Lecture 08 - MCMC Convergence

Recap: Helpful convergence visualizations

- Traceplot of the sampling chain
- Histograms over subsets of the chain
- Traceplots of multiple chains with different random start points.

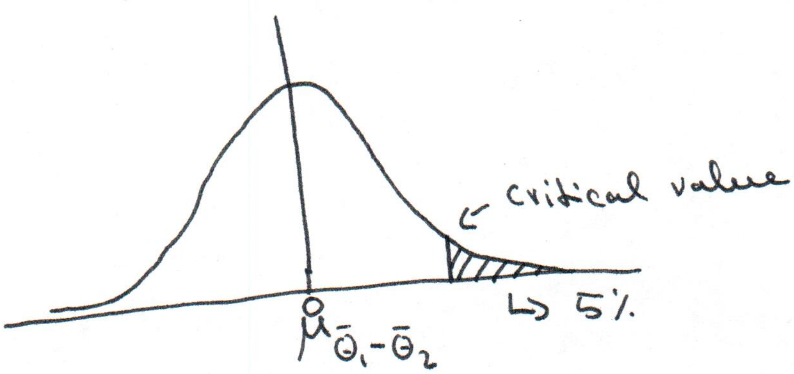
A bit more formal testing for convergence:

Geweke: Take two non-overlapping samples and compare the means.

$\Rightarrow$ Hypothesis test for difference of means

$$H_0: \mu_{\theta_1} - \mu_{\theta_2} = 0 \quad \Rightarrow \quad \mu_{\bar{\theta}_1 - \bar{\theta}_2} = 0$$

↳ mean of the distribution of the differences



what is the standard deviation of this distribution?

$$\sigma_{\bar{\theta}_1 - \bar{\theta}_2} = \sqrt{\frac{var(\theta_1)}{n_1} + \frac{var(\theta_2)}{n_2}}$$

Pavlos rule of thumb:

$$-2 < \frac{\bar{\Theta}_1 - \bar{\Theta}_2}{\sigma_{\bar{\Theta}_1 - \bar{\Theta}_2}} < 2 \quad \Rightarrow \quad 2 \cdot \sigma_{\bar{\Theta}_1 - \bar{\Theta}_2} > \bar{\Theta}_1 - \bar{\Theta}_2$$

↳ this is the 68, 95, 99,7 rule

⇒ If the null hypothesis is correct there is only ~ 5% chance of the mean difference being larger than $2 \cdot \sigma_{\bar{\Theta}_1 - \bar{\Theta}_2}$.

Gelman-Rubin test

- uses multiple chains
- Compares between chain variance and within chain variance.
- Large deviation ⇒ have not converged yet.

Assuming m chains of length n:

within chain variance:

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2 \qquad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\Theta_{ij} - \bar{\Theta}_j)^2$$

↳ mean of the chain variances

↳ variance of the jth chain.

Between chain variance:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\bar{\theta}})^2 \quad \text{with } \bar{\bar{\theta}} = \frac{1}{m} \sum_{j=1}^{m} \bar{\theta}_j$$

↳ mean of the chain means

⇒ variance of the chain means multiplied by the number of samples in each chain.

$$\text{var}(\theta | y) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B$$

⇒ just a weighted average

potential scale reducing factor:

$$\hat{R} = \sqrt{\frac{\text{var}(\theta | y)}{W}}$$

⇒ should be close to one when we converge.

If $\hat{R} \gg 1$ then between chain variance is greater than within chain variance.

Note: The starting points should be overdispersed to make this analysis useful.

Maize example:

Data:

Length: L

Diameter: D

Weight: W

Goal:

Estimate a simple linear relationship between these variables.

How do we find our model?

⇒ start simple and make it more complex if you need to

⇒ All models are wrong, but some models are useful. (George P. Box)

⇒ Maize kind of looks like a cylinder:

$$W \propto L \cdot D^2$$

$$W = \beta_0 + \beta_1 \cdot L \cdot D^2 + \underbrace{\epsilon}_{\text{error}} \qquad \epsilon \sim N(0, \sigma_\epsilon)$$

Now we need our likelihood:

$$P(\text{Data} \mid \Theta) = P(W_i, \underbrace{L_i, D_i}_{\text{fixed input, no stochasticity}} \mid \beta_0, \beta_1, \sigma_\epsilon)$$

↳ stochasticity caused by measurement error.

$P(w_i, L_i, D_i \mid \beta_0, \beta_1, \sigma_\epsilon)$

$$\propto \prod_i P(w_i \mid \beta_0, \beta_1, \sigma_\epsilon, x_i)$$

↳ measurement input L_i and D_i

$$\propto \prod_i e^{\frac{-(w_i - (\beta_0 + \beta_i x_i))^2}{2\sigma_\epsilon^2}}$$

$$\propto e^{-\sum_i \frac{(w_i - (\beta_0 + \beta_i x_i))^2}{2\sigma_\epsilon^2}}$$

Now we need to think about priors:

⇒ We don't really know anything about Maize

⇒ Make priors flat and uninformative

⇒ Normal priors are nice, because they are the conjugate prior to our likelihood.

$$P(\beta_0) = N(\mu_{\beta_0}, \sigma_{\beta_0})$$

$$P(\beta_1) = N(\mu_{\beta_1}, \sigma_{\beta_1})$$

What about $P(\sigma_\epsilon^2)$?

Although our likelihood is a Gaussian, it is not normal with respect to σ_ϵ^2!

The conjugate prior for the variance of a normal distribution is an Inverse gamma.

Instead we can also use the precision $\tau_\epsilon = \frac{1}{\sigma_\epsilon^2}$ and the conjugate prior is a gamma distribution.

In all we now have our posterior as:

$$p(\beta_0, \beta_1, \tau_\epsilon \mid w_i, x_i) \propto P(w_i \mid \beta_0, \beta_1, \tau_\epsilon, x_i) \cdot P(\beta_0) \cdot P(\beta_1) \cdot P(\tau_\epsilon)$$

We can sample from this using Metropolis-Hastings.

$\Rightarrow$ We need our proposal distribution.

$q(\theta^* \mid \theta^{(i)})$ for $\theta: \beta_0, \beta_1, \tau_\epsilon$

normal or uniform are generally good choices $\Rightarrow N(\theta^{(i)}, \lambda)$

↳ we need to tune λ (corresponds to our step size).

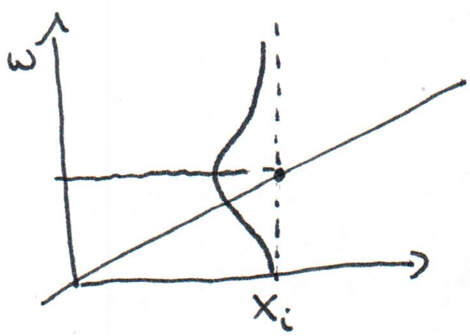
Predictive Probability:

How do we check if our model works?

- Sanity test: visual inspection
- Better: Evaluate model by making data predictions.
  ↳ Simulated data should have the same distribution as actual data.

In the Bayesian view we have probability distributions over our parameters, not value estimates.

⇒ for each input we have a predictive distribution



⇒ we need to average over all possible parameter values weighted by their posterior probability.

predictive probability distribution:

$$p(\omega^* \mid x^*, x, \omega) = \int \underbrace{p(\Theta \mid x, \omega)}_{\text{posterior}} \cdot \underbrace{p(\omega^* \mid x^*, \Theta)}_{\text{likelihood}} \, d\Theta$$

⇒ likelihood gives the likelihood of the new unseen data point for a given choice of parameter.

⇒ posterior says how likely this choice of parameter is.

How do we do this in praxis?

- We have an analytic form for the likelihood (because we designed it)

- We have (or can get) samples from our posterior.

From the homework we remember:

$$\int h(x) \cdot g(x)\, dx \propto \sum_{\substack{i \\ x \sim g(x)}} h(x_i)$$

$$\propto \sum_{\substack{i \\ x \sim u}} h(x_i) \cdot g(x_i)$$

⇒ take the samples from the posterior, plug them into the likelihood and we are good.

$$\Rightarrow \omega_{MAP} = \underset{\omega}{\operatorname{argmax}} \; P(\omega^* \mid x^*, \omega, x)$$