Royal Institute of Technology

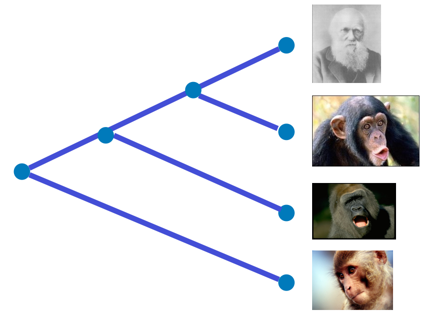
STAT. METH. IN CS – MCMC: GIBBS SAMPLER & METROPOLIS HASTING

Lecture 11

PHYLOGENY

Input: species

Output: tree where proximity correlates with similarity



MR BAYES

BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 12 2003, pages 1572–1574
DOI: 10.1093/bioinformatics/btg180

MrBayes 3: Bayesian phylogenetic inference under mixed models

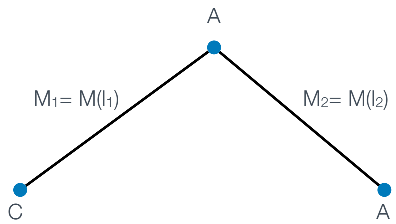
Fredrik Ronquist[1,] and John P. Huelsenbeck[2]*

[1]Department of Systematic Zoology, Evolutionary Biology Centre, Uppsala University, Norbyv. 18D, SE-752 36 Uppsala, Sweden and [2]Section of Ecology, Behavior and Evolution, Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093-0116, USA

Received on December 20, 2002; revised on February 14, 2003; accepted on February 19, 2003

Prof. Entomology

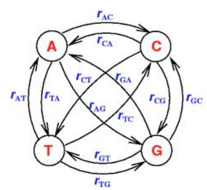
MARKOV MODEL OF SEQUENCE EVOLUTION



The same position

Human

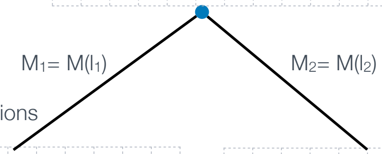
Mouse



M(l)

MARKOV MODEL OF SEQUENCE EVOLUTION

A C A C G T A C T G C G



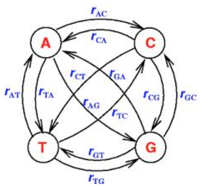
Same gene, same positions

A C A C G T A G T G C C

A A A C G T A C T G C A

Human

Mouse



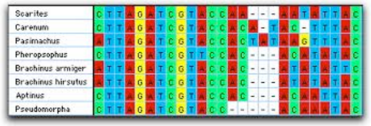
M(l)

MARKOV MODEL OF SEQUENCE EVOLUTION

Human A C A C G T A G T G C C

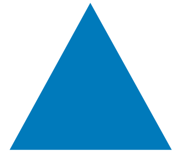
Mouse A A A C G T A C T G C A

In general,



Scarites	C	T	T	A	G	A	T	C	G	T	A	C	C	A	A	-	-	-	A	A	T	A	T	T	A	C
Carenum	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	A	-	T	A	C	-	T	T	T	A	C
Pasimachus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	T	A	T	A	A	G	T	T	T	A	C
Pheropsophus	C	T	T	A	G	A	T	C	G	T	T	C	C	A	C	-	-	-	A	C	A	T	A	T	A	C
Brachinus armiger	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	T	C
Brachinus hirsutus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	A	C
Aptinus	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	C	A	A	T	T	A	C
Pseudomorpha	C	T	T	A	G	A	T	C	G	T	A	C	C	-	-	-	-	-	A	C	A	A	A	T	A	C

Uniform

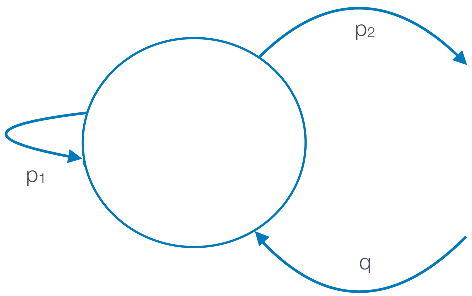


A A C ....T A

MARKOV CHAINS (DISCRETE)

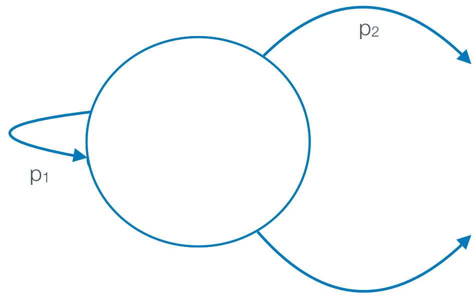
★Directed graph with transition probabilities

★ We observe the sequence of visited vertices



MARKOV CHAINS (DISCRETE)

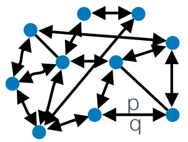
Probabilities on outgoing edges sum to one



$$\sum_{i\in[d]} p_i = 1$$

MCMC

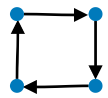
- In order to sample p, set up a MC M
  - select transition probabilities so that p = stationary distribution
  - sample from M



gcd - greatest common divisor

The period of a state i
gcd{ t : p(i → i in t steps) > 0 }

Period 4



M is aperiodic if each states has period 1

M is irreducible if ∀states i,j: p(i→j) > 0

M is recurrent if ∀states i: p(i→i) = 1

MC- EXISTENCE OF STATIONARY

Theorem: Every irreducible, aperiodic, finite state MC has a stationary distribution

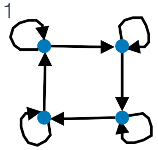
Theorem: Every irreducible, aperiodic, and recurrent MC has a stationary distribution

We want to satisfy these conditions!

gcd - greatest common divisor

The period of a state i
gcd{ t : p(i → i in t steps) > 0 }

Period 1



M is aperiodic if each states has period 1

M is irreducible if ∀states i,j: p(i→j) > 0

M is recurrent if ∀states i: p(i→i) = 1

MC- EXISTENCE OF STATIONARY

Theorem: Every irreducible, aperiodic, finite state MC has a stationary distribution

Theorem: Every irreducible, aperiodic, and recurrent MC has a stationary distribution

We want to satisfy these conditions!

METROPOLIS HASTINGS (MH)

How?

We want to compute p*(x) (typically p(x|D))

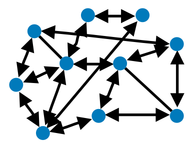
Implicitly construct Markov Chain M with stationary distribution p*(x)

Traverse it and sample every k:th visit

Use good or random starting point

Discard the first l:th samples

The remaining samples $x_1,\ldots,x_S$ is an approximation of p*(x)



$p^*(x) \approx [\sum_i I(x=x_i)]/S$

IMPLICIT CONSTRUCTION OF MC WITH STATIONARY DISTRIBUTION P*(X)

Transition when in state x:

Propose state according to proposal distribution q(x´|x) (conditions later)

Accept x´with probability min(1, α)

$$\alpha = \frac{p^*(x')/q(x'|x)}{p^*(x)/q(x|x')} = \frac{p^*(x')q(x|x')}{p^*(x)q(x'|x)}$$

MC WITH STATIONARY DISTRIBUTION P*(X)=P(X|D)

Transition when in state x:

Propose state according to proposal distribution q(x´|x) (conditions later)

Accept x´with probability min(1, α)

We want and don’t have p*, but it is a ratio that is required!

$$\alpha = \frac{p^*(x')q(x|x')}{p^*(x)q(x'|x)}$$

$$\frac{p^*(x')}{p^*(x)} = \frac{p(x'|\mathcal{D})}{p(x|\mathcal{D})} = \frac{\frac{p(\mathcal{D}|x')p(x')}{p(\mathcal{D})}}{\frac{p(\mathcal{D}|x)p(x)}{p(\mathcal{D})}} = \frac{p(\mathcal{D}|x')p(x')}{p(\mathcal{D}|x)p(x)}$$

MC WITH STATIONARY DISTRIBUTION P*(X)=P(X|D)

MH typically work when we cannot compute p*(x), but p*(x´)/p*(x)

$$\alpha = \frac{p^*(x')q(x|x')}{p^*(x)q(x'|x)}$$

$$\frac{p^*(x')}{p^*(x)} = \frac{p(x'|\mathcal{D})}{p(x|\mathcal{D})} = \frac{\frac{p(\mathcal{D}|x')p(x')}{p(\mathcal{D})}}{\frac{p(\mathcal{D}|x)p(x)}{p(\mathcal{D})}} = \frac{p(\mathcal{D}|x')p(x')}{p(\mathcal{D}|x)p(x)}$$

likelihood

prior

MH

Algorithm 24.2: Metropolis Hastings algorithm

Initialize x^0 ;
for $s = 0, 1, 2, \ldots$ **do**
  Define $x = x^s$;
  Sample $x' \sim q(x'|x)$;
  Compute acceptance probability

$$\alpha = \frac{p^*(x')q(x|x')}{p^*(x)q(x'|x)}$$

  Compute $r = \min(1, \alpha)$;
  Sample $u \sim U(0, 1)$;
  Set new sample to

$$x^{s+1} = \begin{cases} x' & \text{if } u < r \\ x^s & \text{if } u \geq r \end{cases}$$

DETAILED BALANCE EQUATIONS

- A transition matrix, i.e., $A_{ij} = p(i \rightarrow j \text{ in 1 step})$
- A regular if $\forall k,l\ \exists n$ s/t $(A_{k,l})^n > 0$
- Detailed balance equations $\forall k,l\ \ \pi_k A_{kl} = \pi_l A_{lk}$
- Theorem: If MC M with regular transition matrix A that satisfies detailed balance wrt π, then π the stationary distribution of M.
- Proof: Note that

$$\pi_l^{t+1} = \sum_k \pi_k^t A_{kl} = \sum_k \pi_l^t A_{lk} = \pi_l^t \sum_k A_{lk} = \pi_l^t$$

WHY MH WORKS

$p^*(x)$ the distribution we want to estimate

$\alpha(x'|x) = (p^*(x')q(x|x'))/(p^*(x)q(x'|x))$

Let $r(x'|x) = \min(1, \alpha(x'|x))$

The transition probability for $x' \neq x$ ($x' = x$ easy) $p(x'|x) = q(x'|x)\ r(x'|x)$

Assume $p^*(x)q(x'|x) \geq p^*(x')q(x|x')$, so $r(x'|x) = \alpha(x'|x)$ and $r(x|x') = 1$

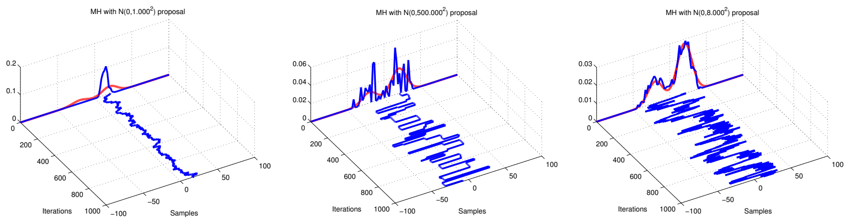
We want $p^*(x)p(x'|x) = p^*(x')p(x|x')$

$p^*(x)p(x'|x) = p^*(x)\ q(x'|x)r(x'|x)$
$= p^*(x)q(x'|x)\ (p^*(x')q(x|x'))/(p^*(x)q(x'|x))$
$= p^*(x')q(x|x')$
$= p^*(x')q(x|x')r(x|x')$
$= p^*(x')p(x|x')$

- Efficiency of proposal distribution
- Burn-in
- Convergence

PRACTICAL CONSIDERATIONS

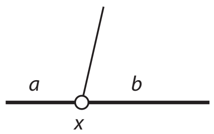
THE CRAFTSMANSHIP OF PROPOSAL DESIGN



- Conservative proposal - hard to move away from local optimum
- Wild proposal - few accepted proposals
- 25-40% acceptance rate considered good (use pilot runs)

MR BAYES PROPOSAL NODE (VERTEX) SLIDER

Node Slider



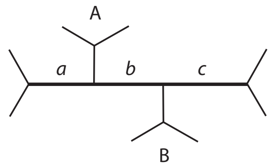
Two adjacent branches *a* and *b* are chosen at random
The length of $a + b$ is changed using a multiplier with tuning paremeter λ
The node *x* is randomly inserted on $a + b$ according to a uniform distribution

Bolder proposals: increase λ
More modest proposals: decrease λ

The boldness of the proposal depends heavily on the uniform reinsertion of *x*, so changing λ may have limited effect

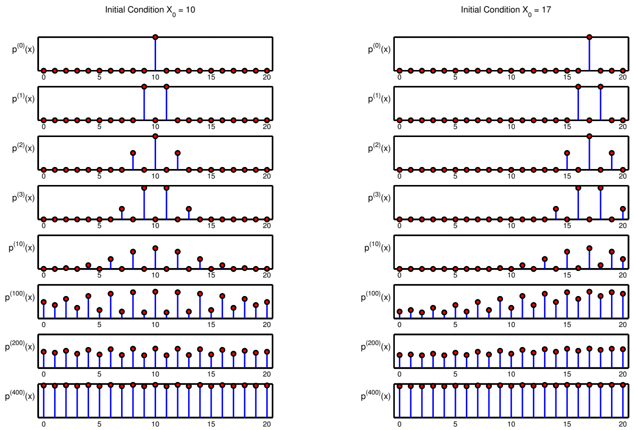
MR BAYES PROPOSAL LOCAL TREE OPERATION

LOCAL

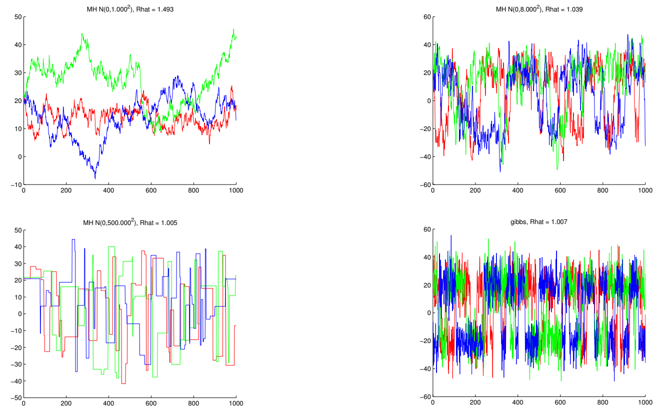


Three internal branches - *a, b,* and *c* - are chosen at random.
Their total length is changed using a multiplier with tuning paremeter λ.
One of the subtrees A or B is picked at random.
It is randomly reinserted on $a + b + c$ according to a uniform distribution

Bolder proposals: increase λ
More modest proposals: decrease λ
Changing λ has little effect on the boldness of the proposal



BURN-IN



CONVERGENCE DIAGNOSTICS - MULTIPLE CHAINS

GIBBS SAMPLER

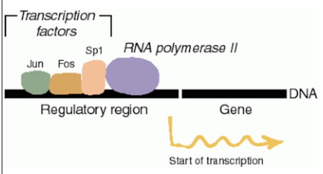
- A way to define transition probabilities
- We seek $p(x_1,\ldots,x_K)$
- States are vectors $(x_1,\ldots,x_K)$
- Transitions possible only between states differing in one position
- $t((x_1,\ldots,x_{i-1},x'_i,x_{i+1},\ldots,x_K) \mid x) = p(x'_i \mid x_{-i}, D)$ (from now D implicit)
  - where $x_{-i}=(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_K)$
- Called *full conditional*

MOTIVATION

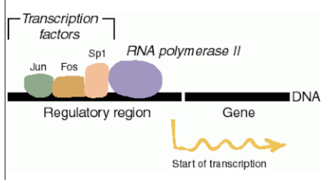
Problem 4 (4p):
The following generative model generates K sequences of length N: $s_1,\ldots,s_K$ where $s_i = s_{i,1},\ldots,s_{i,N}$. All sequences are over the alphabet [M]. Each of these sequences has a "magic" word of length w hidden in it and the rest of the sequence is called background.

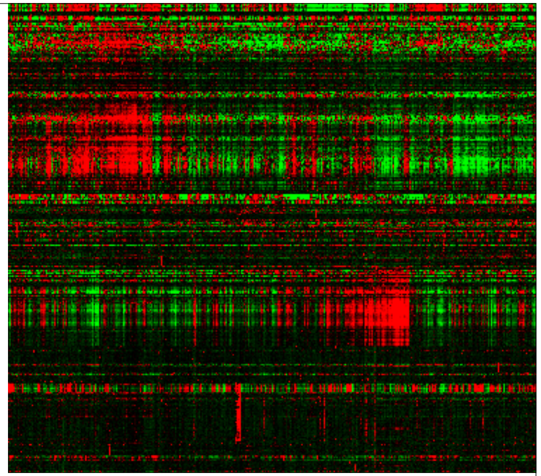
First, for each i, a start position r_i for the magic word is sampled uniformly from [N-w+1]. Then the j:th positions in the magic words are sampled from $q_j(x)$, which is $Cat(x|\theta_j)$ where θ_j has a $Dir(\theta_j|\alpha)$ prior. All other positions in the sequences are sampled from the background distribution $q(x)$, which is $Cat(x|\theta)$ where θ has a $Dir(\theta|\alpha')$ prior.

Describe a Gibbs sampler that can be used for estimating the posterior over start positions after having observed $s_1,\ldots,s_K$. Make the sampler as collapsed as possible. You do know α and α'.



CONSERVED BINDING TRANSCRIPTION FACTORS BINDING SITES





CO-REGULATED VS CO-EXPRESS

GIBBS IS A SPECIAL CASE OF MH

- In Gibbs we sample from the full conditional
- View Gibbs as MH and the full conditional as proposal
- This means that we always accept. Is that correct?

GIBBS IS A SPECIAL CASE OF MH

- Proposal, pick an index i and then

$$q(x'|x) = p(x'_i|x_{-i})I(x'_{-i} = x_{-i})$$

- Acceptance (according to MH) x and x' are neighbours so $x_{-i}=x'_{-i}$

$$\begin{aligned}\alpha &= \frac{p(x')q(x|x')}{p(x)q(x'|x)} \\ &= \frac{p(x'_i|x'_{-i})p(x'_{-i})p(x_i|x'_{-i})}{p(x_i|x_{-i})p(x_{-i})p(x'_i|x_{-i})} \\ &= \frac{p(x'_i|x_{-i})p(x_{-i})p(x_i|x_{-i})}{p(x_i|x_{-i})p(x_{-i})p(x'_i|x_{-i})} = 1\end{aligned}$$

GIBBS SAMPLING

- Pick initial state $x_1=(x_{1,1},\ldots,x_{1,K})$
- For s=1 to S
  - Sample $k \sim_u [K]$
  - Sample $x_{s+1,k} \sim p(x_{s+1,k}| x_{s,-k})$
  - Let $x_{s+1} = (x_{s,1},\ldots,x_{1,k-1}, x_{s+1,k},\ldots, x_{s,K})$
  - If k|s record x_{s+1} (thinning)

Notation

$$\mathcal{D} = (x_1,\ldots,x_N),\;\; H = (z_1,\ldots,z_N),\;\; N_k = \sum_n I(z_i = k)$$

$$\boldsymbol{\pi} = (\pi_1,\ldots,\pi_k),\;\; \boldsymbol{\mu} = (\mu_i,\ldots,\mu_k),\;\; \boldsymbol{\lambda} = (\lambda_i,\ldots,\lambda_k), \text{ and }\;\; \lambda_k = 1/\sigma_k^2$$

Hyperparameters $\boldsymbol{\theta}_0 = (\mu_0,\lambda_0,\lambda_0,\beta_0,\alpha)$

Model

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}),\;\; \mu_k \sim N(\mu_0,\lambda_0),\;\; \lambda_k \sim \mathrm{Ga}(\alpha_0,\beta_0),\;\; z_i \sim \mathrm{Cat}(\boldsymbol{\pi}), \text{ and}$$

$$p(x_n|Z_n = k) = N(\mu_k,\lambda_k)$$

GIBBS SAMPLER FOR GMM

A STATE

$$(H, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda})$$

Hyperparameters $\boldsymbol{\theta}_0 = (\mu_0, \lambda_0, \lambda_0, \beta_0, \alpha)$

Model

$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}),\;\; \mu_k \sim N(\mu_0, \lambda_0),\;\; \lambda_k \sim \mathrm{Ga}(\alpha_0, \beta_0),\;\; z_i \sim \mathrm{Cat}(\boldsymbol{\pi})$, and $p(x_n|Z_n = k) = N(\mu_k, \lambda_k)$

Likelihood

$$\begin{aligned} p(D, H, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}) =& p(D, H|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}, \boldsymbol{\lambda}) \\ =& \prod_{n,k} [\pi_k N(x_n|\mu_k, \lambda_k)]^{I(z_n=k)} \mathrm{Dir}(\pi|\alpha) \\ & \prod_k N(\mu_k|\mu_0, \lambda_0) \mathrm{Ga}(\lambda_k|\alpha_0, \beta_0) \end{aligned}$$

LIKELIHOOD FOR GMM

$$(H, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda})$$

$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}),\;\; \mu_k \sim N(\mu_0, \lambda_0),\;\; \lambda_k \sim \mathrm{Ga}(\alpha_0, \beta_0),\;\; z_i \sim \mathrm{Cat}(\boldsymbol{\pi})$, and $p(x_n|Z_n = k) = N(\mu_k, \lambda_k)$

$$p(z_n|\mathcal{D}, H_{-n}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda})$$

FULL CONDITIONAL ON H

$$(H, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda})$$

$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}),\;\; \mu_k \sim N(\mu_0, \lambda_0),\;\; \lambda_k \sim \mathrm{Ga}(\alpha_0, \beta_0),\;\; z_i \sim \mathrm{Cat}(\boldsymbol{\pi})$, and $p(x_n|Z_n = k) = N(\mu_k, \lambda_k)$

$$p(z_n = k|\mathcal{D}, H_{-n}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \propto p(z_n = k|\boldsymbol{\pi}) N(x_n; \mu_k, \lambda_k)$$

FULL CONDITIONAL ON H

FULL CONDITIONAL ON Π

$$p(\boldsymbol{\pi}|D,H,\boldsymbol{\mu},\boldsymbol{\lambda}) = p(\boldsymbol{\pi}|H) = \mathrm{Dir}(\alpha_1+N_1,\ldots,\alpha_K+N_K)$$

- Categorial with Dirichlet prior has Dirichlet posterior

FULL CONDITIONAL ON THE PRECISION

$$p(\lambda_k|D,H,\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\lambda}_{-k}) \propto \mathrm{Ga}(\lambda_k|\alpha_0,\beta_0)\prod_{n:z_n=k}\pi_k N(x_n|\mu_k,\lambda_k)$$

$$\propto \lambda_k^{\alpha_0-1}e^{-\lambda_k\beta_0}\lambda_k^{N_k/2}e^{-\frac{\lambda_k}{2}\sum_{n:z_n=k}(x_n-\mu_k)^2}$$

$$= \lambda_k^{\alpha_0+N_k/2-1}e^{-\lambda_k(\beta_0+\frac{1}{2}\sum_{n:z_n=k}(x_n-\mu_k)^2)}$$

- So posterior

$$\mathrm{Ga}(\alpha_0+N_k/2,\beta_0+\frac{1}{2}\sum_{n:z_n=k}(x_n-\mu_k)^2)$$

GAUSSIAN

★ If

- g(x) is a p.d.f.
- g(x) $\propto$ exp(-ax^2/2+bx+c)

★ then

- g is Gaussian
- λ = a and μ = b/a

- Easy "trick" when working with Gaussians

GAUSSIAN IS SELF CONJUGATE

$$p(\mu'|D',\lambda',\mu_0,\lambda_0) = N(\mu'|\mu_0,\lambda_0)\prod_{n'=1}^{N'}N(x'_{n'}|\mu',\lambda')$$

$$\propto\sqrt{\lambda_0}e^{-\frac{\lambda_0}{2}(\mu'-\mu_0)^2}(\lambda')^{N'/2}e^{-\frac{\lambda'}{2}\sum_{n'}(x'_{n'}-\mu')^2}$$

- $D' = \{x'_1,\ldots,x'_{N'}\}$
- $p(x'_i) = N(x'_i \mid \mu',\lambda')$ where
  - λ′ is given
  - $p(\mu')$ is $N(\mu' \mid \mu_0,\lambda_0)$

GAUSSIAN IS SELF CONJUGATE

$$\propto \sqrt{\lambda_0} e^{-\frac{\lambda_0}{2}(\mu'-\mu_0)^2} (\lambda')^{N'/2} e^{-\frac{\lambda'}{2}\sum_{n'}(x'_{n'}-\mu')^2}$$

Let $M' = \sum_{n'} x_{n'}$

The log is

$$\propto -\frac{\lambda_0}{2}(\mu'^2 - 2\mu'\mu_0 + \mu_0^2) - \frac{\lambda'}{2}\sum_{n'}(x'^2_{n'} - 2x'_{n'}\mu' + mu'^2)$$

$$= -\frac{1}{2}(\lambda_0 + \lambda' N')\mu'^2 + (\lambda_0\mu_0 + \lambda' M')\mu' + C$$

a constant

GAUSSIAN IS SELF CONJUGATE

The log is

$$\propto -\frac{1}{2}(\lambda_0 + \lambda' N')\mu'^2 + (\lambda_0\mu_0 + \lambda' M')\mu'$$

where

$$M' = \sum_{n'} x_{n'}$$

- If
  - D´={x´₁,…,x´N}
  - p(x´i) = N(x´i | μ´,λ´) where
    - λ´ is given
    - N(μ´ | μ0,λ0)
- then
  - p(μ´ | D´,λ´,μ0,λ0) = N(μ´ | μ, λ) where
  - λ = λ0 + λ´N´
    - μ = (λ0μ0 + λ´M´)/λ

FULL CONDITIONAL ON MEAN

$$p(\mu_k | D, H, \boldsymbol{\pi}, \boldsymbol{\mu}_{-k}, \boldsymbol{\lambda}) \propto N(\mu_k | \mu_0, \lambda_0) \prod_{n:z:n=k} [\pi_k N(x_n | \mu_k, \lambda_k)]$$

$$\propto \sqrt{\lambda_0} e^{-\frac{\lambda_0}{2}(\mu_k-\mu_0)^2} \lambda_k^{N_k/2} e^{-\frac{\lambda_k}{2}\sum_{n:z_n=k}(x_n-\mu_k)^2}$$

RECALL AND APPLY

We had

$$\propto \sqrt{\lambda_0} e^{-\frac{\lambda_0}{2}(\mu'-\mu_0)^2} (\lambda')^{N'/2} e^{-\frac{\lambda'}{2}\sum_{n'}(x'_{n'}-\mu')^2}$$

where

$$M' = \sum_{n'=1}^{N'} x'_{n'}$$

and got N(μ´ | μ, λ) where λ = λ0 + λ´N´ & μ = (λ´μ0 + λ´M´)/λ

We now have

$$\propto \sqrt{\lambda_0} e^{-\frac{\lambda_0}{2}(\mu_k-\mu_0)^2} \lambda_k^{N_k/2} e^{-\frac{\lambda_k}{2}\sum_{n:z_n=k}(x_n-\mu_k)^2}$$

where

$$M_k = \sum_{n:z_n=k} x_n$$

$$N_k = \sum_{n:z_n=k} 1$$

and get N(μk | μ,λ) where λ = λ0 + λ Nk & μ = (λμ0 + λMk)/λ

THE END