

Multilevel longitudinal network estimation using `sienaBayes` principles of Bayesian inference

Johan Koskinen and T.A.B. Snijders

Department of Statistics
Stockholm University

June 24, 2025



Stockholm
University

Fundamentals



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 use the posterior expected value ('average') of the parameter as your point estimate
 - ▶ The amount of posterior uncertainty given information in data is captured by the standard deviation of the parameter
- Prior distribution
 - ▶ In order to obtain a posterior distribution you need a prior distribution
 - ▶ Different priors give different posteriors for the same data



(Statistical) model

Model

We call

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

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 \mathbf{x}

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

Example (SAOM)

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

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

Inference

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

Likelihood

We can use the model We call

$$P(Data | \theta)$$

as a function of θ

$$L(\theta; Data) = \overbrace{P(Data | \theta)}^{const.}$$

now a function of θ

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



Deriving Bayes theorem

Example (Your sock drawer!)

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

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



In total 10 (single) socks

Probability of picking BIG and red sock

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

but also

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

We have equality

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

Deriving Bayes theorem

Example (Your sock drawer!)

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

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



Since

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

we can write **red** given **BIG**

$$P(\text{red} | \text{BIG}) = \frac{P(\text{BIG} | \text{red})P(\text{red})}{P(\text{BIG})}$$

in terms of **BIG** given **red**

Bayes

Bayes theorem

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

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

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

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

We observe: $y = \sum_{i < j} x_{ij} = 2$.

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

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

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

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

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

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

Posterior of tie-probability in Bernoulli graph

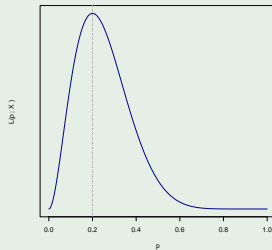
Example (Posterior distribution for p in Bernoulli graph (A))

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



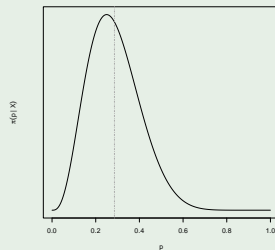
Prior

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



× Likelihood

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

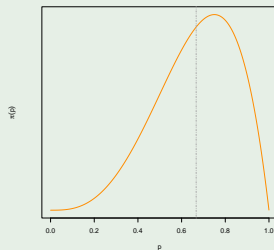


∝ Posterior

Posterior of tie-probability in Bernoulli graph

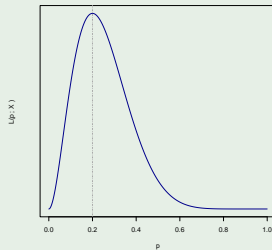
Example (Posterior distribution for p in Bernoulli graph (B))

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



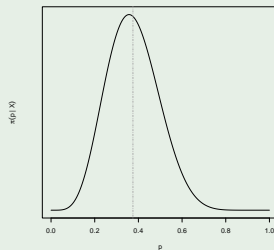
Prior

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



× Likelihood

$$p^{6-1}(1-p)^{10-1}$$



∝ Posterior

Posterior summaries

The posterior distribution of our parameters given data

$$P(\theta \mid \text{Data})$$

fully describes our uncertainty about the parameters given our observed data

We can summarise these using the expected values

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

as **point estimates**, and standard deviations

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

as measures of uncertainty



Posterior of tie-probability in Bernoulli graph

Example (Posterior distribution for p in Bernoulli graph)

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

We observe $y = \sum_{i < j} x_{ij} = 2$ and use $p \sim \text{Beta}(\alpha, \beta)$ prior \Rightarrow the posterior is $\text{Beta}(\alpha^*, \beta^*)$, where $\alpha^* = \alpha + 2$ and $\beta^* = \beta + 8$

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

and

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

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

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

Posterior for parameters in SAOM

For

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

we cannot obtain the posterior distribution

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

so easily

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

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



Convergence of MCMC

to target distribution $\pi(\theta \mid \text{Data})$



What does 'convergence' look like?

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

With $\theta \sim \text{Beta}(1, 1)$, we know $\theta \mid \mathbf{x} \sim \text{Beta}(16, 120)$

MCMC: iteratively update by

- (a) update θ to $\theta^* = \theta + U$
- (b) U is uniform on $(-\text{steplength}, \text{steplength})$
- (c) accept move with probability:

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

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

- (d) starting in $\theta = 1$



What does 'convergence' look like?

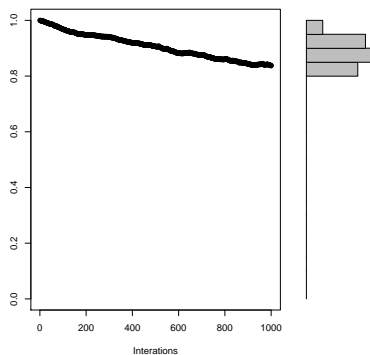


Figure: Steplength: 0.001 too small

What does 'convergence' look like?

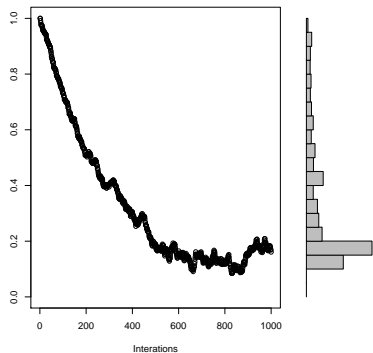


Figure: Steplength: 0.01 still too small

What does 'convergence' look like?

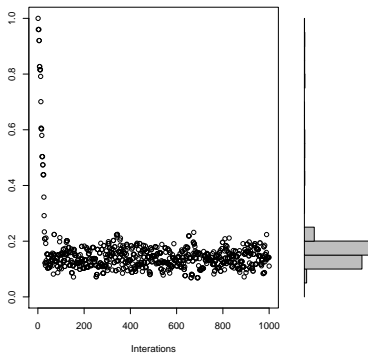


Figure: Steplength: 0.1 looking quite good

What does 'convergence' look like?

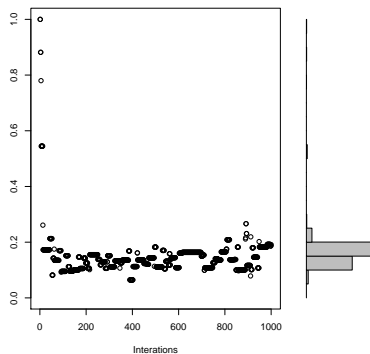
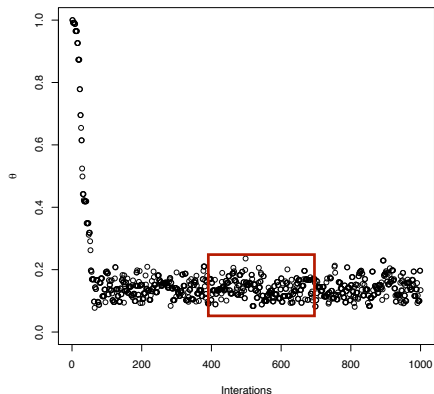
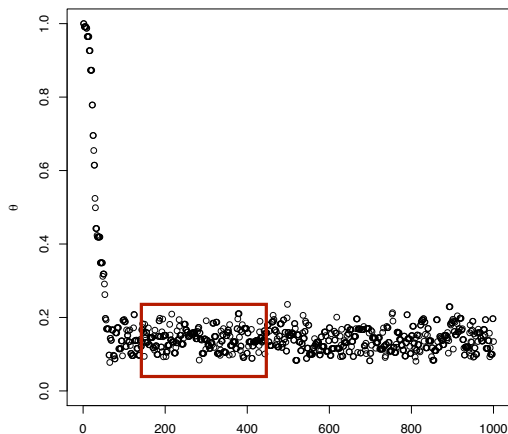


Figure: Steplength: 0.5 maybe too large

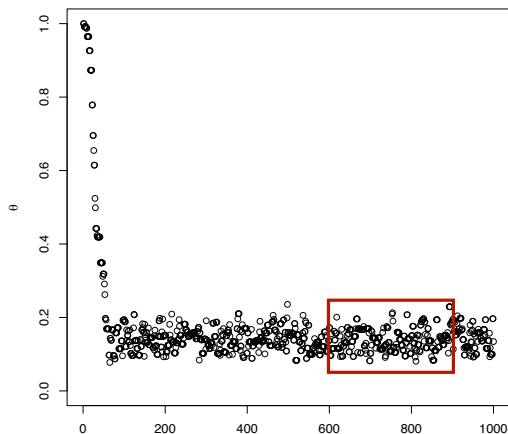
Converged to draws from the *same* distribution?



Converged to draws from the *same* distribution?



Converged to draws from the *same* distribution?



All unobservables have distributions - chain rule

Chain-rule of probability

We can define a model **observable** Data given **unobservable** variables \mathbf{y}

$$P(\text{Data} \mid \mathbf{y})$$

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

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

which gives a joint distribution (chain-rule)

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

The posterior distribution for θ is

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

All unobservables have distributions - chain rule

Example (Hierarchical SAOM)

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

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

and

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

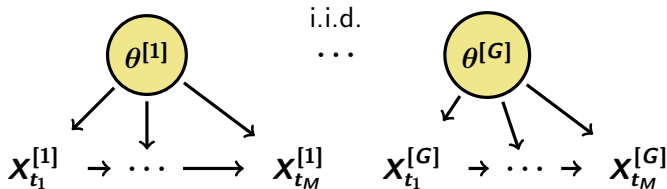
Interpretation: For a value on the **unknown** parameter μ (and Σ), we draw some **unknown** value θ_j , and then generate data from $\sim \text{SAOM}(\theta_j)$.

Aim: Find the 'true' value on μ



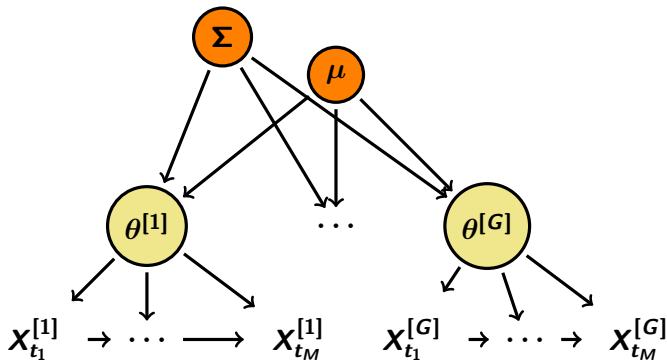
HSAOM (DAG)

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

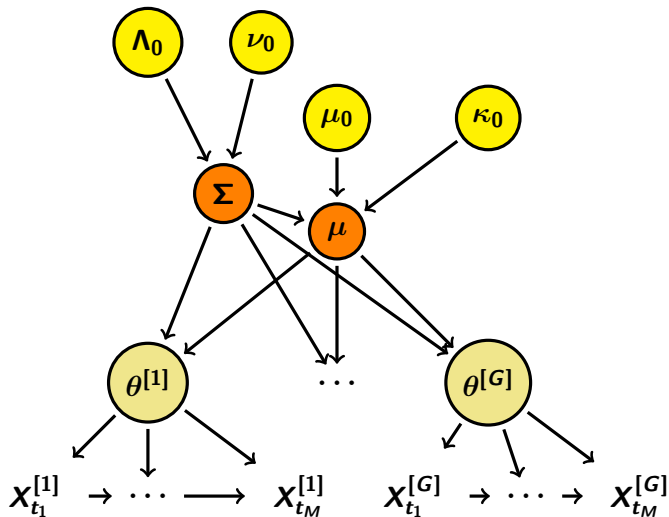


HSAOM (DAG)

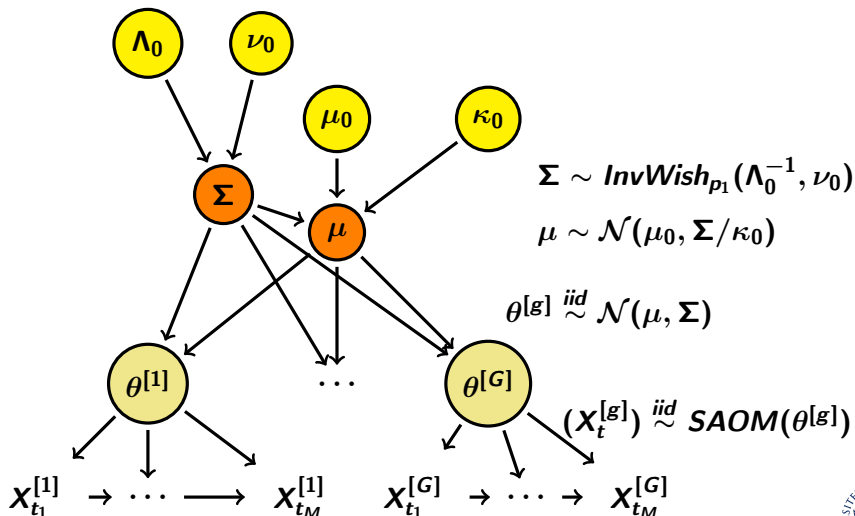
'population' parameters



HSAOM (DAG)



HSAOM (DAG)



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

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!



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 use the posterior expected value ('average') of the parameter as your point estimate
 - ▶ The amount of posterior uncertainty given information in data is captured by the standard deviation of the parameter
- Prior distribution
 - ▶ In order to obtain a posterior distribution you need a prior distribution
 - ▶ Different priors give different posteriors for the same data



Theorem

