Hierarchical Bayesian modeling

Tom Loredo
Cornell Center for Astrophysics and Planetary Science
http://www.astro.cornell.edu/staff/loredo/bayes/

SAMSI ASTRO — 19 Oct 2016
1970 baseball averages

Efron & Morris looked at batting averages of baseball players who had $N = 45$ at-bats in May 1970 — ‘large’ $N$ & includes Roberto Clemente (outlier!)

*Red* = $n/N$ maximum likelihood estimates of true averages  
*Blue* = Remainder of season, $N_{rmdr} \approx 9N$

\begin{center}
\begin{tikzpicture}
    \fill[cyan, opacity=0.2] (0.265, 0.2) rectangle (0.3, 0.4);
    \draw[->, cyan] (0.265, 0.2) -- (0.265, 0.4);
    \node at (0.265, 0.3) {Early season};
    \node at (0.265, 0.27) {Shrinkage};
    \node at (0.3, 0.3) {RMSE = 0.148};
    \node at (0.265, 0.2) {RMSE = 0.277};
\end{tikzpicture}
\end{center}

*Cyan* = James-Stein estimator: nonlinear, correlated, biased  
But *better*!
Theorem (independent Gaussian setting): In dimension $d > 3$, shrinkage estimators always beat independent MLEs in terms of expected RMS error.

"The single most striking result of post-World War II statistical theory" — Brad Efron

Lines show closer estimate
Shrinkage closer 15/18
Theorem (independent Gaussian setting): In dimension $\geq 3$, shrinkage estimators always beat independent MLEs in terms of expected RMS error.

“The single most striking result of post-World War II statistical theory”
— Brad Efron
All 18 players are *humans playing baseball*—they are members of a population, not arbitrary, unrelated binomial random number generators!

In the absence of data about player $i$, we may use the performance of the other players to guide a guess about that player’s performance—they provide *indirect evidence* (Efron) about player $i$

But information that is relevant in the absence of data for $i$ remains relevant when we additionally obtain that data; shrinkage estimators account for this

There is “mustering and *borrowing of strength*” (Tukey) across the population

*Hierarchical Bayesian modeling* is the most flexible framework for generalizing this lesson; *empirical Bayes* is an approximate version with a straightforward frequentist interpretation
Agenda

1 Basic Bayes recap

2 Key idea in a nutshell

3 Going deeper
   Joint distributions and DAGs
   Conditional dependence/independence
   Example: Binomial prediction
   Beta-binomial model
   Point estimation and shrinkage
   Gamma-Poisson model & Stan
   Algorithms
Bayesian inference in one slide

*Probability as generalized logic*

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

**Bayes’s theorem**

\[ p(\text{Hypothesis} \mid \text{Data}) \propto p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis}) \]

Data *change* the support for a hypothesis \( \propto \) ability of hypothesis to *predict* the data

**Law of total probability**

\[ p(\text{Hypotheses} \mid \text{Data}) = \sum p(\text{Hypothesis} \mid \text{Data}) \]

The support for a *compound/composite* hypothesis must account for all the ways it could be true
Bayes’s theorem

\( \mathcal{C} = \text{context, initial set of premises} \)

Consider \( P(H_i, D_{\text{obs}} | \mathcal{C}) \) using the product rule:

\[
P(H_i, D_{\text{obs}} | \mathcal{C}) = P(H_i | \mathcal{C}) P(D_{\text{obs}} | H_i, \mathcal{C})
= P(D_{\text{obs}} | \mathcal{C}) P(H_i | D_{\text{obs}}, \mathcal{C})
\]

Solve for the \textit{posterior probability} (expands the premises!):

\[
P(H_i | D_{\text{obs}}, \mathcal{C}) = P(H_i | \mathcal{C}) \frac{P(D_{\text{obs}} | H_i, \mathcal{C})}{P(D_{\text{obs}} | \mathcal{C})}
\]

Theorem holds for any propositions, but for hypotheses & data the factors have names:

\[\text{posterior} \propto \text{prior} \times \text{likelihood}\]

norm. const. \( P(D_{\text{obs}} | \mathcal{C}) = \text{prior predictive} \)
**Law of Total Probability (LTP)**

Consider exclusive, exhaustive \( \{B_i\} \) (\( C \) asserts one of them must be true),

\[
\sum_i P(A, B_i | C) = \sum_i P(B_i | A, C)P(A | C) = P(A | C)
\]

\[
= \sum_i P(B_i | C)P(A | B_i, C)
\]

If we do not see how to get \( P(A | P) \) directly, we can find a set \( \{B_i\} \) and use it as a “basis”—extend the conversation:

\[
P(A | C) = \sum_i P(B_i | C)P(A | B_i, C)
\]

If our problem already has \( B_i \) in it, we can use LTP to get \( P(A | C) \) from the joint probabilities—*marginalization*:

\[
P(A | C) = \sum_i P(A, B_i | C)
\]
Example: Take $A = D_{\text{obs}}$, $B_i = H_i$; then

$$P(D_{\text{obs}}|C) = \sum_i P(D_{\text{obs}}, H_i|C)$$

$$= \sum_i P(H_i|C) P(D_{\text{obs}}|H_i, C)$$

prior predictive for $D_{\text{obs}} = \text{Average likelihood for } H_i$
(a.k.a. *marginal likelihood*)
Parameter Estimation

**Problem statement**

\( \mathcal{C} = \text{Model } M \text{ with parameters } \theta \text{ (+ any add’l info)} \)

\( H_i = \text{statements about } \theta; \text{ e.g. } “\theta \in [2.5, 3.5],” \text{ or } “\theta > 0” \)

Probability for any such statement can be found using a probability density function (PDF) for \( \theta \):

\[
P(\theta \in [\theta, \theta + d\theta] | \cdots) = f(\theta) d\theta
\]

\[
= p(\theta | \cdots) d\theta
\]

**Posterior probability density**

\[
p(\theta | D, M) = \frac{p(\theta | M) \mathcal{L}(\theta)}{\int d\theta \ p(\theta | M) \mathcal{L}(\theta)}
\]
Summaries of posterior

- “Best fit” values:
  - Mode, \( \hat{\theta} \), maximizes \( p(\theta|D, M) \)
  - Posterior mean, \( \langle \theta \rangle = \int d\theta \theta p(\theta|D, M) \)

- Uncertainties:
  - Credible region \( \Delta \) of probability \( C \):
    \[
    C = P(\theta \in \Delta|D, M) = \int_{\Delta} d\theta p(\theta|D, M)
    \]
  - Highest Posterior Density (HPD) region has \( p(\theta|D, M) \) higher inside than outside
  - Posterior standard deviation, variance, covariances

- Marginal distributions
  - Interesting parameters \( \phi \), nuisance parameters \( \eta \)
  - Marginal dist’n for \( \phi \): \( p(\phi|D, M) = \int d\eta p(\phi, \eta|D, M) \)
Many Roles for Marginalization

Eliminate nuisance parameters

\[ p(\phi|D, M) = \int d\eta \ p(\phi, \eta|D, M) \]

Propagate uncertainty

Model has parameters \( \theta \); what can we infer about \( F = f(\theta) \)?

\[ p(F|D, M) = \int d\theta \ p(F, \theta|D, M) = \int d\theta \ p(\theta|D, M) \ p(F|\theta, M) \]

\[ = \int d\theta \ p(\theta|D, M) \ \delta[F - f(\theta)] \quad \text{[single-valued case]} \]

Prediction

Given a model with parameters \( \theta \) and present data \( D \), predict future data \( D' \) (e.g., for experimental design):

\[ p(D'|D, M) = \int d\theta \ p(D', \theta|D, M) = \int d\theta \ p(\theta|D, M) \ p(D'|\theta, M) \]
Model comparison

Marginal likelihood for model $M_i$:

$$Z_i \equiv p(D|M_i) = \int d\theta_i \ p(\theta_i|M) \ L_i(\theta_i)$$

Bayes factor $B_{ij} \equiv Z_i/Z_j$

Can write $Z_i = L_i(\hat{\theta}_i) \cdot \Omega_i$ with Ockham factor

$$\Omega_i \approx \delta\theta/\Delta\theta = \text{(posterior volume)}/\text{(prior volume)}$$

Hierarchical modeling, aka...

- Graphical models — Hierarchical and other structures
- Multilevel models — In regression, linear model settings
- Bayesian networks (Bayes nets) — In AI/ML settings
Agenda

1. Basic Bayes recap

2. Key idea in a nutshell

3. Going deeper
   - Joint distributions and DAGs
   - Conditional dependence/independence
   - Example: Binomial prediction
   - Beta-binomial model
   - Point estimation and shrinkage
   - Gamma-Poisson model & Stan
   - Algorithms
Motivation: Measurement error in surveys

BATSE GRB peak flux estimates

- **Selection effects** (truncation, censoring) — *obvious* (usually)
  Typically treated by “correcting” data
  Most sophisticated: product-limit estimators

- **“Scatter” effects** (measurement error, etc.) — *insidious*
  Typically ignored (average out??? — *No!* )
Suppose $f(x|\theta)$ is a distribution for an observable, $x$ (scalar or vector, $\vec{x} = (x, y, \ldots)$); and $\theta$ is unknown.

From $N$ precisely measured samples, $\{x_i\}$, we can infer $\theta$ from

\[ \mathcal{L}(\theta) \equiv p(\{x_i\}|\theta) = \prod_i f(x_i|\theta) \]

\[ p(\theta|\{x_i\}) \propto p(\theta)\mathcal{L}(\theta) = p(\theta, \{x_i\}) \]
But what if the $x$ data are noisy, $D_i = \{x_i + \epsilon_i\}$?

$\{x_i\}$ are now uncertain (latent/hidden/incidental) parameters.

We should somehow incorporate $\ell_i(x_i) = p(D_i|x_i)$

The joint PDF for everything is

$$p(\theta, \{x_i\}, \{D_i\}) = p(\theta) p(\{x_i\}|\theta) p(\{D_i\}|\{x_i\})$$

$$= p(\theta) \prod_i f(x_i|\theta) \ell_i(x_i)$$

The conditional (posterior) PDF for the unknowns is

$$p(\theta, \{x_i\}|\{D_i\}) = \frac{p(\theta, \{x_i\}, \{D_i\})}{p(\{D_i\})} \propto p(\theta, \{x_i\}, \{D_i\})$$
\[
p(\theta, \{x_i\}|\{D_i\}) \propto p(\theta, \{x_i\}, \{D_i\}) \\
= p(\theta) \prod_i f(x_i|\theta) \ell_i(x_i)
\]

**Marginalize over** \(\{x_i\}\) **to summarize inferences for** \(\theta\)

**Marginalize over** \(\theta\) **to summarize inferences for** \(\{x_i\}\)

**Key point:** Maximizing over \(x_i\) (i.e., just using best-fit \(\hat{x}_i\)) and integrating over \(x_i\) **can give very different results**!

(See Loredo (2004) for tutorial examples)
To estimate $x_1$:

$$p(x_1|\{x_2, \ldots \}) = \int d\theta \, p(\theta) f(x_1|\theta) \ell_1(x_1) \times \prod_{i=2}^{N} \int dx_i \, f(x_i|\theta) \ell_i(x_i)$$

$$= \ell_1(x_1) \int d\theta \, p(\theta) f(x_1|\theta) \mathcal{L}_{m,\bar{1}}(\theta)$$

$$\approx \ell_1(x_1) f(x_1|\hat{\theta}_{\bar{1}})$$

with $\hat{\theta}_{\bar{1}}$ determined by the remaining data

$f(x_1|\hat{\theta}_{\bar{1}})$ behaves like a “prior” that shifts the $x_1$ estimate away from the peak of $\ell_1(x_1)$; each member’s prior depends on all of the rest of the data → shrinkage

[For astronomers: This generalizes the corrections derived by Eddington, Malmquist and Lutz-Kelker (sans selection effects)]
Agenda

1. Basic Bayes recap

2. Key idea in a nutshell

3. Going deeper
   - Joint distributions and DAGs
   - Conditional dependence/independence
   - Example: Binomial prediction
   - Beta-binomial model
   - Point estimation and shrinkage
   - Gamma-Poisson model & Stan
   - Algorithms
Joint and conditional distributions

Bayesian inference is largely about the interplay between *joint* and *conditional* distributions for related quantities.

Ex: Bayes’s theorem relating hypotheses and data ($|\|C\|$):

$$P(H_i|D) = \frac{P(H_i)P(D|H_i)}{P(D)} = \frac{P(H_i,D)}{P(D)} = \text{joint for everything}$$

$$= \frac{P(D)}{\text{marginal for knowns}}$$

The usual form identifies an available factorization of the joint.

Express this via a *directed acyclic graph* (DAG):
Joint distribution structure as a graph

- Graph = *nodes/vertices* connected by *edges/links*
- Circular/square nodes/vertices = a priori uncertain quantities (gray/square = becomes known as data)
- Directed edges specify conditional dependence
- Absence of an edge indicates conditional *independence* → *the most important edges are the missing ones*

\[
P(H_i, D) = P(H_i) \times P(D|H_i)
\]
\[ p(x, y, z) \]

\[
p(x)p(y|x)p(z|x, y) \quad p(y)p(x|y)p(z|y, x) \quad p(z)p(x|z)p(y|z, x)
\]

\[
p(x)p(z|x)p(y|x, z) \quad p(y)p(z|y)p(x|y, z) \quad p(z)p(y|z)p(x|z, y)
\]
Cycles not allowed

\[ p(x|z) \times p(y|x) \times p(z|y) \]
Conditional independence

Suppose for the problem at hand $z$ is independent of $x$ when $y$ is known:

$$p(z|x, y) = p(z|y)$$

"$z$ is conditionally independent of $x$, given $y$": $z \perp \perp x \mid y$

Absence of an edge indicates conditional independence

Missing edges indicate simplification in structure

→ the most important edges are the missing ones
DAGs with missing edges

Conditional independence

\[ p(x) p(y|x) p(z|y) \]

\[ x \rightarrow y \rightarrow z \]

\[ z \perp x \mid y \]

“Causal chain”

\[ p(x) p(y|x) p(z|x) \]

\[ x \rightarrow y \rightarrow z \]

\[ z \perp y \mid x \]

“Common cause”

Conditional dependence

\[ p(x) p(y) p(z|x, y) \]

\[ x \rightarrow y \rightarrow z \]

“Common effects”
Conditional vs. complete independence

“z is *conditionally* independent of x, given y”

≠

“z is independent of x”

(Complete) independence between z and x (“z ⊥ x”) would imply:

\[ p(z|x) = p(z) \] (i.e., not a function of x)

Conditional independence *given y* (“z ⊥ x | y”) is weaker:

\[
\begin{align*}
p(z|x) &= \int dy \ p(z, y|x) \\
&= \int dy \ p(y|x)p(z|x, y) \\
&= \int dy \ p(y|x)p(z|y) \quad \text{since } z \perp x \mid y
\end{align*}
\]

Although \( x \) drops out of the last factor, \( x \) dependence remains in \( p(y|x) \)

\( x \) *does* provide information about \( z \), but it only does so through the information it provides about \( y \) (which directly influences \( z \))
Bayes’s theorem with IID samples

For model with parameters $\theta$ predicting data $D = \{x_i\}$ that are IID given $\theta$:

$$p(\theta, D) = p(\theta)p(\{x_i\}|\theta) = p(\theta) \prod_{i=1}^{N} p(x_i|\theta)$$

To find the posterior for the unknowns ($\theta$), divide the joint by the marginal for the knowns ($\{x_i\}$):

$$p(\theta|\{x_i\}) = \frac{p(\theta) \prod_{i=1}^{N} p(x_i|\theta)}{p(\{x_i\})} \quad \text{with} \quad p(\{x_i\}) = \int d\theta \; p(\theta) \prod_{i=1}^{N} p(x_i|\theta)$$
Binomial counts

\[ n_1 \text{ heads in } N \text{ flips} \]

\[ n_2 \text{ heads in } N \text{ flips} \]

Suppose we know \( n_1 \) and want to predict \( n_2 \).
Predicting binomial counts — known $\alpha$

Success probability $\alpha \rightarrow p(n|\alpha) = \frac{N!}{n!(N-n)!} \alpha^n (1 - \alpha)^{N-n} \mid N$

Consider two successive runs of $N = 20$ trials, known $\alpha = 0.5$

$p(n_2|n_1, \alpha) = p(n_2|\alpha) \mid N$

$n_1$ and $n_2$ are *conditionally independent*
DAG for binomial prediction — known $\alpha$

Knowing $\alpha$ lets you predict each $n_i$, independently
Predicting binomial counts — uncertain $\alpha$

Consider the same setting, but with $\alpha$ uncertain

Outcomes are *physically* independent, but $n_1$ tells us about $\alpha \rightarrow$ outcomes are *marginally dependent* (see Lec 12 for calculation):

$$p(n_2|n_1, N) = \int d\alpha \, p(\alpha, n_2|n_1, N) = \int d\alpha \, p(\alpha|n_1, N) \, p(n_2|\alpha, N)$$

Flat prior on $\alpha$

Prior: $\alpha = 0.5 \pm 0.1$
DAG for binomial prediction

\[ p(\alpha, n_1, n_2) = p(\alpha)p(n_1|\alpha)p(n_2|\alpha) \]

From joint to conditionals:

\[ p(\alpha|n_1, n_2) = \frac{p(\alpha, n_1, n_2)}{p(n_1, n_2)} = \frac{p(\alpha)p(n_1|\alpha)p(n_2|\alpha)}{\int d\alpha \ p(\alpha)p(n_1|\alpha)p(n_2|\alpha)} \]

\[ p(n_2|n_1) = \frac{\int d\alpha \ p(\alpha, n_1, n_2)}{p(n_1)} \]

Observing \( n_1 \) lets you learn about \( \alpha \)
Knowledge of \( \alpha \) affects predictions for \( n_2 \) \( \rightarrow \) dependence on \( n_1 \)
A population of coins/flippers

Each flipper+coin flips different number of times

- What do we learn about the population of coins—the distribution of $\alpha$s?
- How does population membership effect inference for a single coin’s $\alpha$?
Terminology: $\theta$ are hyperparameters, $\pi(\theta)$ is the hyperprior
A simple multilevel model: beta-binomial

Goals:

- Learn a population-level “prior” by pooling data
- Account for population membership in member inferences

Qualitative

\[ p(\theta, \{\alpha_i\}, \{n_i\}) = \pi(\theta) \prod_i p(\alpha_i|\theta) \, p(n_i|\alpha_i) \]

\[ = \pi(\theta) \prod_i p(\alpha_i|\theta) \, \ell_i(\alpha_i) \]

Quantitative

\[ \theta = (a, b) \text{ or } (\mu, \sigma) \]

\[ \pi(\theta) = \text{Flat}(\mu, \sigma) \]

\[ p(\alpha_i|\theta) = \text{Beta}(\alpha_i|\theta) \]

\[ p(n_i|\alpha_i) = \binom{N_i}{n_i} \alpha_i^{n_i} (1 - \alpha_i)^{N_i-n_i} \]
Generating the population & data

- Beta distribution (mean, conc'n)
- Binomial distributions

\[ p(n_i | \alpha) \]

\[ \alpha = 0.21, \quad N = 80, \quad n = 20 \]
\[ \alpha = 0.34, \quad N = 40, \quad n = 16 \]
\[ \alpha = 0.36, \quad N = 10, \quad n = 1 \]
\[ \alpha = 0.45, \quad N = 20, \quad n = 11 \]
\[ \alpha = 0.54, \quad N = 160, \quad n = 79 \]
Likelihood function for one member’s $\alpha$

$N=20$

$n=11$
Learning the population distribution

\[ p(\alpha) \]

\[ \mathcal{L}(\alpha_i, p(\alpha_i|D)) \]

\( n=20 \quad N=80 \)
\( n=16 \quad N=40 \)
\( n=1 \quad N=10 \)
\( n=11 \quad N=20 \)
\( n=79 \quad N=160 \)
Lower level estimates

Two approaches

• Hierarchical Bayes (HB): Calculate marginals

\[ p(\alpha_j | \{n_i\}) \propto \int d\theta \pi(\theta) \prod_{i \neq j} \int d\alpha_i p(\alpha_i | \theta) p(n_i | \alpha_i) \]

• Empirical Bayes (EB): Plug in an optimum \(\hat{\theta}\) and estimate \(\{\alpha_i\}\)

View as approximation to HB, or a frequentist procedure that estimates a prior from the data
Lower level estimates

Bayesian outlook

- Marginal posteriors are *narrower* than likelihoods
- Point estimates tend to be closer to true values than MLEs (averaged across the population)
- Joint distribution for \(\{\alpha_i\}\) is *dependent*
Frequentist outlook

- Point estimates are biased
- Reduced variance $\rightarrow$ estimates are closer to truth on average (lower MSE in repeated sampling)
- Bias for one member estimate depends on data for all other members

Lingo

- Estimates *shrink* toward prior/population mean
- Estimates “muster and *borrow strength*” across population (Tukey’s phrase); increases accuracy and precision of estimates
- Efron* describes shrinkage as a consequence of accounting for *indirect evidence*

Beware of point estimates!

Population and member estimates

\[ p(\alpha) \]

\[
\begin{array}{c}
\text{True} \\
\text{ML} \\
\text{EB pts} \\
\text{EB}
\end{array}
\]

\[ \text{RMSE} = 0.096 \]

\[ \text{RMSE} = 0.057 \]
Competing data analysis goals

“Shrunken” member estimates provide improved & reliable estimate for population member properties

But they are *under-dispersed* in comparison to the true values $\rightarrow$ not optimal for estimating *population* properties*

*No point estimates of member properties are good for all tasks!*

We should view population data tables/catalogs as providing *descriptions of member likelihood functions*, not “estimates with errors”

*Louis (1984); Eddington noted this in 1940!
Measurement error perspective

If the data provided *precise* \( \{\alpha_i\} \) values (coin measurements, flip physics), we could easily model them as points drawn from a (beta) population PDF with params \( \theta \):

\[
D = \{\alpha_i\}
\]

\[
p(D|\theta) = \prod_i p(\alpha_i|\theta) = \prod_i \text{Beta}(\alpha_i|\theta)
\]

(A binomial point process)
Here the finite number of flips provide *noisy measurements of each* $\alpha_i$, described by the member likelihood functions $\ell_i(\alpha_i)$;

$$D = \{n_i\}$$

$$p(D|\theta) = \prod_i \int d\alpha_i \ p(D, \{\alpha_i\}|\theta)$$

$$= \prod_i \int d\alpha_i \ p(\alpha_i|\theta) \ p(n_i|\theta)$$

$$= \prod_i \int d\alpha_i \ \text{Beta}(\alpha_i|\theta) \ \text{Binom}(n_i|\theta)$$

This is a prototype for *measurement error problems*
Another conjugate MLM: Gamma-Poisson

Goal: Learn a rate dist’n from count data
(E.g., learn a star or galaxy brightness dist’n from photon counts)

Qualitative

Quantitative

\[ \theta = (\alpha, s) \text{ or } (\mu, \sigma) \]

\[ \pi(\theta) = \text{Flat}(\mu, \sigma) \]

\[ p(F_i | \theta) = \text{Gamma}(F_i | \theta) \]

\[ p(n_i | F_i) = \text{Pois}(n_i | \epsilon_i F_i) \]
Simulations: $N = 60$ sources from gamma with $\langle F \rangle = 100$ and $\sigma_F = 30$; exposures spanning dynamic range of $\times16$
Consider the posterior PDF for $\theta$ and $\{\alpha_i\}$ in the beta-binomial MLM:

$$p(\theta, \{\alpha_i\}|\{n_i\}) \propto \pi(\theta) \prod_{i=1}^{N_{\text{mem}}} \text{Beta}(\alpha_i|\theta) \text{Binom}(n_i|\alpha_i)$$

For each member, the $\text{Beta} \times \text{Binom}$ factor is $\propto$ a beta distribution for $\alpha_i$; but as a function of $\theta$ (e.g., $(a, b)$ or $(\mu, \sigma)$) it is not simple.

The full posterior has a product of $N_{\text{mem}}$ such factors specifying its $\theta$ dependences $\Rightarrow$ **even for a conjugate model for the lower levels, the overall model is typically analytically intractable**

Two approaches exploit *conditional independence of lower-level parameters*
**Member marginalization**

- Analytically or numerically integrate over \( \{x_i\} \) to explore the reduced-dimension marginal for \( \theta \) via MCMC:
  \[ \{\theta_i\} \sim p(\theta|D) \]

- If \( x_i \) are of interest, sample them from their conditionals, conditioned on \( \theta_i \):
  - Pick a \( \theta \) from \( \{\theta_i\} \)
  - Draw \( \{x_i\} \) by *independent* sampling from their conditionals (given \( \theta \))
  - Iterate

GPUs can accelerate this for application to large datasets

Only useful for low-dimensional latent parameters \( x_i \)
**Metropolis-within-Gibbs algorithm**

Block the full parameter space:

- Block of $m$ population parameters, $\theta$
- $N$ blocks of lower level (latent) parameters, $x_i$

Get posterior samples by iterating back and forth between:

- $m$-D Metropolis-Hastings sampling of $\theta$ from $p(\theta|\{x_i\}, D)$
  
  This requires a problem-specific proposal distribution

- $N$ independent samples of $x_i$ from the conditional $p(x_i|\theta, D_i)$
  
  This can often exploit conjugate structure

  E.g., Beta-binomial: $\alpha_i \sim \text{Beta}(\alpha_i|\theta) \text{ Binom}(n_i|\alpha_i)$, which is just a Beta for $\alpha_i$

MWG explicitly displays the feedback between population and member inference