\kappa_n}; \nu_n, \sigma_n^2\right)$ distribution, where

$$\mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n}, \qquad \kappa_n = \kappa_0 + n,$$

$$\nu_n = \nu_0 + n, \qquad \nu_n\sigma_n^2 = \nu_0\sigma_0^2 + \sum_{i=1}^n (y_i - \bar{y})^2 + \frac{\kappa_0 n(\bar{y} - \mu_0)^2}{\kappa_0 + n}.$$

- Likelihood $p(y_i \mid \mu) \sim MVN(\mu, \Sigma)$, where Σ is known. If prior $p(\mu) \sim MVN(\mu_0, \Lambda_0)$, then posterior $p(\mu \mid y) \sim MVN(\mu_n, \Lambda_n)$, where

$$\mu_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}), \qquad \Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}.$$

- Likelihood $p(y_i \mid \mu) \sim MVN(\mu, \Sigma)$, where both μ and Σ are unknown. If prior is $\mathcal{N}\text{-Inv-Wishart}\left(\mu_0, \dfrac{\Lambda_0}{\kappa_0}; \nu_0, \Lambda_0\right)$, i.e. $\Sigma \sim \text{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1})$ and $\mu \mid \Sigma \sim \mathcal{N}(\mu_0, \Sigma/\kappa_0)$, then the posterior is $\mathcal{N}\text{-Inv-Wishart}\left(\mu_n, \dfrac{\Lambda_n}{\kappa_n}; \nu_n, \Lambda_n\right)$, where

$$\mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n}, \qquad \kappa_n = \kappa_0 + n,$$

$$\nu_n = \nu_0 + n, \qquad \Lambda_n = \Lambda_0 + \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T.$$

4 Bayesian Computation

For Bayesian computation, often we may not have the posterior $p(\theta \mid y)$. Instead, we might have $q(\theta \mid y) = p(y \mid \theta)p(\theta)$, which is an un-normalized version of the posterior. With just q, we can get $\dfrac{p(\theta' \mid y)}{p(\theta \mid y)} = \dfrac{q(\theta' \mid y)}{q(\theta \mid y)}$.

- If we can get S samples $\theta^s \sim p(\theta \mid y)$, then $\mathbb{E}[h(\theta) \mid Y = y] = \displaystyle\int h(\theta)p(\theta \mid y)d\theta \approx \frac{1}{S}\sum_{s=1}^{S} h(\theta^s)$.
- $\theta \in \Theta \subseteq \mathbb{R}^d$ for small d, we can use a grid. Then $\mathbb{E}[h(\theta) \mid Y] \approx \displaystyle\sum_{s \in \text{ grid}} p(\theta^s \mid y)h(\theta^s)$. If we just have un-normalized density q, we can do
$$\mathbb{E}[h(\theta) \mid Y] \approx \frac{\displaystyle\sum_{s \in \text{ grid}} q(\theta^s \mid y)h(\theta^s)}{\displaystyle\sum_{s \in \text{ grid}} q(\theta^s \mid y)}.$$
- Often computing the posterior directly by multiplying the prior and likelihood is numerically unstable. If this is the case, it is easier to work with the logarithms of these objects.

Acceptance-rejection sampling (version 1)

Choose distribution g with $\dfrac{p(\theta \mid y)}{g(\theta)} \leq M$ for all θ. (M can just be $\sup_\theta \dfrac{p(\theta \mid y)}{g(\theta)}$.)

1. Sample θ_i from $g(\theta)$. Draw $y_i \sim Mg(\theta) \cdot \text{Unif}(0,1)$. Plot the points (t_i, y_i).
2. Keep the points which lie under the curve $y = p(\theta \mid y)$.
3. Use the x-coordinates of the points kept.

- The algorithm above can be run with q instead of p.
- $\mathbb{P}(\text{Acceptance}) = \dfrac{\text{area under } q(\theta \mid y)}{\text{area under } Mg(\theta)}$. Hence, for this algorithm to work well, we need a tight fit enclosing $q(\theta \mid y)$. (This is hard to do in high dimensions.)

Acceptance-rejection sampling (version 2)

Choose distribution g with $\dfrac{p(\theta \mid y)}{g(\theta)} \leq M$ for all θ. (M can just be $\sup_\theta \dfrac{p(\theta \mid y)}{g(\theta)}$.)

1. Sample θ_i from $g(\theta)$.
2. Independently generate $U \sim \text{Unif}(0,1)$.
3. If $U < \dfrac{p(\theta \mid y)}{Mg(\theta)}$, accept sample. If not, reject.

It'll take on average M iterations to get one sample.

Importance sampling

Choose distribution g such that $g(\theta) > 0$ whenever $p(\theta \mid y)h(\theta) \neq 0$. Take S samples $\theta^s \sim g$. Since $\displaystyle\int h(\theta)p(\theta \mid y)d\theta = \int \frac{h(\theta)p(\theta \mid y)}{g(\theta)} g(\theta)d\theta$, we can approximate $\displaystyle\mathbb{E}[h(\theta) \mid y] \approx \frac{1}{S}\sum_{s=1}^{S} \frac{h(\theta^s)p(\theta^s \mid y)}{g(\theta^s)}$.

- If we only have q, we can still do **self-normalized importance sampling**:

$$\mathbb{E}[h(\theta) \mid y] \approx \frac{\dfrac{1}{S}\displaystyle\sum_{s=1}^{S} \frac{h(\theta^s)q(\theta^s \mid y)}{g(\theta^s)}}{\dfrac{1}{S}\displaystyle\sum_{s=1}^{S} \frac{q(\theta^s \mid y)}{g(\theta^s)}}.$$

- Advantage over acceptance-rejection sampling: don't need $\displaystyle\sup_\theta \frac{q(\theta \mid y)}{g(\theta)} < \infty$.
- Good choices of g: $g \mathrel{\dot\propto} p$. Better: $g \mathrel{\dot\propto} ph$.
- Importance sampling has a problem with light-tailed g (get enormous ratios because $g(\theta^s)$ is in the denominator. (For p approximately normal, can use $g \sim t_\nu$.)
- We can write the self-normalized importance sampling formula as $\displaystyle\mathbb{E}[h(\theta) \mid y] \approx \sum_s w_s h(\theta^s)$, where

$$w_s = \frac{q(\theta^s \mid y)/g(\theta^s)}{\sum_{t=1}^{S} q(\theta^t \mid y)/g(\theta^t)}.$$

- Importance sampling doesn't work well if there is a large $q(\theta^s \mid y)/g(\theta^s)$ relative to the rest.
- **Effective sample size** $n_{eff} := \dfrac{(\sum w_s)^2}{\sum w_s^2}$.

Defensive sampling

If we can compute $p(\theta \mid y)$ and sample $p(\theta \mid y)$, then we can sample $\alpha p(\theta \mid y) + (1-\alpha)g(\theta)$.

In this case $\dfrac{p(\theta \mid y)}{\alpha p(\theta \mid y) + (1-\alpha)g(\theta)} \leq \dfrac{1}{\alpha}$, so we won't have a problem of large q/g.

Markov chains

- **Stationarity:** $\displaystyle\sum_{x \in \Omega} \pi(x)P(x \to y) = \pi_y$ for all y.
- **Balance:** "probability flowing into y" = "probability flowing out of y" for all y.
- **Detailed balance:** $\pi(x)P(x \to y) = \pi(y)P(y \to x)$ for all x, y.
  - Detailed balance implies balance.
  - A doubly stochastic matrix satisfies detailed balance with the uniform distribution on the state space.
  - If P and Q both have detailed balance, then PQP does as well.

Metropolis-Hastings

We want to sample according to distribution π (possibly un-normalized). **Idea:** Try to construct a Markov chain with stationary distribution π. We do this by constructing a transition probability matrix which satisfies the detailed balance condition.

Say we have a proposal: Given that we are at x_i, we move to y with probability $Q(x_i \to y)$. The algorithm is as follows:

1. Say we are at x_i. Generate a proposal y according to distribution $Q(x_i \to y)$.
2. Accept with probability $A(x \to y)$ (in which case $x_{i+1} = y$, else reject (in which case $x_{i+1} = x_i$).

- From the above, we have $p(x \to y) = Q(x \to y)A(x \to y)$. In order to achieve detailed balance, we need $\pi(x)Q(x \to y)A(x \to y) = \pi(y)Q(y \to x)A(y \to x)$.
- **Metropolis-Hastings acceptance probability:** $A(x \to y) = \min\left(1, \dfrac{\pi(y)Q(y \to x)}{\pi(x)Q(x \to y)}\right)$.
- Metropolis-Hastings works for un-normalized π.
- **Theorem (Peskun):** Let P and $\tilde{P}$ are 2 irreducible chains with detailed balance for π such that $\tilde{P}(x \to y) \leq P(x \to y)$ for all $x \neq y$. Let $X_i \sim P$ starting at x_0 and let $\tilde{X}_i \sim \tilde{P}$ starting at $\tilde{x}_0$. Then $\lim_{n\to\infty} n\text{Var}\left(\dfrac{1}{n}\sum f(X_i)\right) \leq \lim_{n\to\infty} n\text{Var}\left(\dfrac{1}{n}\sum f(\tilde{X}_i)\right)$.
- **Random Walk Metropolis:** Propose $y_i = x_i + Z_i$, where $Z_i \sim \mathcal{N}(0, \sigma^2 I_d)$ or $Z_i \sim \text{Unif}[-\Delta, \Delta]^d$, or some other distribution (usually symmetric).
  - For symmetric distributions, the acceptance probability becomes $A(x \to y) = \min\left(1, \dfrac{\pi(y)}{\pi(x)}\right)$.
  - In the normal case, what is a good σ to use? If σ too small, we'll almost always accept, and will not move around the space much. If σ too large, we'll almost always reject, and will end up with a small number of potentially large jumps.
  - Good σ maximizes mean squared jumping distance.
  - If $\pi \sim \mathcal{N}(\mu, \Sigma)$ and $y \sim \mathcal{N}(x, \lambda\Sigma)$, take $\lambda \approx \dfrac{2.38}{\sqrt{d}}$ to maximize mean squared jumping distance. For $d \geq 5$, acceptance probability is around 0.234. For $d = 1$, it is around 0.44. Efficiency vs. i.i.d. sampling is $\approx \dfrac{0.33}{d}$.
- **Independence Sampler:** Propose $y_i \sim Q$, i.e. ignore x_i. Acceptance probability is $A(x \to y) = \min\left(1, \dfrac{w(y)}{w(x)}\right)$, where $w = \dfrac{\pi}{Q}$.
  - w is called the **importance ratio**, telling us how under-represented a value is. (If it is more unrepresented, the sampler is more likely to take it.)
  - Both π and Q can be unnormalized.
  - Q must sample any region that π does (if not bias is introduced). Q should have heavier tails than π, although the closer Q is to π, the better.

Gibbs sampling

Let $x_i = \big(x_{i,1}, \dots, x_{i,d}\big) \in \mathbb{R}^d$. Suppose we can sample from the **full conditional**, i.e. $x_{i,j} \mid x_{i,-j}$.

- **Random scan:** Pick x_0. For $i = 1, \dots, N$ (no. of samples), pick $j \sim \text{Unif}\{1, 2, \dots, d\}$, then pick $z \sim \pi_{j|-j}(\cdot \mid x_{i,-j})$. Set $x_{i,-j} \leftarrow x_{i-1,-j}$, and $x_{i,j} \leftarrow z$.
- **Systematic/Fixed scan:** Pick x_0. For $i = 1, \dots, N$ (no. of samples), let $\ell = i - 1 \mod d$, let $j = \ell + 1$. Pick $z \sim \pi_{j|-j}(\cdot \mid x_{i,-j})$. Set $x_{i,-j} \leftarrow x_{i-1,-j}$, and $x_{i,j} \leftarrow z$.
- We can view the conditional distribution for x_j as a transition matrix P_j. Each P_j is a Metropolis Hastings which accepts.
- **Examples of Gibbs sampling:** Truncated normal, mixture of binomials, hierarchical normal.
- **Sampling from a truncated distribution:** Sampling from $F \mid [a, b]$ (i.e. F truncated for the interval $[a, b]$) is easy: $F^{-1}[F(a) + \text{Unif}(0, 1) \cdot [F(b) - F(a)]] \sim F \mid [a, b]$.
- **Metropolis within Gibbs:** If we can't sample $x_j \mid x_{-j}$ for some j, take a Metropolis step instead: propose value y and accept/reject it by Metropolis Hastings.

Other MCMC

- **Hit and run algorithm:** For sampling points within a convex Ω uniformly. Pick a random direction, draw line to hit edges of Ω, then sample uniformly from that line.

  To compute $P(A)$ for $A \subseteq \Omega$, can perform the hit and run algorithm and get $\hat{P}(A) = \dfrac{\#\text{ pts in } A}{\#\text{ pts in } \Omega}$.
- **CLT for Markov chains:** Say our Markov chain gives values $x_1, x_2, \dots$, and we want to evaluate some function f. If the chain is stationary, irreducible, has detailed balance and $\displaystyle\int f^2(x)\pi(x)dx < \infty$, then

  $$\frac{\frac{1}{n}\sum_{i=1}^n f(x_i) - \int f(x)\pi(x)dx}{\sqrt{\sigma_f^2/n}} \xrightarrow{d} \mathcal{N}(0, 1),$$

  where $\displaystyle\sigma_f^2 = \text{Var}_\pi(f) + 2\sum_{\ell=1}^{\infty} \text{Cov}_\pi\left(f(X_0), f(X_\ell)\right)$.
- **Effective sample size** $n_{eff} = \dfrac{n}{1 + 2\sum_{\ell=1}^\infty \rho_\ell}$, where $\rho_\ell = \text{Corr}\left(f(X_0), f(X_\ell)\right)$.
- **Confidence interval for** $\displaystyle\int f(x)\pi(x)dx$**:** Could do $\displaystyle\frac{1}{n}\sum f(x_i) \pm \frac{1.96}{\sqrt{n}}\hat{\sigma}_f$. (Problem: Could be hard to estimate $\hat{\sigma}_f$.)

  Alternative: Split the data into k blocks and assume that the blocks are essentially independent. Then if we let the means of the k blocks be $\bar{Y}_1, \dots, \bar{Y}_k$ and the grand mean be $\bar{Y}$, we can take as CI $\bar{Y} \pm \dfrac{1.96s}{\sqrt{k}}$, where $\displaystyle s^2 = \frac{1}{k-1}\sum_{r=1}^k (\bar{Y}_r - \bar{Y})^2$.
- One way to avoid MCMC bias is to introduce **burn-in**, i.e. throw away the first x% of data. We could also run the chain until it looks like it has achieved stationarity, then throw away existing data and restart the chain there.

- If we know where the high posterior probability region is, we could start there.
- **Diagnostics:** Assume samples $\theta_1, \theta_2, \ldots \in \mathbb{R}^d$.
  - For each j, plot θ_{ij} against i. Hope to see numerous transitions over its range.
  - **Autocorrelation function (ACF):** Autocorrelation at lag k is defined by $\hat{\rho}_k = \dfrac{\frac{1}{n}\sum_{i=1}^{n-k}(\theta_i - \bar{\theta})(\theta_{i+k} - \bar{\theta})}{\frac{1}{n}\sum_{i=1}^{n}(\theta_i - \bar{\theta})^2}$.
    Plot $\hat{\rho}_k$ against k. Hope to see $\hat{\rho}_k$ decay rapidly to 0.
    Common ACF pattern: $\rho_k \approx \rho^k$ (called AR(1) model). For this model, we can compute $n_{eff} = \dfrac{n(1-\rho)}{1+\rho}$.
  - Do multiple starts and plot θ_{ij} against i on the same graph. Hope to see the graphs mix.
  - **Gelman-Rubin diagnostic:** Run m chains for n steps ($j = 1, \ldots, m$, $i = 1, \ldots, n$). Let $\psi_{ij} = \psi(\theta_{ij}) \in \mathbb{R}$ for some function ψ. Do ANOVA for groups $j = 1, 2, \ldots, m$.

$$\text{Within sum of squares } W = \frac{1}{n}\sum_{j=1}^{m} s_j^2, \qquad s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(\psi_{ij} - \bar{\psi}_{\cdot j})^2,$$
$$\text{Between sum of squares } B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\psi}_{\cdot j} - \bar{\psi}_{\cdot\cdot})^2,$$
$$\widehat{\text{Var}}^{+}(\psi \mid y) = \frac{n-1}{n}W + \frac{1}{n}B.$$

    If the chains don't mix, then B will be large but W will be small. For good mixing, $\hat{R} := \sqrt{\dfrac{\widehat{\text{Var}}^{+}(\psi \mid y)}{W}}$ should be close to 1.
- **Hamiltonian MCMC:** Write $p(\theta \mid y) = e^{-H(\theta)}$ or $p(\theta \mid y) = e^{-H(\theta)/T}$. H called the **Hamiltonian**, T called the temperature. Introduce a momentum parameter ϕ independent of θ and look at $p(\theta \mid y)p(\phi)$.

5 Bayesian Regression

- Model: $Y_i = \beta^T x_i + \varepsilon_i$, with $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Likelihood is

$$p(y_1, \ldots y_n \mid x_1, \ldots x_n, \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta^T x_i)^2\right]$$
$$= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\underbrace{SSR(\beta)}_{\text{sum of square residuals}}\right].$$

  In matrix notation, $y \mid X, \beta, \sigma^2 \sim \text{MVN}(X\beta, \sigma^2 I)$.
- Prior specification: $\beta \sim MVN(\beta_0, \Sigma_0)$. Then posterior is also multivariate normal, with

$$\mathbb{E}[\beta \mid y, X, \sigma^2] = \left(\Sigma_0^{-1} + \frac{X^T X}{\sigma^2}\right)^{-1}\left(\Sigma_0^{-1}\beta_0 + \frac{X^T y}{\sigma^2}\right), \qquad \text{Var}\,[\beta \mid y, X, \sigma^2] = \left(\Sigma_0^{-1} + \frac{X^T X}{\sigma^2}\right)^{-1}.$$

  Let $\gamma = \sigma^{-2}$, and take prior $\gamma \sim \dfrac{\text{Gam}(\nu_0/2)}{\nu_0\sigma_0^2/2}$. Then posterior is $\sigma^2 \mid y, X, \beta \sim \text{InvGam}\left(\dfrac{\nu_0 + n}{2}, \dfrac{\nu_0\sigma_0^2 + SSR(\beta)}{2}\right)$.

- **Ridge regression:** Take $\beta_0 = 0$, $\Sigma_0 = \tau^2 I$. (We can take $(\Sigma_0)_{11}$ to be zero if we don't want to penalize the intercept. We can also have Σ_0 be a general diagonal matrix with positive entries.)

- **Unit information prior:** Contains the same amount of information as a single observation. This sets $\Sigma_0^{-1} = \dfrac{X^T X}{n\sigma^2}$, and $\beta_0 = \hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$.

- **Invariance:** Let H be some $p \times p$ matrix, $\tilde{X} = XH$. The principle of invariance says that if we get the posterior distributions of β ($\tilde{\beta}$ resp.) from y and X (y and $\tilde{X}$ resp.), then the posterior distributions of β and $H\tilde{\beta}$ should be the same.

  To achieve this, we need $\beta_0 = 0$ and $\Sigma_0 = k(X^T X)^{-1}$ for any positive k.

- **Zellner's g-prior:** In the invariance set-up above, if we further take $k = g\sigma^2$, we get Zellner's g-prior. (If we set $g = n$, we get the unit information prior.) Under this prior, $\beta \mid y, X, \sigma^2$ still multivariate normal, with Var $[\beta \mid y, X, \sigma^2] = \dfrac{g}{g+1}(X^T X)^{-1}\sigma^2$, $\mathbb{E}[\beta \mid y, X, \sigma^2] = \dfrac{g}{g+1}(X^T X)^{-1} X^T Y$.

  With Zellner's g-prior, if we let $\gamma = \sigma^{-2}$ and set the prior $\gamma \sim \text{Gam}(\nu_0/2, \nu_0\sigma_0^2/2)$, then the posterior distribution is $\sigma^2 \mid y, X \sim \text{InvGam}\left(\dfrac{\nu_0 + n}{2}, \dfrac{\nu_0\sigma_0^2 + SSR_g}{2}\right)$, where $SSR_g = y^T\left(I - \dfrac{g}{g+1} X(X^T X)^{-1} X^T\right) y$.

- **Hierarchical regression:** Say we want to run regression for m different groups which are different but somewhat related. For each group j, we have the within-group sampling model

  $$Y_{i,j} = \beta_j^T x_{i,j} + \varepsilon_{i,j}, \qquad \varepsilon_{i,j}, \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

  We can set up a between group sampling model, e.g. $\beta_1, \ldots, \beta_m \overset{iid}{\sim} MVN(\theta, \Sigma)$. Typical priors for this set-up:

  $$\begin{aligned} \theta &\sim MVN(\mu_0, \Lambda_0), \\ \Sigma &\sim \text{InvWishart}(\eta_0, S_0^{-1}), \\ \sigma^2 &\sim \text{InvGam}(\nu_0/2, \nu_0\sigma_0^2/2). \end{aligned}$$

- **Ordered probit regression:** Response variable Y is related to predictors X through a latent variable:

  $$\begin{aligned} \varepsilon_1, \ldots, \varepsilon_n &\overset{iid}{\sim} \mathcal{N}(0, 1), \\ Z_i &= \beta^T x_i + \varepsilon_i, \\ Y_i &= g(Z_i), \end{aligned}$$

  where g is usually taken to be non-decreasing. If Y can only take on K values, then set thresholds $-\infty = g_0 < g_1 < \cdots < g_K = \infty$, and have $Y_i = j$ if $g(Z_j) \in (g_{j-1}, g_j)$.

  If we use normal prior distributions, the joint posterior of $\{\beta, g_1, \ldots, g_K, Z_1, \ldots, Z_n\}$ given Y can be approximated using a Gibbs sampler (see Hoff p212).

6 Bayesian Model Selection

- If we believe that many of the regression coefficients are potentially equal to zero, then we come up with a prior distribution that reflects this possibility.

- **Spike and slab prior:** Mix an atom at $\{0\}$ (or $U[-\varepsilon, \varepsilon]$ or $\mathcal{N}(0, \varepsilon^2)$) with a diffuse distribution (e.g. $U[-M, M], \mathcal{N}(0, M^2)$). (We could put a prior on the proportion of each component as well.)

- **Alternative:** Can write $\beta_j = z_j b_j$, where $z_j \in \{0, 1\}$. Each value of $z = (z_1, \dots, z_p)$ corresponds to a different model.
    - Possible prior: Say z has p_z non-zero entries. Let X_z be the $n \times p_z$ matrix corresponding to the variables with $z_j = 1$, and let β_z be the $p_z \times 1$ vector consisting of β_j for which $z_j = 1$. Modified g-prior for β is $\beta_j = 0$ if $z_j = 0$, and $\beta_z \mid X_z, \sigma^2 \sim \text{MVN}(0, g\sigma^2(X_z^T X_z)^{-1})$.
    - Let $\gamma = \sigma^{-2}$, and give γ a Gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$ prior. Then the conditional density of y given X and z is

      $$p(y \mid X, z) = \frac{\pi^{-n/2}\Gamma([\nu_0 + n]/2)(1+g)^{-p_z/2}}{\Gamma(\nu_0/2)} \frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_0\sigma_0^2 + SSR_g^z)^{(\nu_0+n)/2}},$$

      where $SSR_g^z = y^T \left(I - \frac{g}{g+1} X_z (X_z^T X_z)^{-1} X_z^T \right) y$.
    - Assume that we further set $g = n$ and use the unit information prior for σ^2 for each model z (i.e. $\nu_0 = 1$, σ_0^2 the estimated residual variance under the least squares estimate for model z). To compare 2 models z_a and z_b, we may look at

      $$\frac{p(y \mid X, Z_a)}{p(y \mid X, Z_b)} = (1+n)^{(p_{z_b} - p_{z_a})/2} \left(\frac{s_{z_a}^2}{s_{z_b}^2} \right)^{1/2} \left(\frac{s_{z_b}^2 + SSR_g^{z_b}}{s_{z_a}^2 + SSR_g^{z_a}} \right)^{(n+1)/2}.$$

      There is a balance between model complexity and goodness of fit: A large value of p_{z_b} penalizes model z_b, but a large value of $SSR_g^{z_a}$ penalizes model z_a.
    - After setting up a prior for z, we can run the Bayesian machinery to get a posterior distribution for z, which is a posterior probability for each of the models.
    - **Prediction:** We could get a prediction from each of the models, then weight according to posterior probabilities.
- **Bayesian model averaging:** If we sample the posterior distribution of β S times, then the Bayesian model averaged estimate of β is $\hat{\beta}_{bma} = \frac{1}{S}\sum_{s=1}^{S} \beta^{(s)}$.

7 Special Topics

See notes for the material in this section.

- **Bayesian testing:** For 2 models M_1 and M_2,

  $$\underbrace{\frac{p(M_1 \mid y)}{p(M_2 \mid y)}}_{\text{posterior ratio}} = \underbrace{\frac{p(y \mid M_1)}{p(y \mid M_2)}}_{\text{Bayes factor}} \cdot \frac{p(M_1)}{p(M_2)}.$$

  The Bayes factor is also denoted by $B_{1,2}$. $\log B_{1,2}$ is called **evidence**.
- Jeffreys: "substantial evidence" if $\log_{10} B \in (1/2, 1)$, "strong evidence" if $\log_{10} B \in (1, 2)$, "decisive evidence" if $\log_{10} B > 2$.