# BTRY 4840/6840, CS 4775
# Computational Genetics and Genomics

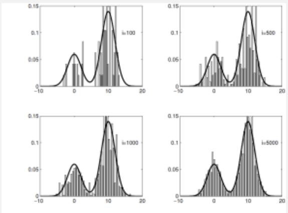


*Figure 6.* Target distribution and histogram of the MCMC samples at different iteration points.

October 25, 2018

Andrieu *et al.* (2003)

## Announcements

- Problem set 4 due today
- Problem set 5 out today
- Final project proposals due Tuesday (Oct 30)
- Reading on MCMC posted on website
  - A bit advanced: no need to understand it, but a good resource for learning about this topic

## Today's lecture

- More on MCMC
  - General properties, reversibility
  - Gibbs sampling and reversibility
  - Gibbs sampling for motif finding
  - STRUCTURE
  - Convergence

## MCMC background: Sampling

- Suppose we can draw samples $x^{(1)}, x^{(2)}, \dots, x^{(N)}$ from a pdf of interest $p(x)$
- Can then estimate several properties such as:

$$E[f(x)] = \int f(x)p(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f\left(x^{(i)}\right) \qquad [f \text{ any function}]$$

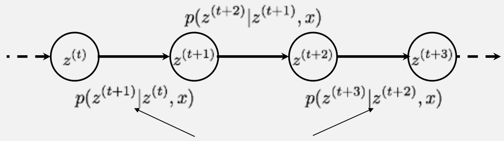
$$p(\tilde{x}_j, \dots, \tilde{x}_k) = \int p(x_j, \dots, x_k) I(x_j = \tilde{x}_j, \dots, x_k = \tilde{x}_k)dx$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} I\left(x_j^{(i)} = \tilde{x}_j, \dots, x_k^{(i)} = \tilde{x}_k\right)$$

- Key idea: each sample $x^{(i)}$ obtained in proportion to its probability, approximations implicitly include $p(x)$

## MCMC typically used to infer hidden variables

- Often with MCMC we:
  - Have observed data $x$ (often multivariate)
  - Want to infer unobserved variables $z$ (or parameters) related to $x$ via a model
  - That is, we wish to sample from $p(z|x)$, even if this is complex
  - We do so using a Markov chain


$p(z^{(t+2)}|z^{(t+1)}, x)$
$z^{(t)}$
$z^{(t+1)}$
$z^{(t+2)}$
$z^{(t+3)}$
$p(z^{(t+1)}|z^{(t)}, x)$
$p(z^{(t+3)}|z^{(t+2)}, x)$

Transition probabilities

## Markov chain Monte Carlo (MCMC) basics

- Want to sample from some very complex $p(x)$ that is:
  - Hard to sample from directly
  - Quick to compute $p(x)$ for given $x$
- MCMC in a nutshell:
  - Aim: construct a Markov chain with each state $x^{(i)}$ a sample from the distribution of interest $p(x)$
  - How? Define transition probabilities between states such that the stationary distribution = $p(x)$
    - Initial samples often *not* from stationary distribution: called burn-in
    - Informally: after huge number of iterations, samples are from $p(x)$
  - When $x^{(i)} \sim p(x)$, get unbiased sample of distribution
- Questions: How do we define transition probabilities? Will the chain converge to the stationary distribution?

## Ensuring stationary distribution of chain is distribution we want to sample from

- Markov chain is reversible wrt distribution $p\left(z^{(t)} \middle| x\right)$ if:

$$p\left(z^{(t)} \middle| x\right) p\left(z^{(t+1)} \middle| z^{(t)}, x\right) = p\left(z^{(t+1)} \middle| x\right) p\left(z^{(t)} \middle| z^{(t+1)}, x\right)$$

  - Also called detailed balance
    - ➢ Transition $z^{(t)} \rightarrow z^{(t+1)}$ or reverse with equal probability
- Let $\pi\left(z^{(t)} \middle| x\right)$ be stationary distribution of chain, then:

$\pi\left(z^{(t+1)} \middle| x\right) = \sum_{z^{(t)}} \pi\left(z^{(t)} \middle| x\right) p\left(z^{(t+1)} \middle| z^{(t)}, x\right)$ [by def of stationarity]

  - ➢ Multiplying sample from stationary distribution by transition probabilities produces a value from the stationary distribution
  - – Synonyms: stationary, invariant, equilibrium distribution
- In fact, reversibility wrt $p\left(z^{(t)} \middle| x\right)$ implies it is stationary dist.

$$\sum_{z^{(t)}} p\left(z^{(t)} \middle| x\right) p\left(z^{(t+1)} \middle| z^{(t)}, x\right) = \sum_{z^{(t)}} p\left(z^{(t+1)} \middle| x\right) p\left(z^{(t)} \middle| z^{(t+1)}, x\right)$$
$$= p\left(z^{(t+1)} \middle| x\right) \sum_{z^{(t)}} p\left(z^{(t)} \middle| z^{(t+1)}, x\right) = p\left(z^{(t+1)} \middle| x\right)$$

## Will obtain stationary distribution under specific conditions

- We must define transition equations such that the stationary distribution they imply is the distribution we wish to sample
  - In practice, we leverage reversibility to accomplish this
- To ensure convergence to the stationary distribution, transition equations must also be:
  - Aperiodic: the chain doesn't revisit states in periodic fashion
  - Irreducible: for any state, there is a positive probability of visiting all other states. (That is, the transition graph is connected.)
- Given these conditions, the MCMC will converge to stationary distribution at some point
  - Initially it will generally not sample from stationary distribution, but will eventually

## Gibbs sampling overview

- Gibbs sampling: applies to multivariate distributions

$$P(z_1, z_2, \ldots, z_n)$$

  - One of many MCMC approaches
  - Examples:
    - Motif finding: variables are start positions in $t$ sequences
    - STRUCTURE method (will discuss more later)
- General Gibbs sampling algorithm:

```
1. Randomly initialize starting point $z_j^{(0)}$, for $1 \le j \le n$
2. For $i = 0$ to $N - 1$:
    • Sample $z_1^{(i+1)} \sim p\left(z_1 \middle| z_2^{(i)}, z_3^{(i)}, \ldots, z_n^{(i)}\right)$
    • Sample $z_2^{(i+1)} \sim p\left(z_2 \middle| z_1^{(i+1)}, z_3^{(i)}, \ldots, z_n^{(i)}\right)$
      ⋮
    • Sample $z_n^{(i+1)} \sim p\left(z_n \middle| z_1^{(i+1)}, z_2^{(i+1)}, \ldots, z_{n-1}^{(i+1)}\right)$
```

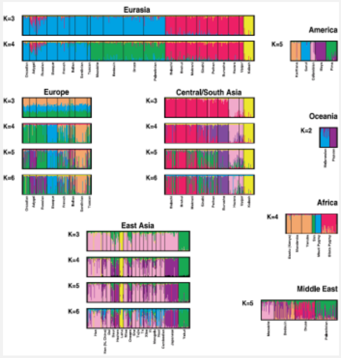
## Gibbs samplers are reversible by construction

- Gibbs sampling idea:
  - Successively update single variable, treating all others as known
    - "Known" values are most recent samples of other variables
  - Variable to be updated is sampled conditional on others
- Must show transition probabilities imply reversibility wrt $P(z_1, \ldots, z_n)$
  - Dropping super scripts for simplicity:
  - Want to show that:

$$P\left(z_1, \ldots, z_{j-1}, z_j, z_{j+1}, \ldots, z_n\right) P\left(\tilde{z}_j \middle| z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_n\right) =$$
$$P\left(z_1, \ldots, z_{j-1}, \tilde{z}_j, z_{j+1}, \ldots, z_n\right) P\left(z_j \middle| z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_n\right)$$

  - By definition of conditional probability, both sides equal to $\frac{P\left(z_1, \ldots, z_{j-1}, z_j, z_{j+1}, z_n\right) P\left(z_1, \ldots, z_{j-1}, \tilde{z}_j, z_{j+1}, \ldots, z_n\right)}{P\left(z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_n\right)}$, so reversibility does hold
- Irreducibility and aperiodicity also necessary:
  - Is specific to each problem; can be simple to demonstrate

## STRUCTURE uses Gibbs sampling



Rosenberg *et al.* (2002)

## STRUCTURE algorithm overview

- Given $X$: genotypes for $n$ samples; $K$: # populations
- Want to estimate:
  - Matrix $Z$ : ancestral population $1, \ldots, K$ of each allele
    - $z_l^{(i,a)}$: ancestral population of allele $a \in \{1,2\}$ at locus $l$ in sample $i$
  - Matrix $P$ containing allele frequencies in each population
    - $p_{klj}$: frequency of allele $j$ at locus $l$ in population $k$
  - Matrix Q containing admixture proportions for each sample
    - $q_k^{(i)}$: proportion of genome of sample $i$ that originated in $k$

1. Initialize $Z^{(0)}$ by random sample from uniform prior
2. Iterate:
   1. Sample $P^{(m)}, Q^{(m)}$ from $P\left(P, Q \middle| X, Z^{(m-1)}\right)$
   2. Sample $Z^{(m)}$ from $P\left(Z \middle| X, P^{(m)}, Q^{(m)}\right)$

Forms a Gibbs sampler

## STRUCTURE Gibbs sampling

- $p_{klj}$ values are sampled from Dirichlet distribution at each locus, parameters for each allele set to $\lambda_j + n_{klj}$, where
  - $n_{klj}$ = counts of allele $j$ in population $k$, locus $l$ (from $Z$)
- $q_k^{(i)}$ values are sampled from Dirichlet for each sample, parameters for each population set to $\alpha + m_k^{(i)}$, where
  - $m_k^{(i)}$ = counts of alleles in sample $i$ from population $k$ (from $Z$)
- $z_l^{(i,a)}$ are sampled from: [$x_l^{(i,a)}$ is allele in sample $i$, locus $l$]

$$P\left(z_l^{(i,a)} = k \middle| X, P, Q\right) = \frac{q_k^{(i)} P\left(x_l^{(i,a)} \middle| P, z_l^{(i,a)} = k\right)}{\sum_{k'} q_{k'}^{(i)} P\left(x_l^{(i,a)} \middle| P, z_l^{(i,a)} = k'\right)}$$

where $P\left(x_l^{(i,a)} \middle| P, z_l^{(i,a)} = k\right) = p_{klx_l^{(i,a)}}$

## Dirichlet distribution

- Multivariate distribution where values of each variant sum to 1
  - Example: frequencies of $J$ alleles at a given locus
  - (Generalization of Beta distribution [which is univariate])
- Uses concentration parameters $\boldsymbol{\alpha}$ corresponding to each variable

$$f(x_1, \ldots, x_J; \alpha_1, \ldots, \alpha_J) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{J} x_i^{\alpha_i - 1}$$

$$\boxed{\mathrm{E}[X_i] = \frac{\alpha_\mathrm{i}}{\sum_j \alpha_j}}$$

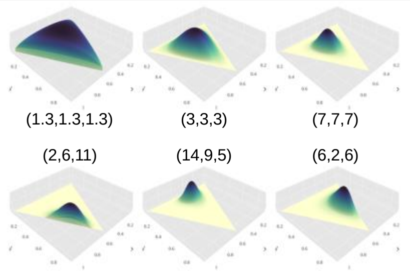
Intuition: on average gives fraction of $\alpha$s contributed by $i$

$$\mathrm{Var}[X_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}, \qquad \alpha_0 = \sum_{i=1}^{J} \alpha_i$$

## Concentration parameters and variance

- Distribution more strongly peaked for larger concentration parameters

Example 3-variant Dirichlet, $\boldsymbol{\alpha}$ shown


(1.3,1.3,1.3) (3,3,3) (7,7,7)

(2,6,11) (14,9,5) (6,2,6)

Plot: Empetrisor (wikipedia)

## Gibbs sampling for motif finding

Algorithm:

1. Choose random start sites $s^{(i)}$ for all $i = 1 \ldots t$
2. For $i = 1 \ldots t$, do:
   1. Compute position weight matrix using other starts $s^{(j)}, j \neq i$
      - Should include pseudocounts if using (good idea to do so)
   2. Compute likelihood of motif start at all positions $1 \ldots n - k$ in $x^{(i)}$
      - For a given start position $r$, likelihood is $\sum_{j=1}^{k} w_{j, x_{r+j-1}^{(i)}}$
   3. Sample new start position $s^{(i)}$ in proportion to likelihood
      - Note: normalizing likelihoods (summing over all possible starts) gives the probability of motif starting at each site
- Repeat for $N$ (large) iterations OR
  Until $w_{j,\cdot}$ doesn't change much for a pass over $t$

## Likelihood of motif positions

- Given sequences $X = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ of length $n$, start positions $\boldsymbol{s} = (s_1, \ldots, s_t)$, background model $\boldsymbol{\theta}$, and motif model $\boldsymbol{\pi}$, motif length $k$
- Likelihood is:

$$p(X, \boldsymbol{s} | \boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{t} p(s_i) \prod_{j=1}^{n} p\left(x_{i,j} \middle| s_i, \boldsymbol{\theta}, \boldsymbol{\pi}\right)$$

$$= \prod_{i=1}^{t} \frac{1}{n-k+1} \left[\prod_{j=1}^{s_i - 1} p\left(x_{i,j} \middle| \boldsymbol{\theta}\right)\right] \left[\prod_{j=s_i}^{s_i+k-1} p\left(x_{i,j} \middle| \boldsymbol{\pi}_{j-s_i+1}\right)\right] \left[\prod_{j=s_i+k}^{n} p\left(x_{i,j} \middle| \boldsymbol{\theta}\right)\right]$$

$$= \prod_{i=1}^{t} \frac{1}{n-k+1} \left[\prod_{j=1}^{n} p\left(x_{i,j} \middle| \boldsymbol{\theta}\right)\right] \left[\prod_{j=s_i}^{s_i+k-1} \frac{p\left(x_{i,j} \middle| \boldsymbol{\pi}_{j-s_i+1}\right)}{p\left(x_{i,j} \middle| \boldsymbol{\theta}\right)}\right]$$

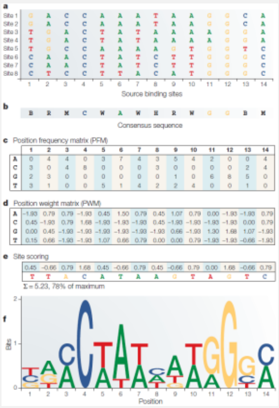
## Probability of start positions

- Posterior probability of start position for sequence $i$ is

$$p(s_i = p | \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{\pi}) = \frac{p(s_i = p, \boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\pi})}{p(\boldsymbol{x}_i | \boldsymbol{\theta}, \boldsymbol{\pi})}$$

$$= \frac{\frac{1}{L-k+1}\left[\prod_{j=1}^{n} p\left(x_{i,j} \middle| \boldsymbol{\theta}\right)\right]\left[\prod_{j=p}^{p+k-1} \frac{p\left(x_{i,j} \middle| \boldsymbol{\pi}_{j-p+1}\right)}{p\left(x_{i,j} \middle| \boldsymbol{\theta}\right)}\right]}{\frac{1}{L-k+1}\left[\prod_{j=1}^{n} p\left(x_{i,j} \middle| \boldsymbol{\theta}\right)\right]\left[\sum_{s_i=1}^{n-k+1} \prod_{j=s_i}^{s_i+k-1} \frac{p\left(x_{i,j} \middle| \boldsymbol{\pi}_{j-s_i+1}\right)}{p\left(x_{i,j} \middle| \boldsymbol{\theta}\right)}\right]}$$

$$= \frac{\prod_{j=p}^{p+k-1} \frac{p\left(x_{i,j} \middle| \boldsymbol{\pi}_{j-p+1}\right)}{p\left(x_{i,j} \middle| \boldsymbol{\theta}\right)}}{\sum_{s_i=1}^{n-k+1} \prod_{j=s_i}^{s_i+k-1} \frac{p\left(x_{i,j} \middle| \boldsymbol{\pi}_{j-s_i+1}\right)}{p\left(x_{i,j} \middle| \boldsymbol{\theta}\right)}}$$



|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Site 1 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 2 | G | A | C | C | A | A | A | T | A | A | G | G | C | A |
| Site 3 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 4 | T | G | A | C | T | A | T | A | A | A | A | G | G | A |
| Site 5 | T | G | C | C | A | A | A | A | G | T | G | G | T | C |
| Site 6 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 7 | C | A | A | C | T | A | T | C | T | T | G | G | G | C |
| Site 8 | C | T | C | C | T | T | A | C | A | T | G | G | G | C |

Source binding sites

| B | R | M | C | W | A | W | H | R | W | G | G | B | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

Consensus sequence

Position frequency matrix (PFM)

|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 4 | 4 | 0 | 3 | 7 | 4 | 3 | 5 | 4 | 2 | 0 | 0 | 4 |
| C | 3 | 0 | 4 | 8 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 4 |
| G | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 6 | 8 | 5 | 0 |
| T | 3 | 1 | 0 | 0 | 5 | 1 | 4 | 2 | 2 | 4 | 0 | 0 | 1 | 0 |

Position weight matrix (PWM)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | −1.93 | 0.79 | 0.79 | −1.93 | 0.45 | 1.50 | 0.79 | 0.45 | 1.07 | 0.79 | 0.00 | −1.93 | −1.93 | 0.79 |
| C | 0.45 | −1.93 | 0.79 | 1.68 | −1.93 | −1.93 | −1.93 | 0.45 | −1.93 | −1.93 | −1.93 | −1.93 | 0.00 | 0.79 |
| G | 0.00 | 0.45 | −1.93 | −1.93 | −1.93 | −1.93 | −1.93 | −1.93 | 0.66 | −1.93 | 1.30 | 1.68 | 1.07 | −1.93 |
| T | 0.15 | 0.66 | −1.93 | −1.93 | 1.07 | 0.66 | 0.79 | 0.00 | 0.00 | 0.79 | −1.93 | −1.93 | −0.66 | −1.93 |

Site scoring

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.45 | −0.66 | 0.79 | 1.68 | 0.45 | −0.66 | 0.79 | 0.45 | −0.66 | 0.79 | 0.00 | 1.68 | −0.66 | 0.79 |
| T | T | A | C | A | T | A | A | G | T | A | G | T | C |

Σ = 5.23, 78% of maximum

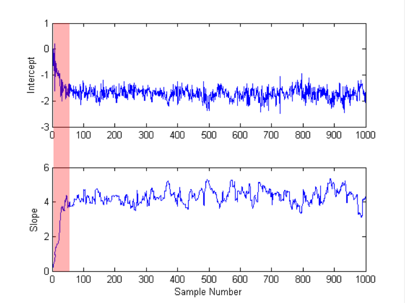
## Thinning in MCMC

- After convergence, MCMC samples from stationary distribution, however…
- Sample $z^{(i)}$ is conditioned on previous sample $z^{(i-1)}$
  - Note: in Gibbs sampling, $\boldsymbol{z}$ is multivariate
- Successive samples $z^{(i-1)}, z^{(i)}$ are *not* independent
  - ➢ Samples separated by some number of iterations are (effectively) independent
  - ➢ Can address this by thinning: keep only every $k$th sample

## Convergence of MCMC (1)

- How do we know if the MCMC has converged?
- Simple approach: plot log likelihood in each iteration, visually inspect
  - At stationarity and with good mixing, likelihood will have roughly constant mean and variance
- Plots for parameters being estimated will also have roughly constant mean and variance
- Can make informed guess about burn-in length $B$ from these plots

## Visually inspection: trace plots



Plot: iteration number vs. value of a parameter being estimated
Top plot has higher variance: "mixes" better than bottom
Relatively constant mean and variance *suggests* stationarity

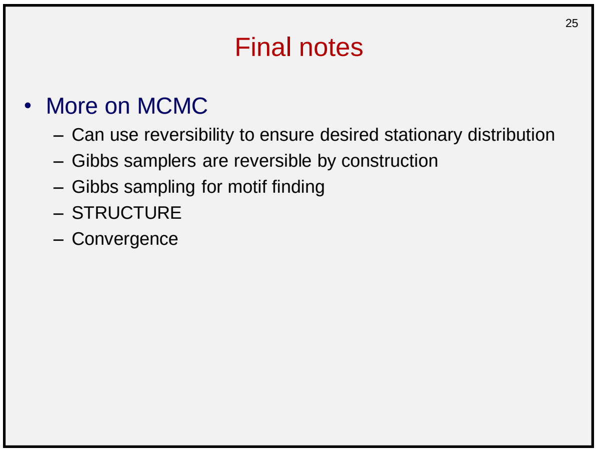
## Convergence of MCMC (2)

- Best to run more than one chain at different places and ensure that they converge to the same distribution
- More principled:
  - Run ≥ 2 chains from high variance (overdispersed) starting values
  - Compare within-chain and between-chain variance of estimated parameters
  - Expect similar variance for the comparison (i.e., within-chain and between-chain should be roughly the same)
    - If not, should run longer
  - [Above based on Gelman-Rubin diagnostic; this diagnostic gives recommendation for when to run chain longer]
- Other principled approaches exist (beyond scope here)

## Gibbs sampling algorithm with burn-in, thinning

- … with $B$ burn-in iterations, $S$ sampling iterations, thinning interval $k$

```
1. Initialize $\boldsymbol{z}^{(0)}$ such that $p(\boldsymbol{z}^{(0)}|\boldsymbol{x}) > 0$
2. For $i = 1 \ldots (B+S)$:
    1. For $j = 1 \ldots |\boldsymbol{z}|$:
        • Sample $\boldsymbol{z}_j^{(i)}$ from $p\left(\boldsymbol{z}_j^{(i)} \middle| z_1^{(i+1)}, \ldots, z_{j-1}^{(i+1)}, z_{j+1}^{(i)}, \ldots, z_n^{(i)}, x\right)$
    2. If $(i > B$ and $(i \bmod k) = 0)$ retain sample $z^{(i)}$
```



25

## Final notes

- More on MCMC
  - Can use reversibility to ensure desired stationary distribution
  - Gibbs samplers are reversible by construction
  - Gibbs sampling for motif finding
  - STRUCTURE
  - Convergence