Why does the Gibbs sampler work on hierarchical models?

Krys Łatuszyński (Warwick)

joint work with

Omiros Papaspiliopoulos (Barcelona)
Natesh Pillai (Harvard)
Gareth Roberts (Warwick)

(Why) does the Gibbs sampler work on hierarchical models?

Krys Łatuszyński (Warwick)

joint work with

Omiros Papaspiliopoulos (Barcelona)
Natesh Pillai (Harvard)
Gareth Roberts (Warwick)

Many years ago, when models were simple and people didn't even try understanding MCMC...

Many years ago, when models were simple and people didn't even try understanding MCMC...

Y

Many years ago, when models were simple and people didn't even try understanding MCMC...

$$\theta \longrightarrow X \longrightarrow Y$$

Many years ago, when models were simple and people didn't even try understanding MCMC...

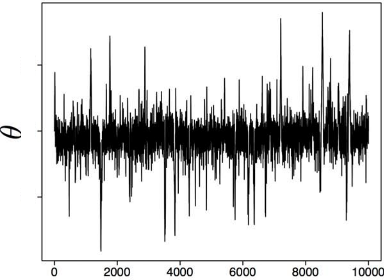
$\theta \longrightarrow X \longrightarrow Y$

Gibbs sampler to infer about θ

Many years ago, when models were simple and people didn't even try understanding MCMC...

$$\theta \longrightarrow X \longrightarrow Y$$

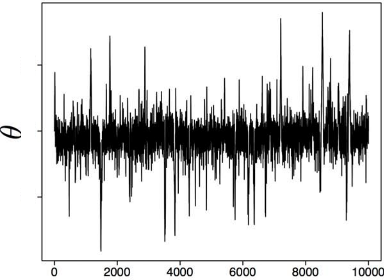
Gibbs sampler to infer about θ



Many years ago, when models were simple and people didn't even try understanding MCMC...

$\theta \longrightarrow X \longrightarrow Y$

Gibbs sampler to infer about θ



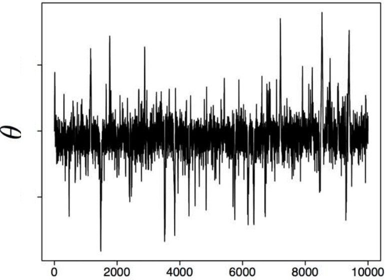
lucky → paper

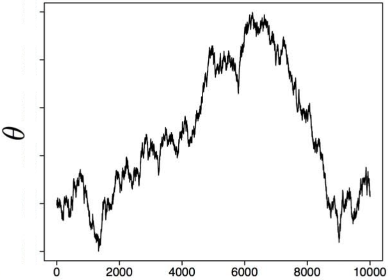
HAPPY END!

Many years ago, when models were simple and people didn't even try understanding MCMC...

$\theta \longrightarrow X \longrightarrow Y$

Gibbs sampler to infer about θ





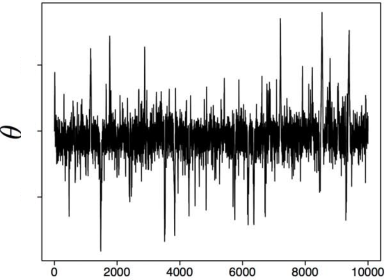
lucky → paper

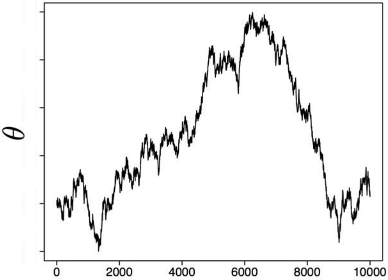
HAPPY END!

Many years ago, when models were simple and people didn't even try understanding MCMC...

$\theta \longrightarrow X \longrightarrow Y$

Gibbs sampler to infer about θ





lucky $\rightarrow$ paper

HAPPY END!

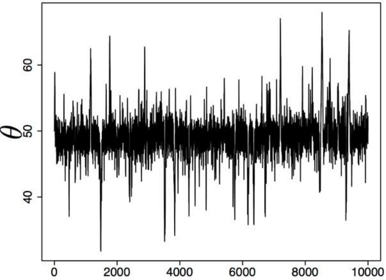
unlucky $\rightarrow$ no paper

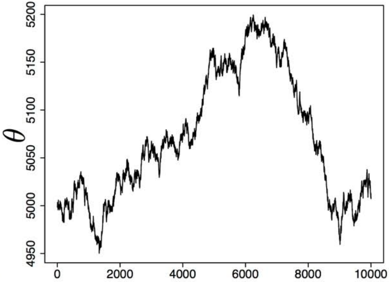
NO HAPPY END!

Many years ago, when models were simple and people didn't even try understanding MCMC...

$\theta \longrightarrow X \longrightarrow Y$

Gibbs sampler to infer about θ





lucky → paper

HAPPY END!

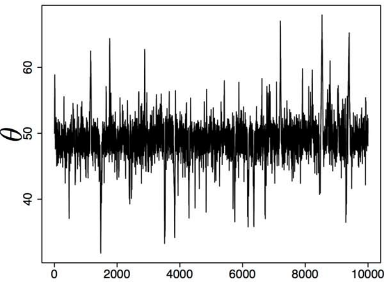
unlucky → no paper

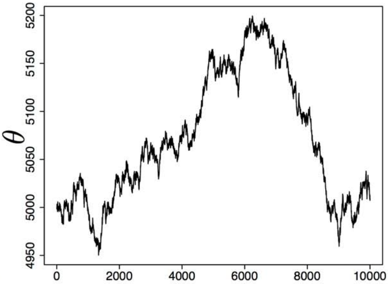
NO HAPPY END!

Many years ago, when models were simple and people didn't even try understanding MCMC...

$\theta \longrightarrow X \longrightarrow Y$

Gibbs sampler to infer about θ





unlucky → wrong paper (possibly)

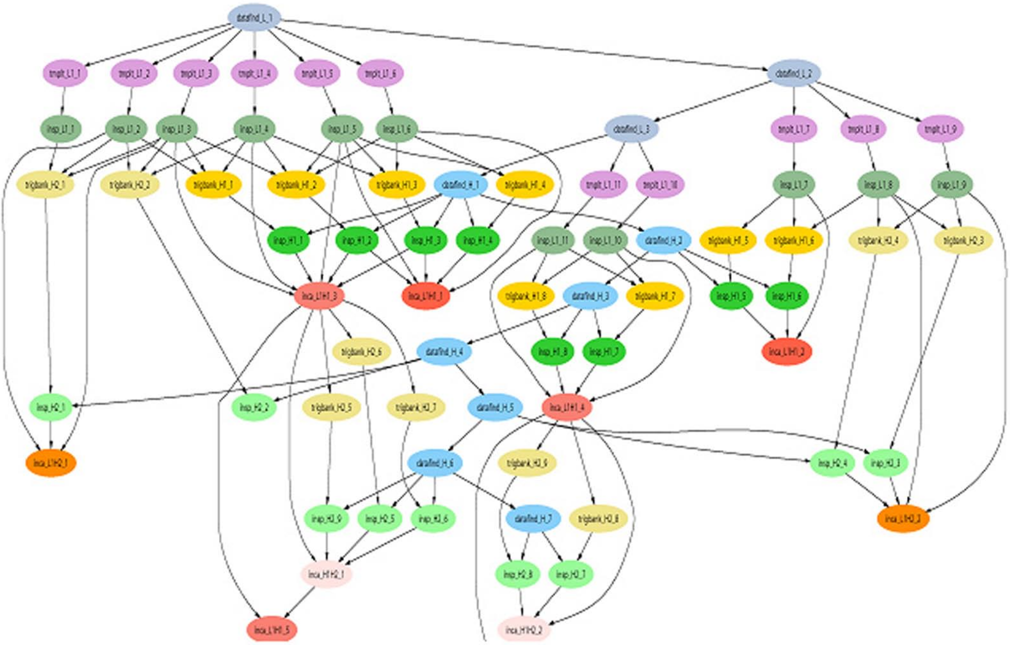
NO HAPPY END!

unlucky → no paper

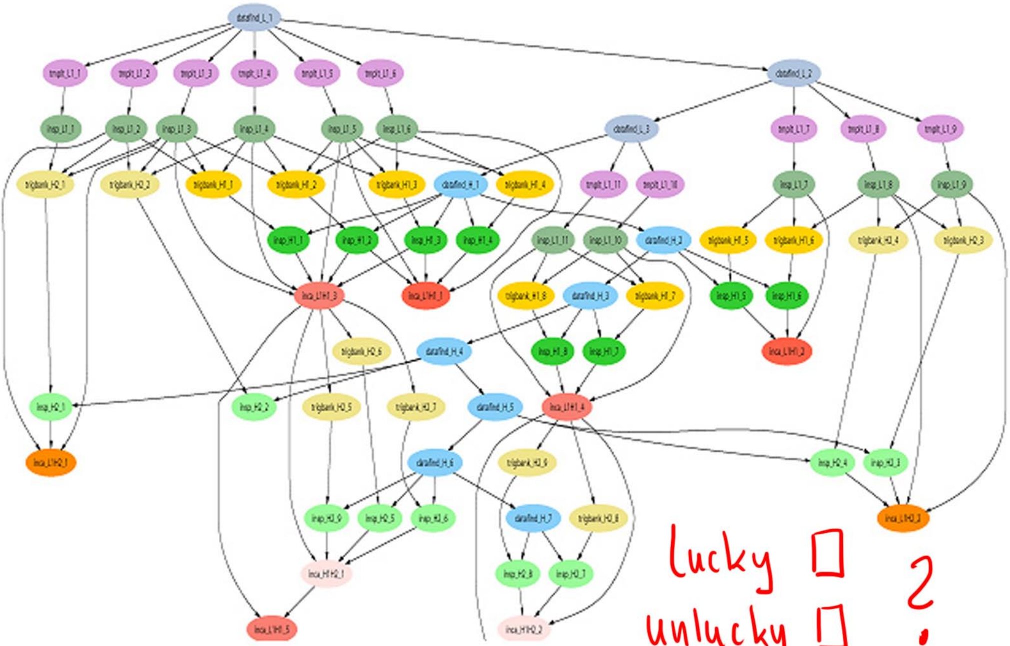
NO HAPPY END!

More recently, there was a statistician ...

More recently, there was a statistician ...



More recently, there was a statistician ...



lucky ☐

unlucky ☐ ?

Why hierarchical models?

Why hierarchical models?

- Natural for model building
- Local interpretation
- Local computation, often accesible to Gibbs sampler

Why hierarchical models?

- Natural for model building
- Local interpretation
- Local computation, often accesible to Gibbs sampler
- Often surprisingly good convergence of the Gibb sampler!

Why hierarchical models?

- Natural for model building
- Local interpretation
- Local computation, often accesible to Gibbs sampler
- Often surprisingly good convergence of the Gibb sampler! WHY?

Why hierarchical models?

- Natural for model building
- Local interpretation
- Local computation, often accesible to Gibbs sampler
- Often surprisingly good convergence of the Gibb sampler! WHY?

how to tell?

Convergence of MCMC

$$\| P^n(x,\cdot) - \pi(\cdot) \| \leq \rho^n V(x)$$

Convergence of MCMC

$$\| P^n(x,\cdot) - \pi(\cdot) \| \leq \rho^n V(x)$$

the vate of convergence < 1

Convergence of MCMC

the rate of convergence < 1

$$\| P^n(x,\cdot) - \pi(\cdot) \| \le \rho^n V(x)$$

GOOD NEWS: for many MCMC samplers we know ρ EXACTLY

Convergence of MCMC

the rate of convergence < 1

$$\| P^n(x,\cdot) - \pi(\cdot) \| \le \rho^n V(x)$$

GOOD NEWS: for many MCMC samplers we know ρ EXACTLY

THE BAD NEWS: for almost all of these $\rho = 1$

Convergence of MCMC

$$\| P^n(x,\cdot) - \pi(\cdot) \| \le \rho^n V(x)$$

the rate of convergence < 1

Coarse classification:

Convergence of MCMC

the rate of convergence ≤ 1

$$\| P^n(x,\cdot) - \pi(\cdot) \| \leq \rho^n V(x)$$

__Coarse classification:__

- uniformly ergodic (UE) if V bounded and $\rho < 1$

Convergence of MCMC

$$\| P^n(x,\cdot) - \pi(\cdot) \| \leq \rho^n V(x)$$

the rate of convergence ≤ 1

Coarse classification:

- uniformly ergodic (UE) if V bounded and $\rho < 1$
- geometrically ergodic (GE) if V unbounded and $\rho < 1$

Convergence of MCMC

$$\| P^n(x,\cdot) - \pi(\cdot) \| \leq \rho^n V(x)$$

the rate of convergence ≤ 1

Coarse classification:

- uniformly ergodic (UE) if V bounded and $\rho < 1$
- geometrically ergodic (GE) if V unbounded and $\rho < 1$
- not geometrically ergodic (N) otherwise

Convergence of MCMC

$$\| P^n(x,\cdot) - \pi(\cdot) \| \leq \rho^n V(x)$$

the rate of convergence ≤ 1

Coarse classification:

- uniformly ergodic (UE) if V bounded and $\rho < 1$
- geometrically ergodic (GE) if V unbounded and $\rho < 1$
- not geometrically ergodic (N) otherwise

GE and reversibility imply that CLTs hold for all $L^2(\pi)$ functions.

Convergence of MCMC

the rate of convergence <1

$$\| P^n(x,\cdot) - \pi(\cdot) \| \leq \rho^n V(x)$$

Coarse classification:

- uniformly ergodic (UE) if V bounded and $\rho < 1$
- geometrically ergodic (GE) if V unbounded and $\rho < 1$
- not geometrically ergodic (N) otherwise

GE and reversibility imply that CLTs hold for all $L^2(\pi)$ functions.

GE is essentially a necessary condition for this to hold.

Consider a version of the model

$$\Theta \rightarrow X \rightarrow Y$$

Consider a version of the model

$\Theta \longrightarrow X \longrightarrow Y$, where $\Theta \propto 1$

$$X = \Theta + \tilde{X} \qquad \tilde{X} \sim N(0, \sigma_X^2)$$

$$Y = X + Z \qquad Z \sim N(0, \sigma_Z^2)$$

Consider a version of the model

$\Theta \rightarrow X \rightarrow Y$, where

$$\Theta \propto 1$$
$$X = \Theta + \tilde{X} \qquad \tilde{X} \sim N(0, \sigma_X^2)$$
$$Y = X + Z \qquad Z \sim N(0, \sigma_Z^2)$$

Posterior is bivariate Gaussian.

The Gibbs sampler

$W = (\Theta, X)$

Consider a version of the model

$\Theta \to X \to Y$, where $\Theta \propto 1$

$X = \Theta + \tilde{X} \qquad \tilde{X} \sim N(0, \sigma_X^2)$

$Y = X + Z \qquad Z \sim N(0, \sigma_Y^2)$

Posterior is bivariate Gaussian.

The Gibbs sampler

$W = (\Theta, X)$ is a Gaussian autoregression

$$W_{t+1} = B W_t + \text{error}$$

is GE with convergence rate

$$\rho_c = \frac{\sigma_Y^2}{\sigma_X^2 + \sigma_Y^2}$$

(Roberts, Sahu 1997)

$$\Theta \propto 1$$

$$X = \Theta + \tilde{X} \qquad \tilde{X} \sim N(0, \sigma_x^2)$$

$$Y = X + Z \qquad Z \sim N(0, \sigma_y^2)$$

If we reparametrize

$$\begin{tikzcd} \Theta \arrow[dr] & & \\ & X \arrow[r] & Y \\ \tilde{X} \arrow[ur] & & \end{tikzcd}$$

$$\Theta \propto 1$$

$$X = \Theta + \tilde{X} \qquad \tilde{X} \sim N(0, \sigma_x^2)$$

$$Y = X + Z \qquad Z \sim N(0, \sigma_y^2)$$

If we reparametrize

$\theta \propto 1$

$X = \theta + \tilde{X} \qquad \tilde{X} \sim N(0, \sigma_X^2)$

$Y = X + Z \qquad Z \sim N(0, \sigma_Y^2)$

$$\begin{tikzcd}
\theta \arrow[dr] & & \\
 & X \arrow[r] & Y \\
\tilde{X} \arrow[ur] & &
\end{tikzcd}$$

and take the Gibbs sampler for $(\theta, \tilde{X})$, then it is GE with convergence rate

$$\rho_{NC} = \frac{\sigma_X^2}{\sigma_Y^2 + \sigma_X^2}$$

If we reparametrize

$\theta \propto 1$

$X = \theta + \tilde{X} \qquad \tilde{X} \sim N(0, \sigma_X^2)$

$Y = X + Z \qquad Z \sim N(0, \sigma_Y^2)$

$$
\begin{tikzcd}
\theta \arrow[dr] & & \\
& X \arrow[r] & Y \\
\tilde{X} \arrow[ur] & &
\end{tikzcd}
$$

and take the Gibbs sampler for $(\theta, \tilde{X})$, then it is GE with convergence rate

$$\rho_{NC} = \frac{\sigma_X^2}{\sigma_Y^2 + \sigma_X^2}$$

Non-centered parametrization (vs. centered)

Heuristic: centering works well for informative data

If we reparametrize $\theta \propto 1$

$$X = \theta + \tilde{X} \qquad \tilde{X} \sim N(0, \sigma_X^2)$$

$$Y = X + Z \qquad Z \sim N(0, \sigma_Y^2)$$

$$\begin{tikzcd}
\theta \arrow[dr] & & \\
& X \arrow[r] & Y \\
\tilde{X} \arrow[ur] & &
\end{tikzcd}$$

and take the Gibbs sampler for $(\theta, \tilde{X})$, then it is GE with convergence rate

$$\rho_{NC} = \frac{\sigma_X^2}{\sigma_Y^2 + \sigma_X^2}$$

Non-centered parametrization (vs. centered)

Heuristic: centering works well for informative data

non-centering for non-informative data

Consider the model

$$Y = X + Z$$

$$X = \theta + \tilde{X}$$

$$\theta \propto 1$$

Consider the model

$$Y = X + Z \leftarrow \text{Cauchy}$$

$$X = \theta + \tilde{X} \leftarrow \text{Normal}$$

$$\theta \propto 1$$

Consider the model

$Y = X + Z$ ← Cauchy

$X = \theta + \tilde{X}$ ← Normal

$\theta \propto 1$

Joint posterior $\propto$

$$\frac{e^{-(x-\theta)^2/2}}{1+(y-x)^2}$$

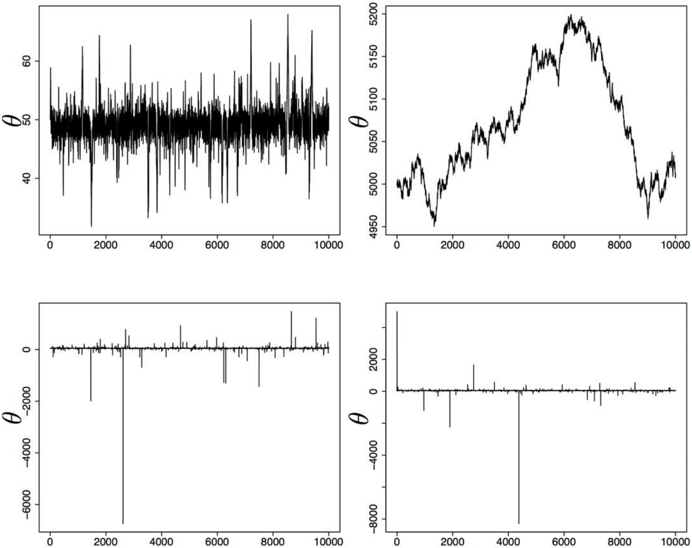
Consider the model

$$Y = X + Z \leftarrow \text{Cauchy}$$

$$X = \theta + \tilde{X} \leftarrow \text{Normal}$$

Joint posterior $\propto$

$$\frac{e^{-(x-\theta)^2/2}}{1+(y-x)^2}$$



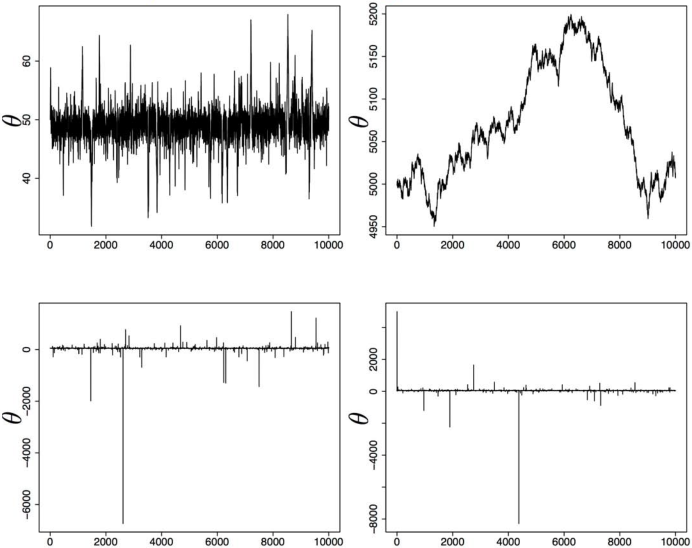
Consider the model

$$Y = X + Z \leftarrow \text{Cauchy}$$

$$X = \theta + \tilde{X} \leftarrow \text{Normal}$$

Joint posterior $\propto$

$$\frac{e^{-(x-\theta)^2/2}}{1+(y-x)^2}$$



← CENTERED ALGORITHM (N)

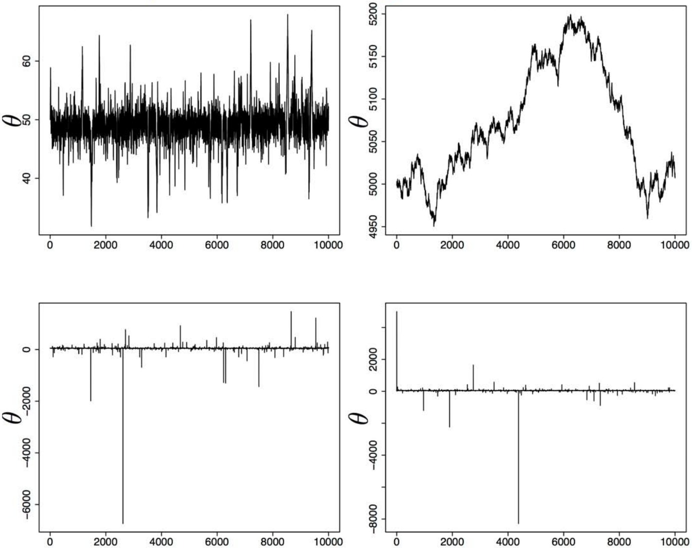
Consider the model

$$Y = X + Z \leftarrow \text{Cauchy}$$

$$X = \theta + \tilde{X} \leftarrow \text{Normal}$$

Joint posterior $\propto$

$$\frac{e^{-(x-\theta)^2/2}}{1+(y-x)^2}$$



$\leftarrow$ CENTERED ALGORITHM (N)

$\leftarrow$ NON-CENTERED ALGORITHM (UE)

More generally $\quad Y = X + Z \quad$ observation eqn

$X = \theta + \tilde{X} \quad$ hidden eqn

More generally $Y = X + Z$ observation eqn

$X = \theta + \tilde{X}$ hidden eqn

Error distributions for $Z, \tilde{X}$ are

(C) Cauchy, (N) Normal, (E) Double Exponential

and (L) Light tailed $e^{-|x|^{\beta}}$, $\beta > 2$

More generally $\quad Y = X + Z \quad$ observation eqn

$\quad X = \theta + \tilde{X} \quad$ hidden eqn

Error distributions for $Z, \tilde{X}$ are

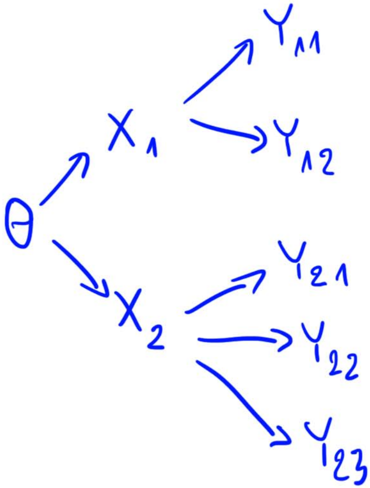
(C) Cauchy, (N) Normal, (E) Double Exponential

and (L) Light tailed $e^{-|x|^\beta}$, $\beta > 2$

Then the centered algorithm (θ, X) is

		Observation equation			
		C	E	G	L
Hidden eqn	C	U	U	U	U
	E	N	G/U	G	G
	G	N	G	G	G
	L	N	G	G	G

The results generalize:



Logistic regression with random effects.

Logistic regression with random effects.

$$Y_i \sim \text{Binom}(n_i, L(X_i)) \qquad 1 \le i \le m$$

$$X_i = \theta + Z_i$$

Logistic regression with random effects.

$$Y_i \sim \mathrm{Binom}(n_i, L(X_i)) \qquad 1 \le i \le m$$

$$X_i = \theta + Z_i$$

flat prior, symmetric about 0, $\quad L(x) = e^x / (1+e^x)$

Logistic regression with random effects.

$$Y_i \sim \mathrm{Binom}(n_i, L(X_i)) \qquad 1 \le i \le m$$

$$X_i = \theta + Z_i$$

flat prior, symmetric about 0, $L(x) = e^x / (1+e^x)$

Z	GIBBS SAMPLER		
	UE	GE	N
C			
E			
G			

Logistic regression with random effects.

$$Y_i \sim \mathrm{Binom}(n_i, L(X_i)) \qquad 1 \le i \le m$$

$$X_i = \theta + Z_i$$

flat prior, symmetric about 0, $L(x) = e^x / (1+e^x)$

Z	GIBBS SAMPLER		
	UE	GE	N
C	$\#\{Y_i > 0\} \geq m/2$ and $\#\{Y_i < n_i\} \geq m/2$	never	otherwise
E	$\#\{Y_i > a\} \geq m/2$ and $\#\{n_i - Y_i > a\} \geq m/2$	otherwise	never
G	never	always	never

Probit regression with random effects.

$$Y_i \sim \text{Binom}(n_i, L(X_i)) \qquad 1 \le i \le m$$

$$X_i = \theta + Z_i$$

flat prior, symmetric about 0, $L(x) = \Phi(x)$

Z	GIBBS SAMPLER		
	UE	GE	N
C	$\#\{Y_i > 0\} \geq m/2$ and $\#\{Y_i < n_i\} \geq m/2$	never	otherwise
E	$\#\{Y_i > 0\} \geq m/2$ and $\#\{n_i - Y_i > 0\} \geq m/2$	otherwise	never
G	never	always	never

Longer Hierarchies Centered parametrization

for $\quad \Theta_0 \to \Theta_1 \to \Theta_2 \cdots \Theta_k \to Y$

Longer Hierarchies Centered parametrization

for $\Theta_0 \rightarrow \Theta_1 \rightarrow \Theta_2 \cdots \Theta_k \rightarrow Y$ with $\epsilon_i \sim N(0, \sigma_i)$

Longer Hierarchies

Centered parametrization

for $\Theta_0 \to \Theta_1 \to \Theta_2 \cdots \Theta_k \to Y$ with $\varepsilon_i \sim N(0, \sigma_i)$

in particular for the RSGS

Longer Hierarchies Centered parametrization

for $\Theta_0 \to \Theta_1 \to \Theta_2 \cdots \Theta_k \to Y$ with $E_i \sim N(0, \sigma_i)$

in particular for the RSGS

$$E(\Theta^{(1)}) = \frac{k}{k+1}\Theta^{(0)} + \frac{1}{k+1} A \Theta^{(0)}$$

Longer Hierarchies   Centered parametrization

for $\Theta_0 \to \Theta_1 \to \Theta_2 \cdots \Theta_k \to Y$ with $\varepsilon_i \sim N(0, \sigma_i)$

in particular for the RSGS

$$E(\Theta^{(1)}) = \frac{k}{k+1}\Theta^{(0)} + \frac{1}{k+1} A\Theta^{(0)}$$ where

$$A = \begin{pmatrix} 0 & 1 & \dots & & \\ 1-\rho_1 & 0 & \rho_1 & \dots & \\ 0 & 1-\rho_2 & 0 & \rho_2 & \dots \\ \dots & \dots & \dots & \dots & \dots \\ & \dots & 1-\rho_{k-1} & 0 & \rho_{k-1} \\ & & \dots & 1-\rho_k & 0 \end{pmatrix}$$

where $\rho_i = \sigma_i^2 / (\sigma_i^2 + \sigma_{i+1}^2)$

- The principal eigenvalue of A determines the RSGS convergence rate

- The principal eigenvalue of A determines the RSGS convergence rate
- The principal (left) eigenvector a normalised has the interpretation as the quasi-stationary vector of a Markov chain with transition matrix A and absorption from K.

- The principal eigenvalue of A determines the RSGS convergence rate
- The principal (left) eigenvector a normalised has the interpretation as the quasi-stationary vector of a Markov chain with transition matrix A and absorption from K.
- A Lyapunov drift condition is of the form $V(\theta) = \| \sum_{i=0}^{k} a_i \theta_i \|^2 + 1$

More general errors:

suppose $E_i \sim f_i(\cdot)$

where $f_i(x) \propto \exp\{-|x|^{\beta_i}\}$

and $2 \leq \beta_1 \leq \beta_2 \leq \dots \leq \beta_K$

Then the RSGS is GE

More general errors:

suppose $E_i \sim f_i(\cdot)$

where $f_i(x) \propto \exp\{-|x|^{\beta_i}\}$

and $2 \leq \beta_1 \leq \beta_2 \leq \dots \leq \beta_K$

Then the RSGS is GE

Remark: We generally obtain GE if the tails of the errors are lightest „close to the data"

Comments on proofs:

Comments on proofs:

- We collapse the analysis on a bivariate Gibbs sampler, e.g $\Theta \rightarrow X \rightarrow Y$

Comments on proofs:

- We collapse the analysis on a bivariate Gibbs sampler, e.g $\Theta \to X \to Y$
- We conclude the properties of the bivariate Gibbs by looking at one component + Markov de-initializing processes argument.

Comments on proofs:

- We collapse the analysis on a bivariate Gibbs sampler, e.g $\Theta \to X \to Y$
- We conclude the properties of the bivariate Gibbs by looking at one component + Markov de-initializing processes argument.
- To deal with the one component of the bivariate Gibbs sampler, we develop a general theory

Comments on proofs:

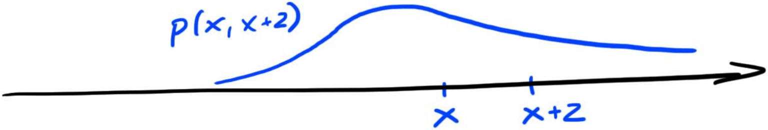
- We collapse the analysis on a bivariate Gibbs sampler, e.g $\Theta \to X \to Y$
- We conclude the properties of the bivariate Gibbs by looking at one component + Markov de-initializing processes argument.
- To deal with the one component of the bivariate Gibbs sampler, we develop a general theory of random walk like tail behaviour of Markov chains

Random walk like behaviour in the tails

P - transition kernel $\quad p(x,y)$ - transition density

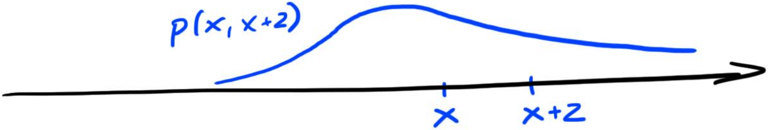
Random walk like behaviour in the tails

P - transition kernel $\quad p(x,y)$ - transition density



Random walk like behaviour in the tails

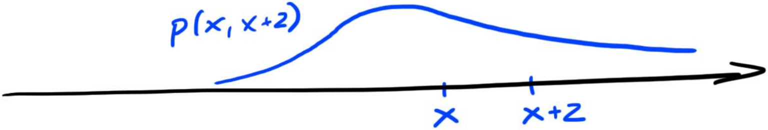
P - transition kernel $\quad p(x,y)$ - transition density



$$\lim_{x\to\infty} p(x, x+z) =: q(z)$$

Random walk like behaviour in the tails

P - transition kernel $p(x,y)$ - transition density

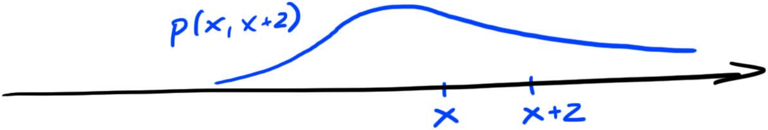


$$\lim_{x\to\infty} p(x, x+z) =: q(z)$$

← provides a full characterization of the Markov chain!

Random walk like behaviour in the tails

P - transition kernel $\quad p(x,y)$ - transition density



$$\lim_{x\to\infty} p(x, x+z) =: q(z)$$

← provides a full characterization of the Markov chain!

For one component of a bivariate Gibbs sampler on a graphical model,

q typically exists!!!

The characterization via q

The characterization via q

Thm 1 If P is reversible and q exists, then

$$q(z) = e^{-cz} f(z)$$

The characterization via q

<u>Thm 1</u> If P is reversible and q exists, then

$$q(z) = e^{-cz} f(z)$$

constant ≥ 0

symmetric function

The characterization via q

Thm 1 If P is reversible and q exists, then

$$q(z) = e^{-cz} f(z)$$

constant ≥ 0 symmetric function

or $q(z) = 0$ for all $z > 0$.

The characterization via q

Thm 1 If P is reversible and q exists, then

$$q(z) = e^{-cz} f(z)$$

constant ≥ 0 symmetric function

or $q(z) = 0$ for all $z > 0$.

(We do allow for $\int q(z)\,dz < 1$)

The characterization via q

Thm 1 If P is reversible and q exists, then

$$q(z) = e^{-cz} f(z)$$

constant ≥ 0 symmetric function

or $q(z) = 0$ for all $z > 0$.

(We do allow for $\int q(z)dz < 1$)

Thm 2
$\int q(z)dz > 0$.
+reversibility

$c = 0 \iff \pi$ has heavy tails

$c > 0 \iff \pi$ has exponential tails

$q(z) = 0 \iff \pi$ has light tails

$$q(z) = e^{-cz} f(z)$$

<u>Thm 3</u> If P is reversible, $c = 0$, $\int q = 1$

then P is NOT geometrically ergodic.

$$q(z) = e^{-cz} f(z)$$

Thm 3 If P is reversible, $c=0$, $\int q = 1$ then P is NOT geometrically ergodic.

Thm 4 Define $Q \sim q/m_q$ where $m_q = \int q(z)\,dz$

If $EQ \in [-\infty, 0)$ then P is geometrically ergodic.

(+ regularity conditions)

$$q(z) = e^{-cz} f(z)$$

Thm 3 If P is reversible, $c=0$, $\int q = 1$ then P is NOT geometrically ergodic.

Thm 4 Define $Q \sim q/m_q$ where $m_q = \int q(z)\,dz$

If $EQ \in [-\infty, 0)$ then P is geometrically ergodic.

(+ regularity conditions)

Cor If $q(z)$ is not symmetric then P is geometrically ergodic

$$q(z) = e^{-cz} f(z)$$

Thm 3 If P is reversible, $c=0$, $\int q = 1$ then P is NOT geometrically ergodic.

Thm 4 Define $Q \sim q/m_q$ where $m_q = \int q(z)dz$

If $EQ \in [-\infty, 0)$ then P is geometrically ergodic. (+ regularity conditions)

Cor If $q(z)$ is not symmetric then P is geometrically ergodic

Thm 5 If $m_q = \int q(z)dz < 1$ (+ regularity cond) then P is geometrically ergodic.

Conclusions:

- There are very good theoretical reasons why Gibbs sampling for the models considered in early 90s was so robust

Conclusions:

- There are very good theoretical reasons why Gibbs sampling for the models considered in early 90s was so robust
- Conditional independence structure and particular error distributions play crucial role in the stability.

Conclusions:

- There are very good theoretical reasons why Gibbs sampling for the models considered in early 90s was so robust
- Conditional independence structure and particular error distributions play crucial role in the stability.
- Not all models will be GE, but the theory tells us when there is a problem and how to solve it.

Conclusions:

- There are very good theoretical reasons why Gibbs sampling for the models considered in early 90s was so robust
- Conditional independence structure and particular error distributions play crucial role in the stability.
- Not all models will be GE, but the theory tells us when there is a problem and how to solve it.
- There is much work still to be done!