Something on Bayes and sienaBayes

Johan Koskinen

Joint work with: T.A.B. Snijders, (RsienaTest: `sienaBayes`)

Sep, 2018, Groningen

Content

- Bayes v MoM
- Intro to Bayes
- Bayes + SAOM = HSAOM, more detail
- Something on priors
- ‘convergence’
- Helpful tips

To Bayes or not to Bayes - difference in estimation

✓ *Bayes rules*

. . .

- For MoM Phase 2 tries to optimise θ so that the model matches data
- For MoM the accuracy of matching is checked in Phase 3
- Principle is the same for MLE
- For Bayes target is to generate a range of probable values of θ
  - ✓ are we drawing from the right distribution?
  - ✓ are our summaries (mean, std, etc) good reflections of distribution?

To Bayes or not to Bayes - why are MLE and Bayes so hard

✓ *Bayes rules*

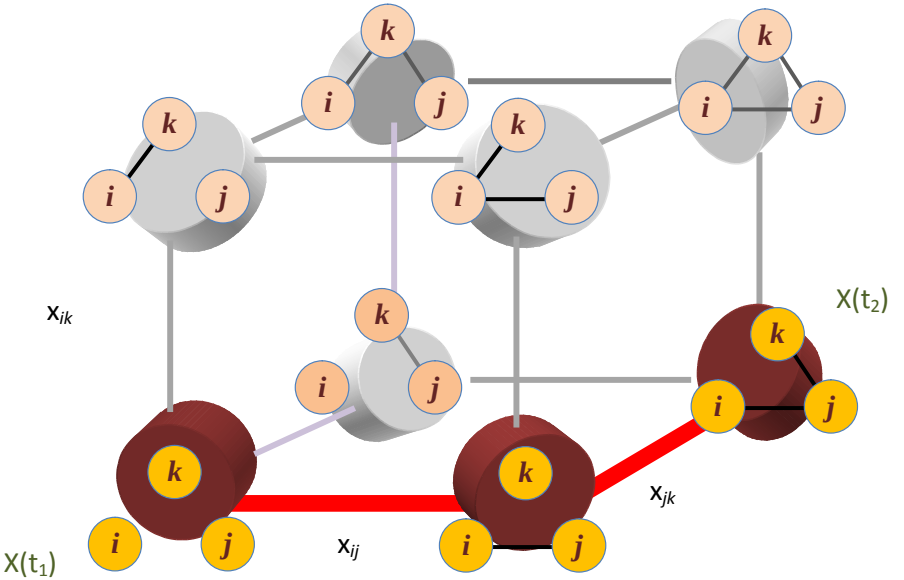
...

- MoM simulates a network $x_{\text{obs}}(t_0) \rightarrow X(t_1)$
  - ✓ this is simulating the SAOM
- MLE and Bayes simulate paths $x_{\text{obs}}(t_0) \rightarrow x_{\text{obs}}(t_1)$
  - ✓ this is NOT simulating the SAOM

likelihood-based

chain must start in $x_{\text{obs}}(t_1)$

and end in $x_{\text{obs}}(t_2)$



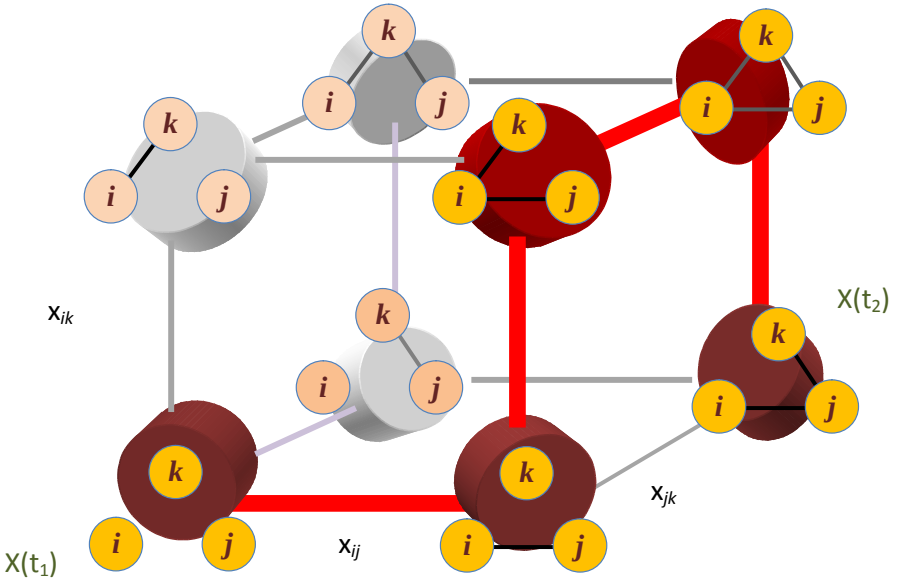
MoM starts in $x_{\text{obs}}(t_1)$

. . . and can end up in any state

likelihood-based

chain must start in $x_{\text{obs}}(t_1)$

and end in $x_{\text{obs}}(t_2)$



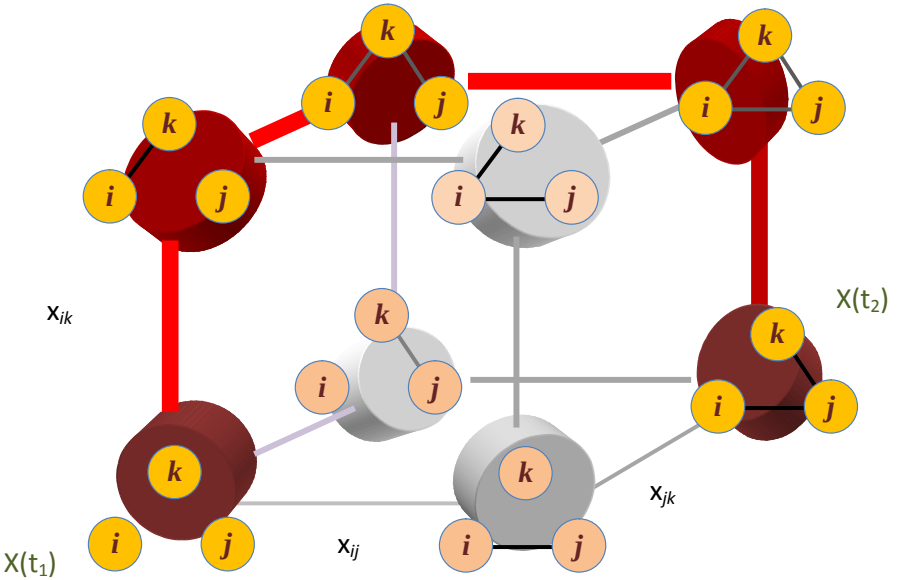
MoM starts in $x_{\text{obs}}(t_1)$

. . . and can end up in any state

likelihood-based

chain must start in $x_{\text{obs}}(t_1)$

and end in $x_{\text{obs}}(t_2)$



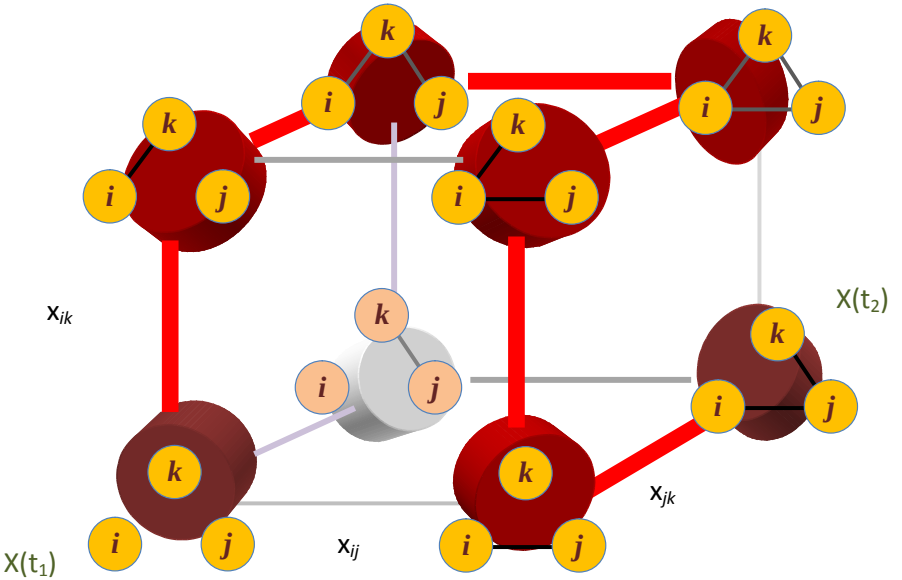
MoM starts in $x_{\text{obs}}(t_1)$

. . . and can end up in any state

likelihood-based

chain must start in $x_{\text{obs}}(t_1)$

and end in $x_{\text{obs}}(t_2)$



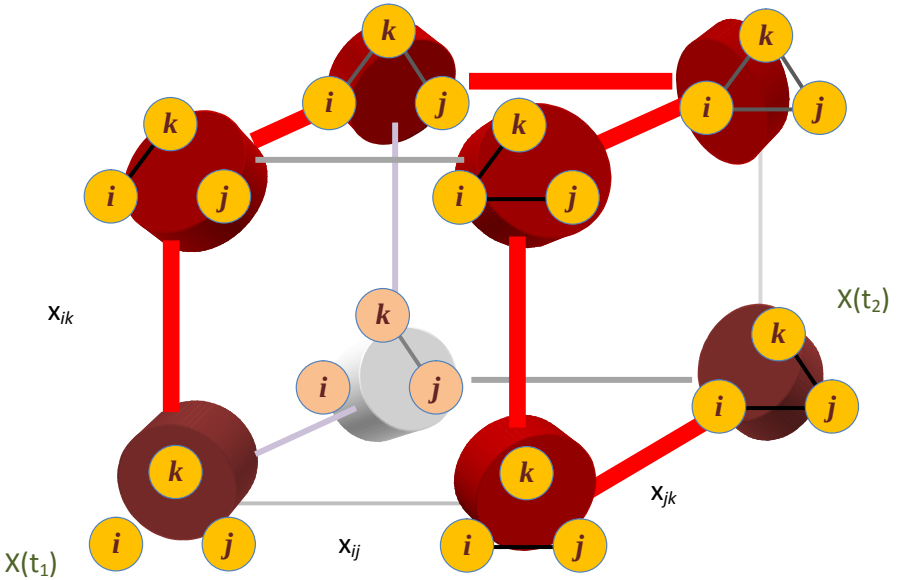
MoM starts in $x_{\text{obs}}(t_1)$

. . . and can end up in any state

likelihood-based

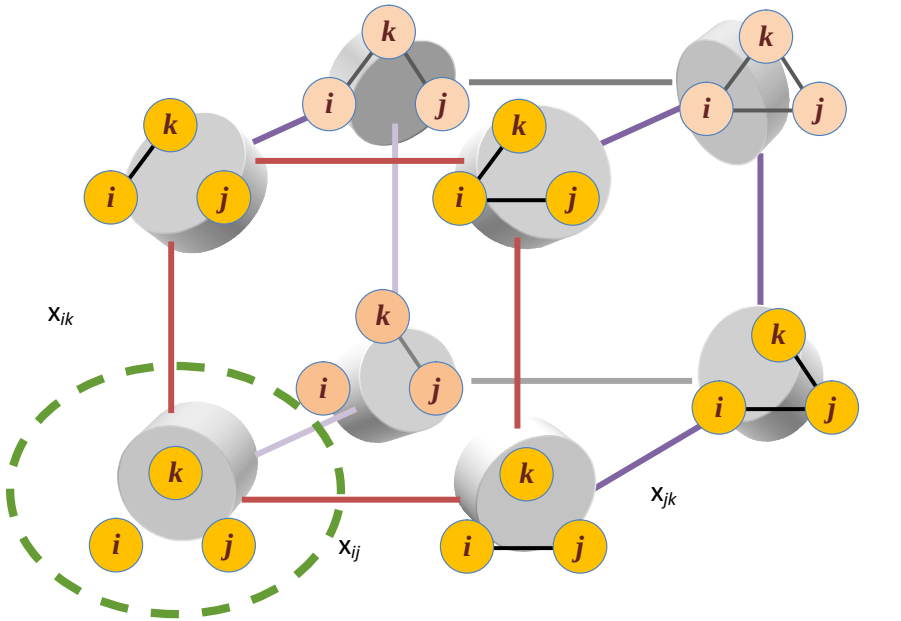
chain must start in $x_{\text{obs}}(t_1)$

and end in $x_{\text{obs}}(t_2)$



MoM starts in $x_{\text{obs}}(t_1)$

. . . and can end up in any state



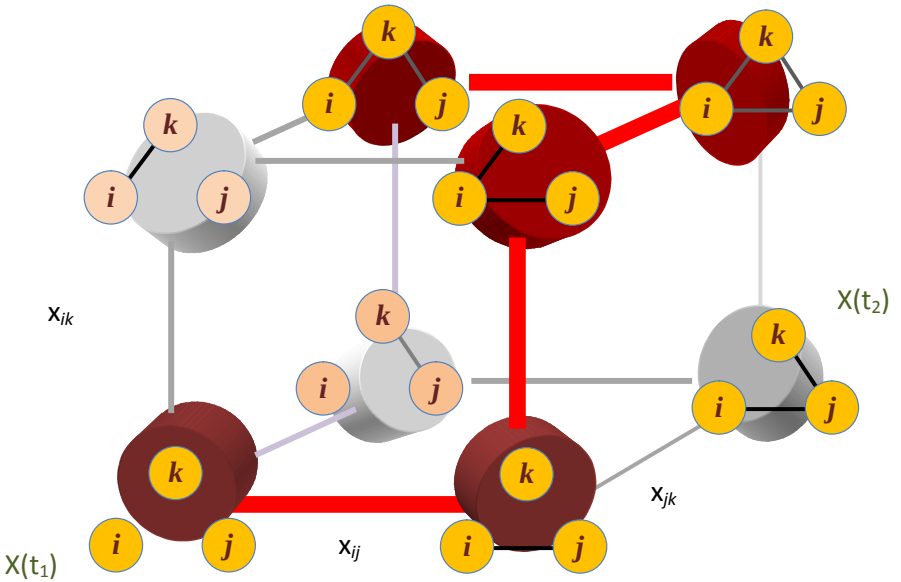
likelihood-based

chain must start in $x_{\text{obs}}(t_1)$

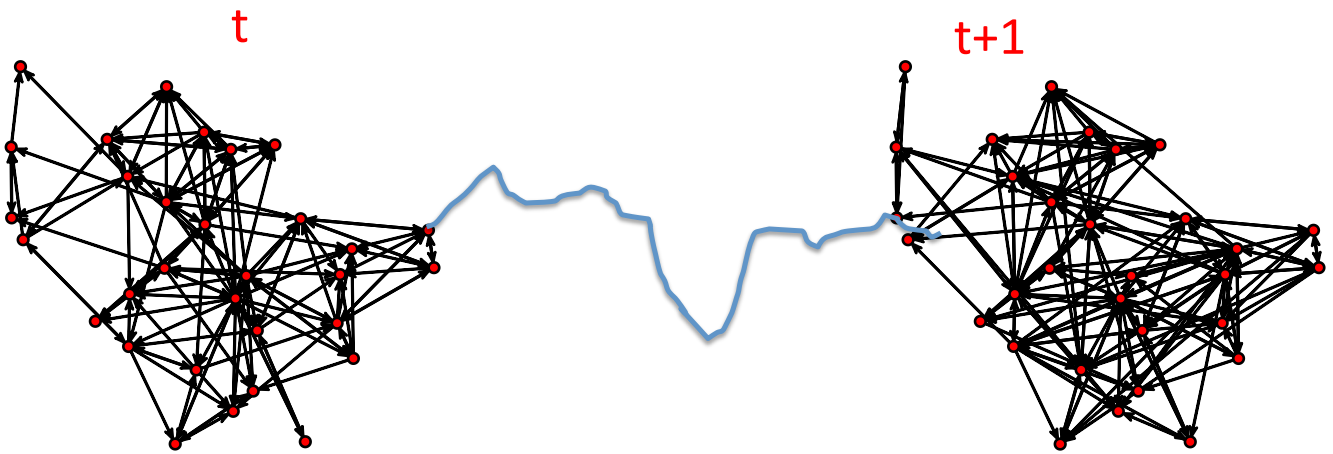
and end in $x_{\text{obs}}(t_2)$

MoM starts in $x_{\text{obs}}(t_1)$

. . . and can end up in any state

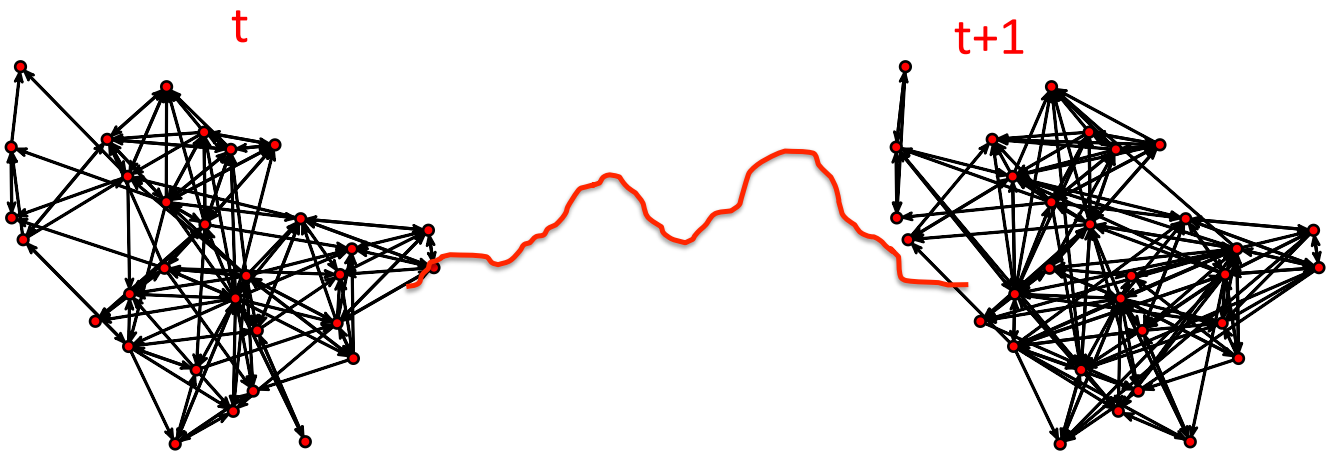


Likelihood-based



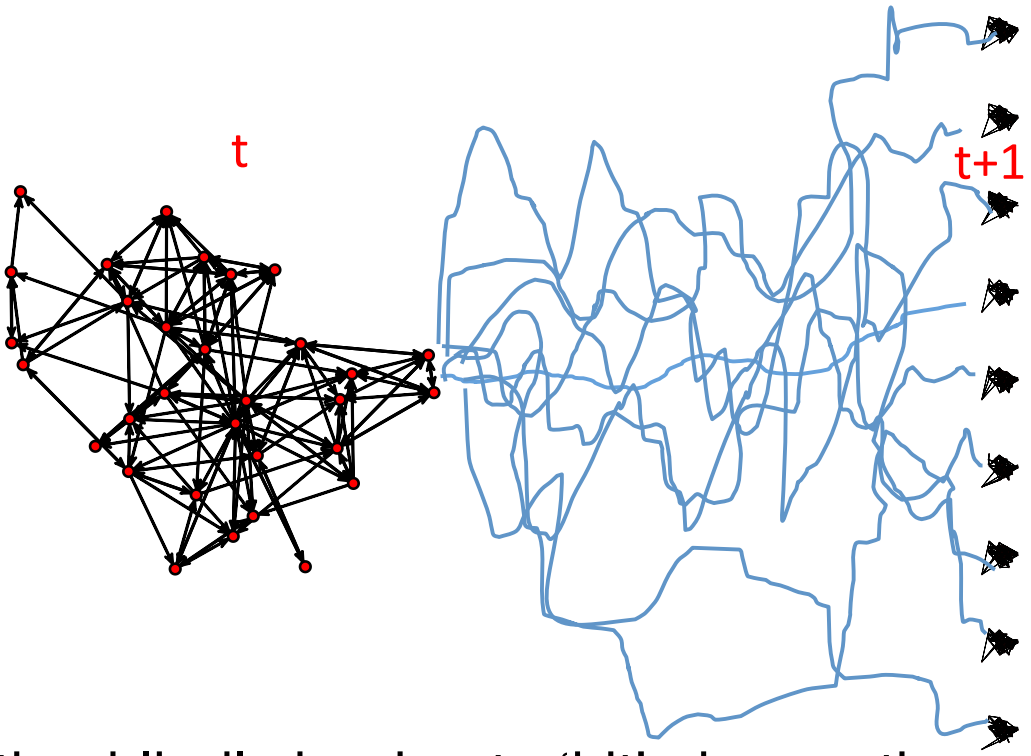
generating likely paths connecting observations

Likelihood-based



generating likely paths connecting observations

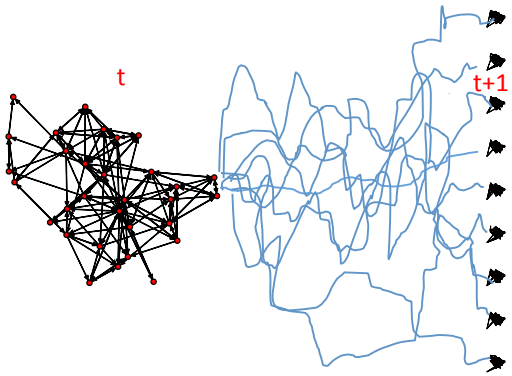
MoM

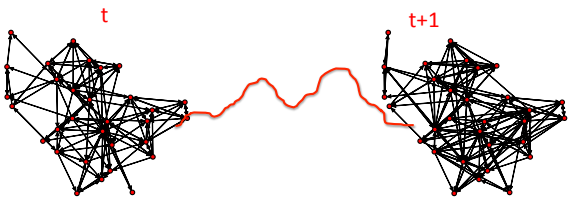


simulating blindly hoping to ‘hit’ observation

MoM/GMM v. MLE/Bayes

blah





- ✓ robust
- ✓ fast (forward simulation)
- × unknown properties

- × only as good as its MCMC
- × conditional draws slow
- ✓ efficient (uses likelihood)
- ✓ rich inference (particular Bayes)

Bayes

The principle of Bayesian inference and Bayes' theorem

A parameter indexes a family of models

P(G|A) = 0 P(G|B) = 2/5 P(G|C) = 4/5

A B C

Parameters represent different states of the world

proportion in population that supports Trump

Trump->no one?

Trump->two in five?

Trump->four in five?

proportion in population that supports Trump

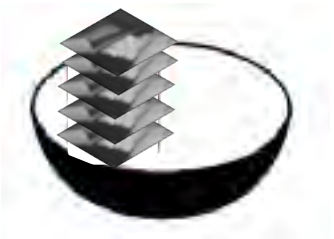
Trump->no one?

Trump->everyone?

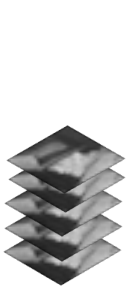
Disposable income in the US

Income distribution in the population

Disposable income in the US



Disposable income in the US



Likelihood ➔ Data

different for different states of the world

p(Data|A)

P(G|A) = 0 P(G|B) = 2/5 P(G|C) = 4/5

A B C

states of the world: A, B, or C

If state of the world is either A or A^c (not A)
and we know $P(Data \mid A)$, $P(Data \mid A^c)$, and $P(A)$
We have

Probability of Data and A

Derived from definition of 'conditional probability'

$$P(A|Data) = \frac{P(Data \cap A)}{P(Data)} = \frac{P(Data|A)P(A)}{P(Data|A)P(A)+P(Data|A^c)(1-P(A))}$$

Probability of Data regardeless of A or A^c

Law of total probability

Simple example

Consider a Bernoulli graph G where $X_{ij} \overset{iid}{\sim} Bern(\theta)$:

$$P(G) = \theta^L (1-\theta)^{M-L}$$

where $L = \sum x_{ij}$ and $M = n(n-1)/2$.

Posterior:

$$\pi(\theta|G) \propto \theta^L (1-\theta)^{M-L} \pi(\theta)$$

Whye 'Simple' example

Consider a Bernoulli graph G where $X_{ij} \overset{iid}{\sim} Bern(\theta)$:

$$P(G) = \theta^{L}(1-\theta)^{M-L}$$

because:

$$\Pr(G) = \Pr(X_{12} = x_{12}, \ldots, X_{n,n-1} = x_{n-1})$$

and independence:

$$\Pr(X_{12} = x_{12}, \ldots, X_{n,n-1} = x_{n-1}) = \Pr(X_{12} = x_{12}) \times \cdots \times \Pr(X_{n,n-1} = x_{n-1})$$

and $\Pr(X_{ij} = x_{ij}) = \theta^{x_{ij}}(1-\theta)^{1-x_{ij}}$. Hence
$\prod \Pr(X_{ij} = x_{ij}) = \theta^{L}(1-\theta)^{M-L}$

With prior:

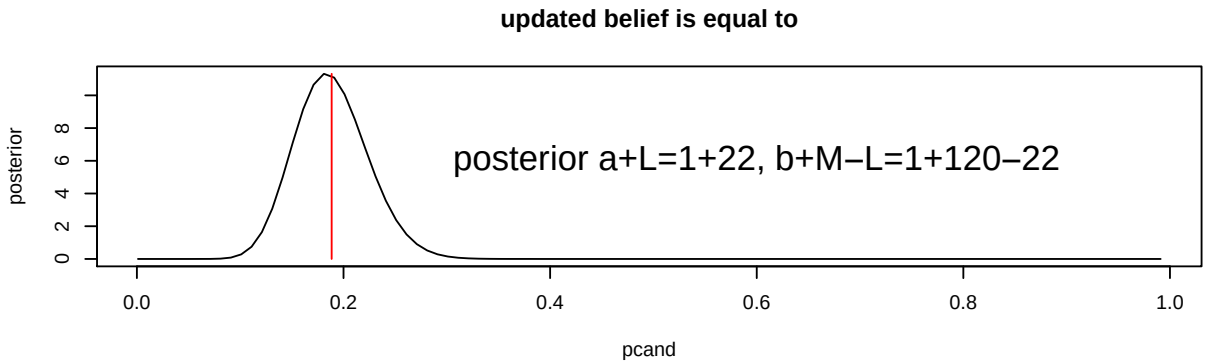
$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

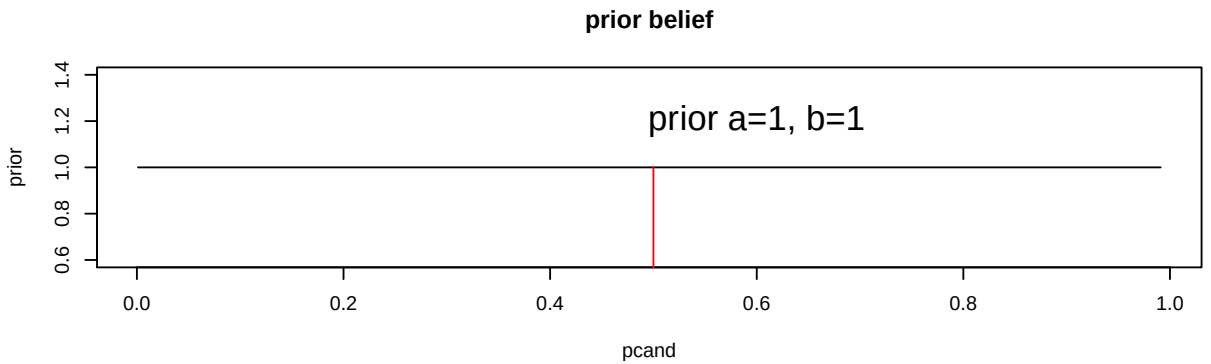
Posterior:

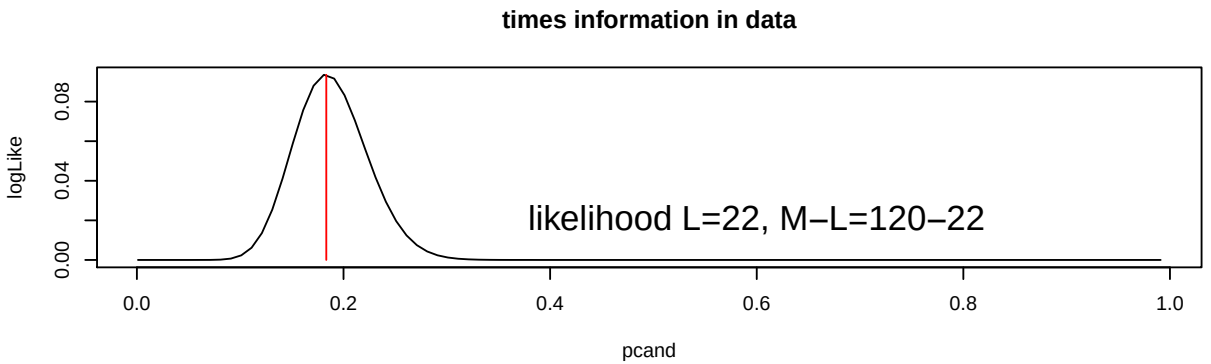
$$\pi(\theta|G) \propto \theta^{L+\alpha-1}(1-\theta)^{\beta-1+M-L}$$

α and β hypothetical prior 'observations'

IntroToBayes.R







What about the fish?

The symbol $\propto$ reads 'i proportional to':

$$\pi(\theta|c) = c \times \theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

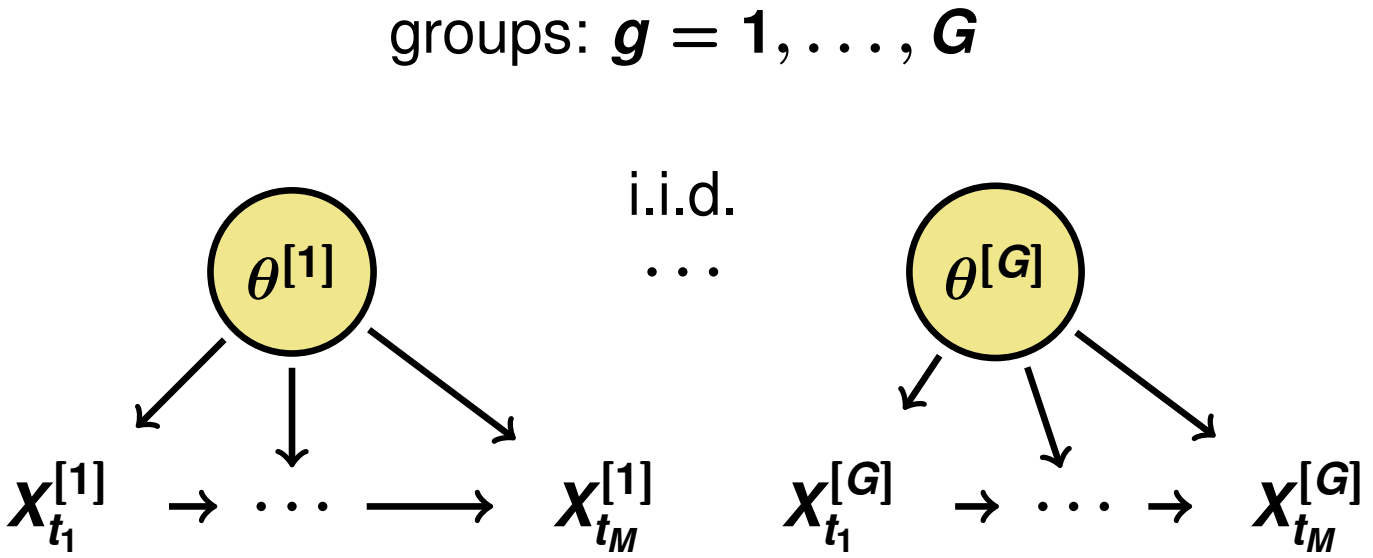
means that

$$\frac{\pi(\theta|c)}{\pi(\theta^*|c)} = \frac{\pi(\theta|c^*)}{\pi(\theta^*|c^*)}$$

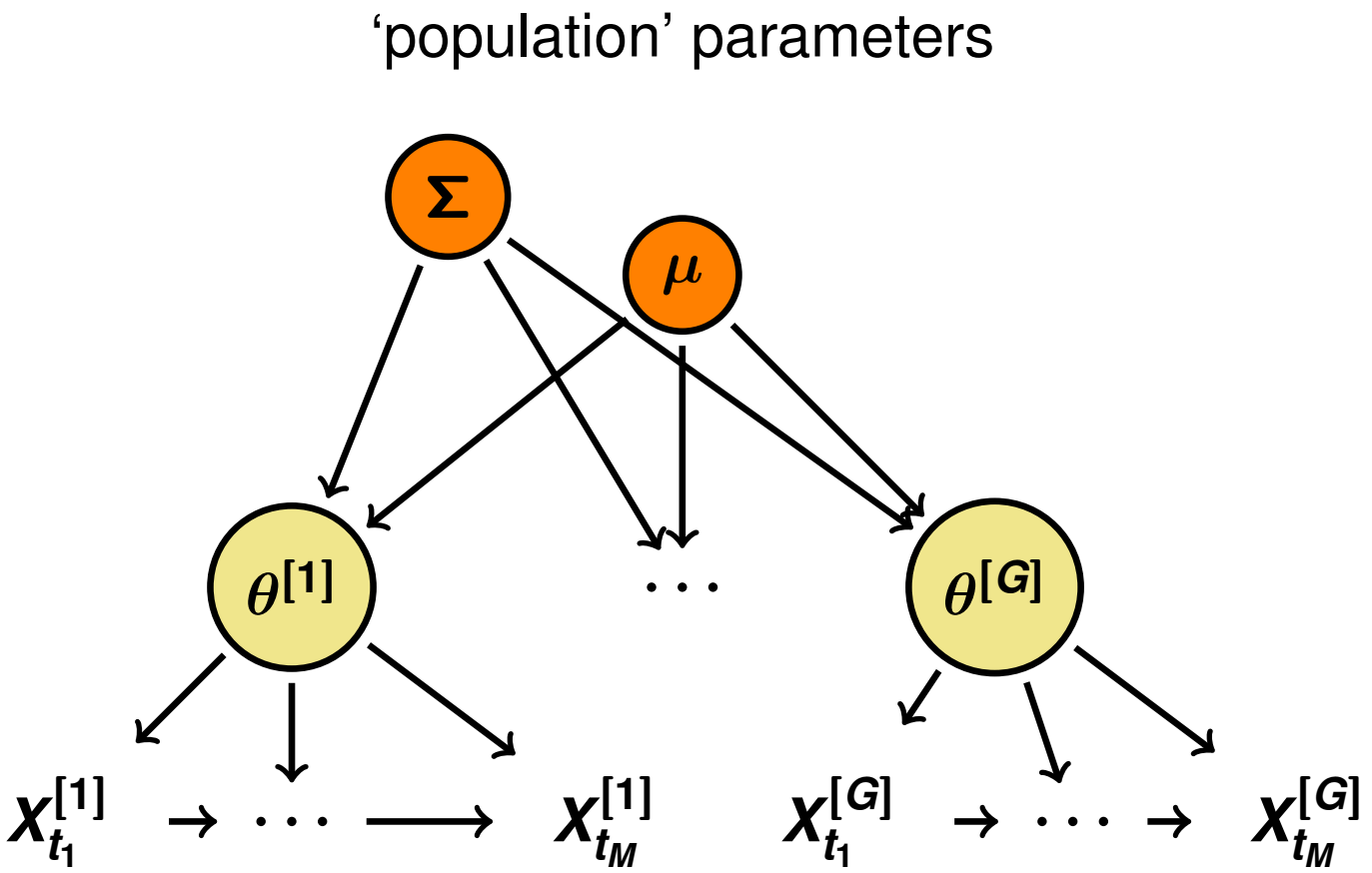
for any choice of constants c and c^* - the constant does not inform us of the 'shape' of the curve

Hierarchical SAOM Defining the conditional hierarchical dependencies

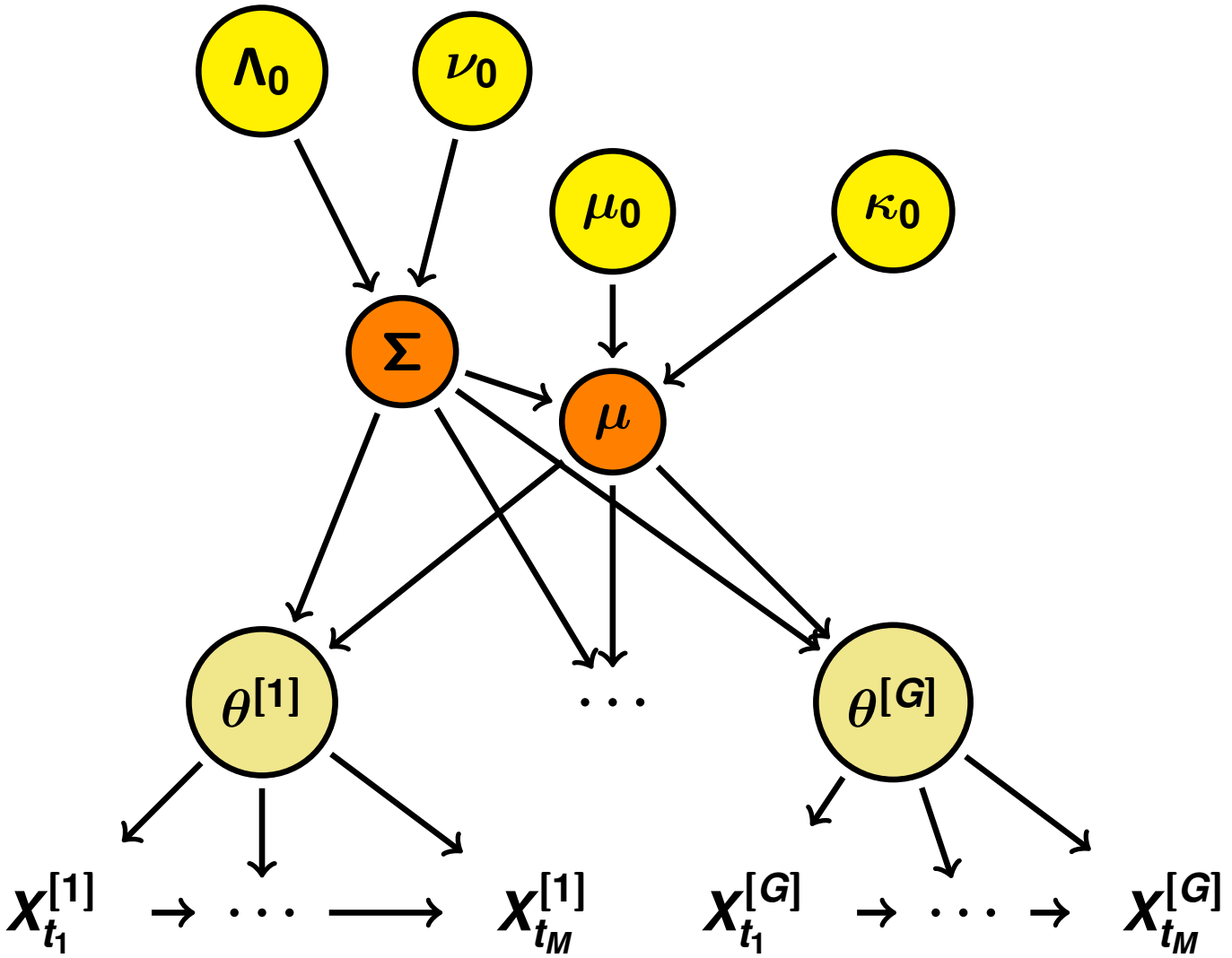
HSAOM (DAG)


groups: $g = 1, \ldots, G$
i.i.d.
$\ldots$
$\theta^{[1]}$
$\theta^{[G]}$
$X_{t_1}^{[1]} \rightarrow \cdots \longrightarrow X_{t_M}^{[1]}$
$X_{t_1}^{[G]} \rightarrow \cdots \rightarrow X_{t_M}^{[G]}$

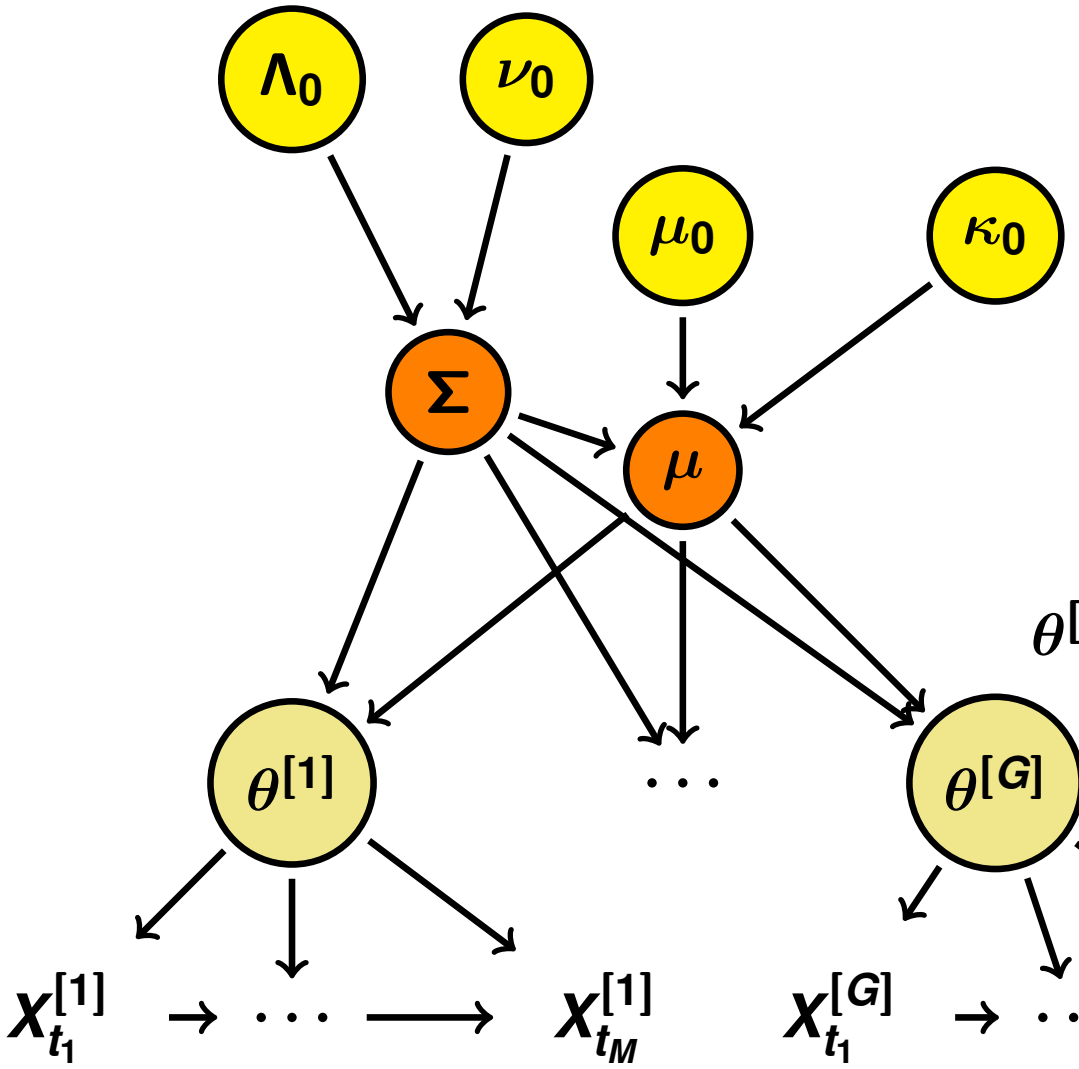
HSAOM (DAG)



HSAOM (DAG)


Λ_0
ν_0
μ_0
κ_0
Σ
μ
$\theta^{[1]}$
$\cdots$
$\theta^{[G]}$
$X_{t_1}^{[1]} \rightarrow \cdots \longrightarrow X_{t_M}^{[1]}$
$X_{t_1}^{[G]} \rightarrow \cdots \rightarrow X_{t_M}^{[G]}$

HSAOM (DAG)



$$\Sigma \sim \mathit{InvWish}_{p_1}(\Lambda_0^{-1}, \nu_0)$$

$$\mu \sim \mathcal{N}(\mu_0, \Sigma/\kappa_0)$$

$$\theta^{[g]} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$$

$$(X_t^{[g]}) \overset{iid}{\sim} \mathit{SAOM}(\theta^{[g]})$$

To Bayes or not to Bayes

✓ *Bayes rules*

. . .

- MoM and MLE are algorithms for producing 2 numbers (point est & s.e.) for each parameter
- Once convergence, MoM or Likelihood equation satisfied
- Bayes wants to give you access to the distribution of parameters implied by your data
- . . . we can only simulate from this distribution of parameters

INFLUENCE OF PRIORS ON μ

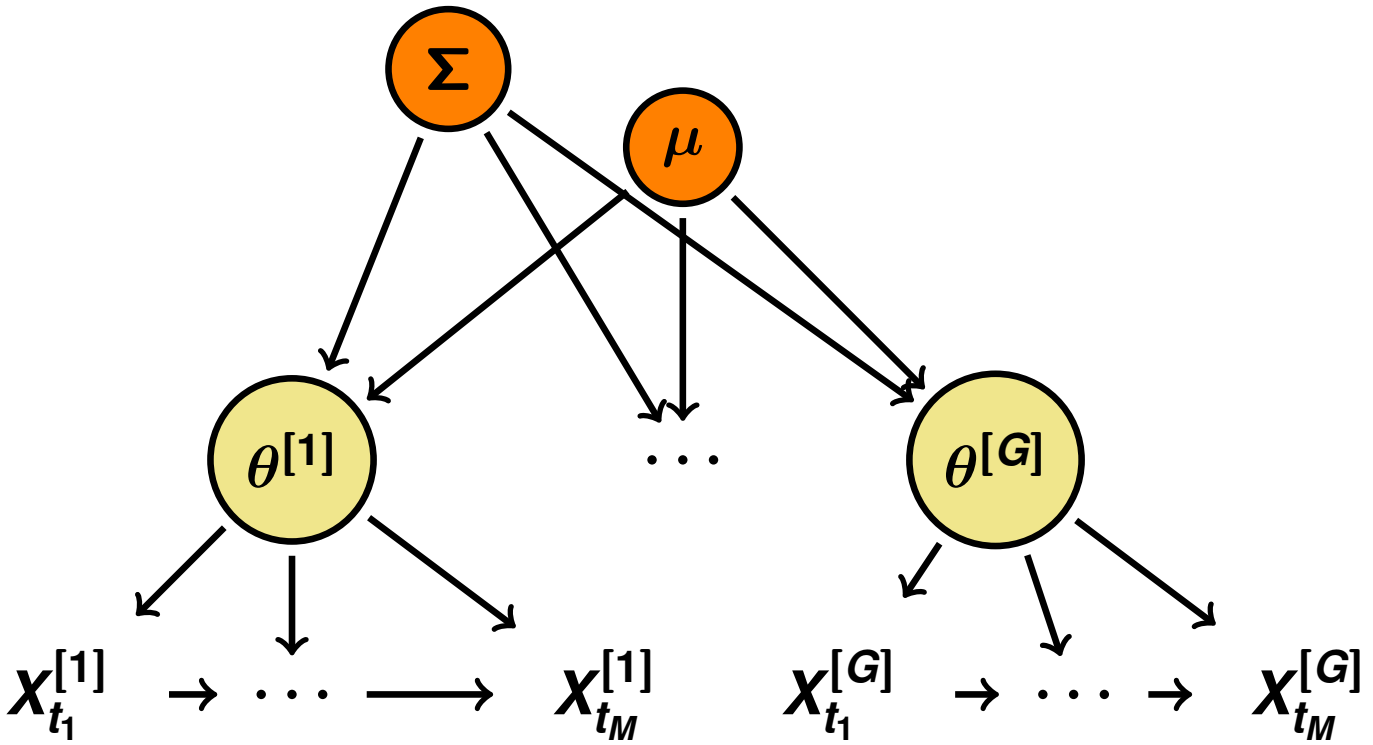
GIGO (garbage in garbage out)

anyone worried about information about θ v μ?

What information do have for population parameters?

you could argue that with little prior information

$$\mu \approx \frac{1}{G} \sum_{g=1}^{G} \theta^{[g]}$$



de Finetti nel 1928

De Finetti, (1970) afferma: *La probabilità non esiste*

While you could argue that with little prior information

$$\mu \approx \frac{1}{G} \sum_{g=1}^{G} \theta^{[g]}$$

default prior

$$\mu \sim \mathcal{N}(\mu_0, \Sigma/\kappa_0)$$

is conditional on Σ (ask me why), for which

$$\Sigma \sim \mathit{InvWish}_{p_1}(\Lambda_0^{-1}, \nu_0)$$

As for $\Sigma \sim \mathit{InvWish}_p(\Lambda^{-1}, \nu)$

$$\mathrm{E}\{\Sigma\} = \frac{1}{\nu - p - 1}\Lambda$$

the posterior is only proper if $\nu_0 + G > p + 1$

But surely we can chose Λ_0 wisely?

Non-informative prior

– Simpson, *A popular prior for Σ is the inverse-Wishart distribution, but there are some problems . . . using the standard "noninformative" version of the inverse-Wishart prior, which makes the marginal distribution of the correlations uniform, large standard deviations are associated with large absolute correlations. This isn't exactly noninformative . . .*

`http://www.themattsimpson.com/2012/08/20/`

`prior-distributions-for-covariance-matrices-the-scaled-inverse-wishart-prior/`

In `sienaBayes`

$\Lambda_0 = \sigma_0^2 I$

let's see how prior uncertainty propagates to posterior uncertainty (21 Dutch school classes)

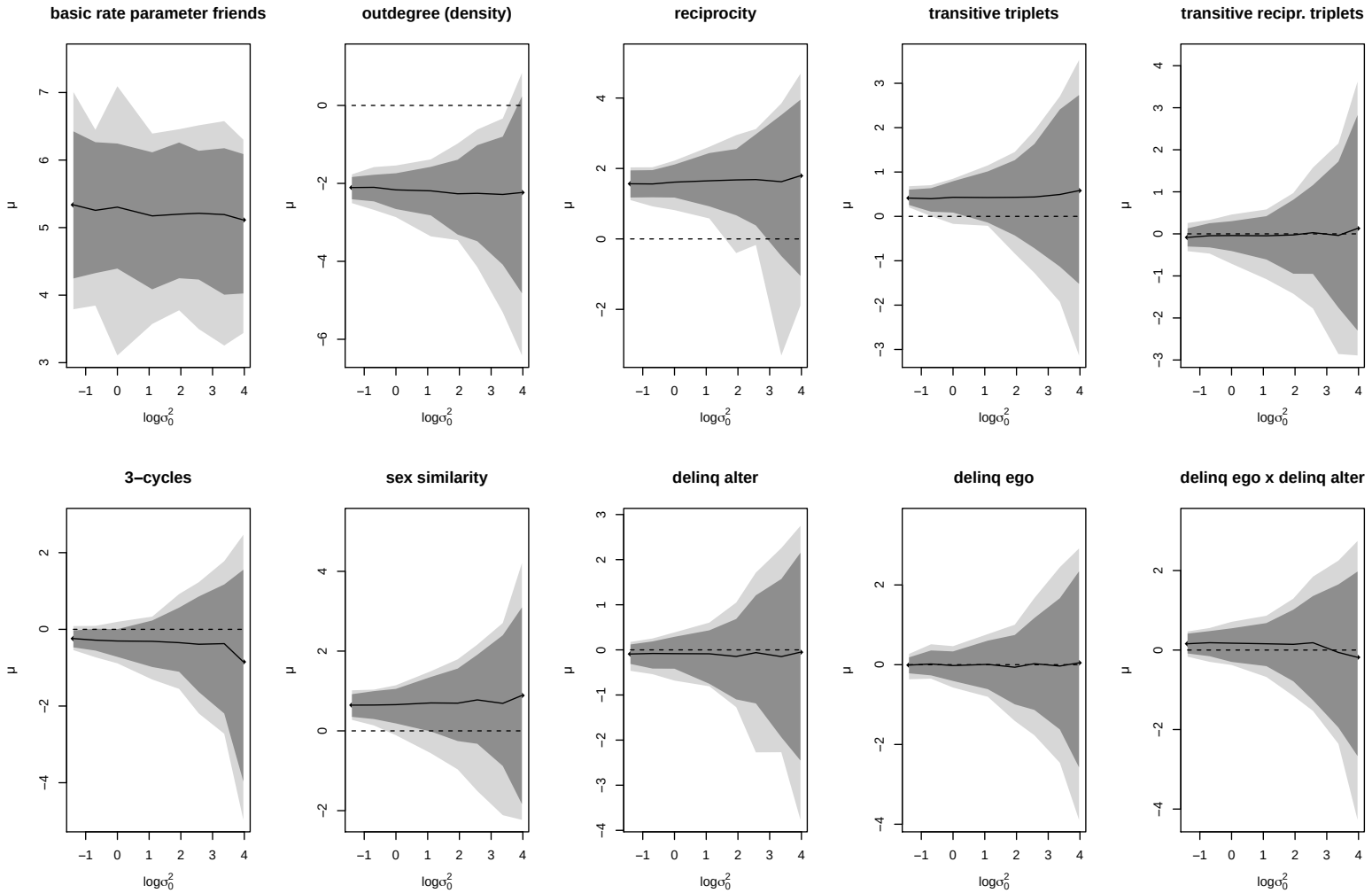


Figure : Ints. 95% /99% for μ in dark/light grey w. default priors $\nu_0 = 12$, and $\Lambda_0 = \sigma_0^2 I$, for different values of σ_0^2

So, large prior uncertainty $\Rightarrow$ large posterior uncertainty

about μ

does it affect inference about $\theta^{[1]}, \ldots, \theta^{[G]}$?

So, large prior uncertainty $\Rightarrow$ large posterior uncertainty about μ

does it affect inference about $\theta^{[1]}, \ldots, \theta^{[G]}$?

So, large prior uncertainty $\Rightarrow$ large posterior uncertainty about μ

does it affect inference about $\theta^{[1]}, \ldots, \theta^{[G]}$?

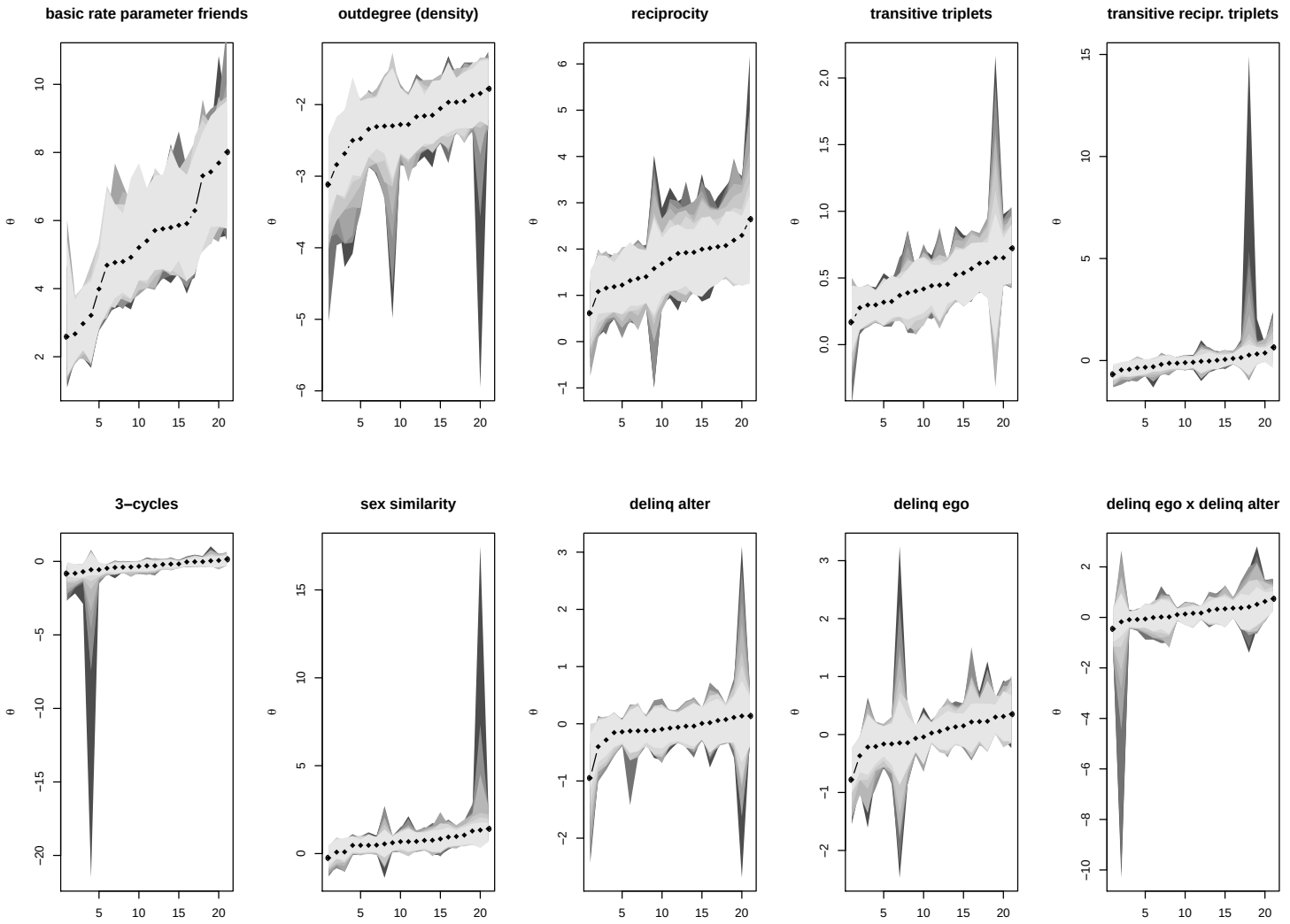


Figure : Post. pred. for $\theta^{[g]}$. Equal tail 95% from $\sigma_0^2 = 1/4$ to $\sigma_0^2 = 113$, light grey to dark grey. Groups ordered according to posterior predictive mean

In our case prior for Λ_0 contributes 'little information' if distributions for $\theta^{[1]}, \ldots, \theta^{[G]}$ under HSAOM are 'similar' to distributions for $\theta^{[1]}, \ldots, \theta^{[G]}$ under independent SAOMs with

$$\pi(\theta^{[g]}|x^{[g]}) \propto p_{\mathrm{SAOM}}(x^{[g]} \mid \theta^{[g]})\pi(\theta^{[g]})$$

and $\pi(\theta^{[g]}) \propto c$

Do we:

- fit independent SAOMS, and
- search for σ_0^2 that match?

Jeffrey's prior

The Normal-Inverse-Wishart Jeffrey's prior implies using (Gelman et al., 2004:88)

$$p(\mu, \Sigma) \propto |\Sigma|^{-(p_1+1)/2}$$

For the conjugate model this corresponds to $\kappa_0 \to 0$, $\nu_0 \to 0$, and letting the determinant of Λ_0 tend to 0.
Reference analysis:

- try Jeffrey's
- if it works, you know HSAOM and independent SAOMs close
- if it fails, you know that you *need* HSAOM

Reference analysis for Andrea's 21 Dutch schools

Result:

fail

Diagnose using independent SAOM

Reference analysis for Andrea's 21 Dutch schools

Result:

fail

Diagnose using independent SAOM

Reference analysis for Andrea's 21 Dutch schools

Result:

fail

Diagnose using independent SAOM

Reference analysis for Andrea's 21 Dutch schools

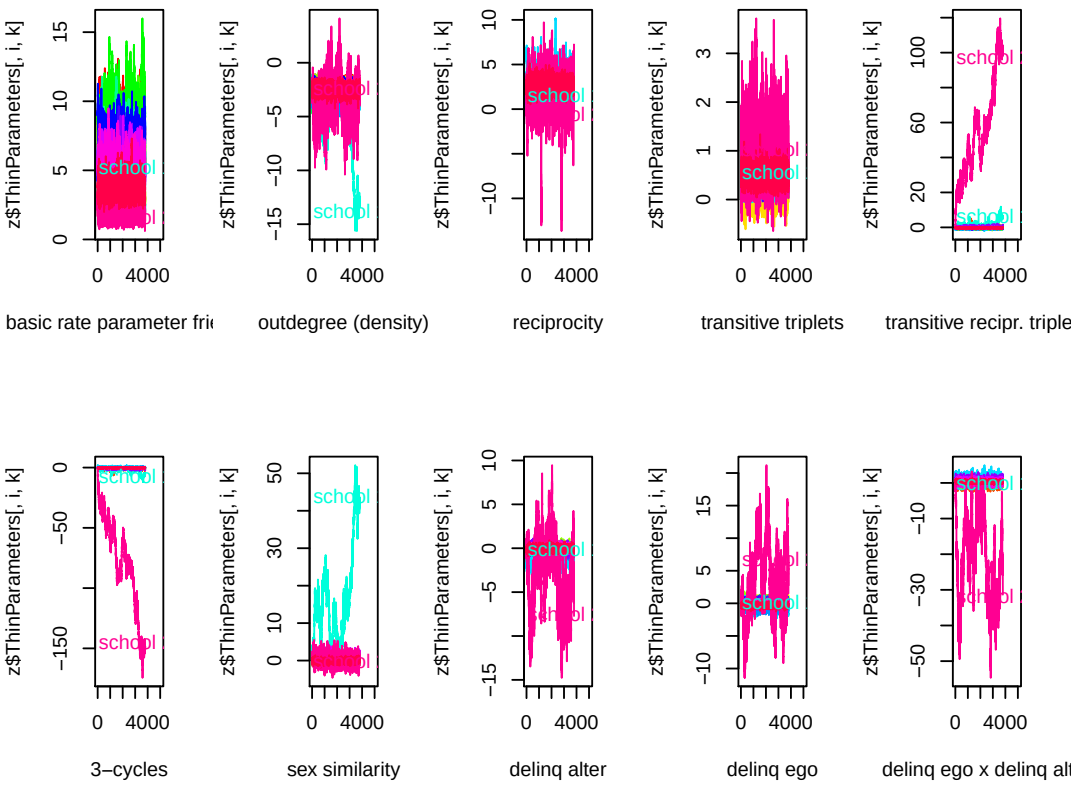


Figure : Posteriors for $\theta^{[g]}$ independently fitted using $\pi(\theta^{[g]}) \propto c$

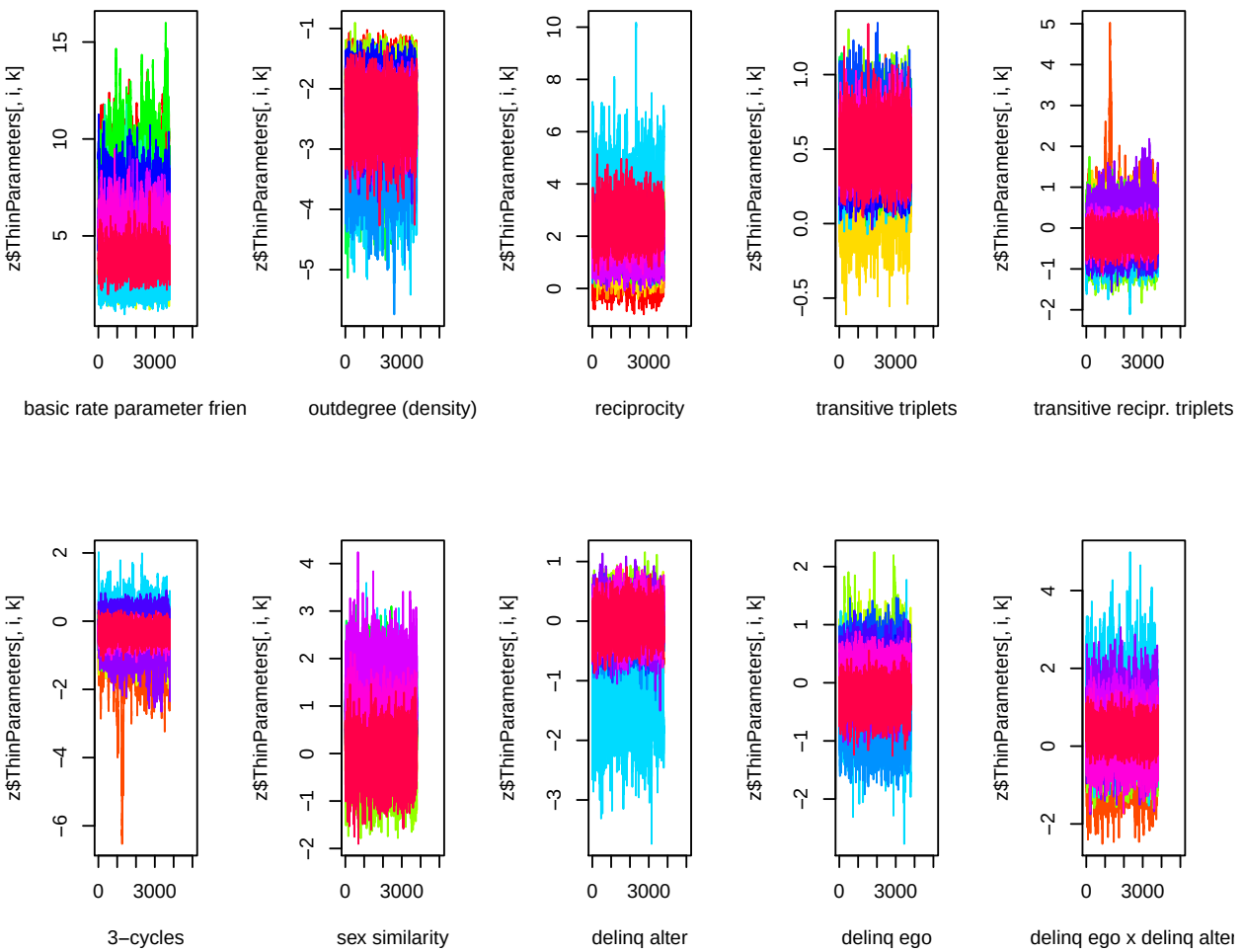


Figure : Posteriors for $\theta^{[g]}$ independently fitted using $\pi(\theta^{[g]}) \propto c$ (excluding rogue $g = 10, 20$)

Marginal indep SAOM (excluding rogue $g = 10, 20$)

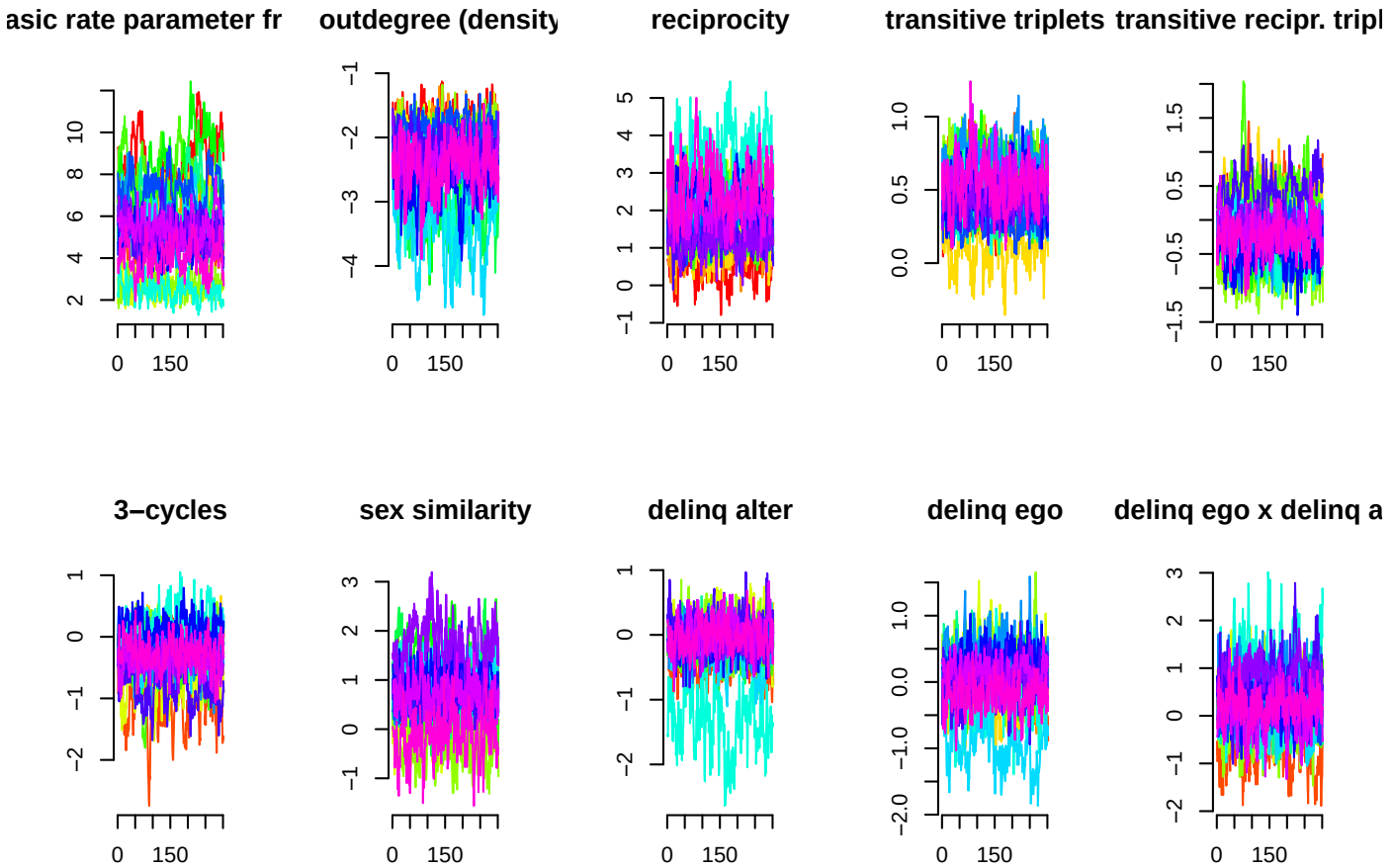


Figure : Posteriors for $\theta^{[g]}$ fitted using Jeffrey's prior

Marginal indep SAOM (excluding rogue $g = 10, 20$)

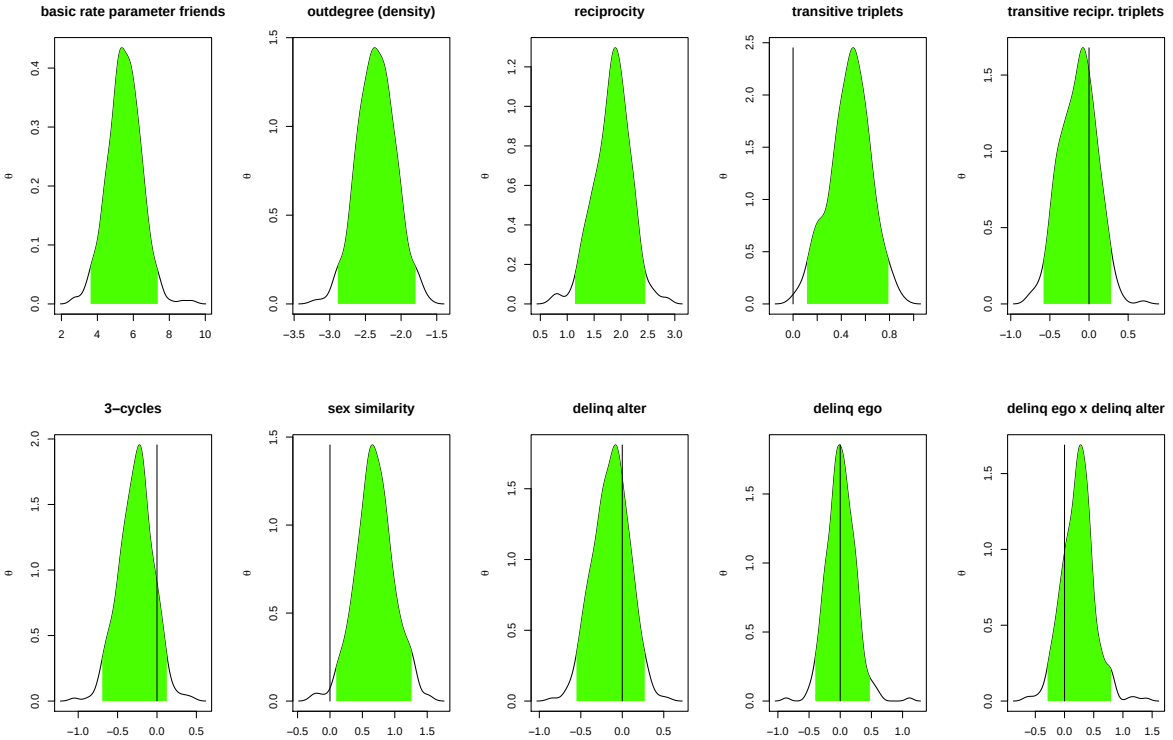


Figure : Posteriors for μ fitted using Jeffrey's prior

Convergence

What does 'convergence' look like?

For Previous example $G \sim Bern(\theta)$, where $n = 16$ and $L = 16$.

MCMC: iteratively update by

(a) update θ to $\theta^* = \theta + U$

(b) U is uniform on (-steplength,steplength)

(c) accept move with probability:

$$\frac{\pi(\theta^*)}{\pi(\theta)} = \frac{\theta^{*L+\alpha-1}(1-\theta^*)^{M-L+\beta-1}}{\theta^{L+\alpha-1}(1-\theta)^{M-L+\beta-1}}$$

or 1 if $\pi(\theta^*)/\pi(\theta) > 0$

(d) starting in $\theta = 1$

A Note on the 'fish' $\propto$

Note that in:

(c) accept move with probability:

$$\frac{\pi(\theta^*)}{\pi(\theta)} = \frac{\theta^{*L+\alpha-1}(1-\theta^*)^{M-L+\beta-1}}{\theta^{L+\alpha-1}(1-\theta)^{M-L+\beta-1}}$$

we can multiply π by any constant as:

$$\frac{\pi(\theta^*)c}{\pi(\theta)c} = \frac{\pi(\theta^*)c^*}{\pi(\theta)c^*} = \frac{\pi(\theta^*)}{\pi(\theta)}$$

What does ‘convergence’ look like?

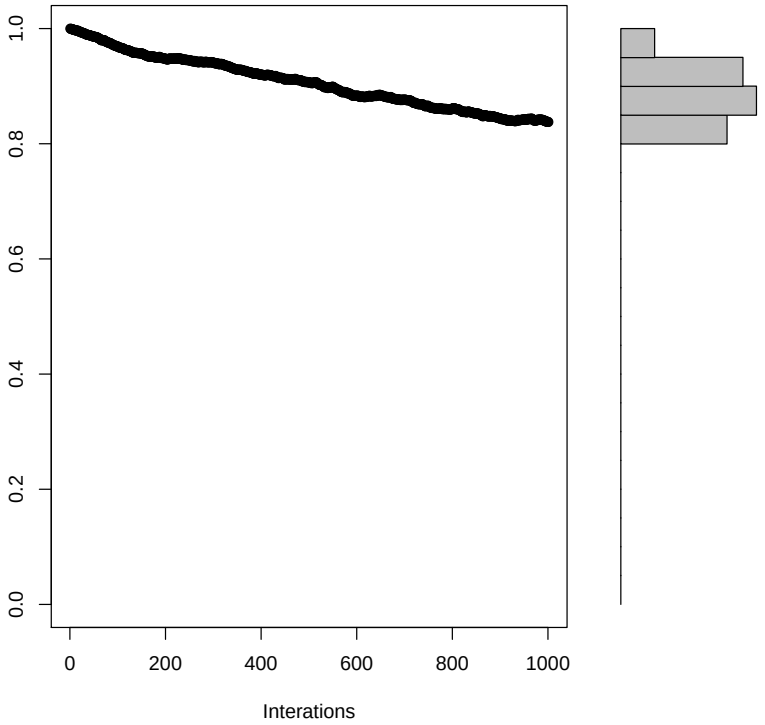


Figure : Steplength: 0.001 too small

What does ‘convergence’ look like?

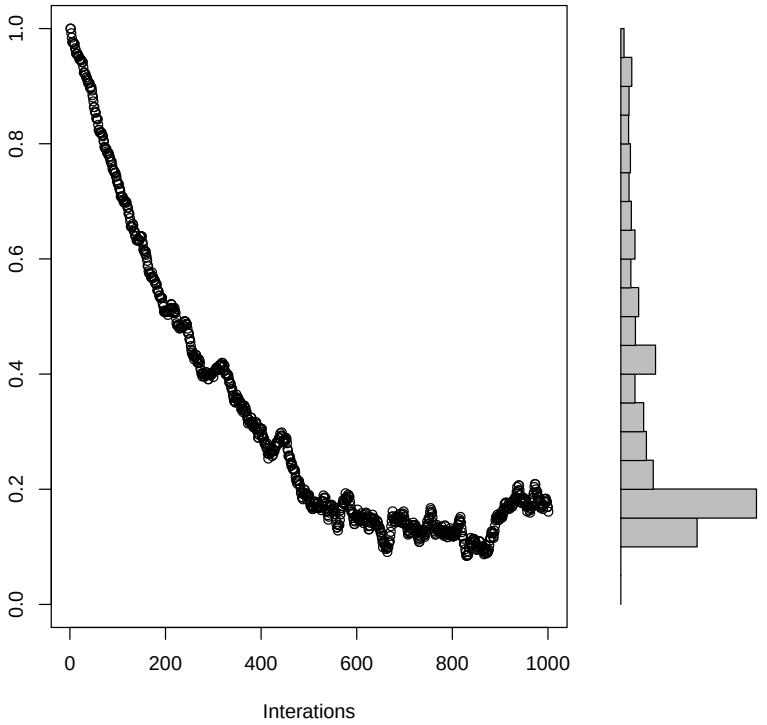


Figure : Steplength: 0.01 still too small

What does ‘convergence’ look like?

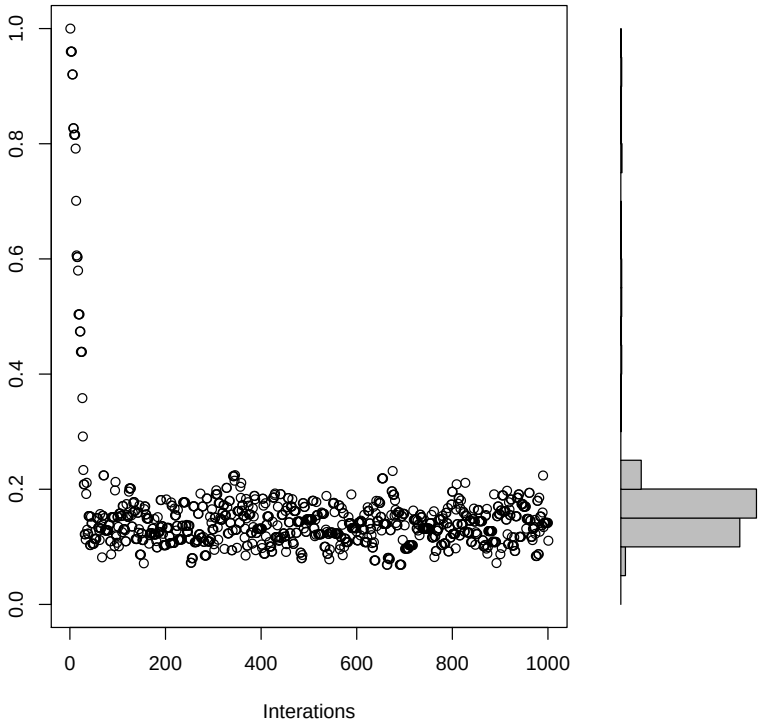


Figure : Steplength: 0.1 looking quite good

What does ‘convergence’ look like?

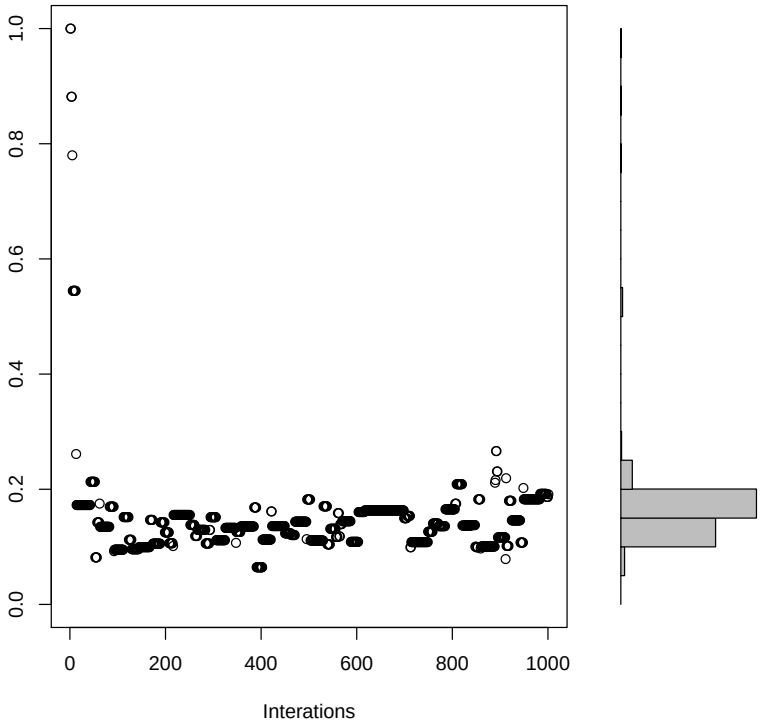
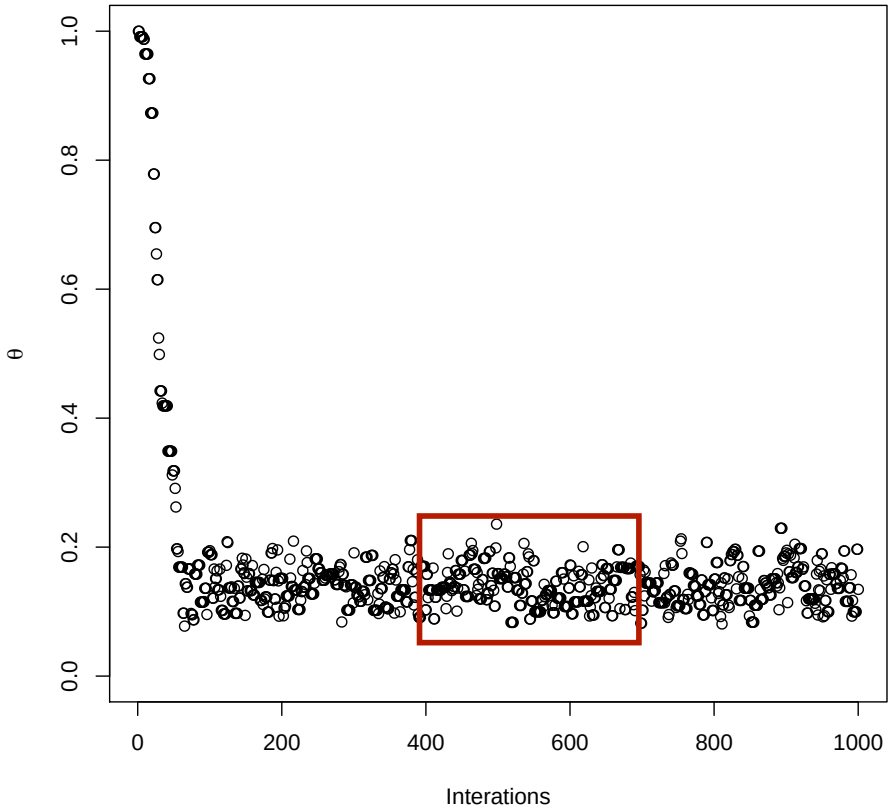
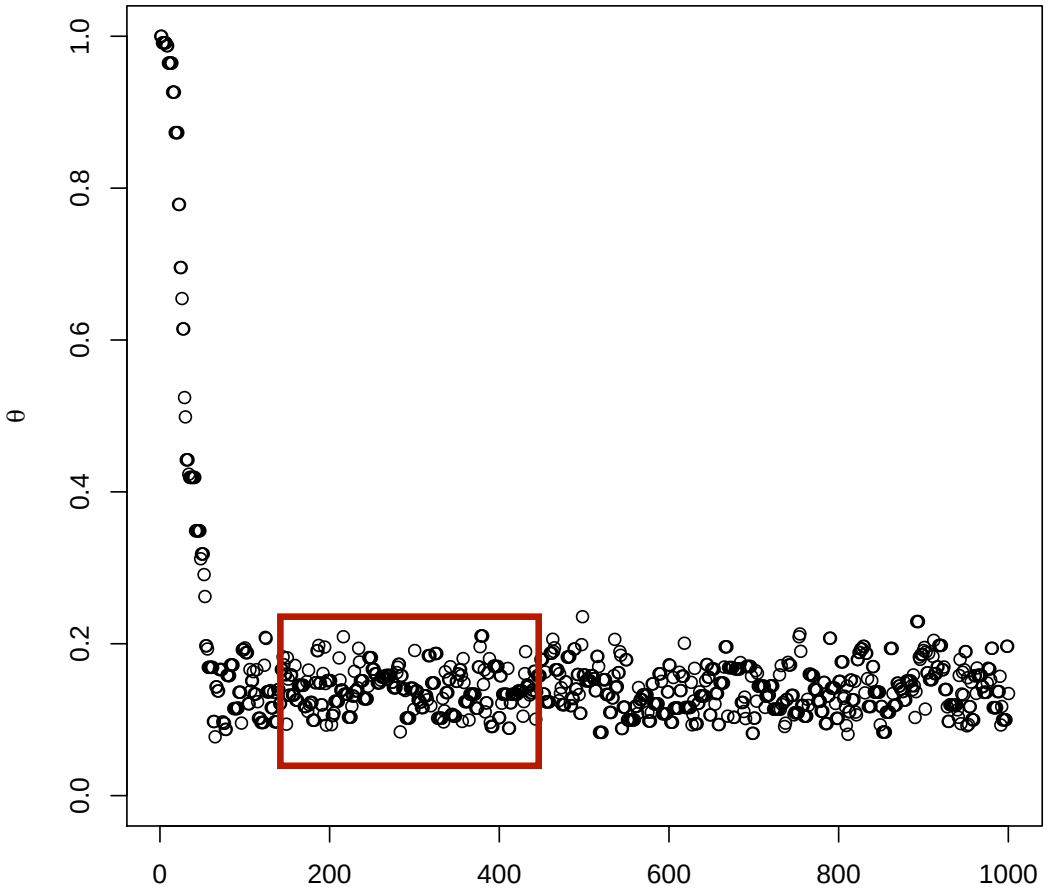


Figure : Steplength: 0.5 maybe too large

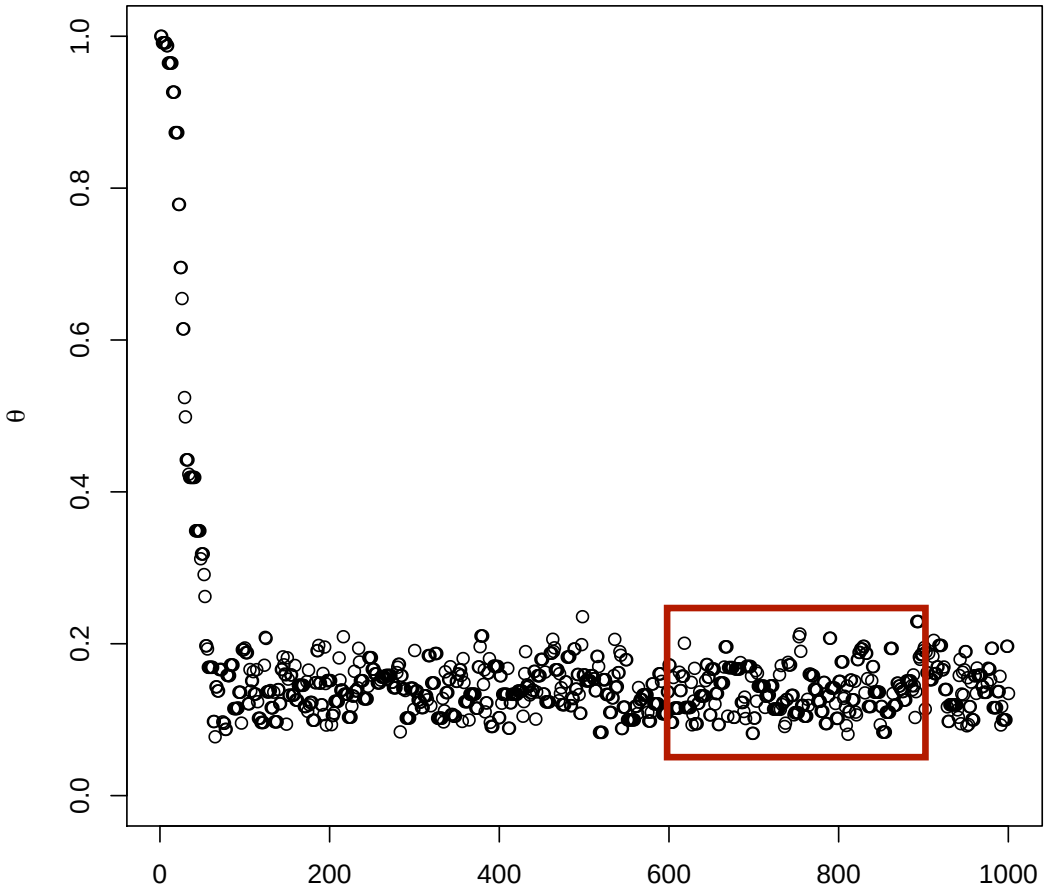
Converged to draws from the *same* distribution?



Converged to draws from the *same* distribution?



Converged to draws from the *same* distribution?



Final Tips

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Thought experiment:

'Sarah and Peter analyse the same data set

Sarah uses the sarah-prior

Peter uses the pete-prior.

Sarah and Peter arrive at different conclusions'

Who is right?

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Thought experiment:

'Sarah and Peter analyse the same data set

Sarah uses the sarah-prior

Peter uses the pete-prior.

Sarah and Peter arrive at different conclusions'

Who is right?

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Thought experiment:

'Sarah and Peter analyse the same data set

Sarah uses the sarah-prior

Peter uses the pete-prior.

Sarah and Peter arrive at different conclusions'

Who is right?

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Thought experiment:

'Sarah and Peter analyse the same data set

Sarah uses the sarah-prior

Peter uses the pete-prior.

Sarah and Peter arrive at different conclusions'

Who is right?

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Thought experiment:

'Sarah and Peter analyse the same data set

Sarah uses the sarah-prior

Peter uses the pete-prior.

Sarah and Peter arrive at different conclusions'

Who is right?

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Thought experiment:

> *'Sarah and Peter analyse the same data set*
>
> *Sarah uses the sarah-prior*
>
> *Peter uses the pete-prior.*
>
> *Sarah and Peter arrive at different conclusions'*

Who is right?

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Thought experiment:

> *'Sarah and Peter analyse the same data set*
>
> *Sarah uses the sarah-prior*
>
> *Peter uses the pete-prior.*
>
> *Sarah and Peter arrive at different conclusions'*

Who is right?

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Possible default priors:

a. standard: a 'non-informative' prior that works for standard cases (N large-ish and $n^{[h]}$ not too small)

b. null: absolutely NO prior information

c. density-dependent: a little like Tom's heterodox approach:
   - if $\bar{x}$ is the average degree of the *first* observation
   - $\mu_{den} = \log \frac{\bar{x}}{(n-1-\bar{x})}$

d. Jeffrey's

e. A prior so that $\theta^{[g]} \approx \eta$

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Possible default priors:

a. standard: a 'non-informative' prior that works for standard cases (N large-ish and $n^{[h]}$ not too small)

b. null: absolutely NO prior information

c. density-dependent: a little like Tom's heterodox approach:
   - if $\bar{x}$ is the average degree of the *first* observation
   - $\mu_{den} = \log \frac{\bar{x}}{(n-1-\bar{x})}$

d. Jeffrey's

e. A prior so that $\theta^{[g]} \approx \eta$

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Possible default priors:

a. standard: a 'non-informative' prior that works for standard cases (N large-ish and $n^{[h]}$ not too small)

b. null: absolutely NO prior information

c. density-dependent: a little like Tom's heterodox approach:
   - if $\bar{x}$ is the average degree of the *first* observation
   - $\mu_{den} = \log \frac{\bar{x}}{(n-1-\bar{x})}$

d. Jeffrey's

e. A prior so that $\theta^{[g]} \approx \eta$

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Possible default priors:

a. standard: a 'non-informative' prior that works for standard cases (N large-ish and $n^{[h]}$ not too small)

b. null: absolutely NO prior information

c. density-dependent: a little like Tom's heterodox approach:
   - if $\bar{x}$ is the average degree of the *first* observation
   - $\mu_{den} = \log \frac{\bar{x}}{(n-1-\bar{x})}$

d. Jeffrey's

e. A prior so that $\theta^{[g]} \approx \eta$

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Possible default priors:

a. standard: a 'non-informative' prior that works for standard cases (N large-ish and $n^{[h]}$ not too small)

b. null: absolutely NO prior information

c. density-dependent: a little like Tom's heterodox approach:
   - if $\bar{x}$ is the average degree of the *first* observation
   - $\mu_{den} = \log \frac{\bar{x}}{(n-1-\bar{x})}$

d. Jeffrey's

e. A prior so that $\theta^{[g]} \approx \eta$

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Possible default priors:

a. standard: a 'non-informative' prior that works for standard cases (N large-ish and $n^{[h]}$ not too small)

b. null: absolutely NO prior information

c. density-dependent: a little like Tom's heterodox approach:
   - if $\bar{x}$ is the average degree of the *first* observation
   - $\mu_{den} = \log \frac{\bar{x}}{(n-1-\bar{x})}$

d. Jeffrey's

e. A prior so that $\theta^{[g]} \approx \eta$

When use HSAOM?

- hierarchical data
  - some groups small (borrow strength)
  - many groups with heterogeneity
- intervention: treatment on class-room-level
- network too large: can you decompose network in a natural way? C.p. settings model
- many waves: time heterogeneity potentially with time-covariate - (t_1, t_2),(t_2, t_3), etc, different 'groups'

Parting shots

Random $\theta^{[g]} \sim N(\mu, \Sigma)$ or fixed parameters η?

- Are differences between groups
  - random or
  - meaningful (i.e. non-random)

Is there a correct prior distribution?

- NO!

Can I say 'there is a 0.95 probability that there is an influence effect'?

YES - you should!

Parting shots

Random $\theta^{[g]} \sim N(\mu, \Sigma)$ or fixed parameters η?

- Are differences between groups
  - random or
  - meaningful (i.e. non-random)

Is there a correct prior distribution?

- NO!

Can I say 'there is a 0.95 probability that there is an influence effect'?

YES - you should!

Parting shots

Random $\theta^{[g]} \sim N(\mu, \Sigma)$ or fixed parameters η?

- Are differences between groups
  - random or
  - meaningful (i.e. non-random)

Is there a correct prior distribution?

- NO!

Can I say 'there is a 0.95 probability that there is an influence effect'?

YES - you should!

Parting shots

Random $\theta^{[g]} \sim N(\mu, \Sigma)$ or fixed parameters η?

- Are differences between groups
  - random or
  - meaningful (i.e. non-random)

Is there a correct prior distribution?

- NO!

Can I say 'there is a 0.95 probability that there is an influence effect'?

YES - you should!

Parting shots

Random $\theta^{[g]} \sim N(\mu, \Sigma)$ or fixed parameters η?

- Are differences between groups
  - random or
  - meaningful (i.e. non-random)

Is there a correct prior distribution?

- NO!

Can I say 'there is a 0.95 probability that there is an influence effect'?

YES - you should!