BAYESIAN HIERARCHICAL MODELS

Tim Mensinger

University of Bonn

1. *Bayesian Thinking*

2. *Hierarchical Models*

Bayesian Thinking

Probabilistic Modeling

Data: $Z_i = (y_i, X_i) \in \mathcal{Z} \quad (= \mathbb{R} \times \mathbb{R}^K)$

Probabilistic Modeling

Data: $Z_i = (y_i, X_i) \in \mathcal{Z} \quad (= \mathbb{R} \times \mathbb{R}^K)$

Model: Probability distribution p on $\mathcal{Z}$

Probabilistic Modeling

Data: $Z_i = (y_i, X_i) \in \mathcal{Z} \quad (= \mathbb{R} \times \mathbb{R}^K)$

Model: Probability distribution p on $\mathcal{Z}$

In Practice: Restrict attention to $p \in \mathcal{M}$

Probabilistic Modeling

Data: $Z_i = (y_i, X_i) \in \mathcal{Z} \quad (= \mathbb{R} \times \mathbb{R}^K)$

Model: Probability distribution p on $\mathcal{Z}$

In Practice: Restrict attention to $p \in \mathcal{M}$

Parameterize: $\mathcal{M} \leftrightarrow \Theta \subset \mathbb{R}^d$

Probabilistic Modeling

Data: $Z_i = (y_i, X_i) \in \mathcal{Z} \quad (= \mathbb{R} \times \mathbb{R}^K)$

Model: Probability distribution p on $\mathcal{Z}$

In Practice: Restrict attention to $p \in \mathcal{M}$

Parameterize: $\mathcal{M} \leftrightarrow \Theta \subset \mathbb{R}^d$

Example: Family of normal distributions

Probabilistic Modeling

Data: $Z_i = (y_i, X_i) \in \mathcal{Z} \quad (= \mathbb{R} \times \mathbb{R}^K)$

Model: Probability distribution p on $\mathcal{Z}$

In Practice: Restrict attention to $p \in \mathcal{M}$

Parameterize: $\mathcal{M} \leftrightarrow \Theta \subset \mathbb{R}^d$

Example: Family of normal distributions
(See blackboard)

Probabilistic Modeling

Data: $Z_i = (y_i, X_i) \in \mathcal{Z} \quad (= \mathbb{R} \times \mathbb{R}^K)$

Model: Probability distribution p on $\mathcal{Z}$

In Practice: Restrict attention to $p \in \mathcal{M}$

Parameterize: $\mathcal{M} \leftrightarrow \Theta \subset \mathbb{R}^d$

Example: Family of normal distributions
(See ~~blackboard~~ whiteboard)

Schools of Thought

True Model: $\mathcal{M}_{\theta_0}$ for $\theta_0 \in \Theta$

Schools of Thought

True Model: $\mathcal{M}_{\theta_0}$ for $\theta_0 \in \Theta$
Frequentist: θ_0 fixed quantity

Schools of Thought

True Model: $\mathcal{M}_{\theta_0}$ for $\theta_0 \in \Theta$

Frequentist: θ_0 fixed quantity

Bayesian: θ_0 fixed quantity

Schools of Thought

True Model: $\mathcal{M}_{\theta_0}$ for $\theta_0 \in \Theta$

Frequentist: θ_0 fixed quantity

Bayesian: θ_0 fixed quantity, but model uncertainty by imposing a prob. distr. on Θ

Bayes' Theorem

$$p(\theta \mid \text{data}) = \frac{p(\text{data} \mid \theta)p(\theta)}{p(\text{data})} \propto p(\text{data} \mid \theta)p(\theta)$$

Bayes' Theorem

$$p(\theta \mid \text{data}) = \frac{p(\text{data} \mid \theta)p(\theta)}{p(\text{data})} \propto p(\text{data} \mid \theta)p(\theta)$$

$$\textit{posterior} = \frac{\textit{likelihood} \times \textit{prior}}{\textit{evidence}} \propto \textit{likelihood} \times \textit{prior}$$

Solving for the Posterior Analytically

Setting: $\{y_i : i = 1, \ldots, n\}$ with $y_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$
and σ^2 known

Solving for the Posterior Analytically

Setting: $\{y_i : i = 1, \ldots, n\}$ with $y_i \stackrel{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$
and σ^2 known

Likelihood: $p(y \mid \mu) = \prod_i p(y_i \mid \mu)$

Solving for the Posterior Analytically

Setting: $\{y_i : i = 1, \ldots, n\}$ with $y_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$
and σ^2 known

Likelihood: $p(y \mid \mu) = \prod_i p(y_i \mid \mu)$

Prior: $p(\mu)$

Solving for the Posterior Analytically

Setting: $\{y_i : i = 1, \dots, n\}$ with $y_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$
and σ^2 known

Likelihood: $p(y \mid \mu) = \prod_i p(y_i \mid \mu)$

Prior: $p(\mu)$

Posterior: $p(\mu \mid y) \propto p(y \mid \mu) p(\mu)$

Solving for the Posterior Analytically

Setting: $\{y_i : i = 1, \ldots, n\}$ with $y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and σ^2 known

Likelihood: $p(y \mid \mu) = \prod_i p(y_i \mid \mu)$

Prior: $p(\mu)$

Posterior: $p(\mu \mid y) \propto p(y \mid \mu) p(\mu)$

Goal: Infer distribution of $\mu \mid y$

Solving for the Posterior Analytically

Setting: $\{y_i : i = 1, \ldots, n\}$ with $y_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$
and σ^2 known

Likelihood: $p(y \mid \mu) = \prod_i p(y_i \mid \mu)$

Prior: $p(\mu)$

Posterior: $p(\mu \mid y) \propto p(y \mid \mu) p(\mu)$

Goal: Infer distribution of $\mu \mid y$
Why?

Uninformative Prior

Let $p(\mu) \propto 1$

Uninformative Prior

Let $p(\mu) \propto 1$
Note that $p(y \mid \mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2\right)$

Uninformative Prior

Let $p(\mu) \propto 1$
Note that $p(y \mid \mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2\right)$
Hence

$$\begin{aligned} p(\mu \mid y) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2/n}(\mu - \bar{y})^2\right) \end{aligned}$$

Uninformative Prior

Let $p(\mu) \propto 1$
Note that $p(y \mid \mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2\right)$
Hence

$$
\begin{aligned}
p(\mu \mid y) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2/n}(\mu - \bar{y})^2\right)
\end{aligned}
$$

$$\implies \mu \mid y \sim \mathcal{N}\left(\bar{y}, \sigma^2/n\right)$$

Conjugate Prior

Let $\mu \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$

Conjugate Prior

Let $\mu \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$
Then

$$p(\mu \mid y) \propto p(y \mid \mu) p(\mu)$$

Conjugate Prior

Let $\mu \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$
Then

$$
\begin{aligned}
p(\mu \mid y) &\propto p(y \mid \mu) p(\mu) \\
&\propto \exp\left(\frac{-n}{2\sigma^2}(\mu - \bar{y})^2\right) \exp\left(\frac{-1}{2\sigma_0^2}(\mu - \mu_0)^2\right)
\end{aligned}
$$

Conjugate Prior

Let $\mu \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$
Then

$$\begin{aligned} p(\mu \mid y) &\propto p(y \mid \mu) p(\mu) \\ &\propto \exp\left(\frac{-n}{2\sigma^2}(\mu - \bar{y})^2\right) \exp\left(\frac{-1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma_\mu^2}(\mu - \bar{\mu})^2\right) \end{aligned}$$

Conjugate Prior

Let $\mu \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$
Then

$$
\begin{aligned}
p(\mu \mid y) &\propto p(y \mid \mu) p(\mu) \\
&\propto \exp\left(\frac{-n}{2\sigma^2}(\mu - \bar{y})^2\right) \exp\left(\frac{-1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma_\mu^2}(\mu - \bar{\mu})^2\right)
\end{aligned}
$$

$$
\implies \mu \mid y \sim \mathcal{N}\left(\bar{\mu}, \sigma_\mu^2\right)
$$

Conjugate Prior

$$\mu \mid y \sim \mathcal{N}\left(\bar{\mu}, \sigma_{\mu}^{2}\right), \text{with}$$

Conjugate Prior

$$\mu \mid y \sim \mathcal{N}\left(\bar{\mu}, \sigma_\mu^2\right) \text{, with}$$

$$\sigma_\mu^2 = \left(\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2}\right)^{-1}$$

Conjugate Prior

$$\mu \mid y \sim \mathcal{N}\left(\bar{\mu}, \sigma_\mu^2\right), \text{with}$$

$$\sigma_\mu^2 = \left(\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2}\right)^{-1}$$

$$\bar{\mu} = \sigma_\mu^2 \left(\frac{1}{\sigma^2/n}\boldsymbol{\bar{y}} + \frac{1}{\sigma_0^2}\boldsymbol{\mu_0}\right)$$

Conjugate Prior

$$\mu \mid y \sim \mathcal{N}\left(\bar{\mu}, \sigma_{\mu}^{2}\right) \text{, with}$$

$$\sigma_{\mu}^{2} = \left(\frac{1}{\sigma^{2}/n} + \frac{1}{\sigma_{0}^{2}}\right)^{-1}$$

$$\bar{\mu} = \sigma_{\mu}^{2}\left(\frac{1}{\sigma^{2}/n}\boldsymbol{\bar{y}} + \frac{1}{\sigma_{0}^{2}}\boldsymbol{\mu_{0}}\right)$$

$$= \alpha\boldsymbol{\bar{y}} + (1-\alpha)\boldsymbol{\mu_{0}}$$

Sampling from the Posterior

Problem: $p(\theta \mid \text{data}) = \textbf{\textit{const.}}\; p(\text{data} \mid \theta) p(\theta)$

Sampling from the Posterior

Problem: $p(\theta \mid \text{data}) = \boldsymbol{\mathit{const.}}\; p(\text{data} \mid \theta)p(\theta)$

Solution: Sampling?

Sampling from the Posterior

Object of interest: $\theta \mid \text{data}$

Sampling from the Posterior

Object of interest: $\theta \mid \text{data}$
Quantity of interest: $\mathbb{E}\left[\, h(\theta) \mid \text{data} \,\right] =: \mathbb{E}_\theta[h]$

Sampling from the Posterior

Object of interest: $\theta \mid \text{data}$

Quantity of interest: $\mathbb{E}\left[\, h(\theta) \mid \text{data} \,\right] =: \mathbb{E}_\theta[h]$

Estimation: Let $\theta^{(1)}, \ldots, \theta^{(n)} \overset{\text{iid}}{\sim} p(\theta \mid \text{data})$, then

$$\frac{1}{\sqrt{n}} \left(\sum_i h(\theta^{(i)}) - \mathbb{E}_\theta[h] \right) \xrightarrow{\text{d}} \mathcal{N}\left(0, \omega\right)$$

Sampling from the Posterior

But how do we sample from $p(\theta \mid \text{data})$?

Sampling from the Posterior

But how do we sample from $p(\theta \mid \text{data})$?
Problems: $p(\theta \mid \text{data})$ might be

Sampling from the Posterior

But how do we sample from $p(\theta \mid \text{data})$?

Problems: $p(\theta \mid \text{data})$ might be

(i.) of unkown form

Sampling from the Posterior

But how do we sample from $p(\theta \mid \text{data})$?

Problems: $p(\theta \mid \text{data})$ might be

(i.) of unkown form

(ii.) very complex

Sampling from the Posterior

But how do we sample from $p(\theta \mid \text{data})$?

Problems: $p(\theta \mid \text{data})$ might be

(i.) of unkown form

(ii.) very complex

(iii.) only known up to an integration const.

Markov Chain Monte Carlo

Algorithm Metropolis-Hastings (1953, 1970)

Input: (π, q, T) = (target, proposal, no. of samples)

1: initialize x_0 in supp q

2: **for** $t = 0, \ldots, T$ **do**

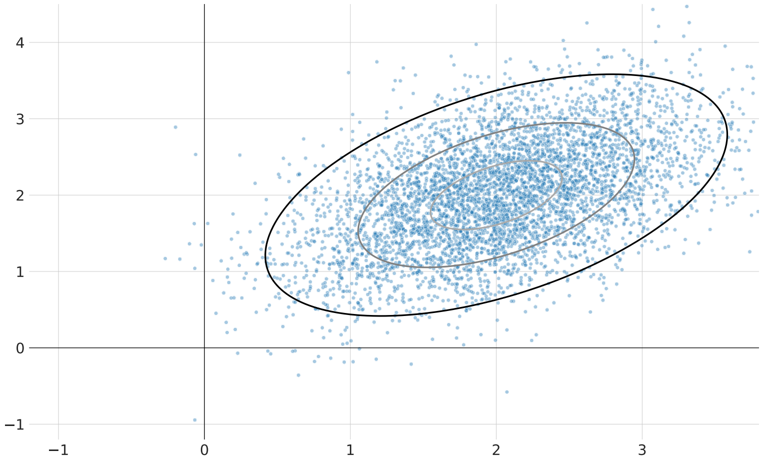
3: candidate: $y \sim q(\cdot \mid x_t)$

4: acceptance prob.: $\mathcal{A} \leftarrow \min\left\{\frac{\pi(y)}{\pi(x_t)}\frac{q(y \mid x_t)}{q(x_t \mid y)}, 1\right\}$

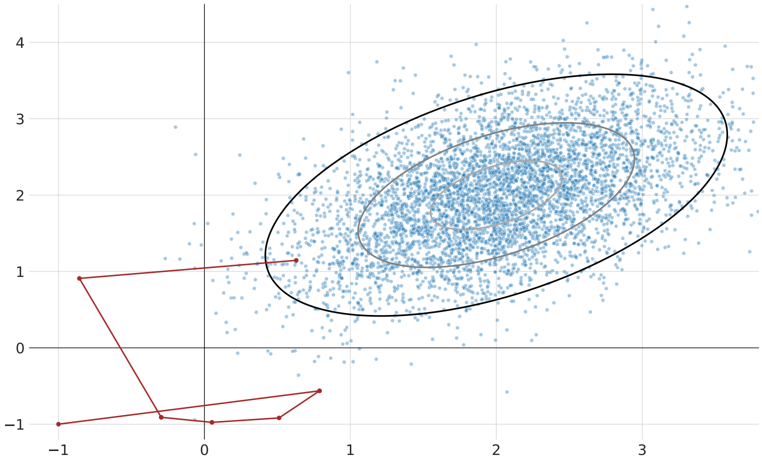
5: update: $x_{t+1} \leftarrow \begin{cases} y & \text{,with prob. } \mathcal{A} \\ x_t & \text{,with remaining prob.} \end{cases}$

6: **return** $\{x_t : t = 1, \ldots, T\}$

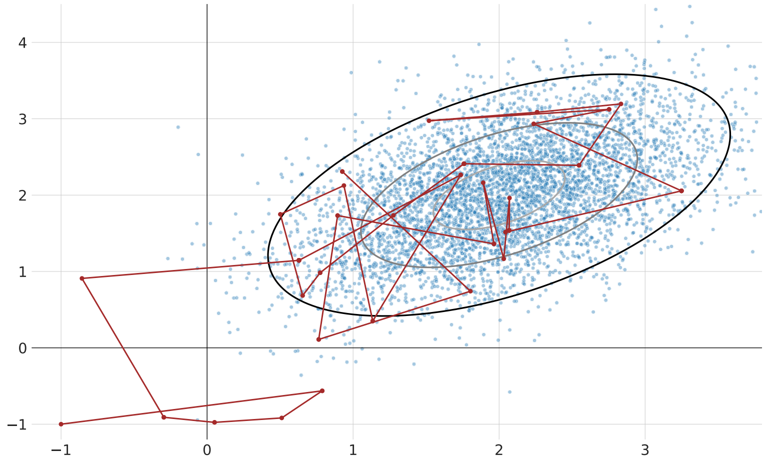
Markov Chain Monte Carlo



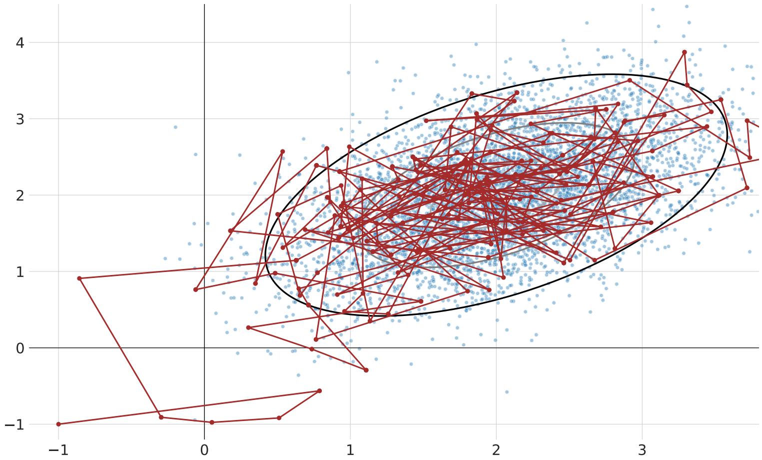
Markov Chain Monte Carlo



Markov Chain Monte Carlo



Markov Chain Monte Carlo



Hierarchical Models

Structure of HM - Setup

Hierarchical Data:

Structure of HM - Setup

Hierarchical Data:

Individual Level: (y_i, x_i) for $i = 1, \ldots, n$

Structure of HM - Setup

Hierarchical Data:

- *Individual Level:* (y_i, x_i) for $i = 1, \ldots, n$
- *Group Level:* u_j for $j = 1, \ldots, J$

Structure of HM - Setup

Hierarchical Data:

Individual Level: (y_i, x_i) for $i = 1, \ldots, n$

Group Level: u_j for $j = 1, \ldots, J$

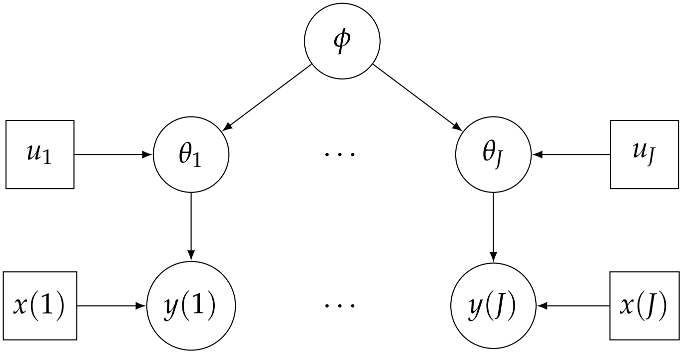
Example:

Test Outcome: y_i

Parental Income: x_i

Num. of Teachers: u_j

Structure of HM



The Prior Revisited

Before:

Model: $p(\text{data} \mid \theta)$

Prior: $p(\theta)$

The Prior Revisited

Before:

Model: $p(\text{data} \mid \theta)$

Prior: $p(\theta)$

Now:

Model: $p(\text{data} \mid \theta, \phi) = p(\text{data} \mid \theta)$

Prior: $p(\theta \mid \phi)$

Hyperprior: $p(\phi)$

The Posterior Revisited

Posterior:

$$p(\theta, \phi \mid \text{data}) \propto p(\text{data} \mid \theta) p(\theta, \phi)$$

The Posterior Revisited

Posterior:

$$
\begin{aligned}
p(\theta, \phi \mid \text{data}) &\propto p(\text{data} \mid \theta)p(\theta, \phi) \\
&\propto p(\text{data} \mid \theta)p(\theta \mid \phi)p(\phi)
\end{aligned}
$$

The Posterior Revisited

Posterior:

$$
\begin{aligned}
p(\theta, \phi \mid \text{data}) &\propto p(\text{data} \mid \theta) p(\theta, \phi) \\
&\propto p(\text{data} \mid \theta) p(\theta \mid \phi) p(\phi) \\
p(\phi \mid \text{data}) &\propto \int p(\theta, \phi \mid \text{data}) \mathrm{d}\theta
\end{aligned}
$$

Varying Slopes, Varying Intercepts

Setup: Individual i in group j

Varying Slopes, Varying Intercepts

Setup: Individual i in group j

Individual Level: $y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$

Varying Slopes, Varying Intercepts

Setup: Individual i in group j

Individual Level: $y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$

Group Level: $\begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} = \gamma_0 + \gamma u_j + \eta_j$

Varying Slopes, Varying Intercepts

Setup: Individual i in group j

Individual Level: $y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$

Group Level: $\begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} = \gamma_0 + \gamma u_j + \eta_j$

Priors on: $\gamma_0, \gamma, \epsilon_i, \eta_j$

Varying Slopes, Varying Intercepts

Setup: Individual i in group j

Individual Level: $y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$

Group Level: $\begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} = \gamma_0 + \gamma u_j + \eta_j$

Priors on: $\gamma_0, \gamma, \epsilon_i, \eta_j$

https://github.com/timmens/
bayesian-hierarchical-models

http://mfviz.com/hierarchical-models/