Convergence Checking and Model Fit Assessment

Wednesday, November 1, 2017 2:51 PM

Questions motivating this lecture:

1. How do I know that the samples of some parameter θ generated from my computational sampling procedure (e.g., Gibbs sampler) are really representative of the unknown distribution $f(\theta)$?

    a. Another way of asking the same question: how do I know that my Gibbs sampler has achieved a limiting distribution equivalent to $f(\theta)$?

    b. We've asked this question before, but now we'll tackle a closely related question: *what can I do about it if I suspect that my sampler is not quickly converging to* $f(\theta)$?

2. Presuming that my sampler is working well, how can I tell whether a model is a good fit to the data set?

    a. Are there reasons to suspect (meaningful) misspecification of the parametric structure of the model?

3. If I have multiple plausible models that might explain the data-generating process, how do I decide which one is the most credible model?

Convergence Diagnostics

Wednesday, November 1, 2017 2:57 PM

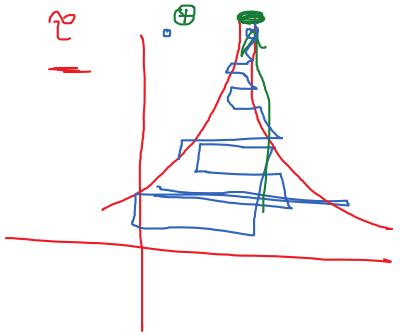
- So you're running a Gibbs sampler…
- How do you assess whether the sampler is producing samples of θ that are representative of $f(\theta)$?
- We've assessed this question before, in the lecture dealing with "Practical MCMC for estimating models"
  - Visual assessment of the Markov Chain
  - Geweke diagnostic (`geweke.diag`)
  - Raftery and Lewis diagnostic (`raftery.diag`)
  - Heidelberger diagnostic (`heidel.diag`)
- There is another diagnostic to consider: the Gelman and Rubin diagnostic (`gelman.diag`), sometimes called $\hat{R}$ or the "potential scale reduction factor"
  - According to Gelman and Hill, "for each parameter, the possible reduction in the width of its confidence interval, were the simulations to be run forever" with a target of less than 1.1
  - Mathematically (according to *BDA3*), for a parameter of interest ψ with m = the number of chains and n = the number of samples per chain, we can write (noting that i indexes iterations in a chain and j indexes chains)…
    - $W = \frac{1}{m}\sum_{j=1}^{m} s_j^2$ within variance
      - $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\psi_{ij} - \bar{\psi}_{\cdot j})^2$ variance in chain j
      - $\bar{\psi}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}$ $E[\psi]$ in chain j
      - $\bar{\psi}_{\cdot\cdot} = \frac{1}{m}\sum_{j=1}^{m}\bar{\psi}_{\cdot j}$
    - $B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\psi}_{\cdot j} - \bar{\psi}_{\cdot\cdot})^2$ between variables
    - $\widehat{var}^{+}(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$
    - $\hat{R} = \sqrt{\frac{\widehat{var}^{+}(\psi|y)}{W}}$ $\frac{W}{W}$
    - Note that as $n \to \infty$, $\hat{R} \to 1$ (because of how $\widehat{var}^{+}(\psi|y)$ is constructed)
- Bad news: sometimes *all* these convergence diagnostics yield misleading results.

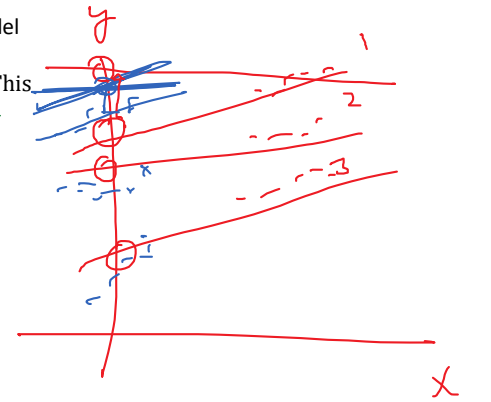
- Bad news: sometimes *all* these convergence diagnostics yield misleading results.
- Let's take a look at an example using the radon dataset often used in Gelman and Hill (*ARM*); the example comes from Thomas Wiecki (twiecki.github.io)

Methods to Speed Convergence

Thursday, November 2, 2017 1:27 PM

- Generic problem: when two parameters in a Bayesian model are very closely correlated, it can cause the sampler to encounter problems exploring the space
- Generic solution: try to break the correlation between the parameters
  - Example: when studying a distribution $f(\theta) \sim \Phi(\mu, \tau)$ for a vector-valued θ (for example, where $\theta_i, i = 1 \ldots m$ is a bunch of random intercepts or slopes *for* m-many units i), it can be the case that the sampler has trouble fully exploring the θ space for very small values of σ / large values of τ (i.e., when precision is very high or variance is very low)
  - Problem: correlation between θ and τ
  - Several ways to break this correlation
  - One idea: Write $\theta_o \sim \Phi(0,1)$, $\sigma_\theta = pow(\tau, -2)$ and $\theta = \mu + \sigma_\theta \theta_o$
  - Another idea: write $f(\theta - \mu) \sim \Phi(0, \tau)$, $\theta_i^a = \theta_i - \sum_{j=1}^{m} \theta_j$ and $\mu^a = \mu + \sum_{j=1}^{m} \theta_j$
- In a hierarchical model, centering various coefficients can help speed the convergence of the model
  - One very simple approach: instead of $\hat{y} = \beta_0 + \eta_i + \beta_1 x$, write $\hat{y} = \beta_0 + \eta_i + \beta_1 (x - \bar{x})$. This breaks the correlation between $\beta_0 / \overline{\eta_i}$ and β_1
  - Instead of writing $\beta_0 + \eta_i(0, \tau_\eta) + \zeta_i(0, \tau_\zeta)$, can write $\eta_i(\beta_0, \tau_\eta) + \zeta_i(0, \tau_\zeta)$





Posterior Predictive Densities

Thursday, November 2, 2017 2:02 PM

- Assuming that your model is sampling properly… how do we determine whether it's the right model?
- One way: check to see that simulated data generated from the model is consistent with the actual data from the data set
    - Idea: simulate data from $f(y|\theta)$ using the samples of θ that you drew, and then compare this to the empirical distribution $f_E(y)$ to assess similarity
    - This is similar to what you might do in a linear regression comparing $\hat{y}$ to the observed y (or, equivalently, $\hat{u}$ to y)

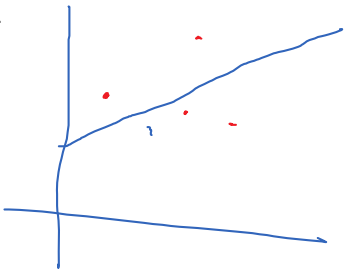
Model Comparison

Thursday, November 2, 2017 2:10 PM

- If I have more than two models, how do I know which model is the best fit to my data?
- Deviance: $\delta = -2 \ln L$, where $L = f(y|\hat{\theta})$ or the likelihood of the data given the Bayes estimates of the parameters $\hat{\theta} = E[\hat{\theta}_i]$
- Deviance Information Criterion:
  - $DIC = 2\delta + 2p_{DIC}$
  - $p_{DIC} = 2(\delta - E_\theta(\ln(f(y|\theta))) = 2\left(\delta - \frac{1}{s}\sum_{s=1}^{S} \ln p(y|\theta^s)\right)$
  - Computed easily in JAGS or WinBUGS/OpenBUGS
  - Lower is better
- Akaike Information Criterion: $AIC = \delta + 2k$ k = the length of θ (typically δ is computed using the MLE $\hat{\theta}$, not the Bayes estimates)
- Bayesian Information Criterion: $BIC = \delta + k \ln n$, where n is the sample size (typically δ is computed using the MLE $\hat{\theta}$, not the Bayes estimates)
- Bayes Factor for model comparison (hard to compute in many cases):
  - $BF = \frac{\int f(\theta_1|M_1) f(y|\theta_1, M_1)}{\int f(\theta_2|M_2) f(y|\theta_2, M_2)}$ $= \frac{f(y|M_1)}{f(y|M_2)}$
- Out-of-sample prediction / Cross-validation
- Gibbs/Stochastic Variable Selection and the "Bayesian LASSO"
  - Double-exponential distribution on the precision of a parameter

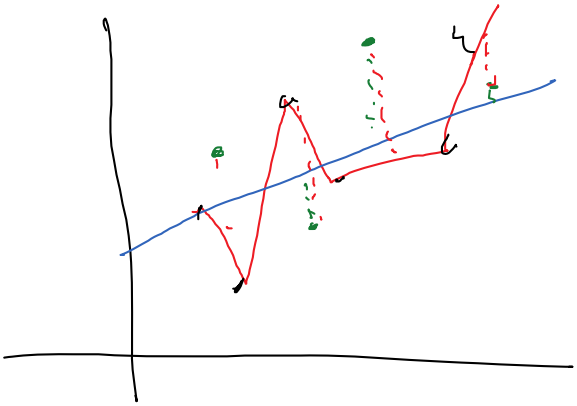
$\hat{\theta}$: "Bayes estimate"

$\int f(\theta|y)\,\theta\,d\theta$ posterior



$pr(M_1|y) \propto pr(y|M_1)\,pr(M_1)$

$\frac{f(y|M_1)}{f(y|M_2)} \times \frac{pr(M_1)}{pr(M_2)} = \frac{pr(M_1|y)}{pr(M_2|y)}$



"test set data"