

Final exam

- Given the current constraints, rather than final presentations we will have a final take-home project. It will be due on the day of the exam.
- I will suggest a specific project nearer the time, but I am also open to you reporting on how you are using the methods of the course in your own research if you prefer to do that instead.

Examinable assignment - week 7 part 1

- Use the IndeptGamma.R code to use MCMC to produce samples from a $\text{Gamma}(2.3, 2.7)$ random variable
- Show how the performance deteriorates when you run the algorithm to produce samples from a $\text{Gamma}(0.1, 0.01)$.
 - Use the Gelman plot to illustrate the deterioration in performance
- Find a proposal kernel (i.e. $q(x \rightarrow x')$, the way of producing new candidate values) that performs more efficiently.
 - Again use the Gelman plots to show how performance has changed.
 - Discuss how the performance has improved
- Due On March 30th, at 1pm.

MCMC: Metropolis-Hastings

1. If at θ , propose move to θ' according to transition kernel $q(\theta \rightarrow \theta')$.
2. Calculate

$$h = \min \left\{ 1, \frac{P(D|\theta')\pi(\theta')q(\theta' \rightarrow \theta)}{P(D|\theta)\pi(\theta)q(\theta \rightarrow \theta')} \right\}$$

3. Move to θ' with prob. h , else remain at θ .
4. Return to 1.

The stationary distribution of this chain will be $P(\theta | D)$.

$P(D)$ disappears, so this can be used even when $P(D)$ cannot be calculated!

MCMC: Metropolis-Hastings

1. If at θ , propose move to θ' according to transition kernel $q(\theta \rightarrow \theta')$.
2. Calculate

$$h = \min \left\{ 1, \frac{P(D|\theta')\pi(\theta')q(\theta' \rightarrow \theta)}{P(D|\theta)\pi(\theta)q(\theta \rightarrow \theta')} \right\}$$

3. Move to θ' with prob. h , else remain at θ .
4. The $q(\theta \rightarrow \theta')$ in step 2. is the prob. (or prob. density) of proposing a move from θ to θ' . It must be calculated using the transition kernel you use, and it does not involve the acceptance probability at all. Similarly for $q(\theta' \rightarrow \theta)$

Th

Examinable Assignment, Week 7 part 2: Code-breaking (due March 30th, 1pm)

- Gzo uclfg gcpo C qhcs okof te Gollk Qoeeth zo rhf slvem ce h LtqqfLtkio Fcqxol Rlhcgz tvgfcso gzo gollhio tu Gzo Sheiolf. Gzo ahlmced qtg hggoesheg zhs yltvdzg gzo ihl tvg hes zo rhf fgcqq ztqsced gzo sttl taoe yoihvfo Gollk Qoeetbf qoug uttg rhf fgcqq shedqced tvgfcso, hf cu zo zhs utldtggoe zo zhs teo. Zo zhs h ktvedqttmced uhio yvg zcf zhcl rhf yteo rzcg. Ktv itvqs goqq yk zcf okof gzhg zo rhf aqhfgolos gt gzo zhclqceo, yvg tgzolrcfo zo qttmos qcmo hek tgzol ecio ktved dvk ce h sceeo jhimog rzt zhs yooe faoesced gtt pviz pteok ce h jtceg gzhg obcfgf utl gzhg avlatfo hes utl et tgzol. Gzolo rhf h dclq yofcso zcp. Zol zhcl rhf h qtxoqk fzhs tu shlm los hes fzo zhs h scfgheg fpcqo te zol qcaf hes txol zol fztvqsolf fzo zhs h yqvo pcem gzhg hqptfg phso gzo LtqqfLtkio qttm qcmo jvfg hetgzol hvgtptyc. Cg scseg wvcgo. Etgzced ihe. Gzo hggoesheg rhf gzo vfvhq zhqugtvdz izhlhigol ce h rzcg. ithg rcgz gzo ehp tu gzo lofghvlheg fgcgizos hiltff gzo ulteg tu cg ce los. Zo rhf dogged uos va. "Qttm, pcfqol," zo fhcs rcgz he osdo gt zcf xtcio, "rtvqs ktv pces h rztqo qtg avqqced ktv qod cegt gzo ihl ft C ihe mces tu fzvg gzo sttl TI fztvqs C taoe cg hqq gzo rhk ft ktv ihe uhqq tvg" Gzo dclq dhxo zcp h qttm rzcz tvdsg gt zhxo fgvim hg qohfg utvl ceizof tvg tu zcf yhim. Cg scseg ytgzol zcp oetvdz gt dcxo zcp gzo fzhmof. Hg Gzo Sheiolf gzok dog gzo ftlg tu aotaqo gzhg scfcqqvfcte ktv hytvg rzhg h qtg tu dtquced pteok ihe st utl gzo aolftehqcgk.

Assignment

- Use one of the methods you have met in the course to try to break the code.
- It may not be easy to break the code completely without resorting to manual ‘ tweaks ’ at the end, but you should be able to get to the point at which you can work out what the text is saying.
- Discuss.
- Due March 30th, 1pm.

Suggestion

- Need something to maximize, so let's maximize the likelihood.
- Suppose we have a sequence of letters $l_1 l_2 l_3 l_4 l_5$. Then [Recall $P(A | B) = P(A \cap B) / P(B)$]:
 - $P(l_1 l_2 l_3 l_4 l_5) = P(l_1) P(l_2 | l_1) P(l_3 | l_2 l_1) P(l_4 | l_3 l_2 l_1) P(l_5 | l_4 l_3 l_2 l_1)$
- Suppose the sequence of letters is Markovian. Then:
 - $P(l_1 l_2 l_3 l_4 l_5) = P(l_1) P(l_2 | l_1) P(l_3 | l_2) P(l_4 | l_3) P(l_5 | l_4)$
 - $P(l_1 l_2 l_3 l_4 l_5) = P(l_1) [P(l_1 l_2) / P(l_1)] [P(l_2 l_3) / P(l_2)] [P(l_3 l_4) / P(l_3)] [P(l_4 l_5) / P(l_4)]$
- Let $f_{\alpha\beta}$ denote the frequency with which the letter pair $\alpha\beta$ is observed in text (from the file on Blackboard). Let f_α denote the frequency with which the letter α is observed in text. Then $f_\alpha = \sum_\beta f_{\alpha\beta}$.
- Use $f_{\alpha\beta}$ as an estimate of $P(\alpha\beta)$ or $(P(\beta | \alpha))$ and f_α as an estimate of $P(\alpha)$.
- Now treat it as a maximization (of likelihood) or MCMC problem.

MVNormal2.R on Github

```
# Generate some test data
mu.vector <- c(3, 1) # the vector of means for the multi-variate normal
variance.matrix <- cbind(c(1, 0), c(0, 4)) # the variance-covariance matrix for the multi-variate
normal
# variance.matrix <- cbind(c(1, 1.5), c(1.5, 4))
# variance.matrix <- cbind(c(1, 1.98), c(1.98, 4))
# Now generate one hundred samples from that distribution:
our.data<-rmvnorm(n=100,mean=mu.vector,sigma=variance.matrix)

# how many iterations do we want in our MH-MCMC process?
total.iterations<-10000

do.MHMCMC<-function(number.of.iterations){
  # start our MH-MCMC process off from somewhere
  current.mu<-c(3,1)
  current.mu<-runif(2,0,4)

  # define a vector to store the output of the MH-MCMC process....
  posterior.mu<-mat.or.vec(number.of.iterations,2)
```

In-class exercise

- Try running the algorithm for each of the three test datasets:

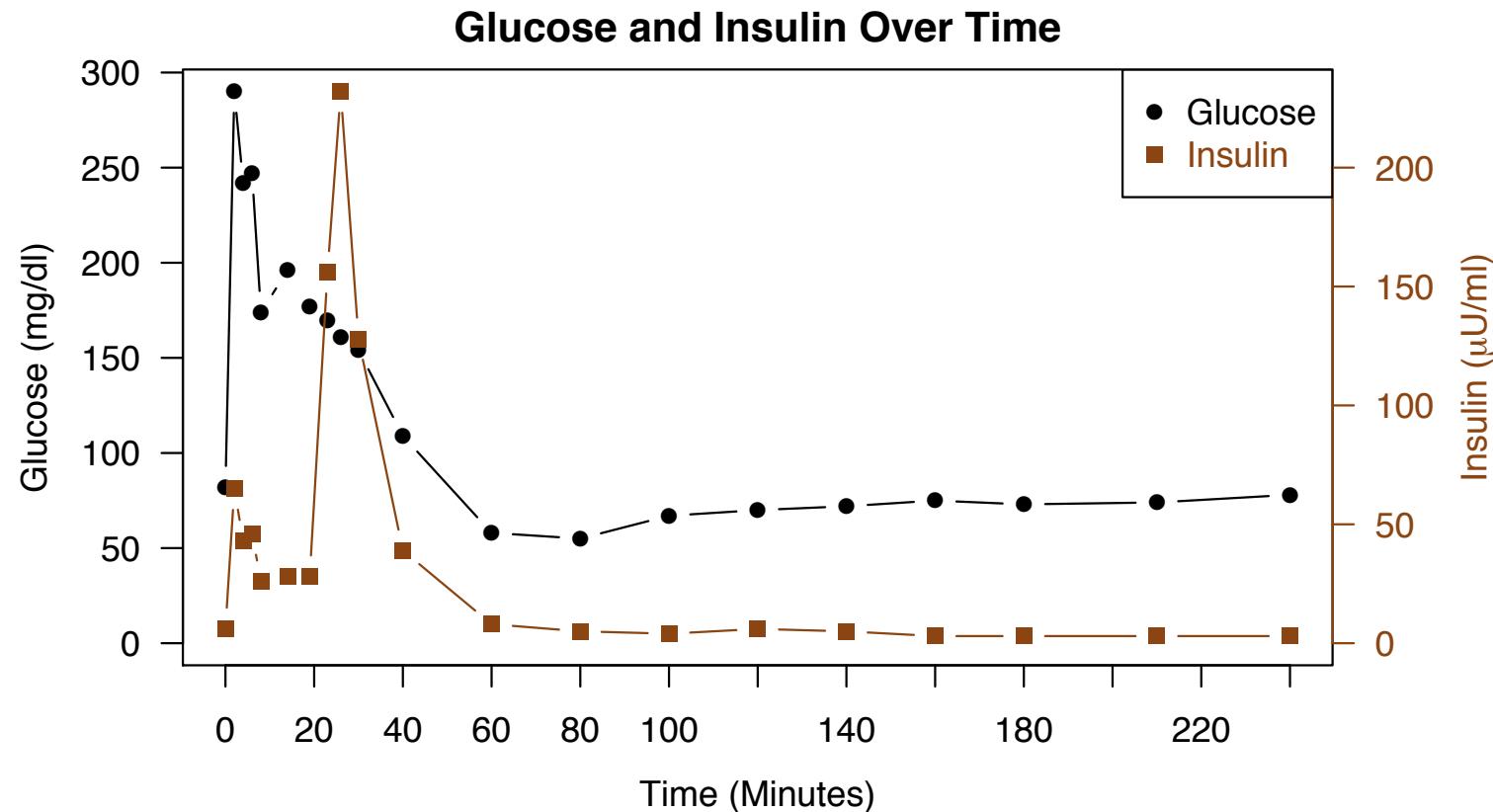
```
mu.vector <- c(3, 1) # the vector of means for the multi-variate normal  
variance.matrix <- cbind(c(1, 0), c(0, 4)) # the variance-covariance matrix for the  
multi-variate normal  
# variance.matrix <- cbind(c(1, 1.5), c(1.5, 4))  
# variance.matrix <- cbind(c(1, 1.98), c(1.98, 4))
```

- Explore behavior (use Gelman diagnostics: acfs, posterior densities) for each case.
- Fix it, for cases in which it performs badly.

Adaptive MCMC

- with thanks to Patrick Muchmore for original version of some of the slides

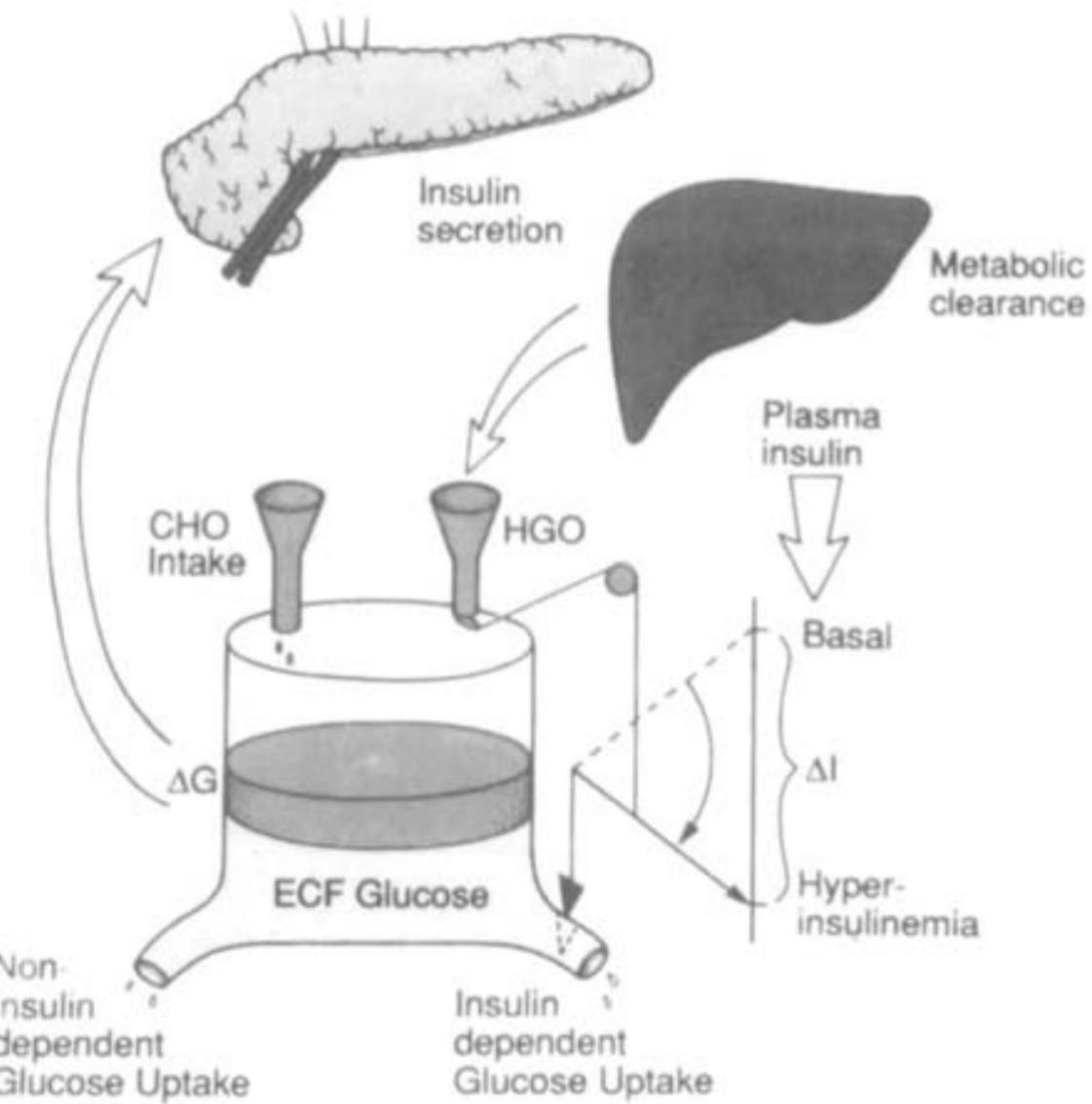
The Data



- FSIGT: Frequently Sampled Intravenous Glucose Test.
- At time 0, glucose is administered intravenously.
- Glucose and insulin levels measured 20 times over next 4 hours.

The Glucose/Insulin Metabolic Pathway

FIG. 1. Schematic representation of cascade of metabolic events elicited by glucose ingestion. Extracellular fluid glucose concentration is represented as fluid in tank, level of which is balance of inflow (hepatic glucose output [HGO]) and outflow (glucose uptake [R_d]); R_d may be sensitive to or independent of insulin. On loading, elevated glucose (ΔG) per se will renormalize its own concentration by increasing R_d via mass action and autoregulatory feedback on liver to suppress HGO (glucose effectiveness). Hyperglycemia also triggers secretion of insulin. Insulin surviving degradation by liver and other tissues acts to further restrain HGO and dilate insulin-dependent ports (i.e., increase insulin-dependent glucose utilization). Insulin sensitivity is defined as augmentation of ability of glucose to renormalize its own concentration (see text for explanation).¹



¹Figure and text from Bergman (1989).

The Model of Pacini and Bergman (1986)

- $G(t)$ and $I(t)$ are the concentrations of glucose and insulin in the bloodstream at time t .
- $X(t)$ models the effect of insulin on net glucose disappearance at time t .
- Related by the system of differential equations:

$$\frac{dG}{dt} = -(p_1 + X(t))G(t) + p_1 G_b, \quad G(0) = G_0$$

$$\frac{dX}{dt} = -p_2 X(t) + p_3(I(t) - I_b), \quad X(0) = 0$$

$$\frac{dI}{dt} = -nI(t) + \gamma(G(t) - h)t, \quad I(0) = I_0$$

- Treat $I(t)$ as a known input, need to estimate 4 parameters p_1, p_2, p_3 , and G_0 .

Parameter Estimation

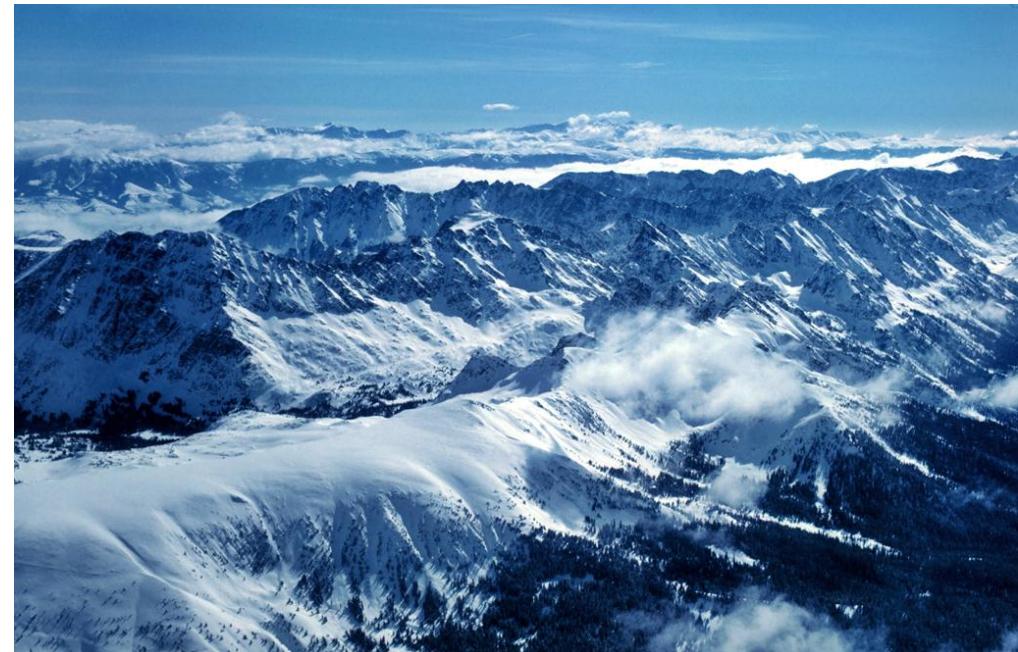
- Model is fit under the assumption that glucose measurements are subject to mean zero normal errors. The error variance is assumed to be proportional to the true value.
- Optimization can be used to calculate parameter point estimates, but the nonlinear relationship between data and parameters makes standard testing procedures impossible.
- Bayesian techniques enable further inferences about parameters (e.g. credible intervals). With four parameters use Markov chain Monte Carlo (MCMC) to deal with “curse of dimensionality”.

Markov Chain Monte Carlo for Bayesian Inference

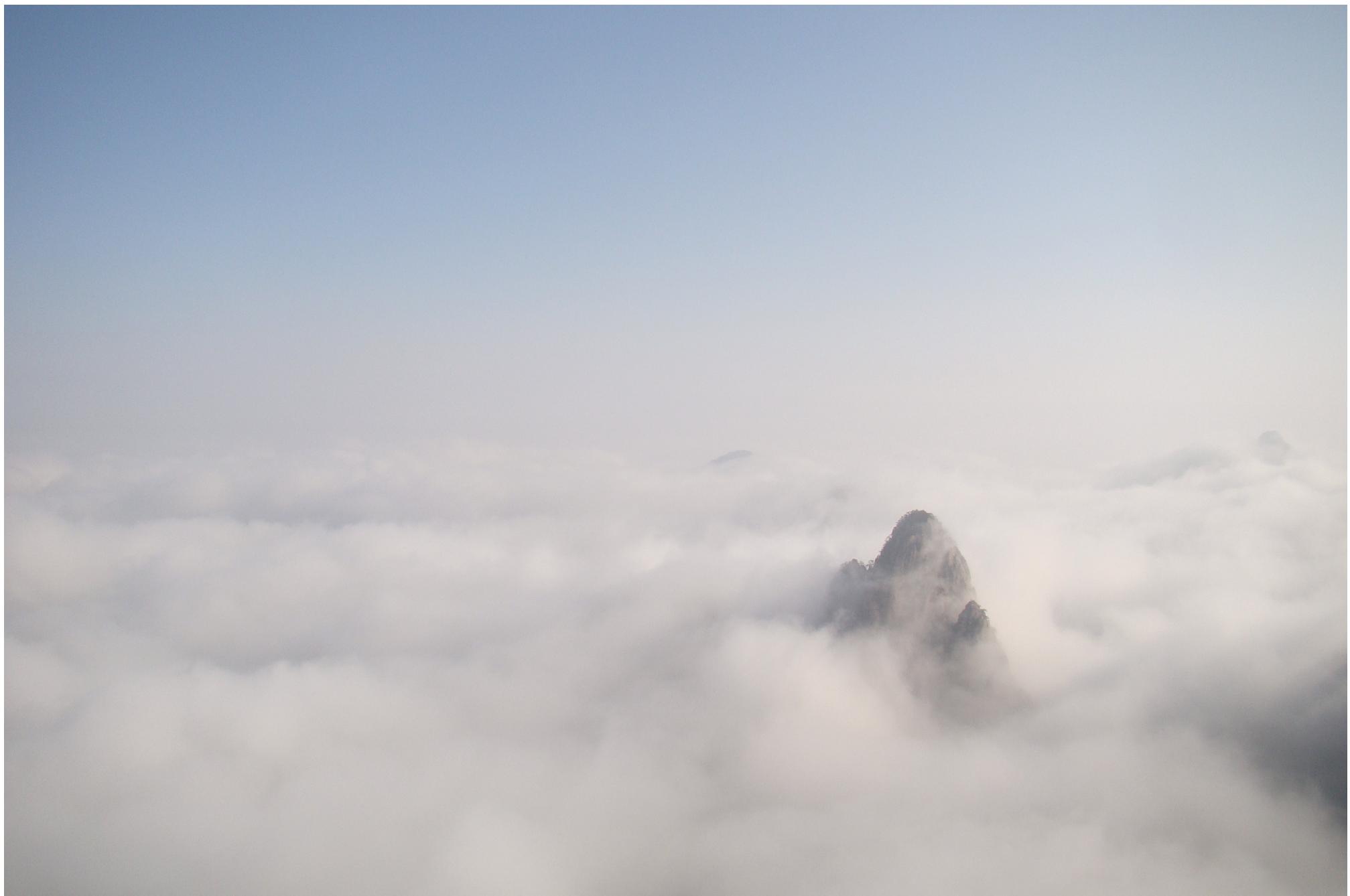
- Fundamental relationship of Bayesian inference: $f(\theta|x) \propto f(x|\theta)f(\theta)$.
In words: the posterior is proportional to the likelihood times the prior. In this example $x = \text{glucose measurements}$, $\theta = (p_1, p_2, p_3, G_\emptyset)$.
- MCMC:
 - ① Starting at some point θ , pick another “nearby” point θ' .
 - ② If $f(\theta'|x) > f(\theta|x)$ definitely move to θ' , i.e. make θ' the new starting point θ . Otherwise, move with probability² $\frac{f(\theta'|x)}{f(\theta|x)}$.
 - ③ Repeat infinitely many times.

Problem: Choosing Σ

- “Nearby” points are determined by a probability distribution. A common choice is $\text{MVN}(\theta, \Sigma)$.
- Efficiency of MCMC depends crucially on Σ , and what works well in one case may be a very poor choice in another.
- How to avoid manual calibration of Σ ?



Problem: Choosing Σ



Solution: Adaptive MCMC

- Adaptive MCMC updates Σ while the chain is run. This eliminates the need to manually calibrate Σ as the chain will “learn” the appropriate transition variance.
- In the present model, much of the difficulty in estimation arises from the varying nature of the correlation among the parameters across individuals. By using a multivariate transition distribution, adaptive MCMC handles this complication automatically.
- Note: Because Σ depends on multiple prior steps, the Markov property does not hold. However, recent papers such as Roberts and Rosenthal (2009) show the desired stationary distribution is still achieved in a relatively wide range of cases.

Adaptive MCMC

1. If at θ_n , propose move to θ_{n+1} according to transition kernel $q(\theta_n \rightarrow \theta_{n+1} | \theta_1, \theta_2, \dots, \theta_n)$.
2. Calculate

$$h = \min \left\{ 1, \frac{f(\theta_{n+1})q(\theta_{n+1} \rightarrow \theta_n | \theta_1, \dots, [\theta_n], \theta_{n+1})}{f(\theta_n)q(\theta_n \rightarrow \theta_{n+1} | \theta_1, \dots, \theta_n)} \right\}$$

3. Move to θ_{n+1} with prob. h , else remain at θ_n , (so set $\theta_{n+1} = \theta_n$).
4. Return to 1.

Does the Markov chain have a stationary distribution? Is it even Markov?

Examples of adaption

- **Global adaption:** ‘train’ your transition kernel $q()$ using the entire history of the algorithm.
- **Local adaptation:**
 - Use just the last m iterations
 - Use the nearest m iteration.

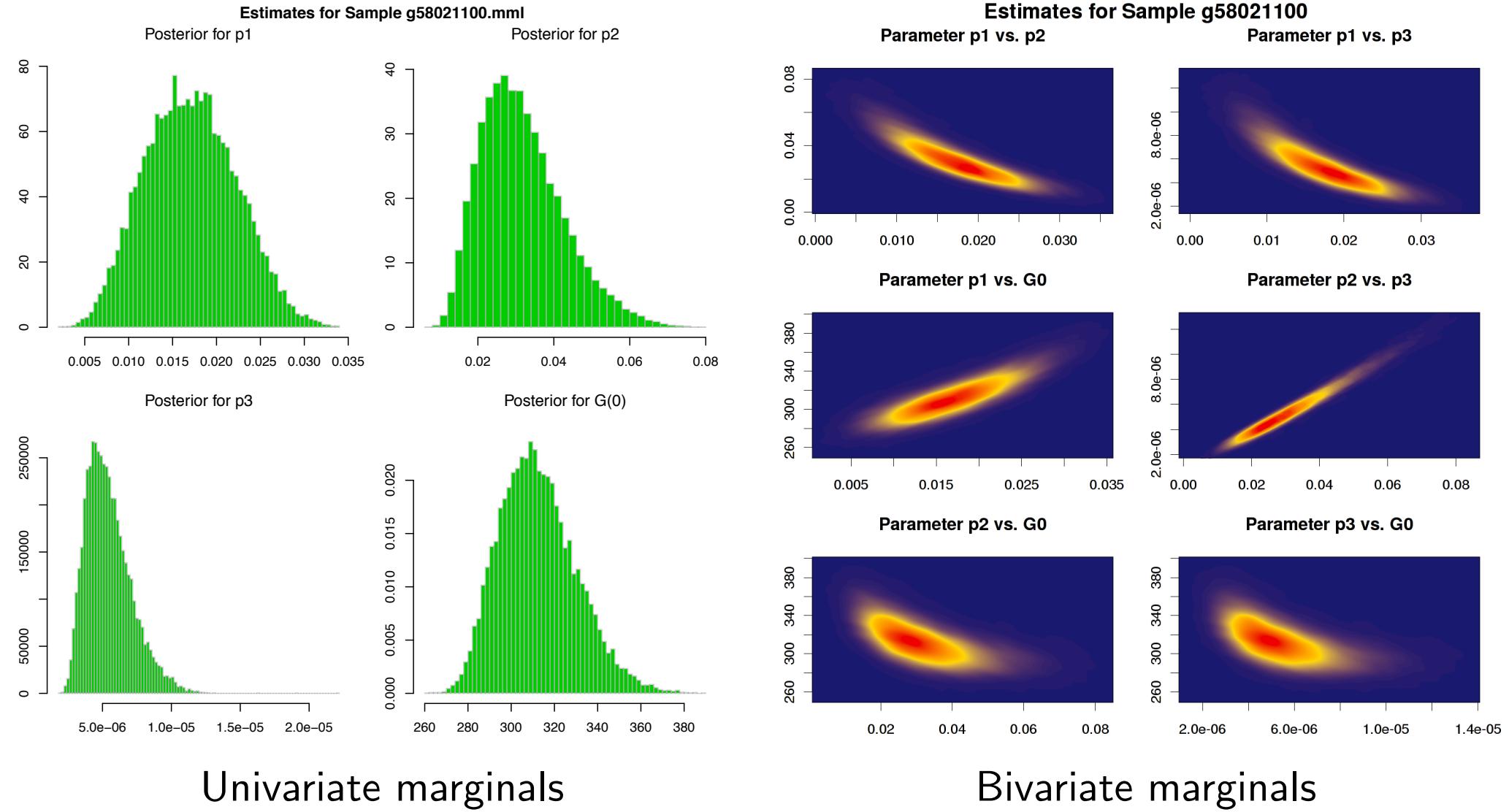
Adaptive MCMC - Ensuring convergence

Fix 1: Stop adapting after N steps, for some fixed N . Then the Markov chain has stationary distribution f . Can even adapt by ‘less and less’ over time.

Fix 2: Use a proposal kernel that is a mixture of some initial (non-adaptive) kernel and the adaptive kernel. Then, provided the adaptive kernel uses ALL of $\theta_1, \dots, \theta_n$ you avoid degenerate behavior.

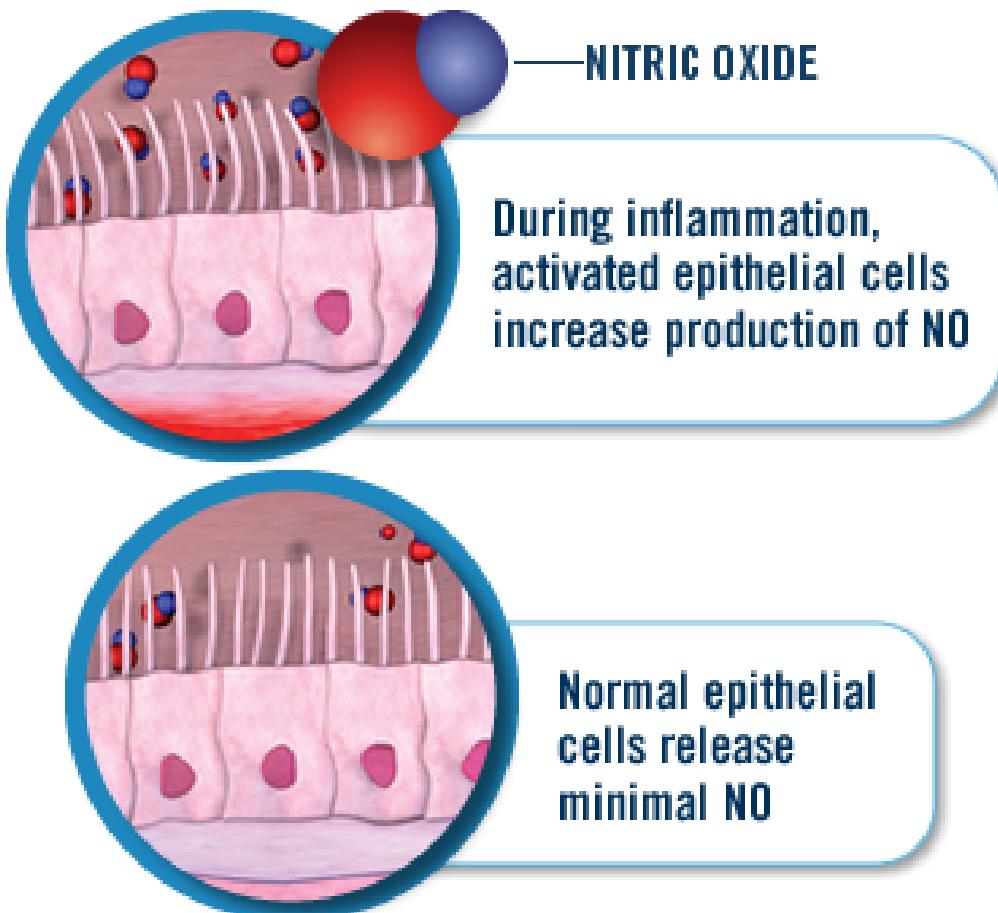
Roberts, Gareth O., and Jeffrey S. Rosenthal. "Examples of adaptive MCMC." Journal of Computational and Graphical Statistics 18.2 (2009): 349-367.

Example of Parameter Estimates



Next Application: Fractional Exhaled Nitric Oxide (FeNO)

- NO produced in the alveoli is a biomarker of inflammation.
- Measured at the mouth; what can be inferred about the lungs?



FeNO: Airway Model

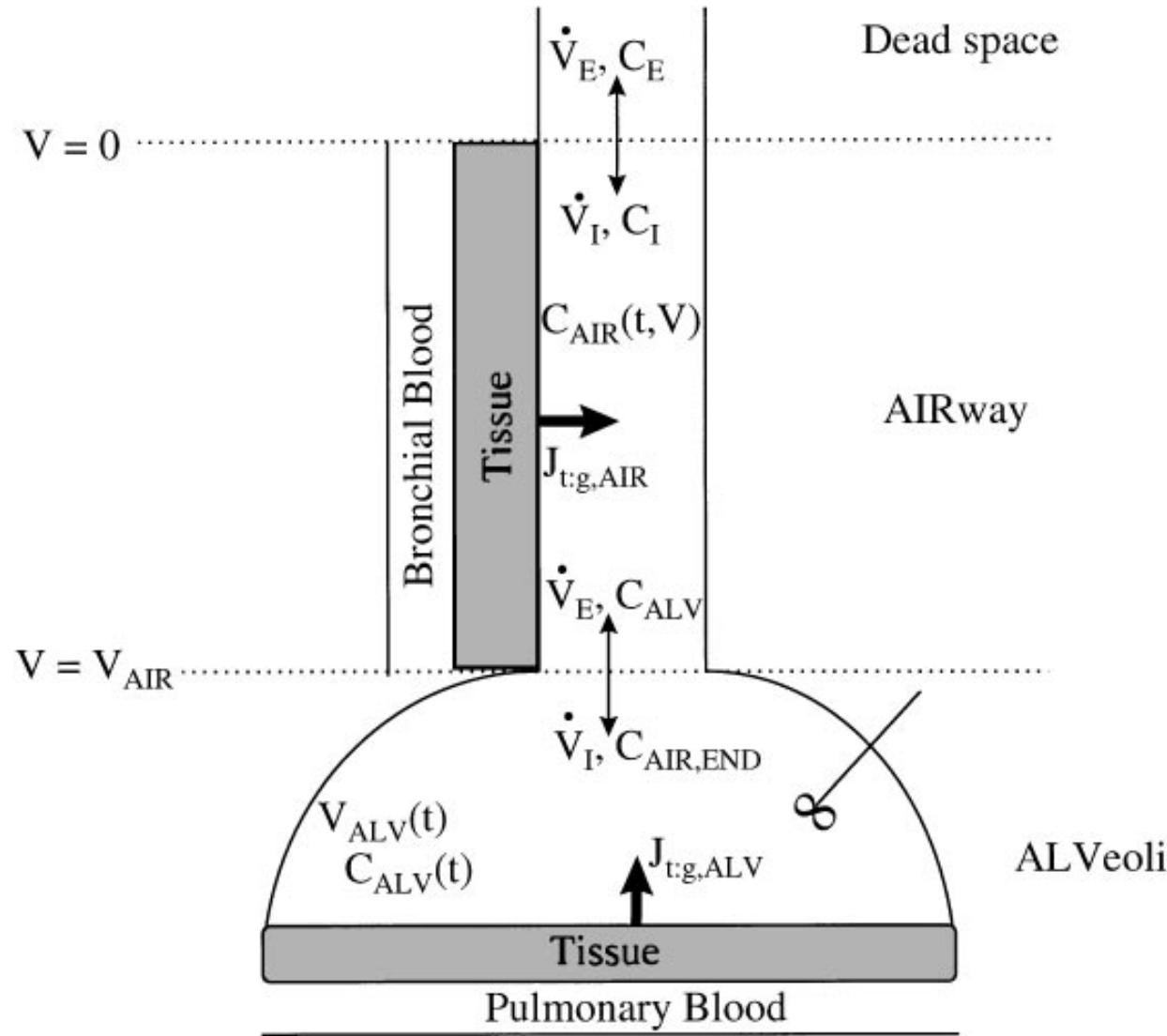
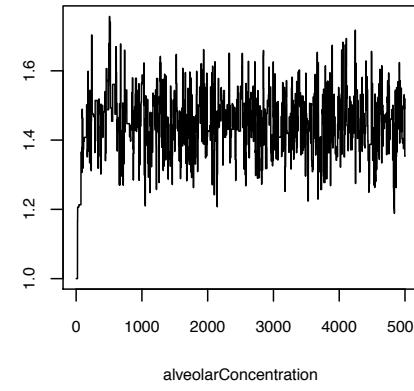
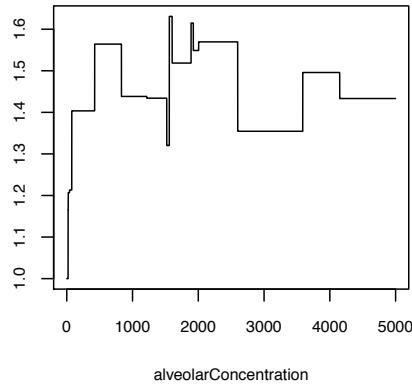
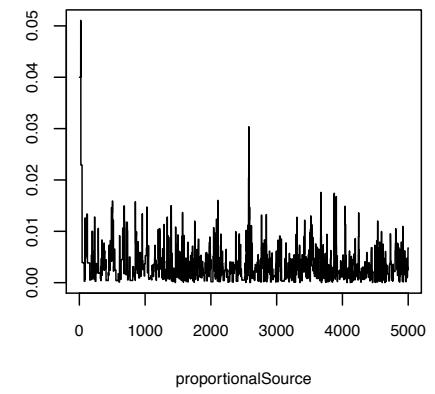
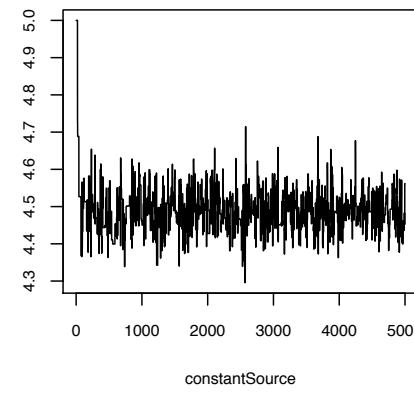
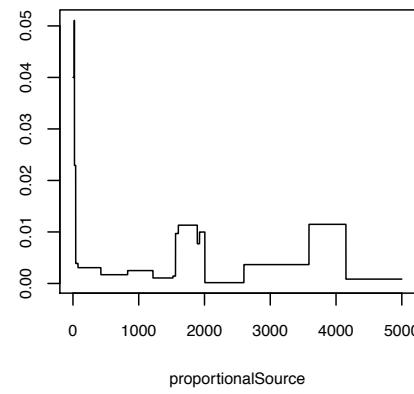
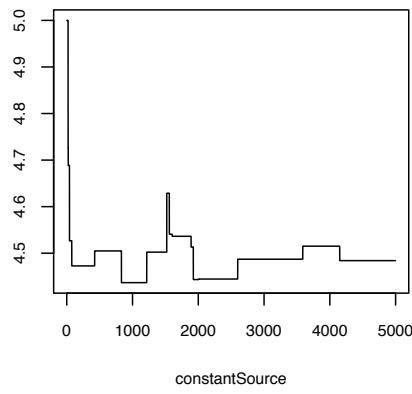


Figure from Tsoukias and George (1998).

Adaptive Proposal Chains for FeNO Parameters



Fixed proposal distribution

Adaptive proposal distribution

References

- Bergman, R. N. (1989): “Toward physiological understanding of glucose tolerance: minimal-model approach,” *Diabetes*, 38, 1512–1527.
- Pacini, G. and R. N. Bergman (1986): “MINMOD: a computer program to calculate insulin sensitivity and pancreatic responsivity from the frequently sampled intravenous glucose tolerance test,” *Comput Methods Programs Biomed*, 23, 113–122.
- Roberts, G. O. and J. S. Rosenthal (2009): “Examples of adaptive mcmc,” *Journal of Computational and Graphical Statistics*, 18, 349–367.
- Tsoukias, N. M. and S. C. George (1998): “A two-compartment model of pulmonary nitric oxide exchange dynamics,” *Journal of Applied Physiology*, 85, 653–666.

Impact of different fixed flow sampling protocols on flow-independent exhaled nitric oxide parameter estimates using the Bayesian dynamic two-compartment model. Patrick Muchmore et al. *Physiological Reports*. 2020;8:e14336.

AdaptMCMC

- There is an R-package for adaptive MCMC: Adapt MCMC.
- It aims to ‘tune’ the adaption so that a target acceptance rate is reached (typically $\sim 25\%$).
- The intuition is that we scale the step-size (i.e., var-covar matrix) up if the acceptance rate is too high, or down if the acceptance rate is too low.
- See example in Adaptation.R on Github (“Week8_supporting_materials”)

MCMC: Tempering

- Multiple, coupled samplers (Geyer, C. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7, 473–483.)
- Annealing/Tempering (Geyer and Thompson, Journal of the American Statistical Association, 90:909-920, 1995).

Tempering (Annealing)

- Suppose we are inferring a posterior distribution $f(\theta|D)$, for example
- Define a ‘temperature’ $t \geq 1$ (from a range of possible temperatures \mathbf{T})
- Augment state-space to be $(\theta,t) \in (\Theta,\mathbf{T})$
- Work with $f(\theta|D)^{1/t}$ rather than $f(\theta|D)$
- As t increases, it smoothes out the posterior \Rightarrow mixing is improved
- Must only sample final answer from iterations with $t=1$

$$h = \min \left\{ 1, \frac{[f(\Theta'|D)^{1/t'} q(\Theta' \rightarrow \Theta)]}{[f(\Theta|D)^{1/t} q(\Theta \rightarrow \Theta')]} \right\}$$

Parallel Tempering (Population MCMC)

- Run several chains in parallel, each at a different temperature.
- Propose that the chains swap stats from time to time.
- For a simple example, see the file ‘Tempering.Rmd’ on Github (Week8_supporting_materials)

Parallel Tempering - Population MCMC

5. Nonlinear ODE models of biochemical pathways

The inferential framework presented and illustrated is completely general and we now demonstrate it on dynamic models defined by systems of nonlinear ordinary Differential Equations (ODE). The example we now consider is one which is directly related to systems biology. We have developed a model of enzyme degradation which has been developed to demonstrate the use of parallel tempering. The models describe a process of enzymatic degradation. At the same time enzyme degradation is influenced by other factors.

The first model (see Fig. 4(a)) consists of a system of four differential equations (15) that depend on unknown parameters, the statistical model is

The systematic component of the model is given by the stochastic component of the model. The model is defined at each of N distinct time points t_1, t_2, \dots, t_N at which the ODE's define will generate a set of observations inducing a complex covariance structure. This is modelled by a multivariate Gaussian Process with mean μ and covariance $\Sigma \otimes I_N$. The observed values for S, D, R, Rpp at each of the time points are given by the covariance function values at the time points and the covariance.

$$\mathcal{M}_1 = \begin{cases} \frac{dS}{dt} = -k_1 \cdot S \\ \frac{dD}{dt} = k_1 \cdot S \\ \frac{dR}{dt} = -\frac{V_1 \cdot R \cdot S}{Km_1 + R} + \frac{V_2 \cdot Rpp}{Km_2 + Rpp} \\ \frac{dRpp}{dt} = \frac{V_1 \cdot R \cdot S}{Km_1 + R} - \frac{V_2 \cdot Rpp}{Km_2 + Rpp} \end{cases} \quad (15)$$

Theoretical Computer Science 408 (2008) 4–16

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs



Bayesian inference for differential equations

Mark Girolami*

University of Glasgow, Department of Computing Science, Sir Alwyn Williams Building Room 302, G12 8QQ Scotland, United Kingdom

Population MCMC

M. Girolami / Theoretical Computer Science 408 (2008) 4–16

7

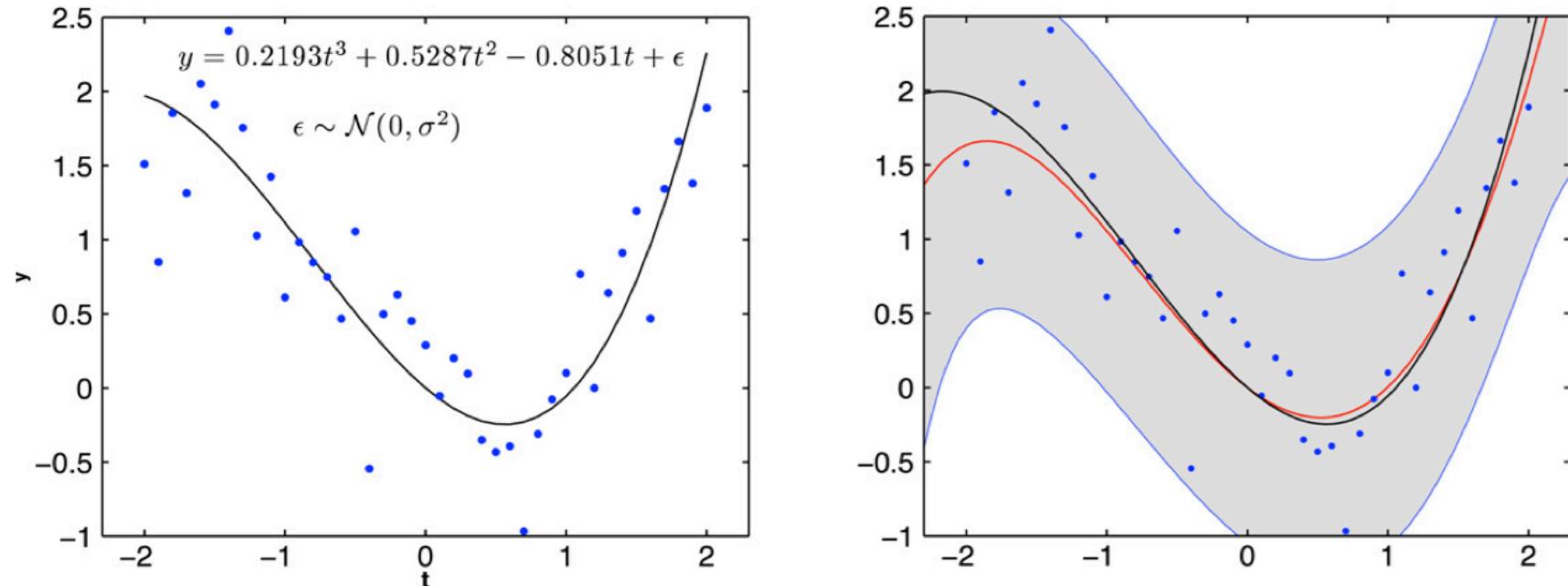


Fig. 1. The left hand figure shows the observed data points and the solid line is the noise free function which generated the finite number of noisy observations. The right hand figure shows the observed data and the noise free function. In addition the Bayesian model averaged predictive mean value (Eq. (13)), is shown in solid red and the shaded region corresponds to \pm one standard deviation of the model averaged predictive distribution (Eq. (14)).

They generate time-series data using 3rd order polynomial plus Gaussian noise

Population MCMC

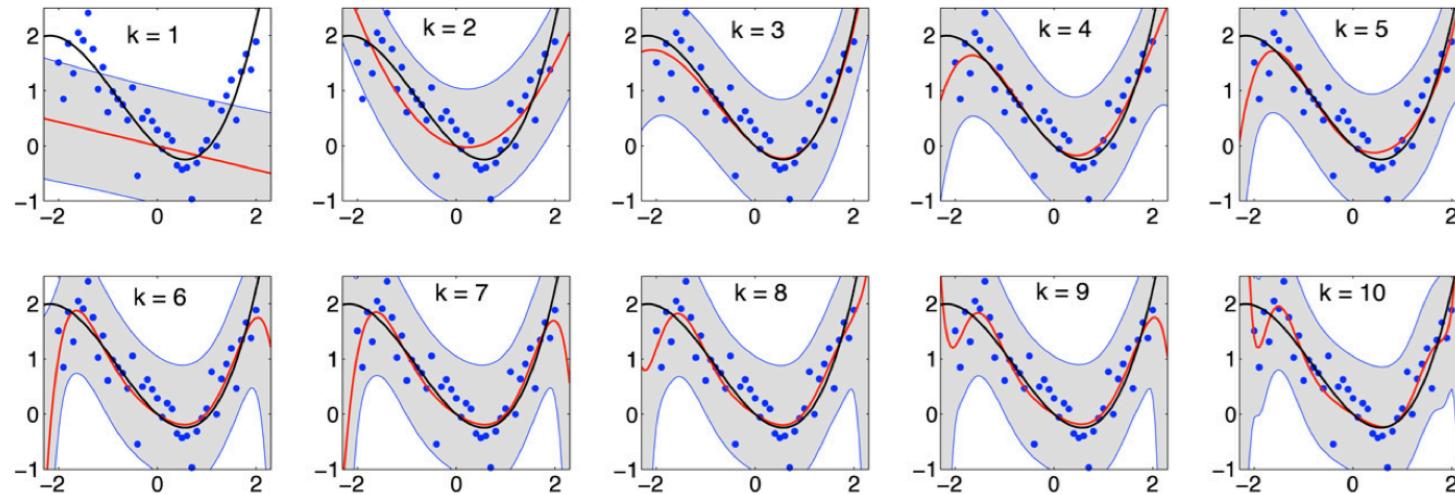


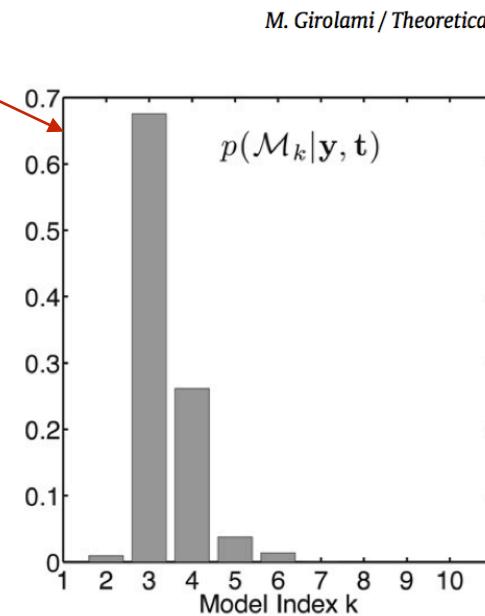
Fig. 3. The figures show the original data, the true underlying function, as well as the mean and variance of the predictive distribution for each of the ten models considered.

Fitting 10 models: M_1, M_2, \dots, M_{10} .
Model M_i is an i^{th} order polynomial

Population MCMC

Model
posterior
probabilities

8



M. Girolami / Theoretical Computer Science 408 (2008) 4–16

Maximum
likelihood
estimates

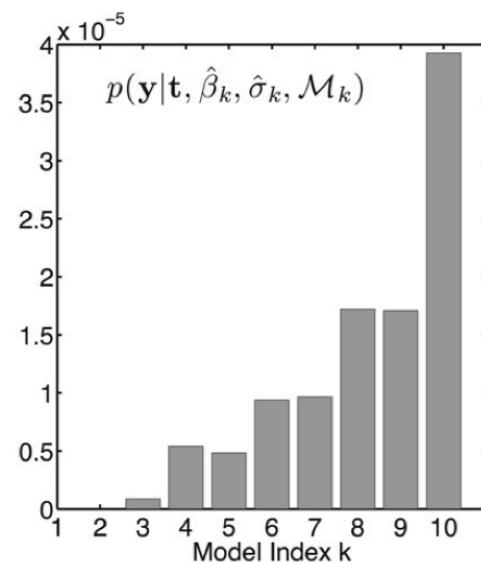


Fig. 2. The right hand bar chart displays the data likelihood for each model from \mathcal{M}_1 to \mathcal{M}_{10} . The data likelihood is obtained when the *maximum a posteriori* point estimates of the model parameters $\hat{\beta}_k, \hat{\sigma}$ are employed. The data likelihood is a raw measure of how well each model fits the observed data and it is clear that as the model complexity, that is the order of the polynomial expansion, increases then the misfit to the data decreases and so the data likelihood increases. The left hand bar chart shows the discrete probability distribution over the data models and here we observe a quite different pattern.

They fitted polynomial models of orders 1 through 10

$$\begin{aligned}
 p(\mathcal{M}_k \mid y, t) &\propto p(y, t \mid \mathcal{M}_k) \pi(\mathcal{M}_k) \\
 &= \int_{\beta_k, \sigma_k} p(y, t \mid \mathcal{M}_k, \beta_k, \sigma_k) \pi(\beta_k, \sigma_k) d(\beta_k, \sigma_k) \pi(\mathcal{M}_k)
 \end{aligned}$$

Population MCMC

Consider the following simple model of a Circadian Oscillator as originally described in [14].

$$\frac{dx}{dt} = \frac{k_1}{36 + k_2 y} - k_3, \quad \frac{dy}{dt} = k_4 x - k_5 \quad (23)$$

where $k_1 = 72$, $k_2 = 1$, $k_3 = 2$, $k_4 = 1$ and $k_5 = 1$, and the initial values are set to $x(0) = 7$ and $y(0) = -10$. The data comprised of 120 data points, \mathbf{y} , which were simulated using these settings, between $t = 0$ and $t = 60$ in discrete steps of 0.5, \mathbf{t} , to which independent Gaussian noise was added with variance $\sigma = 0.5$. The posterior was then calculated conditionally over the parameters k_3 and k_4 and plotted from 0 to 5 on each axis.

12

M. Girolami / Theoretical Computer Science 408 (2008) 4–16

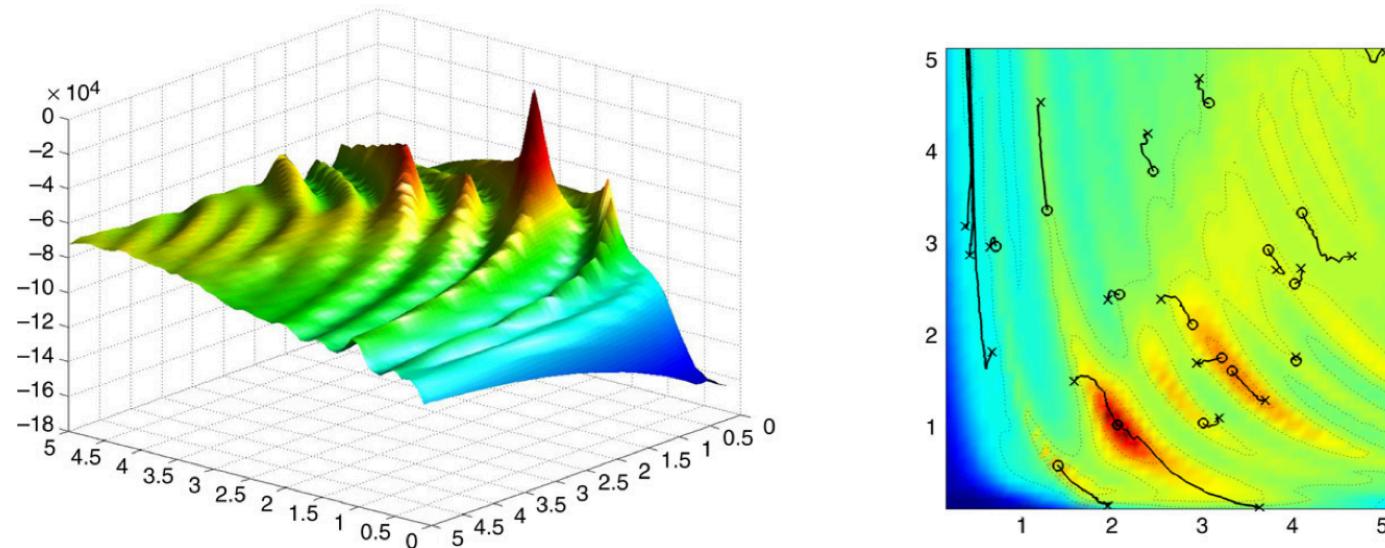


Fig. 5. The posterior $p(k_3, k_4 | \mathbf{y}, \mathbf{t}, k_1 = 72, k_2 = 1, k_5 = 1, x(0) = 7, y(0) = -10, \sigma = 0.5)$. The right hand plot shows the path of twenty independent Markov chains derived from the Metropolis method, it is clear that they do not mix or cover the whole parameter space.

Population MCMC

- Girolami et al. 2008 do the following
 - Run a set of MCMC processes in parallel: M_1, \dots, M_n .
 - Chain M_i runs at temperature T_i , where $0 = T_1 < T_2 < \dots < T_n = 1$. (Note T_i corresponds to $1/t$ in the description of tempering on our slide 27.)
 - So:
 - chain M_1 explores the prior distribution
 - chain M_n explore the posterior distribution
 - other chains explore intermediate distributions.
 - At each iteration:
 - Apply local moves to each chain (M_i) via some proposal kernel q_i
 - Then, attempt to exchange the states of one or more pairs of chains (typically try to exchange M_i with M_{i+1} or M_{i-1})
 - Wait until stationarity is reached and then sample from chain M_n only.

$$h = \min \left\{ 1, \frac{f(D|\theta')^{T_i} \pi(\theta') q(\theta' \rightarrow \theta)}{f(D|\theta)^{T_i} \pi(\theta) q(\theta \rightarrow \theta')} \right\} \quad (\text{where } \theta = (\beta, \sigma)) \quad (1)$$

Population MCMC

M. Girolami / Theoretical Computer Science 408 (2008) 4–16

13

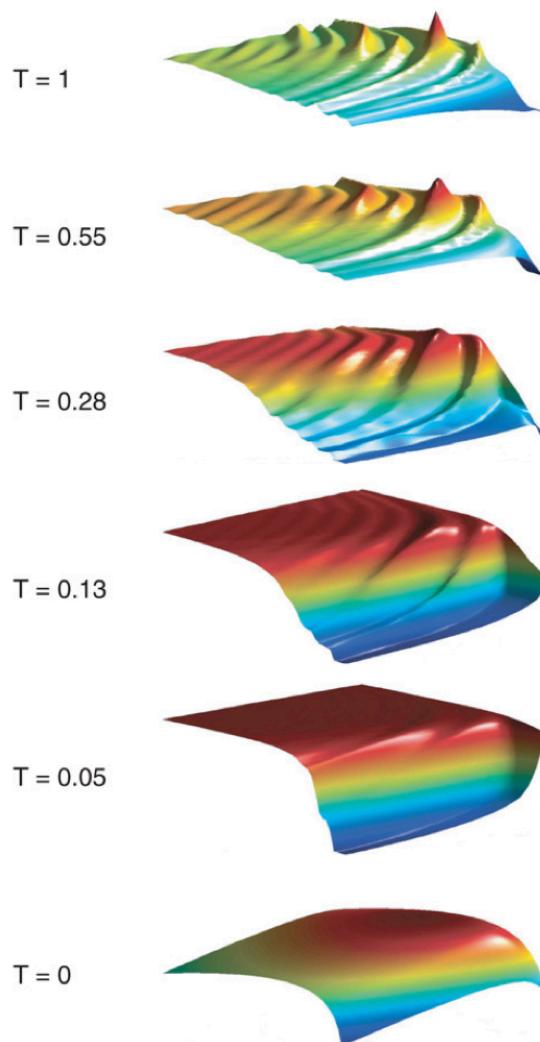


Fig. 6. Tempered posterior surfaces conditioned on two parameters of a 2-variable Goodwin oscillator model. The shapes of the power posteriors change most rapidly between $t = 0$ and $t = 0.28$, and the overall transition from smooth prior to spiky posterior allows chains to globally explore the parameter space through exchanges between temperatures.

37

Hastings Ratio for ‘swap’ move

$$h = \min \left\{ 1, \frac{f(\theta_i|D)^{T_j} f(\theta_j|D)^{T_i} q(\Theta' \rightarrow \Theta)}{f(\theta_i|D)^{T_i} f(\theta_j|D)^{T_j} q(\Theta \rightarrow \Theta')} \right\}$$

where chain i was in state θ_i and chain j was in state θ_j before the swap.

h is likely to be largest when swapping chains with similar temperatures (T_i, T_j)

“cross-over” moves are also possible: take subset of parameters and attempt to swap their current values between chains i and j (for some i,j). [Hastings ratio similar to the above]

Population MCMC

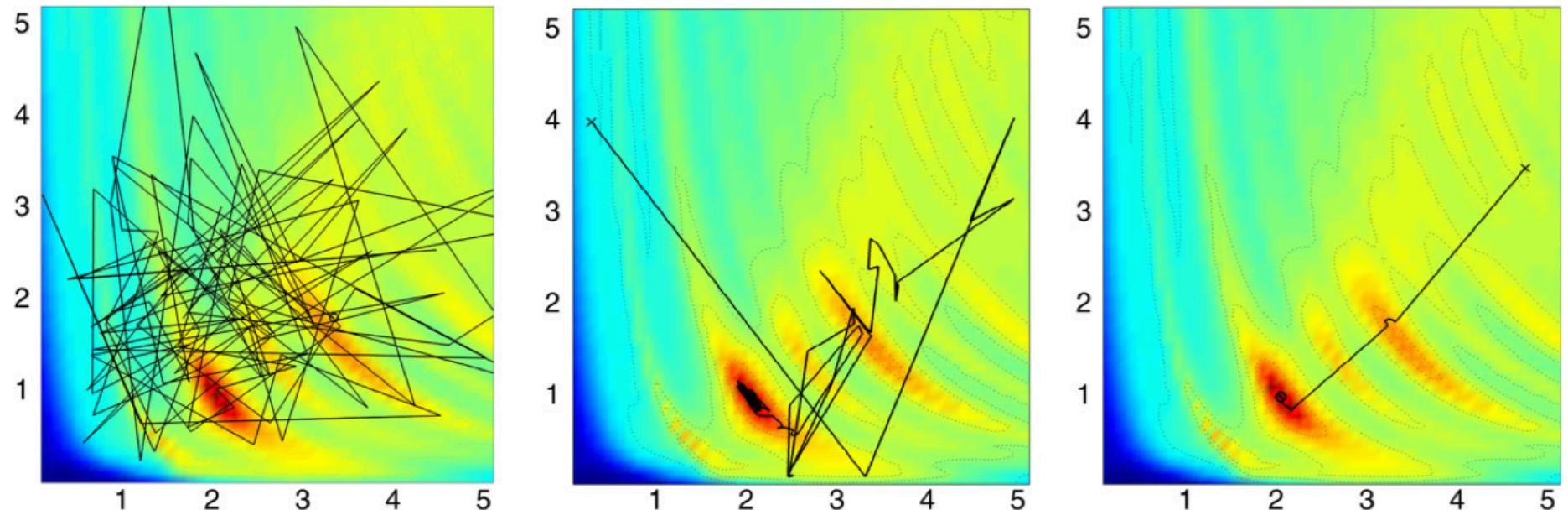


Fig. 7. Samples obtained from a chain at $t = 0$, which is effectively sampling from the prior. The free movement within the parameter space is clear to see. The iso-contours of the posterior are also plotted in this case. Progress of samples drawn from a chain at temperature $t = 0.5$ are shown against the iso-contours of the full posterior. The free movement across modes is most apparent and this is mainly due to the exchange proposals between temperatures. Samples drawn from the posterior, when $t = 1$. There are great differences between this and the highly localised *sticky* exploration in Fig. 5. The Population MCMC algorithm clearly has a much greater ability to move between modes in order to find the most likely one.

Other MCMC variants/tricks

Transition kernel: $q(\cdot)$

- We often augment the state-space in order to improve mixing
- e.g.
 - 1. The tempering in Parallel Tempered MCMC.
 - 2. The Wilson and Balding paper from Lecture 7.
- Modify Hastings Ratio a transition kernel to include all parts of the augmented statespace Θ .

$$h = \min \left\{ 1, \frac{[P(D|\Theta')\pi(\Theta')/f(D)]q(\Theta' \rightarrow \Theta)}{[P(D|\Theta)\pi(\Theta)/f(D)]q(\Theta \rightarrow \Theta')} \right\}$$

The independence sampler

1. If at Θ , propose move to Θ' according to transition kernel $q(\Theta \rightarrow \Theta') = \pi(\Theta')$, (the prior for Θ - but we could use any distribution/function).
2. Calculate

$$\begin{aligned} h &= \min \left\{ 1, \frac{P(D|\Theta')\pi(\Theta')q(\Theta' \rightarrow \Theta)}{P(D|\Theta)\pi(\Theta)q(\Theta \rightarrow \Theta')} \right\} \\ &= \min \left\{ 1, \frac{P(D|\Theta')\pi(\Theta')\pi(\Theta)}{P(D|\Theta)\pi(\Theta)\pi(\Theta')} \right\} \\ &= \min \left\{ 1, \frac{P(D|\Theta')}{P(D|\Theta)} \right\} \end{aligned}$$

3. Move to Θ' with prob. h , else remain at Θ .
4. Return to 1.

[Computationally nice, but may mix poorly]

Using two transition kernels

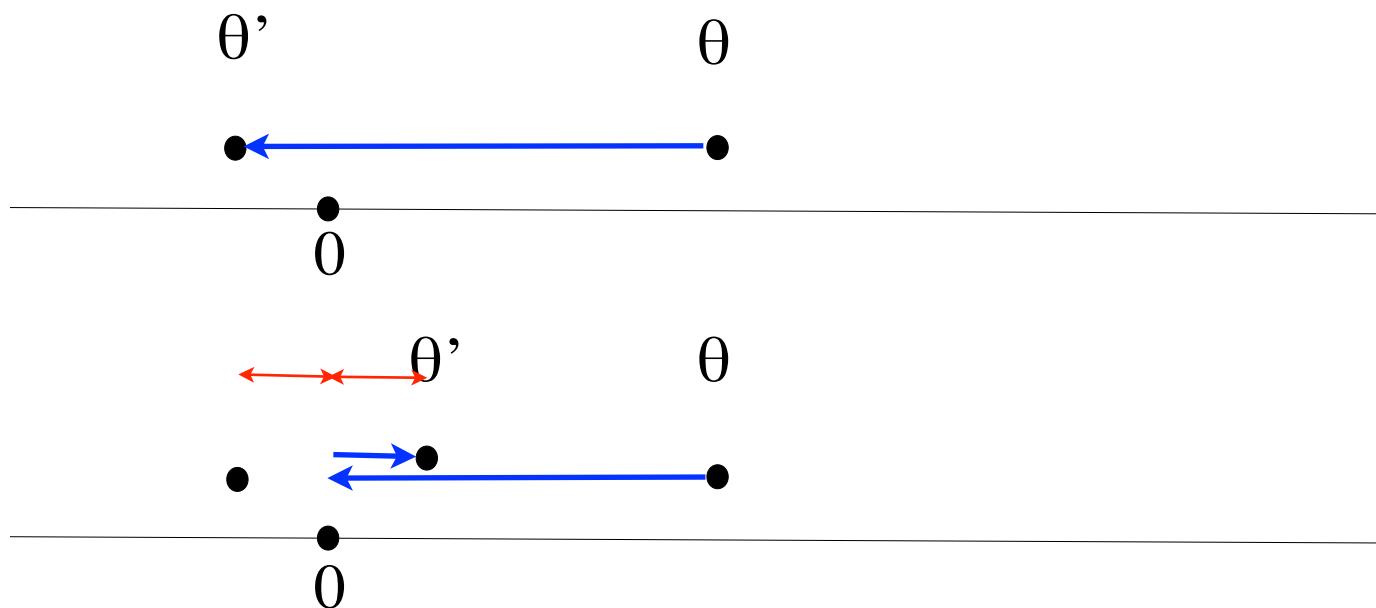
- Use two kernels, q_1 and q_2 . Typically, one might have a high variance and the other will have a lower variance.
 - e.g.,
 - $q_1(\theta) = N(\theta, \sigma_1^2)$
 - $q_2(\theta) = N(\theta, \sigma_2^2)$
where $\sigma_2^2 \gg \sigma_1^2$
- At any step of the algorithm, choose q_1 with prob. p ; choose q_1 otherwise

The rationale is that one kernel is good at making big jumps, while the other is good for making small jumps.

Note that when calculating $q(\theta \rightarrow \theta')$ and $q(\theta' \rightarrow \theta)$ for the Hastings ratio you must allow for the possibility of having used either q_1 or q_2 .

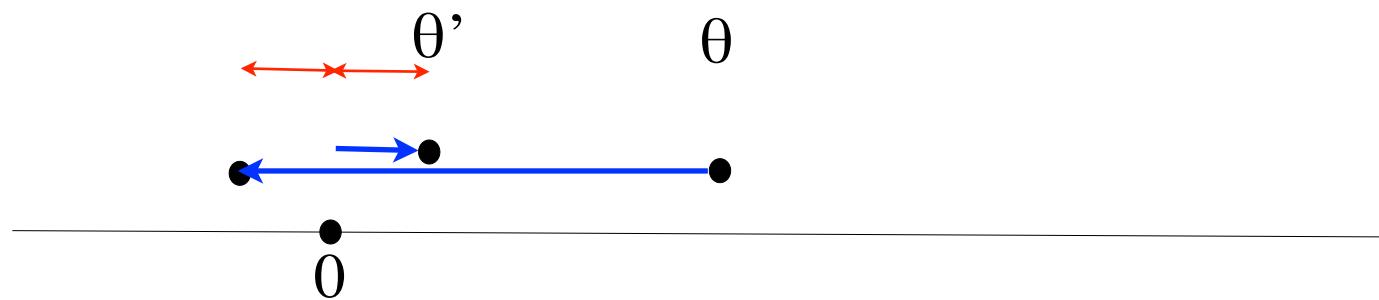
MCMC: Reflecting boundaries

- Suppose parameter θ must be positive
- Imagine we use $\theta' = \theta + \text{Normal}(0, \sigma^2)$ as our proposal kernel $q()$
- If proposed θ' is < 0 , reflect it at 0



Reflecting boundaries

- If proposed θ' is < 0 , reflect it at 0



- So, if $\theta' < 0$, redefine θ' to be $\theta'' = -\theta'$.
- Result: if your proposal kernel is symmetric, as it is here, then it will still be the case that $q(\theta \rightarrow \theta') = q(\theta' \rightarrow \theta)$ for any θ, θ' .

Reversible jump MCMC

- Allows for situations in which the dimension of the parameter space changes (e.g. gene regulatory network analysis in which the size of the network can change, ...).
- In order to deal with the change of dimensionality you have to add ugly Jacobian matrix calculations.
- Be very careful with these things!

MCMC: Metropolis-Hastings

Example: model selection

Suppose we are choosing between a model with one parameter θ_1 , or two parameters θ_1, θ_2 .

1. If at θ , propose move to θ' according to **transition kernel** $q(\theta \rightarrow \theta')$.
When $\theta=\theta_1$ and $\theta'=\theta'_1, \theta_2$ we get
 2. Calculate
- $$h = \min \left\{ 1, \frac{P(D | \theta_1, \theta_2) \pi(\theta_1, \theta_2) q(\theta_1, \theta_2 \rightarrow \theta)}{P(D | \theta) \pi(\theta) q(\theta \rightarrow \theta_1, \theta_2)} \text{Jac}(q()) \right\}$$
3. Move to θ' with prob. h , else remain at θ .
 4. Return to 1.

Similar thing when we go from $\theta'=\theta_1, \theta_2$ to $\theta=\theta_1$

MCMC Example 2: Developmental pathways

Measurement of Expression Level

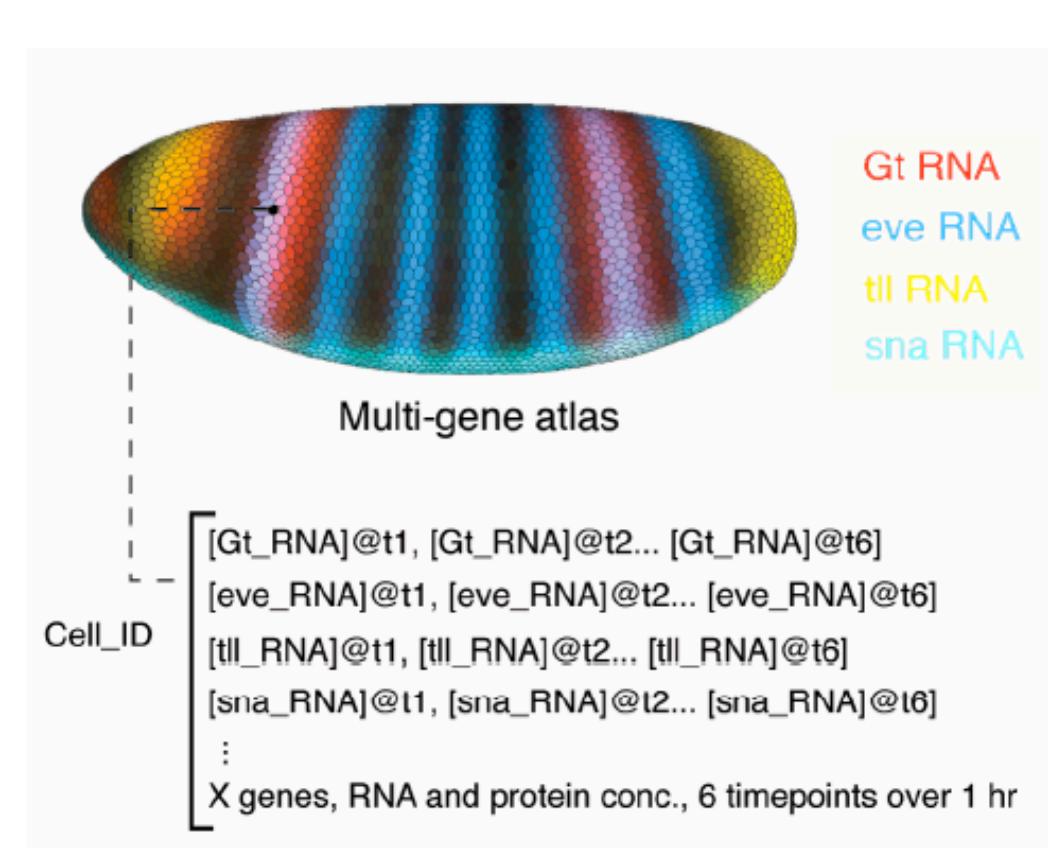
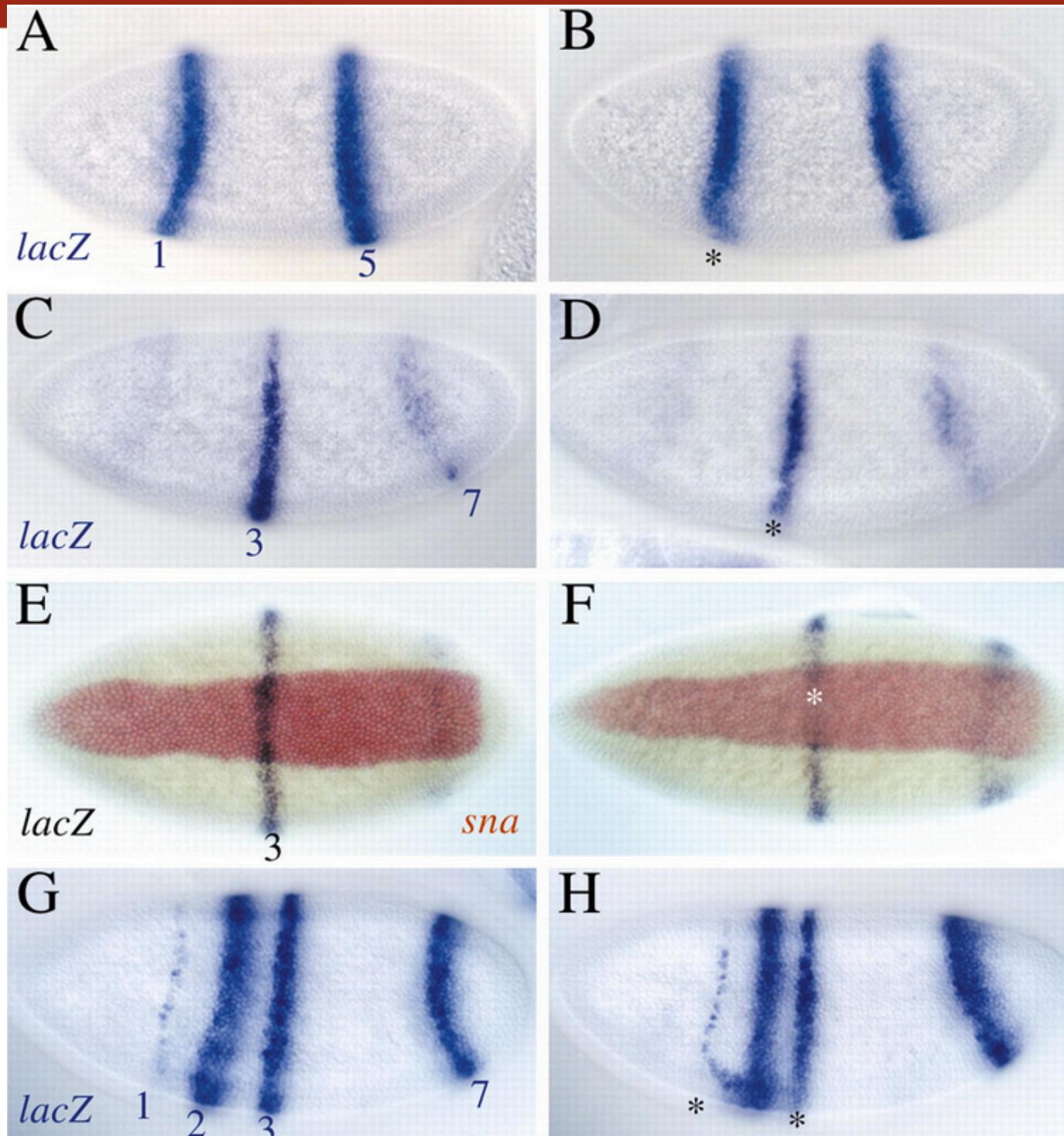
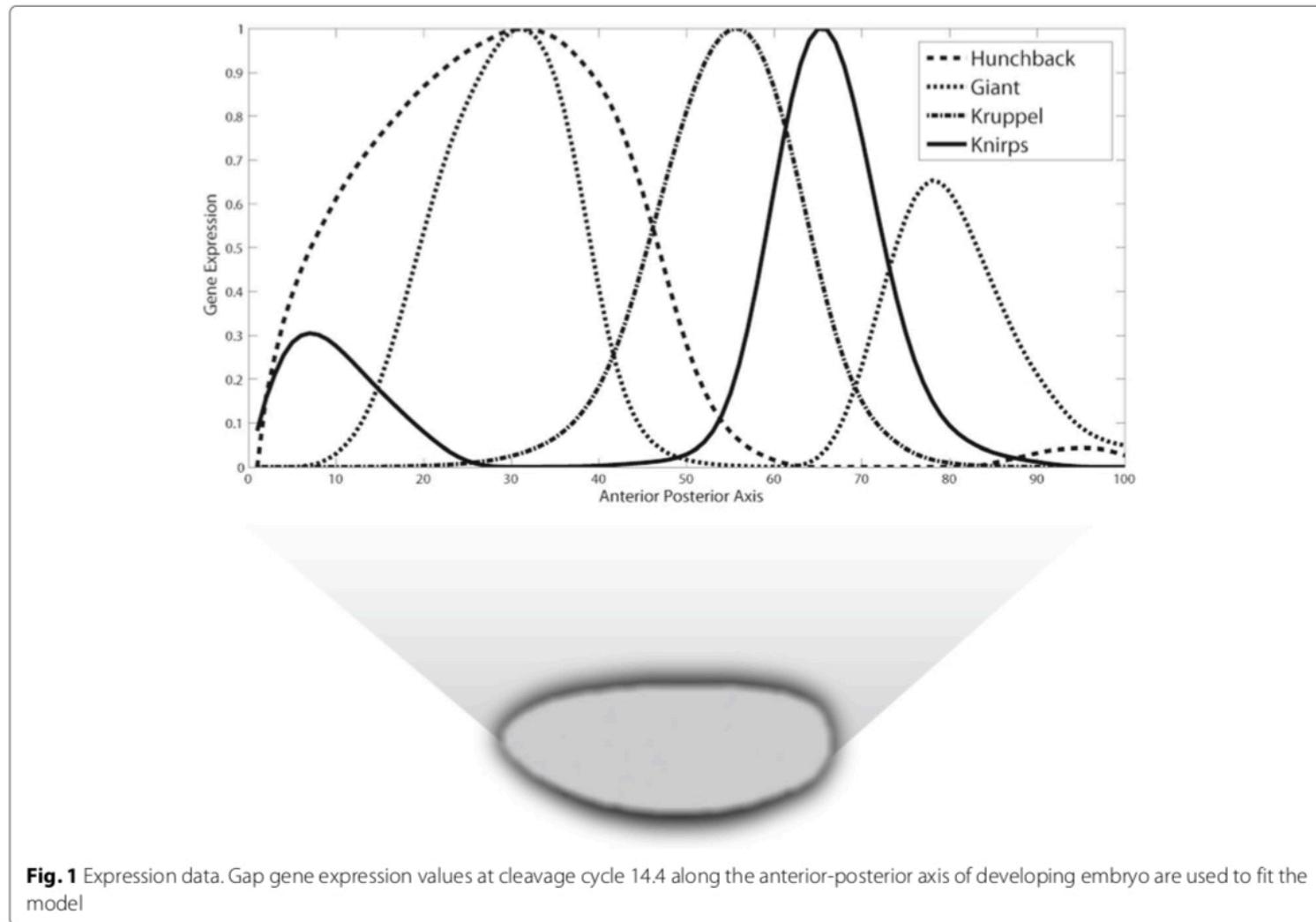


Figure 3: A single morphological framework, containing expression information for multiple genes per cell over time. Data is from *D.mel*, collected by BDTNP.



Gene expression patterns



Zubair, A., Rosen, I. G., Nuzdhn, S. V., & Marjoram, P. (2019). Bayesian model selection for the Drosophila gap gene network. *BMC Bioinformatics*, 20(1), 327.

Developmental pathway

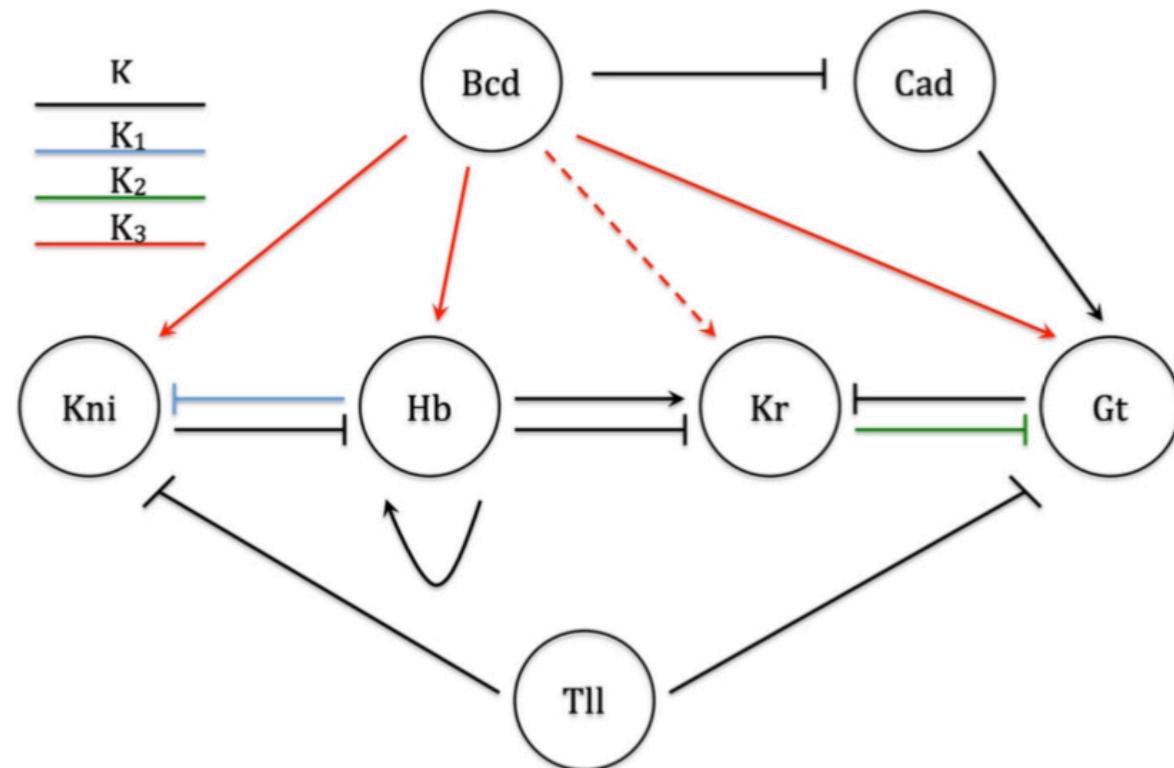


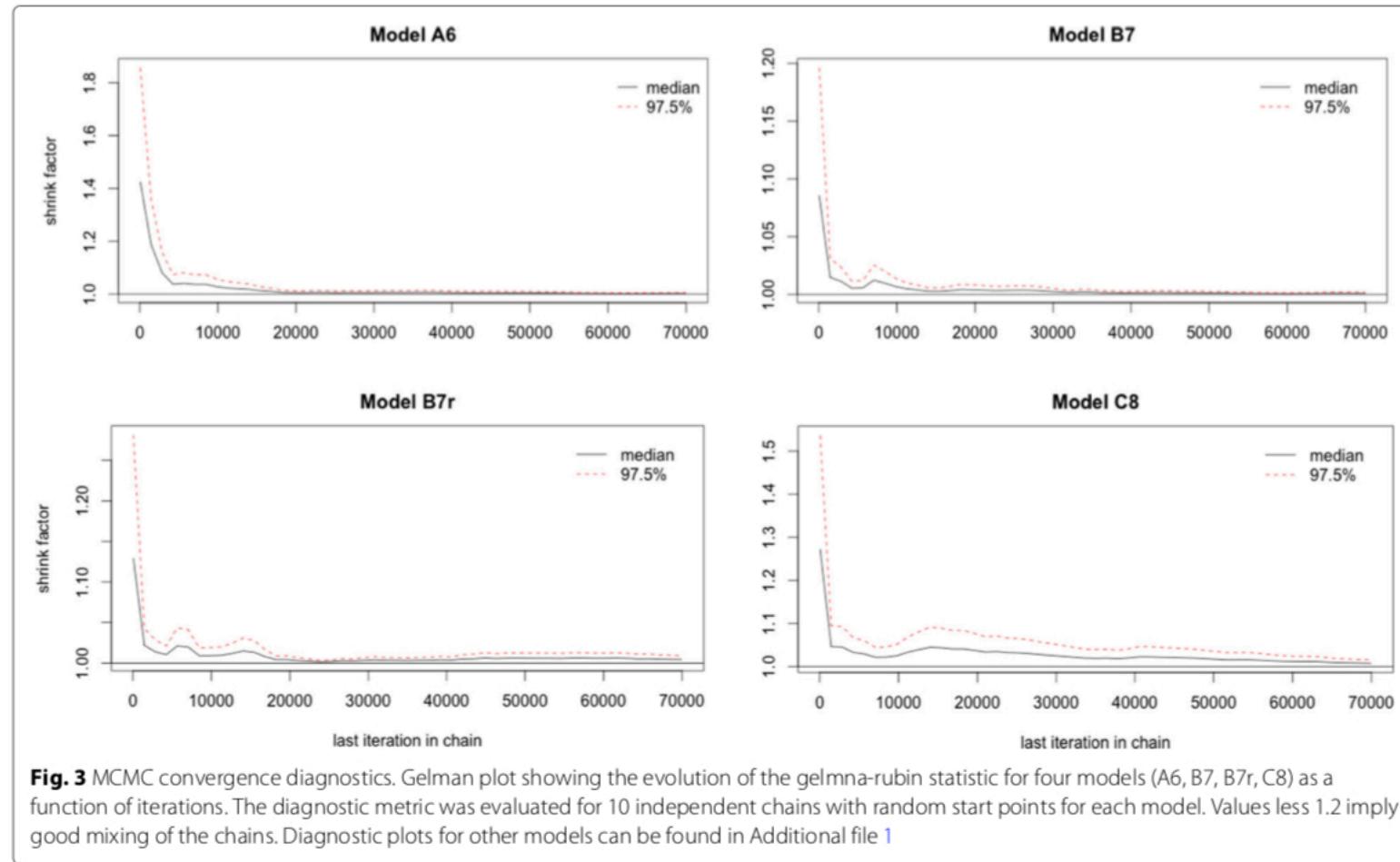
Fig. 2 Gap gene network. Gap gene network showing regulatory interactions between maternal genes, Bicoid (Bcd) & Caudal (Cad), and gap genes (Knirps (Kni), Hunchback (Hb), Kruppel (Kr), Giant (Gt)). Two types of binding affinity parameters are shown - global (K) and edge-specific (K_1, K_2, K_3). We also investigate evidence for Bicoid activation of Kruppel (shown as dashed arrow)

Represented by a set of partial differential equations

MCMC details

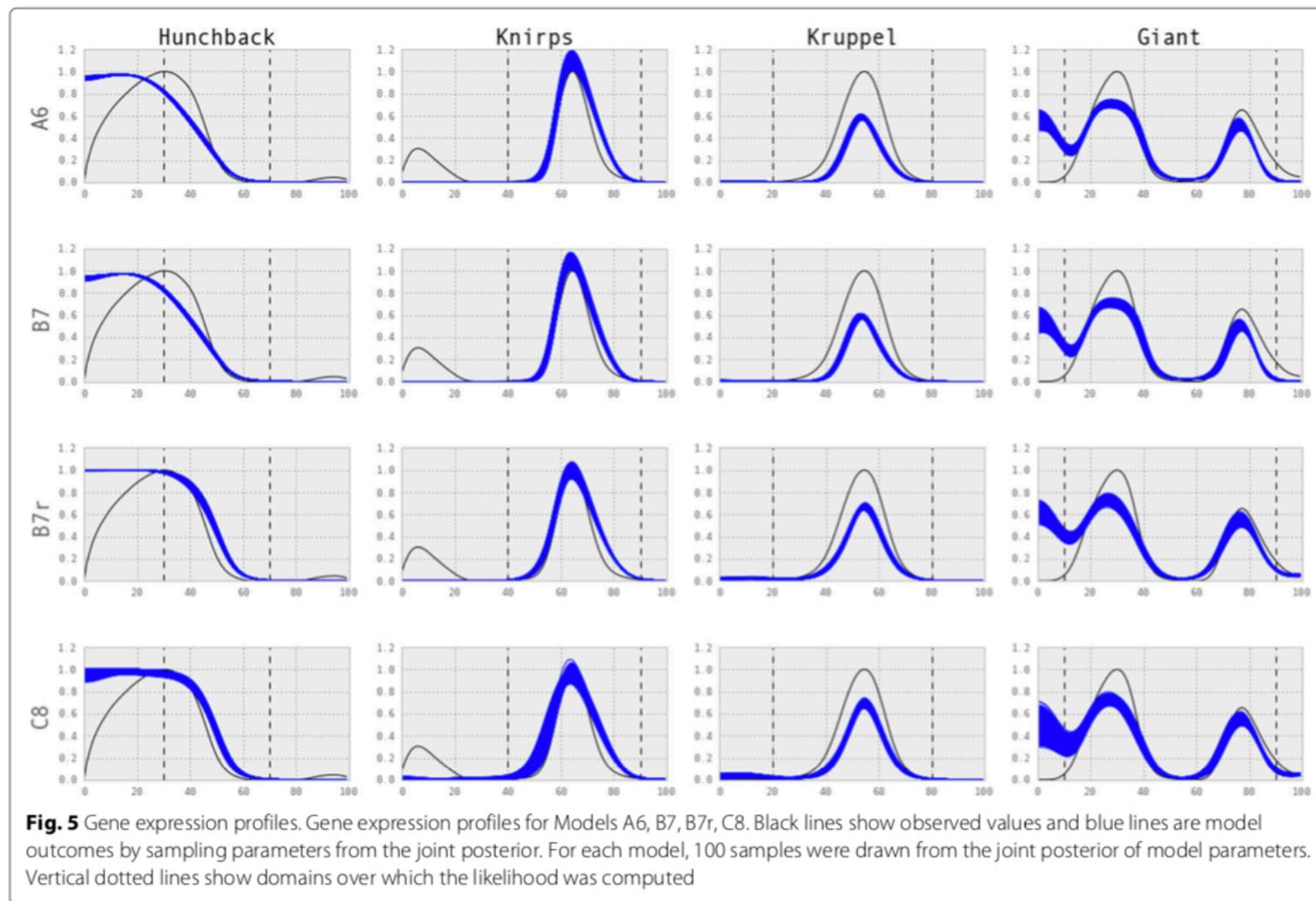
- Simplest analysis: Network topology is assumed to be fixed.
- Statespace: list of parameters. At any given iteration of the MCMC process, each parameter will have an assigned value.
- Transition kernel q will propose changes to parameters. e.g., pick a parameter at random, and add a $\text{Unif}[-0.5,0.5]$ to it. (So the q -term will cancel in the HR if we use reflecting boundaries when needed.)
- Given a set of parameter values, the differential equations tell us the expected level of the measured value of **gene expression** at each **spatial** point. Suppose, the expected value at location t for that set of parameter values is $y(t)$.
- Assume measurement error is normally deviated. So, if the measured value is denoted by $z(t)$, we assume $z(t) \sim \text{Normal}(y(t), \sigma^2)$.
- Assume measured values at each location are independent **conditional on the expected values**.
- So, the likelihood of $z=(z(1), z(2), \dots, z(n))$ is a product of Normal densities.
- Used Parallel Tempered MCMC
- More complex analysis: which network topology is best (should branches be removed/added?)

Example convergence diagnostics



Zubair, A., Rosen, I. G., Nuzdhn, S. V., & Marjoram, P. (2019). Bayesian model selection for the Drosophila gap gene network. *BMC Bioinformatics*, 20(1), 327.

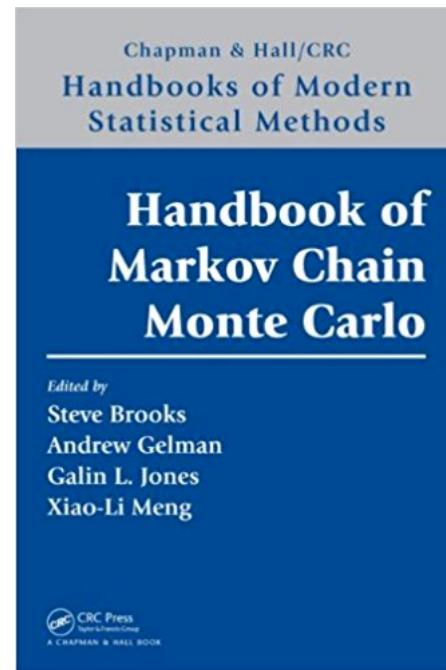
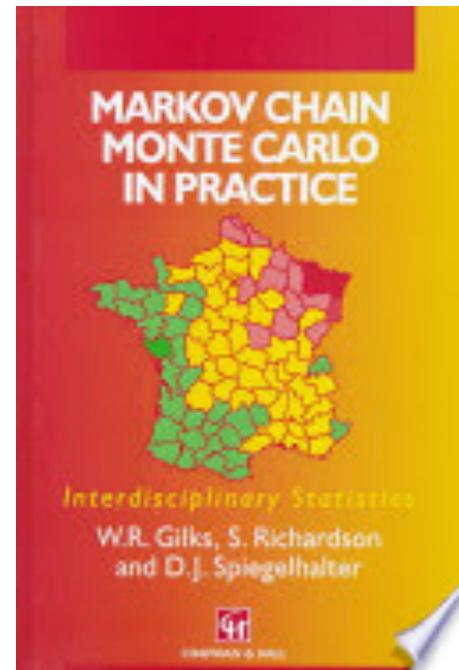
Model fit



Further reading:

Robert and Casella “Monte Carlo Statistical Methods” (Springer) for more on this and other MC methods. (A hard read!)

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2002). An Introduction to MCMC for Machine Learning. *Machine Learning*, 50, 5–43.



Suggested exercises: non-examinable

1. Adapt the example from the MCMC1.Rmd document to run several MCMC chains for that problem and check convergence using the ‘coda’ package.
2. Use the metrop function to implement any of the MCMC exercises we’ve seen so far in this course. the multivariate normal example from earlier today.
3. Try re-writing some of the exercises to use adaptive MCMC or tempering. It general ends up working much better, but at the price of more work.
4. Would either of these methods work for the gamma rv question in Assignment 3?

END