GENOME 541 Section 2
Lecture 2
Gibbs Sampling

Yi Yin (yy224[at]uw.edu)
Shendure Lab
Department of Genome Sciences
University of Washington
Thursday, April 13, 2017

Review

- Bayes' Rule

$$p(\theta|Y) = \frac{p(\theta)L(Y;\theta)}{p(Y)}$$

- One parameter model and conjugacy

  Prior: Beta(a, b)

  Likelihood: Binomial(n, k)

  Posterior: Beta(a+k, b+(n-k))

can be read off directly from the kernel of the distribution

Preview (OK, we can get a posterior distribution, so what?)

- Posterior predictive distribution
- Monte Carlo sampling to obtain posterior summaries
- Going beyond one parameter:
    - Joint posterior is hard to sample from
    - But, full conditional distribution of each parameter is in closed-form
    - e.g., Normal distribution
- MCMC (Markov Chain Monte Carlo)
    - Gibbs sampling
    - Metropolis

Posterior predictive distribution

Let's go back to the Gamma-Poisson example. Remember counts are often modeled by a Poisson likelihood function, # of friends on Facebook, # of barcodes in a sequencing run…

$y_1, \ldots, y_n \mid \theta \sim Poisson(\theta)$, $\theta \sim Gamma(a, b)$. Then $\theta | y_1, \ldots, y_n \sim Ga(\widehat{a}, \widehat{b})$

$$\theta \mid y_1, \ldots, y_n \sim Gamma(a + \sum y_i, b + n)$$

Reminder: $\theta \sim Ga(a, b)$, where the Ga(a,b) pdf is $\frac{b^a}{\Gamma(a)}\theta^{a-1}exp(-b\theta)$

$$E(\theta) = a/b, V(\theta) = a/b^2$$

Now, suppose you did two experiments and collected two sets of counts, $Y^{(1)}$ and $Y^{(2)}$

$Y^{(1)}$: {1, 7, 10, 4, 3, 5, 8, 4, 6, 5} $\widehat{Y^{(1)}} = 5.3$

$Y^{(2)}$: {7, 10, 6, 6, 9, 4, 7, 9, 10, 6} $\widehat{Y^{(2)}} = 7.4$

How to obtain posterior summaries?

Under Ga(1,1) prior (weekly informative, with prior sample size 1), our posterior distributions are

$$\theta_1|Y^{(1)} \sim Ga(54, 11) \text{ with } \widehat{\theta_1}|Y^{(1)} = 4.9$$
$$\theta_2|Y^{(2)} \sim Ga(75, 11) \text{ with } \widehat{\theta_2}|Y^{(2)} = 6.8$$

Note that this is a bit biased, alternatively, you can center your prior mean on the MLE.

But what if we are interested more than a single point estimation like the posterior mean? For example, can we directly obtain a probability of $p(\widehat{\theta_1} < \widehat{\theta_2}|Y^{(1)}, Y^{(2)})$, or, can we alternatively obtain a predictive probability of $p(y_{n+1}^{(1)} < y_{n+1}^{(2)}|Y^{(1)}, Y^{(2)})$? (HW)

Why is it good to have a posterior predictive distribution?

- You always have uncertainty about your point estimate…
- The question is how to account for such uncertainty
- Plugging in your point estimate vastly underestimate such uncertainty

Since we can draw samples from the posterior distribution of the parameter, we can marginalize over the entire posterior distribution and integrate out the unknown parameters to get prediction on y_{n+1}.

Back to the posterior predictive distribution

For the Gamma-Poisson example, the posterior predictive distribution is in closed form, which is a negative binomial distribution.

Let's derive this (for short, we write $y^n=\{y_1, \dots, y_n\}$)

General form: $$p(Y^* \mid Y) = \int p(Y^* \mid \theta) p(\theta \mid Y) d\theta$$

$$\begin{aligned}
f(y|y^n) &= \int \mathrm{Pois}(y;\theta)\mathrm{Ga}(\theta;\widehat{a},\widehat{b})d\theta. \\
&= \frac{\widehat{b}^{\widehat{a}}}{y!\Gamma(\widehat{a})} \int \theta^{\widehat{a}+y-1} \exp\{-\theta(\widehat{b}+1)\} d\theta \\
&= \frac{\widehat{b}^{\widehat{a}}}{y!\Gamma(\widehat{a})} \frac{\Gamma(\widehat{a}+y)}{(1+\widehat{b})^{\widehat{a}+y}} \\
&= \frac{\Gamma(\widehat{a}+y)}{\Gamma(y+1)\Gamma(\widehat{a})} \left(\frac{\widehat{b}}{\widehat{b}+1}\right)^{\widehat{a}} \left(\frac{1}{\widehat{b}+1}\right)^{y},
\end{aligned}$$

which is neg-binomial$(\widehat{a}, \widehat{b}/(1+\widehat{b}))$.

Negative binomial distribution

- Negative binomial distribution is an over-dispersed generalization of the Poisson

- When you marginalize θ out of the Poission (y; θ) likelihood over a gamma distribution, you can obtain a negative-binomial

	pdf	Support	Mean	Var
$NB(\alpha, p)$	$f(x) = \binom{x+\alpha-1}{x} p^\alpha q^x$	$x \in \mathbb{Z}_+$	$\alpha q/p$	$\alpha q/p^2$

- For $(y|y^n) \sim \text{neg-binomial}(\widehat{a}, \widehat{b}/(1+\widehat{b}))$, we have

$$
\begin{aligned}
\mathrm{E}(y|y^n) &= \frac{\widehat{a}}{\widehat{b}} = \mathrm{E}(\theta|y^n) = \text{Posterior mean} \\
\mathrm{V}(y|y^n) &= \frac{\widehat{a}(\widehat{b}+1)}{\widehat{b}^2} = \mathrm{E}(\theta|y^n)\left(\frac{1+\widehat{b}}{\widehat{b}}\right),
\end{aligned}
$$

where variance is larger than the mean by an amount determined by $\widehat{b}$

What happens when n is large?

Note that as the sample size n increases, the posterior density for θ becomes more & more concentrated

$$V(\theta|y^n) = \widehat{a}/\widehat{b}^2 = (a + \sum_i y_i)/(b+n)^2 \approx \overline{y}/n \to 0$$

As we have less uncertainty about θ, inflation factor $(1+\widehat{b})/\widehat{b} \to 1$ and the predictive density $f(y|y^n) \to \text{Poisson}(\overline{y})$

In smaller samples important to inflate our predictive intervals to account for uncertainty in θ

What if you've never seen the negative-binomial distribution, or didn't recognize that this is a negative-binomial distribution?

Use simulation as approximation, which means:
You can still draw samples…

step 1: draw from $\mathrm{Ga}(\theta; \widehat{a}, \widehat{b})$ with $\theta \leftarrow$ rgamma(1, $\widehat{a}, \widehat{b}$)

step 2: plug the θ into Poisson with y $\leftarrow$ rpois(1, θ)

This is Monte Carlo sampling directly from the posterior predictive distribution

Suppose we can sample S values from the posterior distribution of θ, so that

$$\theta^{(1)}, \ldots, \theta^{(S)} \overset{\text{iid}}{\sim} p(\theta \mid Y)$$

for large S.

Law of Large Numbers

$$\frac{1}{S}\sum \theta^{(i)} \rightarrow E[\theta \mid Y]$$
$$\frac{1}{S}\sum g(\theta)^{(i)} \rightarrow E[g(\theta) \mid Y]$$

Sample means converge to their expectations.

1. Does not rely on large n assumptions (asymptotically normal, but restrictive for small sample size)

2. Monte Carlo estimate gives us quantification of uncertainty. Empirical distribution of the sample $\theta^{(1)},\ldots,\theta^{(S)}$ approximates $p(\theta|Y)$. See next slides for visualization

3. Sample moments/quantiles/functions approximate true moments/quantiles/functions, for example, proportions of samples where event $g(\theta^{(i)}) > c$ approximates $p(g(\theta) > c|Y)$

4. Easily manageable even with high dimensional parameters

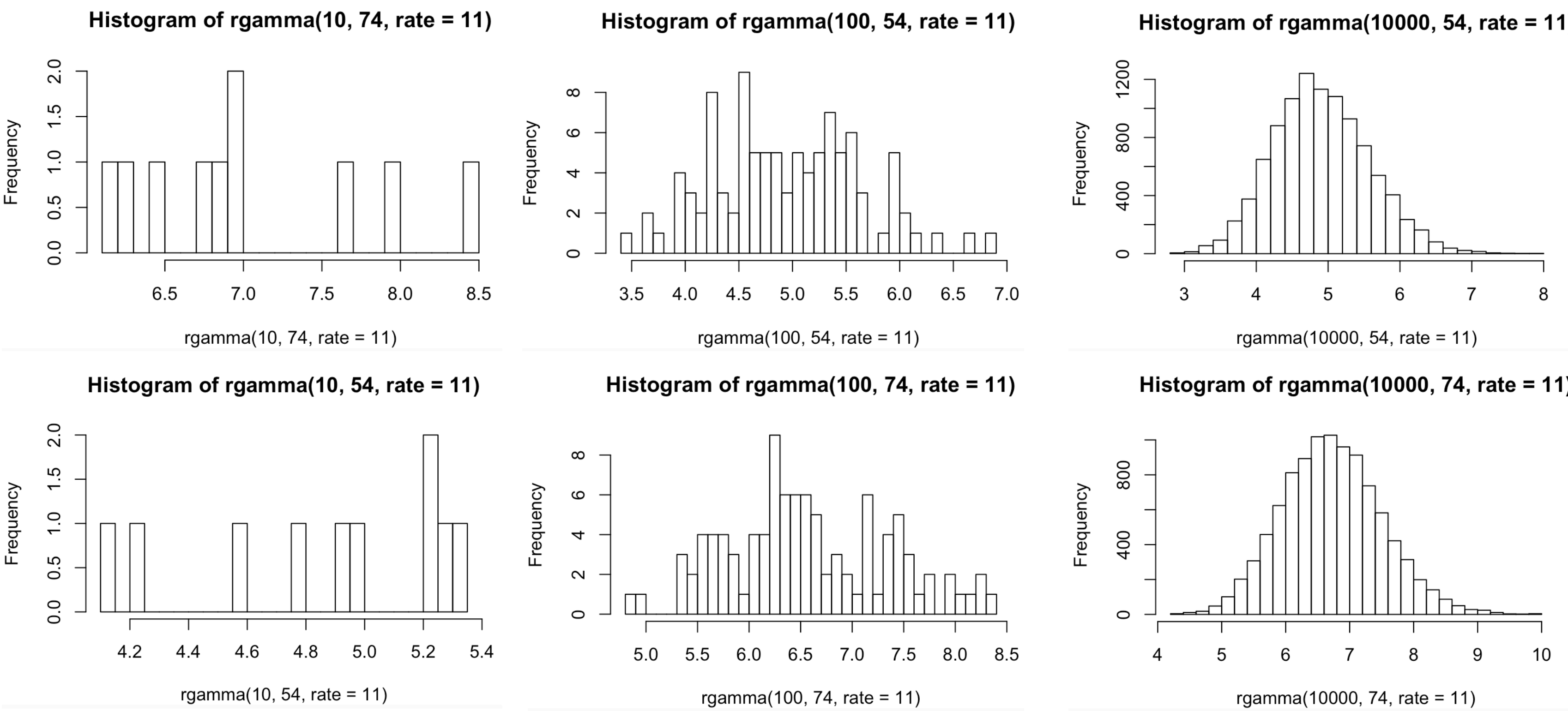
Remember our Gamma-Poisson example:

$$\theta \mid y_1, \ldots, y_n \sim Gamma(a + \sum y_i, b + n)$$

$Y^{(1)}$: {1, 7, 10, 4, 3, 5, 8, 4, 6, 5} $\quad \widehat{Y^{(1)}} = 5.3$

$Y^{(2)}$: {7, 10, 6, 6, 9, 4, 7, 9, 10, 6} $\quad \widehat{Y^{(2)}} = 7.4$

So, $\theta_1 | Y^{(1)} \sim$ Gamma (54, 11) and $\theta_2 | Y^{(2)} \sim$ Gamma (75, 11);



Prior predictive distribution

In addition to posterior predictive distribution, it's often very helpful to plot your prior predictive distribution and see if it looks like your data before doing any posterior computation/simulation etc.

$$f(y) = \int L(y;\theta)p(\theta)d\theta = \int_0^\infty Pois(y;\theta)Ga(\theta;a,b)d\theta$$

Used the likelihood function, but didn't use the data

In practice, should always try to draw samples from prior predictive distribution and examine whether they look like your data!

Summary for one-parameter model

- ✓ Formalize your problem with Bayesian framework: what sampling model (same with frequentist), how to choose a prior…

- ✓ Plot prior predictive distribution and see if it looks like your data

- ✓ Derive the posterior, (make sure it's a proper distribution - we haven't seen an bad example yet), solve the posterior analytically or by simulation

- ✓ Obtain posterior summaries and other quantities of interest: posterior mean, posterior variance, functions of the posterior parameter…

- ✓ You may also want to derive posterior predictive distribution and marginalize out the parameter, if the interest of problem is prediction instead of parameter inference.

Perhaps something more useful: Normal (Gaussian) models

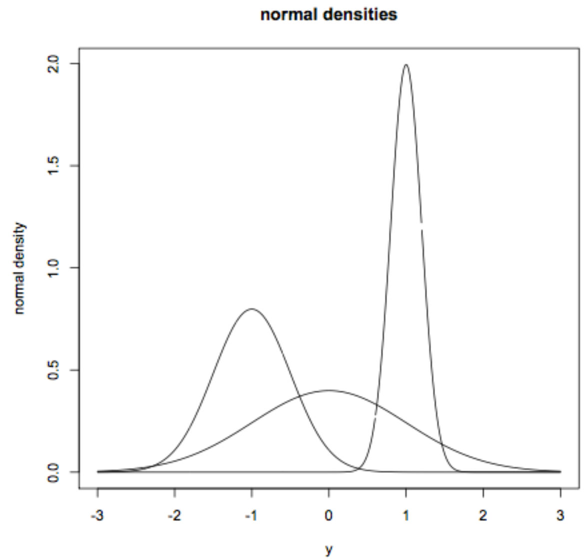
For a random variable $Y \sim N(\mu, \sigma^2)$, the pdf is

$$f(y|\theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\theta)^2\right\}, \quad y \in \Re = (-\infty, \infty).$$

The normal density is symmetric about the mean, median & mode θ

95% probability within $\pm 1.96\sigma$ (approximately two standard deviations) of the mean

rnorm, dnorm, pnorm, qnorm in R take mean and <u>standard deviation</u> σ as arguments



It is amazing how often real data are close to normally distributed

Likely a consequence of central limit theorem - sum or mean of a set of random variables is normally distributed

Occurs under very general conditions

A note on parameterization:

Independent observations $Y = (y_1, y_2, \dots y_n)$

$$y_i \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)$$

unknown parameters μ and σ^2.

Some prefer to work with the *precision*, ϕ, where $\phi = 1/\sigma^2$.

Likelihood:

$$\begin{aligned} \mathrm{L}\,(\mathrm{Y}\,;\mu\,,\phi) \;&\propto\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \phi^{1/2} \exp\left\{ -\frac{1}{2}\phi(y_i - \mu)^2 \right\} \\ &\propto\; \phi^{n/2} \exp\left\{ -\frac{1}{2}\phi \sum_i (y_i - \mu)^2 \right\} \end{aligned}$$

Now we have two parameters μ and ϕ, how should we pick a prior distribution?

You need a joint prior $p(\mu, \phi)$

Conjugate prior (Normal-Gamma)

$$p(\mu, \phi) = p(\mu|\phi)p(\phi)$$

Semi-conjugate prior (Gibbs sampling)

$$p(\mu, \phi) = p(\mu)p(\phi)$$

Prior: The conjugate prior for (μ, ϕ) is Normal-Gamma.

$$\begin{aligned} \mu|\phi &\sim \mathrm{N}(\mu_0, 1/(\kappa_0\phi)) \\ \phi &\sim \mathrm{G}(v_0/2, \mathrm{SS}_0/2) \end{aligned}$$

Note that μ depends on ϕ

where $-\infty < \mu_0 < \infty, \kappa > 0, \mathrm{SS}_0 > 0, v_0 > 0$

This can be expressed as

$$p(\mu, \phi) \propto \phi^{v_0/2-1} \exp\left\{-\phi\frac{\mathrm{SS}_0}{2}\right\}\phi^{1/2}\exp\left\{-\phi\frac{\kappa_0}{2}(\mu-\mu_0)^2\right\}$$

Likelihood:

$$\begin{aligned} L(\mu, \phi|Y) &\propto \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\phi^{1/2}\exp\left\{-\frac{1}{2}\phi(y_i-\mu)^2\right\} \\ &\propto \phi^{n/2}\exp\left\{-\frac{1}{2}\phi\sum_i(y_i-\mu)^2\right\} \end{aligned}$$

Posterior:

$$\mu \mid \phi, Y \sim \mathrm{N}\left(\mu_n, \frac{1}{\kappa_n\phi}\right)$$

$$\phi \mid Y \sim \mathrm{G}\left(\frac{v_n}{2}, \frac{\mathrm{SS}_n}{2}\right)$$

where

$$\begin{aligned} \kappa_n &= \kappa_0 + n \\ \mu_n &= \frac{\phi n\bar{y} + \phi\kappa_0\mu_0}{\phi\kappa_n} \\ v_n &= v_0 + n \\ \mathrm{SS}_n &= \mathrm{SS}_0 + \mathrm{SS} + \frac{n\kappa_0}{\kappa_n}(\bar{y}-\mu_0)^2 \end{aligned}$$

Interpretation:

κ_n: like sample size for estimating μ (precision $= \phi\kappa_n$)

μ_n: expected value for μ is weighted average

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

v_n: degrees of freedom for estimating ϕ

$$\phi \sim G(a/2, b/2) \Leftrightarrow \phi b \sim \chi_a^2 \text{ with degrees of freedom } a$$

$\mathrm{SS}_n = \mathrm{SS}_0 + \mathrm{SS} + \frac{n\kappa_0}{\kappa_n}(\bar{y}-\mu_0)^2$: total posterior variation

- prior variation,
- observed variation (sum of squares),
- variation between prior mean and sample mean

What if we want to model μ and ϕ independently?

Let's explore Gibbs sampler with this example

Prior: $$p(\mu, \phi) = p(\mu)p(\phi) \qquad \begin{aligned} &\mu \sim N(\mu_0, 1/\tau_0) \\ &\phi \sim G(\nu_0/2, SS_0/2) \end{aligned}$$

Likelihood:
$$\begin{aligned} L(\mu, \phi | Y) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \phi^{1/2} \exp\left\{ -\frac{1}{2}\phi(y_i - \mu)^2 \right\} \\ &\propto \phi^{n/2} \exp\left\{ -\frac{1}{2}\phi \sum_i (y_i - \mu)^2 \right\} \end{aligned}$$

Posterior: