Computation III

Statistics 220

Spring 2005

Conditions for Gibbs Sampling to Work

While you can always run the chain, it may not give the answer you want. That is, the realizations may not have the desired stationary distribution.

- One-step transitions: $p(\theta|\theta^0)$
- n-step transitions: $p_n(\theta|\theta^0)$
- Stationary distribution: $\pi(\theta) = \lim_{n\to\infty} p(\theta|\theta^0)$

If the stationary distribution exists, it satisfies

$$\pi(\theta) = \int p(\theta|\phi)\pi(\phi)d\phi$$

A stronger condition which shows that $p(\theta)$ is the density of the stationary distribution is

$$\pi(\theta)p(\phi|\theta) = \pi(\phi)p(\theta|\phi)$$

holds for all θ and ϕ (detailed balance).

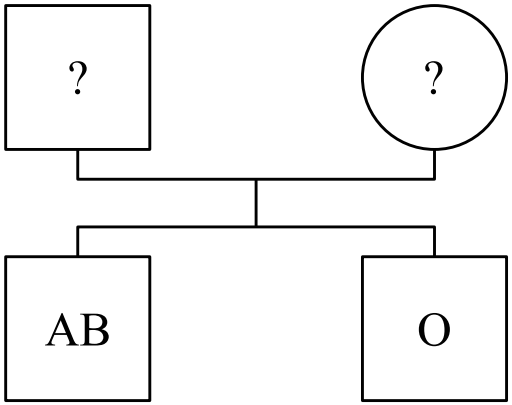
Note that detailed balance $\Longrightarrow$ stationarity but stationarity doesn't imply detailed balance.

If the following two conditions hold, the chain will have the desired stationary distribution.

1. Irreducibility: The chain generated must be irreducible. That is it is possible to get from each state to every other state in a finite number of steps.

   Not all problems lead to irreducible chains.

Example: ABO blood types



The children's data implies that the child with blood type AB must have genotype AB and that the child with blood type O must have genotype OO.

The only possible way for the two children to inherit those genotypes is for one parent to have genotype AO and for the other parent to have genotype BO. However it is not possible to say which parent has which genotype with certainty. By a simple symmetry argument

$$P[\text{Dad} = AO \ \& \ \text{Mom} = BO] = P[\text{Dad} = BO \ \& \ \text{Mom} = AO] = 0.5$$

Lets try running a Gibbs sampler on this data to get the distribution of genotypes for the two parents, by first generating mom's genotype given dad's and then dad's given mom's. Let start the chain with Dad $= AO$.

- Step 1: Generate Mom

$$P[\text{Mom} = AO|\text{Dad} = AO] = 0$$
$$P[\text{Mom} = BO|\text{Dad} = AO] = 1$$

  so Mom $= BO$.

- Step 2: Generate Dad

$$P[\text{Dad} = AO|\text{Mom} = BO] = 1$$
$$P[\text{Dad} = BO|\text{Mom} = BO] = 0$$

  so Dad $= AO$.

This implies that every realization of the chain has Mom $= BO$ & Dad $= AO$.

If the chain is started with Dad $= BO$, every realization of that chain will have Mom $= AO$ & Dad $= BO$.

The reducible chain in this case does not have the correct stationary distribution. (Well reducible chains don't really have stationary distributions anyway). But running the described Gibbs sampler will not correctly the describe the distribution of the mother and father's genotypes.

2. Aperiodicity: Don't want a periodic chain (e.g. certain states can only occur on when t is even say)

   This violates the idea that each state has a long run frequency marginally.

Starting Points

For every chain you need to specify a starting point. There are a number of approaches for choosing this.

1. Prior means

   In pump example, set $\beta^0 = E[\beta] = \frac{\delta}{\gamma}$.

2. Estimate from data

   In pump example, $E[l_i] = \frac{\alpha}{\beta}$, so set $\beta^0 = \frac{\alpha}{\bar{l}}$.

3. Sample from prior (or some other distribution)

   This idea can be combined with other ideas, such as sampling around a data based estimate.

4. Ad hoc choices

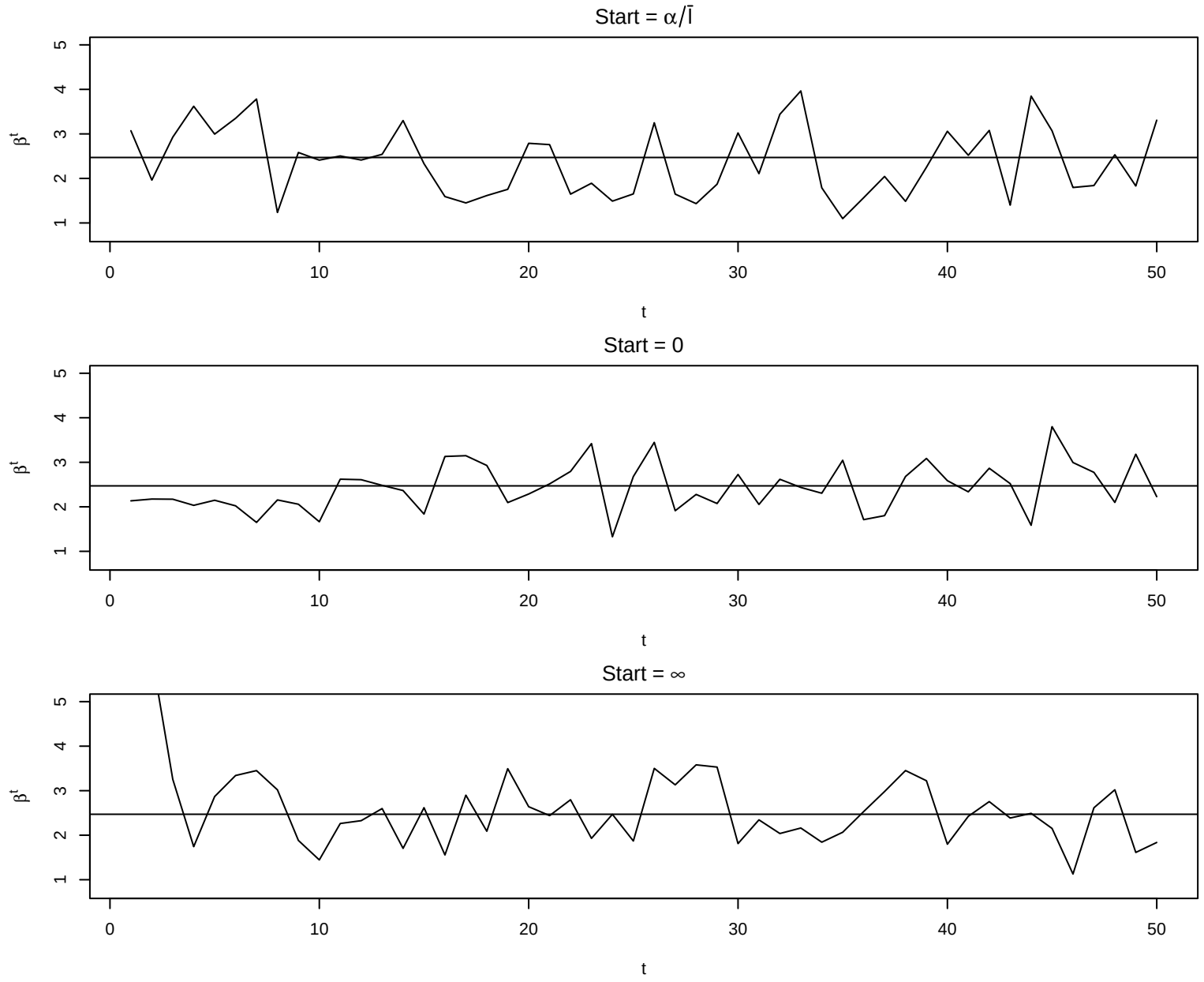
   In pump example, set $\beta^0 = \infty$ or $\beta^0 = 0$

5. Multiple Starting Points

   You do not need to use a single run for your chain. Instead you can run multiple chains with different starting values and then combine the samples for you analysis. `WinBugs` allows to this and Andy Gelman's `RBugs` front end to `WinBugs` requires at least two chains to be run to check for convergence of the chains.

   For simpler problems, it can be useful to start from well dispersed starting points as it is easier to check to see if the chains have converged and adaquately covered the sample space. If the starting points are similar it can be hard to determine whether similar results for each chain are due to convergence of the chain or whether its a slow moving chain and none of the chains have moved far from there starting points.

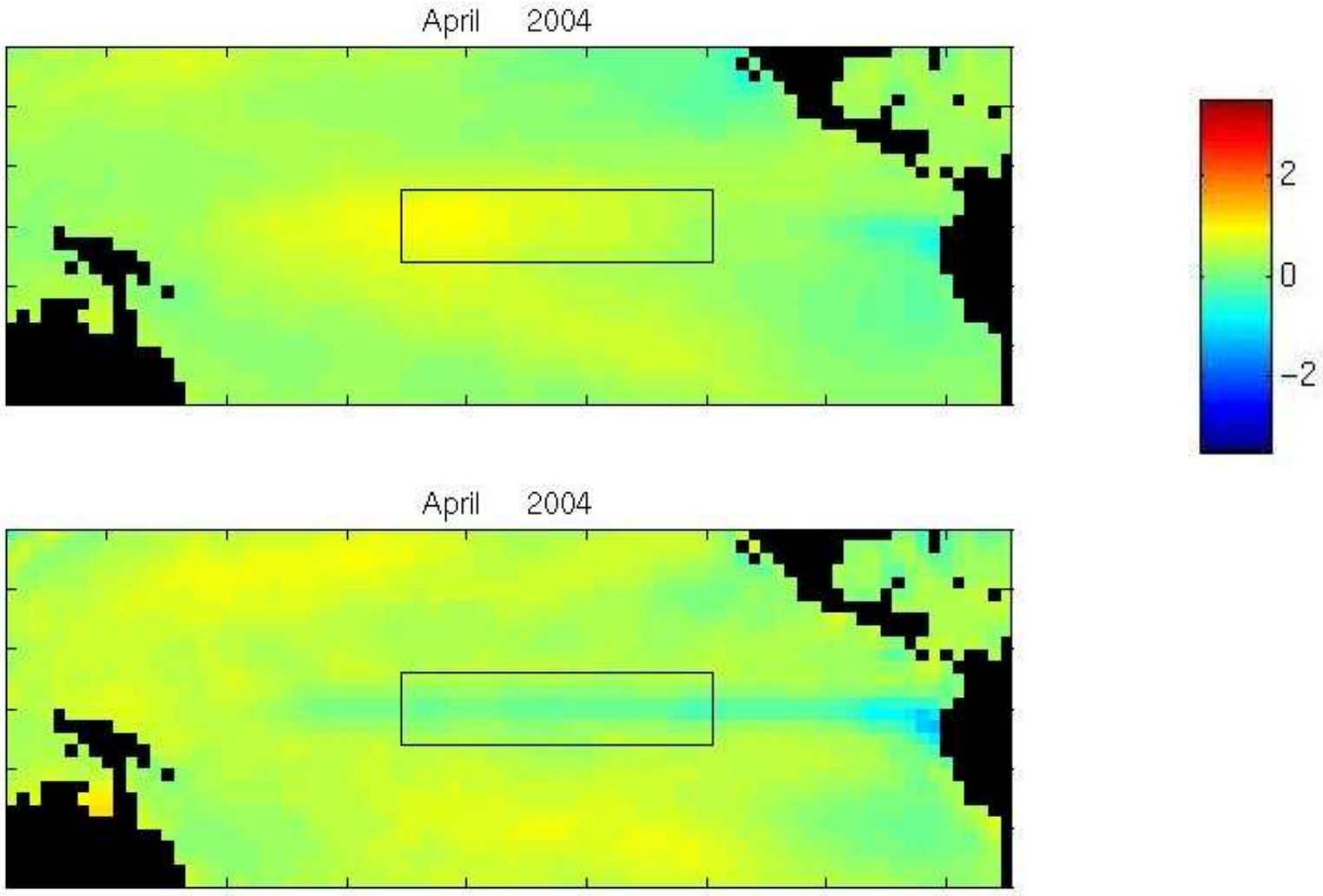
For many problems, the choice of starting values can be important. The stationary distribution is an asymptotic property and it may take a long time for the chain to converge.


Start = α/Ī
Start = 0
Start = ∞
β^t
t
0
10
20
30
40
50
1
2
3
4
5

Starting with $\beta^0 = \infty$ (actually 10^{100}), the initial draws are not consistent with the stationary distribution seen later in the chain.

While for this example, the problem clears up quickly, for other problems it can take a while.

This is more common which larger problems, that might have millions, or maybe billions of variables being sampled in a complete single scan through the data. This can occur with large space time problems, such as the Tropical Pacific sea surface temperature predictions discussed at <http://www.stat.ohio-state.edu/∼sses/collab_enso.php>.



April 2004 anomaly forecast (top) and observed anomaly (bottom) based on Jan 1970 to September 2003 data.

The usual approach to eliminate a poor choice of starting values is to have a "burn-in" period where the initial samples are thrown away since they may not be representative of samples from the stationary distribution.

This was done in the SST example where the first 1000 imputations from a total run of 11000 imputations were discarded.

The following table contains estimates of the posterior means of the 11 parameters in the pump example with 3 different starting points. The first 200 imputations were discarded and then the next 1000 imputations were sampled.

Pump	$\beta^0 = \frac{\alpha}{\bar{l}}$	$\beta^0 = 0$	$\beta^0 = \infty$
1	0.0716	0.0694	0.0692
2	0.1575	0.1508	0.1496
3	0.1037	0.1029	0.1039
4	0.1248	0.1230	0.1220
5	0.6099	0.6273	0.6023
6	0.6162	0.6152	0.6145
7	0.8192	0.8374	0.7907
8	0.8285	0.8301	0.7902
9	1.2651	1.3187	1.2341
10	1.8609	1.8609	1.8105

β	2.5148	2.4363	2.5665

Often the bigger the problem, the longer the burn-in period desired. However those are the problems where time considerations will limit the total number of imputations that can be done. So you do want to think about starting values for your chain.

Gibbs sampling and Bayes - Choice of priors

For Gibbs sampling to be efficient, the draws in each step of the procedure need to be feasible.

That suggests that conjugate distributions need to be used as part of the hierarchical model, as was done in the pump example.

However conjugacy is not required, as rejection sampling with log-concave distributions might be able to be used in some problems.

This idea, and others which probably won't be discussed, are sometimes used in the software package `WinBUGS`. However for some problems the model you want to analyze is not conjugate and the tricks to get around non-conjugacy won't work. For example, lets change model for the pump example to

$$
\begin{aligned}
s_i|\lambda_i &\overset{ind}{\sim} Poisson(\lambda_i t_i) \\
\lambda_i|\mu,\sigma^2 &\overset{iid}{\sim} LogN(\mu,\sigma^2) \\
\mu &\sim Logistic(1,100) \\
\sigma^2 &\sim Weibull(1,100)
\end{aligned}
$$

Good luck on writing down a simple Gibbs sampler on this model (I think). However, `WinBugs` will handle it (and it gives similar answers for the λ's, though there is less shrinkage due the the more diffuse prior).

Other sampling techniques are needed for more complicated problems.

Metropolis - Hastings Algorithm

A general approach for constructing a Markov chain that has the desired stationary distribution $p(\theta)$.

1. Proposal distribution: Assume that a time $t-1$ the chain is at θ^{t-1}. Need to propose a new state θ^* for time t with distribution $J(\theta|\theta^{t-1})$.

2. Calculate the Hastings' ratio

$$r = \frac{p(\theta^*)J(\theta^{t-1}|\theta^*)}{p(\theta^{t-1})J(\theta^*|\theta^{t-1})}$$

3. Acceptance/Reject step

   Generate $u \sim U(0,1)$ and set

$$\theta^t = \begin{cases} \theta^* & \text{if } u \leq min(r,1) \\ \theta^{t-1} & \text{otherwise} \end{cases}$$

Notes:

1. Gibbs sampling is a special case of M-H as for each step,

$$r = \frac{p(\theta^*)J(\theta^{t-1}|\theta^*)}{p(\theta^{t-1})J(\theta^*|\theta^{t-1})} = 1$$

   which implies the relationship also holds for a complete scan through all the variables.

2. The Metropolis (Metropolis et al, 1953) algorithm was based on a symmetric proposal distribution $J(\theta^*|\theta^{t-1}) = J(\theta^{t-1}|\theta^*)$

$$r = \frac{p(\theta^*)}{p(\theta^{t-1})}$$

   So a higher probability state will always be accepted.

3. As with many other sampling procedures, $p(\theta)$ and $J(\theta^*|\theta^{t-1})$ only need to be known up to normalizing constants as they will be cancelled out when calculating the Hastings' ratio.

4. Periodicity isn't a problem usually. For many proposals, $J(\theta^{t-1}|\theta^{t-1}) > 0$. Also if $r < 0$ for some combinations of θ^{t-1} and θ^*, $P[\theta^t = \theta^{t-1}|\theta^{t-1}] > 0$, thus some states have period 1, which implies the chain is aperiodic.

5. Detailed balance is easy to show.

6. The big potential problem is irreducibility. However by setting the proposal J to correspond to a irreducible chain solves this.

Proposal distribution ideas:

1. Approximate the distribution. For example use a normal with similar means and variances. Or use a t with a moderate number of degrees of freedom.

2. Random walk

$$J(\theta^*|\theta^{t-1}) = q(\theta^* - \theta^{t-1})$$

   If there is a continuous state process, you could use

$$\theta^* = \theta^{t-1} + \epsilon; \quad \epsilon \sim q(\cdot)$$

3. Autoregressive chain

$$\theta^* = a + B(\theta^{t-1} - a) + \epsilon; \epsilon \sim q(\cdot)$$

   For the random walk and autoregressive chains, q does not need to correspond to a symmetric distribution (though that is common).

4. Independence sampler

$$J(\theta^*|\theta^{t-1}) = q(\theta^*)$$

For an independence sampler you want q to be similar to p.

$$r = \frac{p(\theta^*)q(\theta^{t-1})}{p(\theta^{t-1})q(\theta^*)}$$

If they are too different, $\frac{q(\theta^{t-1})}{p(\theta^{t-1})}$ could get very small, making it difficult to move from state θ^{t-1}. (The chain mixes slowly).

5. Block at a time

   Deal with variables in blocks like the Gibbs sampler. Sometimes referred to Metropolis within Gibbs.

   Allows for complex problems to be broken down into simpler ones.

   Any M-H style update can be used within each block (e.g. random walk for one block, independence sampler for the next, Gibbs for the one after that).

   Allows for a Gibbs style sampler, but without the worry about conjugate distributions in the model to make sampling easier.

Pump Example:

$$
\begin{aligned}
s_i|\lambda_i &\overset{ind}{\sim} Poisson(\lambda_i t_i) \\
\lambda_i|\mu,\sigma^2 &\overset{iid}{\sim} LogN(\mu,\sigma^2) \\
\mu &\sim N(\delta,\tau^2) \\
\sigma^2 &\sim \text{Inv}-\chi^2(\nu,\gamma)
\end{aligned}
$$

Can perform Gibbs on μ and σ^2 easily, but not on λ, due the non-conjugacy of the Poisson and log Normal distributions.

Step $i, i = 1, \ldots, 10$ (M-H):

Sample λ_i from $\lambda_i|s, \mu, \sigma^2$ with proposal $\lambda_i^* \sim logN(\lambda_i, \theta^2)$ (Multiplicative random walk)

$$
\begin{aligned}
r &= \frac{(\lambda_i^* t_i)^{s_i} e^{-\lambda_i^* t_i} \frac{1}{\lambda_i^* \sigma} \phi\left(\frac{\log \lambda_i^* - \mu}{\sigma}\right)}{(\lambda_i t_i)^{s_i} e^{-\lambda_i t_i} \frac{1}{\lambda_i \sigma} \phi\left(\frac{\log \lambda_i - \mu}{\sigma}\right)} \times \frac{\frac{1}{\lambda_i \theta} \phi\left(\frac{\log \lambda_i - \lambda_i^*}{\theta}\right)}{\frac{1}{\lambda_i^* \theta} \phi\left(\frac{\log \lambda_i^* - \lambda_i}{\theta}\right)} \\
&= \left(\frac{\lambda_i^*}{\lambda_i}\right)^{s_i} e^{-(\lambda_i^* - \lambda_i) t_i} \frac{\phi\left(\frac{\log \lambda_i^* - \mu}{\sigma}\right)}{\phi\left(\frac{\log \lambda_i - \mu}{\sigma}\right)}
\end{aligned}
$$

Step 11 (Gibbs): Sample μ from $\mu|\lambda, \sigma^2, \delta, \tau^2 \sim N(\text{mean}, \text{var})$ where

$$\begin{aligned} \text{mean} &= \text{var}\left(\frac{1}{\sigma^2}\sum \log \lambda_i + \frac{\delta}{\tau^2}\right) \\ \text{var} &= \left(\frac{n}{\sigma^2} + \frac{\delta}{\tau^2}\right)^{-1} \end{aligned}$$

Step 12 (Gibbs): Sample σ^2 from

$$\sigma^2|\lambda, \mu, \nu, \gamma \sim \text{Inv}{-}\chi^2\left(\nu + n, \gamma + \sum(\log \lambda_i - \mu)^2\right)$$