Multilevel longitudinal network estimation using `sienaBayes`

principles of Bayesian inference

Johan Koskinen and T.A.B. Snijders

Department of Statistics
Stockholm University

June 25, 2024

Fundamentals

Take-home points

What we need to know

- Posterior distribution is
  - The distribution of the unknown parameters
  - given the known data

Take-home points

What we need to know

- Posterior distribution is
  - The distribution of the unknown parameters
  - given the known data
- All uncertainty about parameters is described by the posterior distribution
  - The probability that the true parameter lies in the 95% Credibility interval **is** 0.95 (given observed data)
  - You may you use the posterior expected value ('average') of the parameter as your point estimate
  - The amount of posterior uncertainty give information in data is captured by the standard deviation of the parameter

Take-home points

What we need to know

- Posterior distribution is
  - The distribution of the unknown parameters
  - given the known data
- All uncertainty about parameters is described by the posterior distribution
  - The probability that the true parameter lies in the 95% Credibility interval **is** 0.95 (given observed data)
  - You may you use the posterior expected value (‘average’) of the parameter as your point estimate
  - The amount of posterior uncertainty give information in data is captured by the standard deviation of the parameter
- Prior distribbution
  - In order to obtain a posterior distribution you need a prior distribution
  - Different priors give different posteriors for the same data

(Statistical) model

Model

(Statistical) model

Model

We call

$$P(\underbrace{Data}_{observable} \mid \overbrace{\theta}^{un-obs.})$$

a model for *Data* when the model allocates probability to different outcomes *Data* we can observe, indexed by some statistical **parameters** θ

(Statistical) model

Model

We call

$$P(\underbrace{Data}_{observable} \mid \overbrace{\theta}^{\textcolor{red}{un}-obs.})$$

a model for *Data* when the model allocates probability to different outcomes *Data* we can observe, indexed by some statistical **parameters** θ

Example (ERGM)

Data are an adjacency matrix $\boldsymbol{x}$

$$\boldsymbol{x} \sim ERGM(\theta) \text{ , Model: } P(\boldsymbol{x} \mid \theta) = e^{\theta^\top z(\boldsymbol{x}) - \psi(\theta)}$$

(Statistical) model

Model

We call

$$P(\underset{observable}{Data} \mid \overset{\color{red}{un}-obs.}{\theta})$$

a model for *Data* when the model allocates probability to different outcomes *Data* we can observe, indexed by some statistical **parameters** θ

Example (ERGM)

Data are an adjacency matrix $\boldsymbol{x}$

$$\boldsymbol{x} \sim ERGM(\theta) \text{ , Model: } P(\boldsymbol{x} \mid \theta) = e^{\theta^\top z(\boldsymbol{x}) - \psi(\theta)}$$

Example (SAOM)

Data are an adjacency matrix $\boldsymbol{x}(t_1)$, at time t_1, given $\boldsymbol{x}(t_0)$, at time t_0

$$\boldsymbol{x}(t_1) \mid \boldsymbol{x}(t_0) \sim SAOM(\theta)$$

Inference

Given data, we aim to find parameters θ that data gives most evidence for.

Likelihood

Inference

Given data, we aim to find parameters θ that data gives most evidence for.

Likelihood

We can use the model We call

$$P(Data \mid \theta)$$

as a function of θ

$$L(\theta; Data) = \underbrace{P(\overbrace{Data}^{const.} \mid \theta)}_{\text{now a function of } \theta}$$

For different choices of θ, the probability $P(Data \mid \theta)$ will be different!

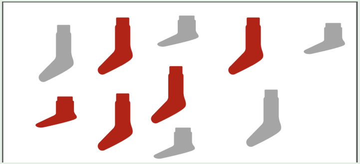
Deriving Bayes theorem

Example (Your sock drawer!)

Probability of picking BIG and red sock

$$\underbrace{P(A,B)}_{4/10=40\%=12/30} \quad =$$

$A = \{\text{sock red}\}$, $\sharp A = 5$
$B = \{\text{sock big}\}$, $\sharp B = 6$



In total 10 (single) socks

Deriving Bayes theorem

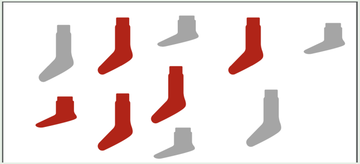
Example (Your sock drawer!)

Probability of picking BIG and red sock

$$\underbrace{P(A, B)}_{4/10=40\%=12/30} = \underbrace{P(A \mid B)}_{4/6=2/3}$$

$A = \{\text{sock red}\}$, $\sharp A = 5$

$B = \{\text{sock big}\}$, $\sharp B = 6$



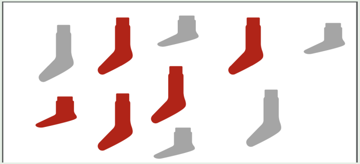
In total 10 (single) socks

Deriving Bayes theorem

Example (Your sock drawer!)

$A = \{\text{sock red}\}$, $\sharp A = 5$

$B = \{\text{sock big}\}$, $\sharp B = 6$



In total 10 (single) socks

Probability of picking BIG and red sock

$$\underbrace{P(A, B)}_{4/10=40\%=12/30} = \underbrace{P(A \mid B)}_{4/6=2/3} \underbrace{P(B)}_{6/10}$$

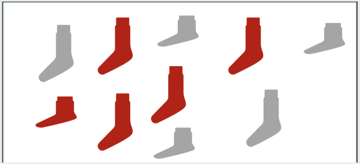
but also

$$\underbrace{P(A, B)}_{4/10=40\%=20/50} =$$

Deriving Bayes theorem

Example (Your sock drawer!)

$A = \{\text{sock } \textcolor{red}{\text{red}}\}$, $\sharp A = 5$
$B = \{\text{sock big}\}$, $\sharp B = 6$



In total 10 (single) socks

Probability of picking BIG and red sock

$$\underbrace{P(A, B)}_{4/10=40\%=12/30} = \underbrace{P(A \mid B)}_{4/6=2/3} \underbrace{P(B)}_{6/10}$$

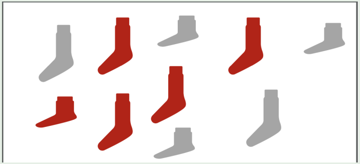
but also

$$\underbrace{P(A, B)}_{4/10=40\%=20/50} = \underbrace{P(B \mid A)}_{4/5}$$

Deriving Bayes theorem

Example (Your sock drawer!)

$A = \{\text{sock red}\}$, $\sharp A = 5$
$B = \{\text{sock big}\}$, $\sharp B = 6$



In total 10 (single) socks

Probability of picking BIG and red sock

$$\underbrace{P(A,B)}_{4/10=40\%=12/30} = \underbrace{P(A \mid B)}_{4/6=2/3}\ \underbrace{P(B)}_{6/10}$$

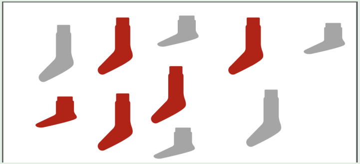
but also

$$\underbrace{P(A,B)}_{4/10=40\%=20/50} = \underbrace{P(B \mid A)}_{4/5}\ \underbrace{P(A)}_{5/10}$$

Deriving Bayes theorem

Example (Your sock drawer!)

$$A = \{\text{sock red}\}, \; \sharp A = 5$$
$$B = \{\text{sock big}\}, \; \sharp B = 6$$



In total 10 (single) socks

Probability of picking BIG and red sock

$$\underbrace{P(A,B)}_{4/10=40\%=12/30} = \underbrace{P(A \mid B)}_{4/6=2/3} \underbrace{P(B)}_{6/10}$$

but also

$$\underbrace{P(A,B)}_{4/10=40\%=20/50} = \underbrace{P(B \mid A)}_{4/5} \underbrace{P(A)}_{5/10}$$

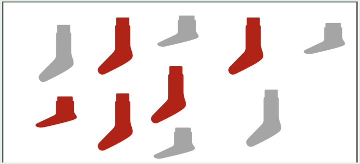
We have equality

$$\underbrace{P(A \mid B)}_{\text{red } \textit{given} \text{ BIG}} \; P(B)$$

Deriving Bayes theorem

Example (Your sock drawer!)

$A = \{\text{sock red}\},\ \sharp A = 5$
$B = \{\text{sock big}\},\ \sharp B = 6$



In total 10 (single) socks

Probability of picking BIG and red sock

$$\underbrace{P(A,B)}_{4/10=40\%=12/30} = \underbrace{P(A \mid B)}_{4/6=2/3}\ \underbrace{P(B)}_{6/10}$$

but also

$$\underbrace{P(A,B)}_{4/10=40\%=20/50} = \underbrace{P(B \mid A)}_{4/5}\ \underbrace{P(A)}_{5/10}$$

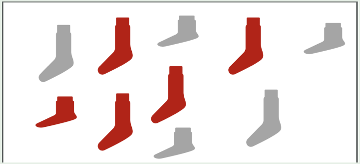
We have equality

$$\underbrace{P(A \mid B)}_{\text{red } \textit{given} \text{ BIG}}\ P(B) = \underbrace{P(B \mid A)}_{\text{red } \textit{given} \text{ BIG}}\ P(A)$$

Deriving Bayes theorem

Example (Your sock drawer!)

$$A = \{\text{sock red}\}$$
$$B = \{\text{sock big}\}$$



Since

$$\underbrace{P(A, B)}_{\text{both A \& B}} = P(A \mid B)P(B) = P(B \mid A)P(A)$$

we can write red *given* BIG

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

in terms of BIG *given* red

Bayes

Bayes theorem

$$P(A \mid B) =$$

Bayes

Bayes theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

where we see that $P(B)$ does not depend on A.

Bayes

Bayes theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

where we see that $P(B)$ does not depend on A. In terms of our model, $P(Data \mid \theta)$

$$P(\theta \mid Data) = \frac{\overbrace{P(Data \mid \theta)}^{L(\theta;Data)}\overbrace{\pi(\theta)}^{prior}}{\underbrace{P(Data)}_{constant}}$$

Bayes

Bayes theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

where we see that $P(B)$ does not depend on A. In terms of our model, $P(Data \mid \theta)$

$$P(\theta \mid Data) = \frac{\overbrace{P(Data \mid \theta)}^{L(\theta;Data)}\overbrace{\pi(\theta)}^{prior}}{\underbrace{P(Data)}_{constant}} \propto L(\theta; Data)\pi(\theta)$$

Bayes

Bayes theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

where we see that $P(B)$ does not depend on A. In terms of our model, $P(Data \mid \theta)$

$$P(\theta \mid Data) = \frac{\overbrace{P(Data \mid \theta)}^{L(\theta;Data)}\overbrace{\pi(\theta)}^{prior}}{\underbrace{P(Data)}_{constant}} \propto L(\theta; Data)\pi(\theta)$$

Ergo, the ‘probability’ for each θ is the likelihood weighted by the a priori ‘probability’

Posterior of tie-probability in Bernoulli graph

Example (Posterior distribution for p in Bernoulli graph)

Assume $\boldsymbol{x} \sim BG(p, n = 5)$, i.e. independently for each $i < j$, $X_{ij} \sim Bern(p)$.

Posterior of tie-probability in Bernoulli graph

Example (Posterior distribution for p in Bernoulli graph)

Assume $\boldsymbol{x} \sim BG(p, n = 5)$, i.e. independently for each $i < j$, $X_{ij} \sim Bern(p)$.
We observe: $y = \sum_{i<j} x_{ij} = 2$.

$$\text{Likelihood} \ : L(p; \boldsymbol{x}) = p^2(1-p)^{10-2}$$

With $p \sim Beta(\alpha, \beta)$ prior

$$\text{Prior} \ : \pi(p) = cp^{\alpha-1}(1-p)^{\beta-1}$$

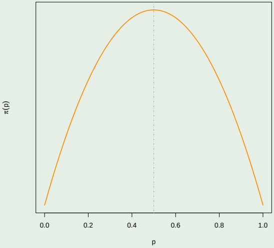
the posterior is $Beta(\alpha^*, \beta^*)$

$$\pi(p \mid \boldsymbol{x}) \propto L(p; \boldsymbol{x})\pi(p) \propto p^{\alpha^*-1}(1-p)^{\beta^*-1}$$

where $\alpha^* = \alpha + 2$ and $\beta^* = \beta + 8$

Example (Posterior distribution for p in Bernoulli graph (A))

$$p^{2-1}(1-p)^{2-1}$$



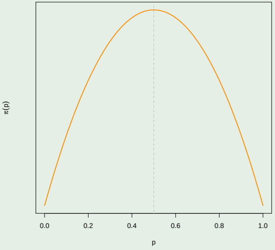
Prior

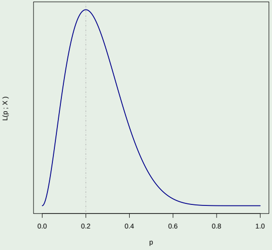
Posterior of tie-probability in Bernoulli graph

Example (Posterior distribution for p in Bernoulli graph (A))

$$p^{2-1}(1-p)^{2-1}$$

$$p^{2}(1-p)^{8}$$





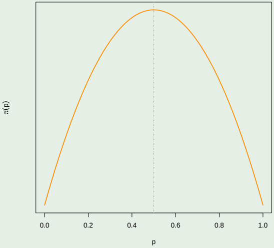
Prior

×Likelihood

Posterior of tie-probability in Bernoulli graph

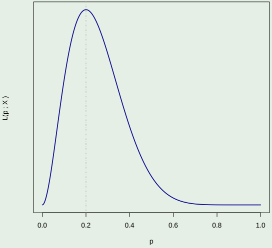
Example (Posterior distribution for p in Bernoulli graph (A))

$$p^{2-1}(1-p)^{2-1}$$



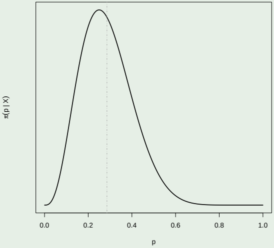
Prior

$$p^{2}(1-p)^{8}$$



$\times$ Likelihood

$$p^{4-1}(1-p)^{10-1}$$

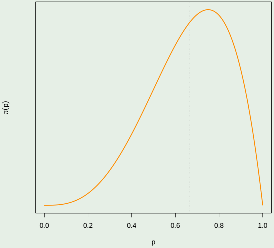


$\propto$ Posterior

Posterior of tie-probability in Bernoulli graph

Example (Posterior distribution for p in Bernoulli graph (B))

$$p^{4-1}(1-p)^{2-1}$$



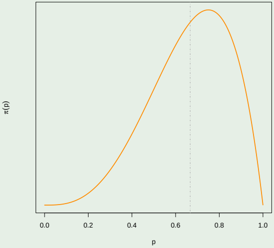
Prior

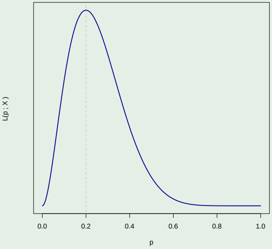
Posterior of tie-probability in Bernoulli graph

Example (Posterior distribution for p in Bernoulli graph (B))

$p^{4-1}(1-p)^{2-1}$

$p^2(1-p)^8$





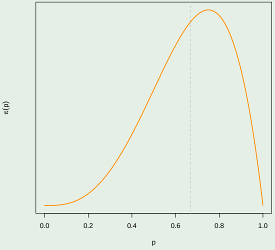
Prior

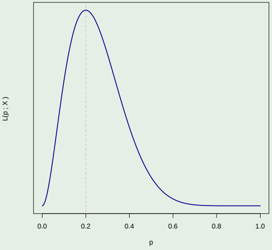
×Likelihood

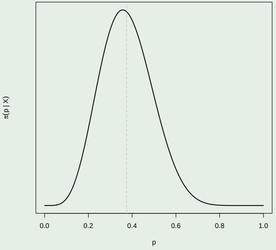
Posterior of tie-probability in Bernoulli graph

Example (Posterior distribution for p in Bernoulli graph (B))

$p^{4-1}(1-p)^{2-1}$ $\qquad$ $p^2(1-p)^8$ $\qquad$ $p^{6-1}(1-p)^{10-1}$







Prior $\qquad$ $\times$Likelihood $\qquad$ $\propto$Posterior

Posterior summaries

The posterior distribution of our parameters given data

$$P(\theta \mid Data)$$

fully describes our uncertainty about the parameters given our observed data

Posterior summaries

The posterior distribution of our parameters given data

$$P(\theta \mid Data)$$

fully describes our uncertainty about the parameters given our observed data

We can summarise these using the expected values

$$\hat{\theta} = E(\theta \mid Data)$$

as **point estimates**,

Posterior summaries

The posterior distribution of our parameters given data

$$P(\theta \mid Data)$$

fully describes our uncertainty about the parameters given our observed data
We can summarise these using the expected values

$$\hat{\theta} = E(\theta \mid Data)$$

as **point estimates**, and standard deviations

$$SD(\theta \mid Data) = \sqrt{E\left[(\theta - E(\theta \mid Data))^2 \mid Data\right]}$$

as measures of uncertainty

Posterior of tie-probability in Bernoulli graph

Example (Posterior distribution for p in Bernoulli graph)

Assume $\boldsymbol{x} \sim BG(p, n = 5)$, i.e. independently for each $i < j$, $X_{ij} \sim Bern(p)$.

Posterior of tie-probability in Bernoulli graph

Example (Posterior distribution for p in Bernoulli graph)

Assume $\boldsymbol{x} \sim BG(p, n = 5)$, i.e. independently for each $i < j$, $X_{ij} \sim Bern(p)$.
We observe $y = \sum_{i<j} x_{ij} = 2$ and use $p \sim Beta(\alpha, \beta)$ prior $\Rightarrow$ the posterior is $Beta(\alpha^*, \beta^*)$, where $\alpha^* = \alpha + 2$ and $\beta^* = \beta + 8$

$$\hat{\theta} = E(\theta \mid Data) = \frac{\alpha^*}{\alpha^* + \beta^*} \text{, point estimate}$$

Posterior of tie-probability in Bernoulli graph

Example (Posterior distribution for p in Bernoulli graph)

Assume $\boldsymbol{x} \sim BG(p, n = 5)$, i.e. independently for each $i < j$, $X_{ij} \sim Bern(p)$.
We observe $y = \sum_{i<j} x_{ij} = 2$ and use $p \sim Beta(\alpha, \beta)$ prior $\Rightarrow$ the posterior is $Beta(\alpha^*, \beta^*)$, where $\alpha^* = \alpha + 2$ and $\beta^* = \beta + 8$

$$\hat{\theta} = E(\theta \mid Data) = \frac{\alpha^*}{\alpha^* + \beta^*} \text{, point estimate}$$

and

$$SD(\theta \mid Data) = \sqrt{\frac{\alpha^* \beta^*}{(\alpha^* + \beta^*)^2(\alpha^* + \beta^* + 1)}} \text{, uncertainty}$$

and with $\alpha = \beta = 2$ a rough 95% credibility interval

$$\hat{\theta} \pm 2SD(\theta \mid Data) = \tfrac{2}{7} \pm 2 \times 0.12 \Rightarrow (0.052, 0.519)$$

Posterior for parameters in SAOM

For

$$\boldsymbol{x}(t_1) \mid \boldsymbol{x}(t_0) \sim SAOM(\theta)$$

we cannot obtain the posterior distribution

$$\pi(\theta \mid \boldsymbol{x}(t_1), \boldsymbol{x}(t_0))$$

so easily

Posterior for parameters in SAOM

For

$$\boldsymbol{x}(t_1) \mid \boldsymbol{x}(t_0) \sim SAOM(\theta)$$

we cannot obtain the posterior distribution

$$\pi(\theta \mid \boldsymbol{x}(t_1), \boldsymbol{x}(t_0))$$

so easily

Using Markov chain Monte Carlo (MCMC) we can simulate/draw from the posterior

$$\theta^0, \theta^1, \dots, \theta^M \overset{approx.iid}{\sim} \pi(\theta \mid \boldsymbol{x}(t_1), \boldsymbol{x}(t_0))$$

Convergence of MCMC

to target distribution $\pi(\theta \mid Data)$

What does ‘convergence’ look like?

For Previous example $\boldsymbol{x} \sim Bern(\theta)$, where $n = 16$ and $L = \sum x_{ij} = 16$.

What does ‘convergence’ look like?

For Previous example $\boldsymbol{x} \sim Bern(\theta)$, where $n = 16$ and $L = \sum x_{ij} = 16$.
With $\theta \sim Beta(1, 1)$, we know $\theta \mid \boldsymbol{x} \sim Beta(16, 120)$

What does ‘convergence’ look like?

For Previous example $\boldsymbol{x} \sim Bern(\theta)$, where $n = 16$ and $L = \sum x_{ij} = 16$.
With $\theta \sim Beta(1,1)$, we know $\theta \mid \boldsymbol{x} \sim Beta(16,120)$
MCMC: iteratively update by

What does 'convergence' look like?

For Previous example $\boldsymbol{x} \sim Bern(\theta)$, where $n = 16$ and $L = \sum x_{ij} = 16$.
With $\theta \sim Beta(1,1)$, we know $\theta \mid \boldsymbol{x} \sim Beta(16,120)$
MCMC: iteratively update by
(a) update θ to $\theta^* = \theta + U$

What does ‘convergence’ look like?

For Previous example $\boldsymbol{x} \sim Bern(\theta)$, where $n = 16$ and $L = \sum x_{ij} = 16$.
With $\theta \sim Beta(1, 1)$, we know $\theta \mid \boldsymbol{x} \sim Beta(16, 120)$
MCMC: iteratively update by
(a) update θ to $\theta^* = \theta + U$
(b) U is uniform on (-steplength,steplength)

What does 'convergence' look like?

For Previous example $\boldsymbol{x} \sim Bern(\theta)$, where $n = 16$ and $L = \sum x_{ij} = 16$.
With $\theta \sim Beta(1,1)$, we know $\theta \mid \boldsymbol{x} \sim Beta(16, 120)$
MCMC: iteratively update by
(a) update θ to $\theta^* = \theta + U$
(b) U is uniform on (-steplength,steplength)
(c) accept move with probability:

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

or 1 if $\pi(\theta^* \mid \boldsymbol{x})/\pi(\theta \mid \boldsymbol{x}) > 0$

What does ‘convergence’ look like?

For Previous example $\boldsymbol{x} \sim Bern(\theta)$, where $n = 16$ and $L = \sum x_{ij} = 16$.
With $\theta \sim Beta(1,1)$, we know $\theta \mid \boldsymbol{x} \sim Beta(16, 120)$
MCMC: iteratively update by
(a) update θ to $\theta^* = \theta + U$
(b) U is uniform on (-steplength,steplength)
(c) accept move with probability:

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

or 1 if $\pi(\theta^* \mid \boldsymbol{x})/\pi(\theta \mid \boldsymbol{x}) > 0$
(d) starting in $\theta = 1$

What does ‘convergence’ look like?

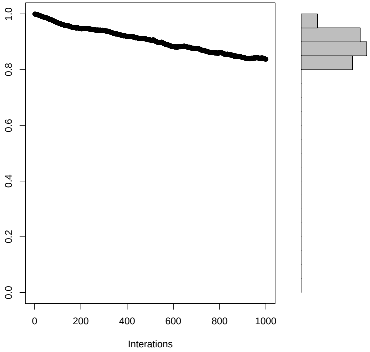


Figure: Steplength: 0.001 too small

What does ‘convergence’ look like?

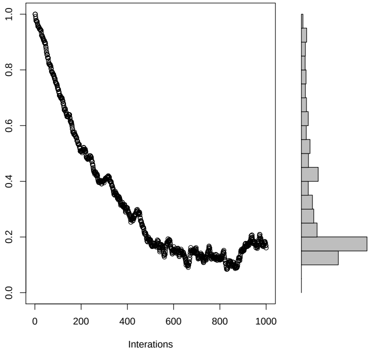


Figure: Steplength: 0.01 still too small

What does ‘convergence’ look like?

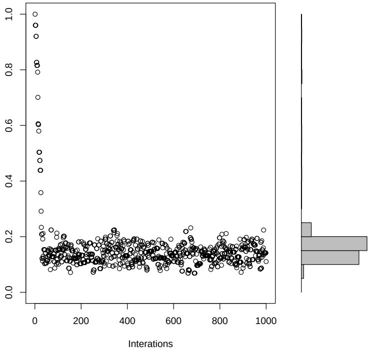


Figure: Steplength: 0.1 looking quite good

What does ‘convergence’ look like?

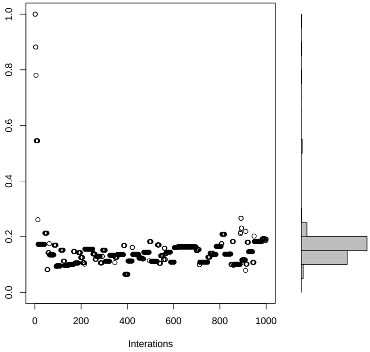
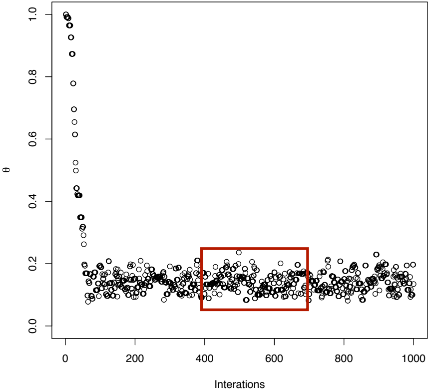
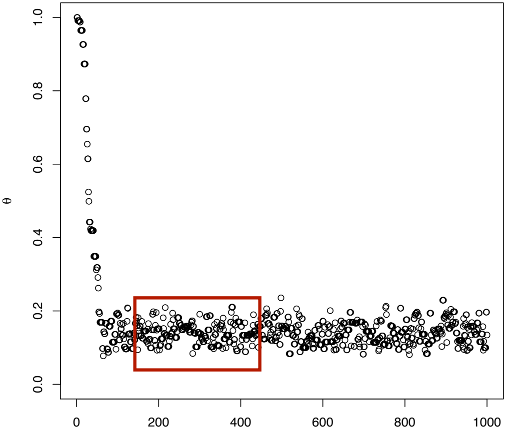


Figure: Steplength: 0.5 maybe too large

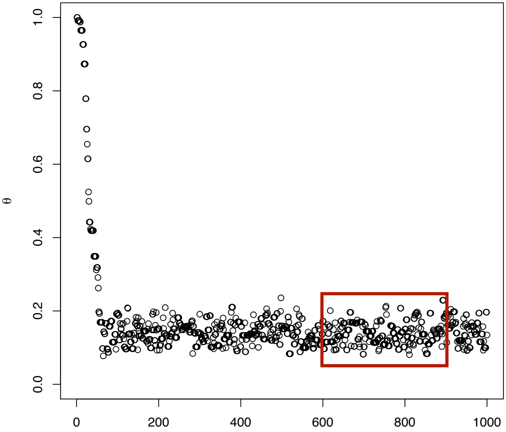
Converged to draws from the *same* distribution?



Converged to draws from the *same* distribution?



Converged to draws from the *same* distribution?



All unobservables have distributions - chain rule

Chain-rule of probability

All unobservables have distributions - chain rule

Chain-rule of probability

We can define a model observable Data given unobservable variables $\boldsymbol{y}$

$$P(Data \mid \boldsymbol{y})$$

All unobservables have distributions - chain rule

Chain-rule of probability

We can define a model observable Data given unobservable variables $\boldsymbol{y}$

$$P(Data \mid \boldsymbol{y})$$

and define a distribution for $\boldsymbol{y}$ given some parameter θ

$$P(\boldsymbol{y} \mid \theta)$$

All unobservables have distributions - chain rule

Chain-rule of probability

We can define a model observable Data given unobservable variables $\boldsymbol{y}$

$$P(Data \mid \boldsymbol{y})$$

and define a distribution for $\boldsymbol{y}$ given some parameter θ

$$P(\boldsymbol{y} \mid \theta)$$

which gives a joint distribution (chain-rule)

$$P(Data, \boldsymbol{y}, \theta) = P(Data \mid \boldsymbol{y}) P(\boldsymbol{y} \mid \theta) \pi(\theta)$$

All unobservables have distributions - chain rule

Chain-rule of probability

We can define a model observable Data given unobservable variables $\boldsymbol{y}$

$$P(Data \mid \boldsymbol{y})$$

and define a distribution for $\boldsymbol{y}$ given some parameter θ

$$P(\boldsymbol{y} \mid \theta)$$

which gives a joint distribution (chain-rule)

$$P(Data, \boldsymbol{y}, \theta) = P(Data \mid \boldsymbol{y}) P(\boldsymbol{y} \mid \theta) \pi(\theta)$$

The posterior distribution for θ is

$$\pi(\theta \mid Data) = \frac{\int P(Data \mid \boldsymbol{y}) P(\boldsymbol{y} \mid \theta) \pi(\theta) d\boldsymbol{y}}{\underbrace{\int \int P(Data \mid \boldsymbol{y}) P(\boldsymbol{y} \mid \theta) \pi(\theta) d\boldsymbol{y} d\theta}}$$

All unobservables have distributions - chain rule

Example (Hierarchial SAOM)

Data for group j an adjacency matrix $\boldsymbol{x}^{[j]}(t_1)$, at time t_1, given $\boldsymbol{x}^{[j]}(t_0)$, at time t_0

$$\boldsymbol{x}^{[j]}(t_1) \mid \boldsymbol{x}^{[j]}(t_0) \sim SAOM(\theta_j)$$

and

$$\theta_j \sim \mathcal{N}(\mu, \Sigma)$$

Interpretation: For a value on the unknown parameter μ (and Σ), we draw some unknown value θ_j, and then generate data from $\sim SAOM(\theta_j)$.

All unobservables have distributions - chain rule

Example (Hierarchial SAOM)

Data for group j an adjacency matrix $\boldsymbol{x}^{[j]}(t_1)$, at time t_1, given $\boldsymbol{x}^{[j]}(t_0)$, at time t_0

$$\boldsymbol{x}^{[j]}(t_1) \mid \boldsymbol{x}^{[j]}(t_0) \sim SAOM(\theta_j)$$

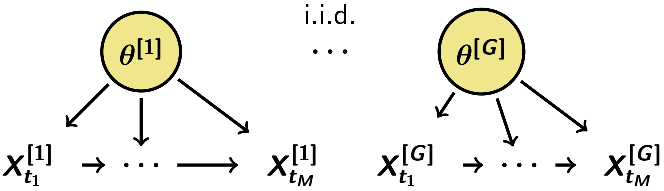
and

$$\theta_j \sim \mathcal{N}(\mu, \Sigma)$$

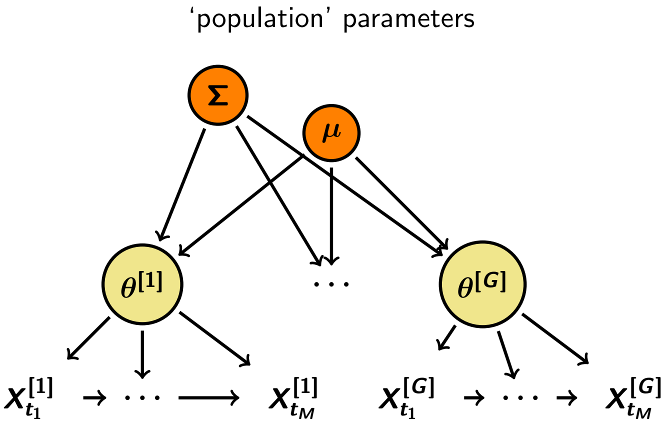
Interpretation: For a value on the unknown parameter μ (and Σ), we draw some unknown value θ_j, and then generate data from $\sim SAOM(\theta_j)$.
Aim: Find the 'true' value on μ

HSAOM (DAG)

groups: $g = 1, \ldots, G$

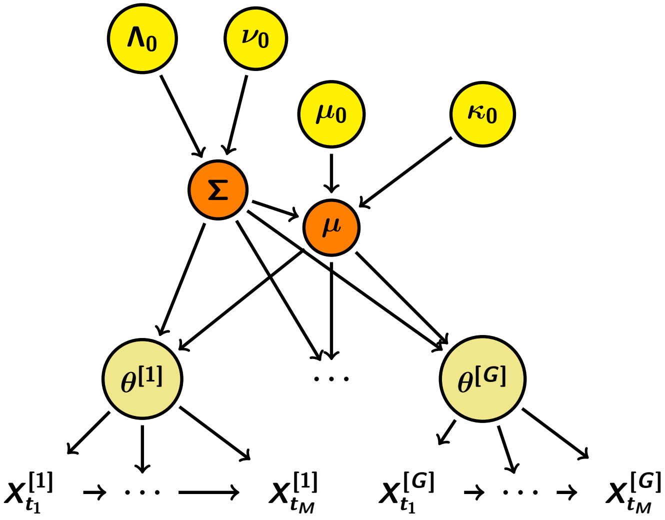


HSAOM (DAG)



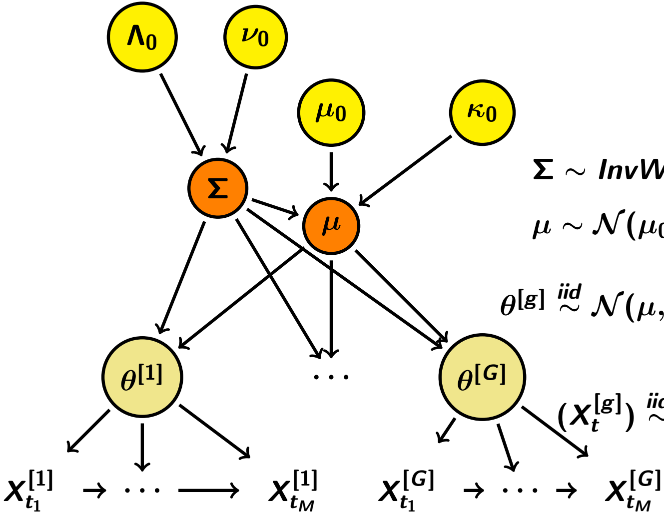
'population' parameters
Σ
μ
$\theta^{[1]}$
$\cdots$
$\theta^{[G]}$
$X_{t_1}^{[1]} \to \cdots \longrightarrow X_{t_M}^{[1]}$
$X_{t_1}^{[G]} \to \cdots \to X_{t_M}^{[G]}$

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)



$$\boldsymbol{\Sigma} \sim InvWish_{p_1}(\Lambda_0^{-1}, \nu_0)$$

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

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

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

Final Tips

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom 

Do you have to decide numbers in priors?

No - not everyone knows as much as Tom 
Thought experiment:

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

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

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.*

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'*

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)

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

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

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
- $\mu_{den} = \log \frac{\bar{x}}{(n-1-\bar{x})}$

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

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)

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?

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!

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’?

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!

Take-home points

What we need to know

- Posterior distribution is
  - The distribution of the unknown parameters
  - given the known data
- All uncertainty about parameters is described by the posterior distribution
  - The probability that the true parameter lies in the 95% Credibility interval **is** 0.95 (given observed data)
  - You may you use the posterior expected value (‘average’) of the parameter as your point estimate
  - The amount of posterior uncertainty give information in data is captured by the standard deviation of the parameter
- Prior distribbution
  - In order to obtain a posterior distribution you need a prior distribution
  - Different priors give different posteriors for the same data

Theorem