METAPSY

VU VRIJE UNIVERSITEIT AMSTERDAM

WHO Collaborating Centre for Research and Dissemination of Psychological Interventions

Bayesian Modelling with JAGS

A Crash Course

Mathias Harrer, Ph.D.
Vrije Universiteit Amsterdam

Agenda

Bayesian Inference

BUGS and JAGS

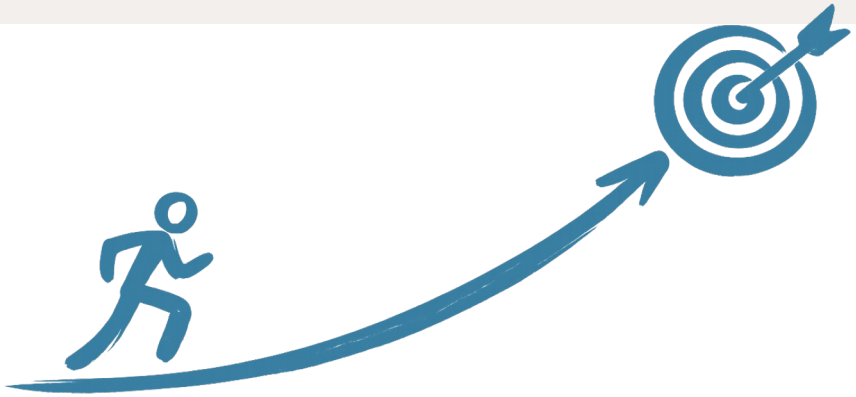
JAGS in Practice

- A First Example in JAGS
- Missing Data Handling
- (IPD) Meta-Analysis
- Network Meta-Analysis

Goals

Basic Introduction into Bayesian Modelling

- Differences to frequentist methods & benefits
- Understanding of basic concepts & notation



Basic Understanding of JAGS

- Learn the basic functions and components of JAGS models
- Run models yourself from R
- Adapt and (re-)write models from existing source code

Beware: Steep Learning Curve!

- Learning JAGS is not easy
- Nor is the statistics behind (IPD)-MA
- See this as first start to your learning journey, no need to understand everything/be perfect right away


DON'T
PANIC

Literature

“The BUGS Book”

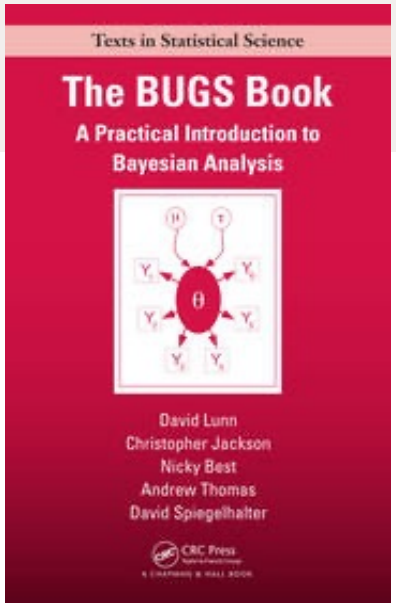
Go-to reference for Bayesian modelling with specific examples in BUGS/JAGS; technical but highly recommended for advanced use.

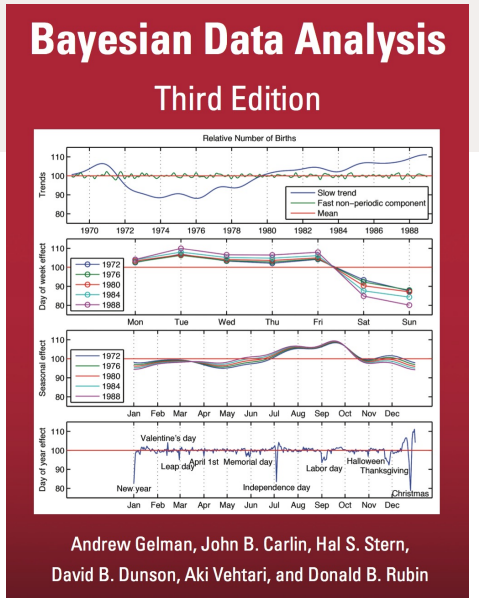
“Bayesian Data Analysis”

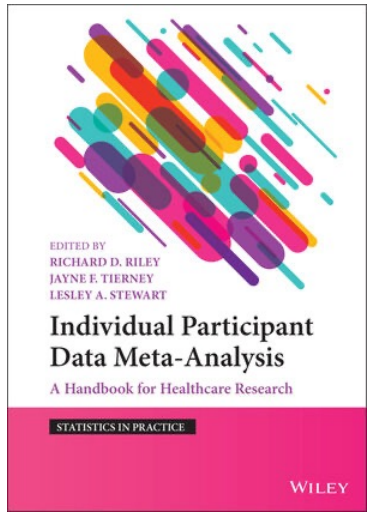
Classic introductory book for Bayesian modelling in general; not specifically for JAGS.

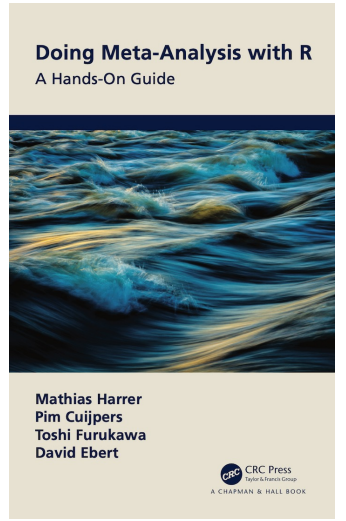
Further Reading

- “Individual Participant Data Meta-Analysis” (Riley et al.; not specifically Bayesian/JAGS)
- ”Doing Meta-Analysis with R” (chapters on NMA, Bayesian MA)









Bayesian Inference

What is Bayesian Inference?

Bayesian inference starts with uncertainty

- In the real world, we **observe some data Y.**
- This data has been generated by some **process** whose **true probability θ** is unknown.

What is Bayesian Inference?

Example: "An argument for divine providence taken from the constant regularity observ'd in the births of both sexes" (1711)

- Yearly **births of females and males** in London during the 17th century.
- John Arbuthnot, a British physician, calculated the **ratio of males or females** each year.
- There seems to be a "true" probability θ that a child is born a boy or a girl.

→ Based on these data, **how certain** can we be that the ratio boys-girls (θ) **lies within a certain range** (e.g., >1)?

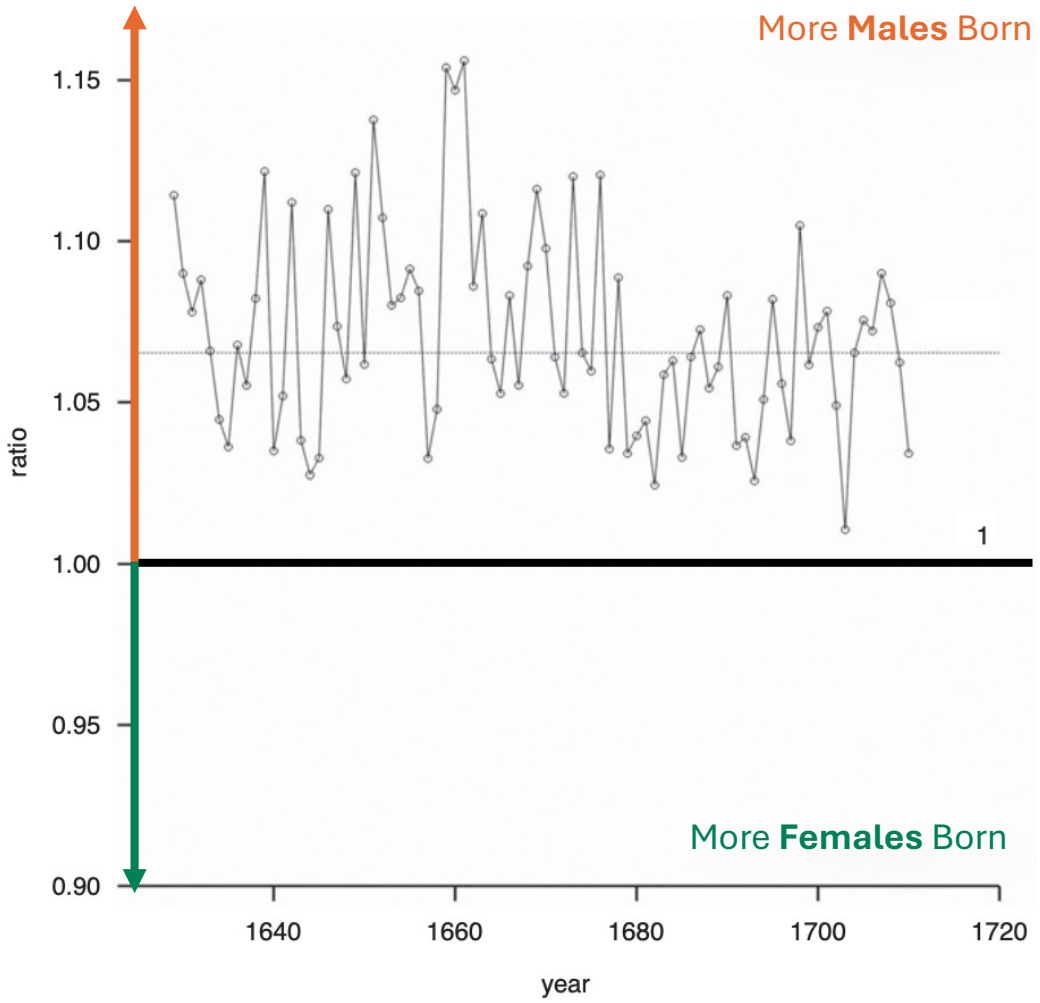


Figure 2.2 Sex ratio (males/females) for Arbuthnot's data. A sex ratio of 1 and the observed median sex ratio of 1.065 are indicated by dashed lines.

Senn, S. (2022). *Dicing with death*. Cambridge University Press.

METAPSY

What is Bayesian Inference?

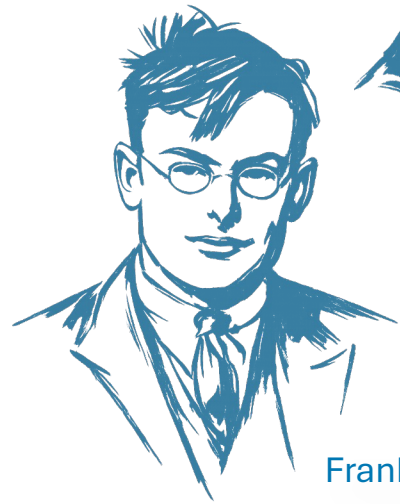
Bayesian inference starts with uncertainty

- **Bayes (1763):** what is the probability that θ lies between two specified values?
- Unknown parameters are treated as uncertain quantities:

$$\theta \sim p(\theta)$$

- In Bayesian statistics, probability represents **degrees of belief** in parameter values, not physical randomness in parameters.
- Core idea behind Bayesian statistics: **probability is "subjective",** quantifies our beliefs and how data can change it (Ramsey, 1926)

Thomas Bayes



Frank Ramsey

METAPSY

The Data Model (Likelihood)

In inferential statistics, we specify a **statistical model**:

$$Y|\theta \sim p(Y|\theta)$$

This statistical model (hopefully) approximates the true process generating the data.

This sampling model for the data is known as the **likelihood**.

→ How likely is each specific data point Y, given the true population values?

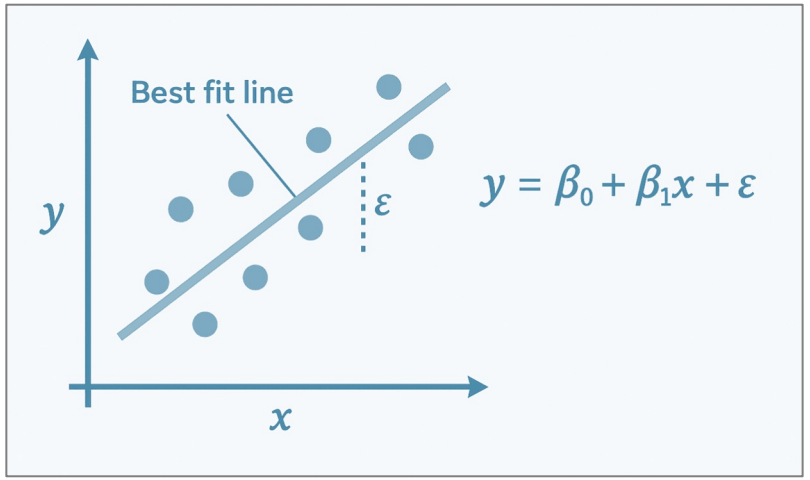
The Data Model (Likelihood)

Example: Linear Regression

In a linear regression with one predictor, we have:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$



Therefore, our model contains three unknown parameters θ:

$$\theta = \quad \beta_0, \beta_1, \sigma !!$$

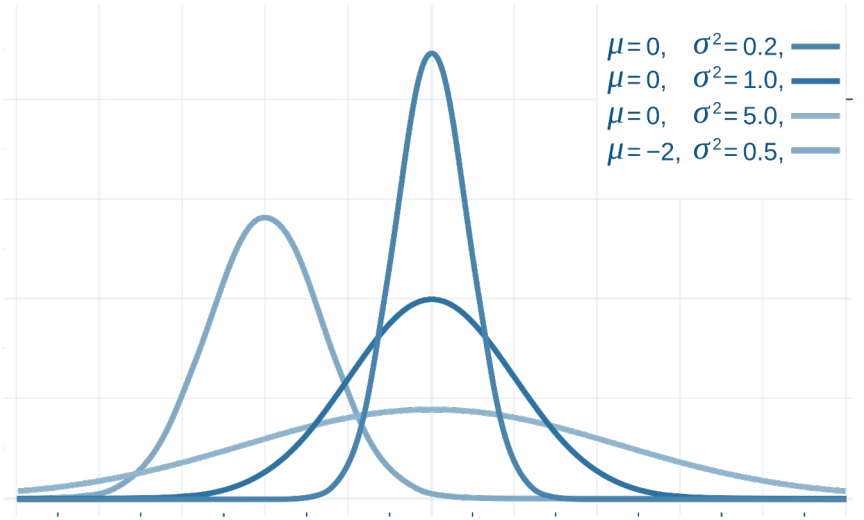
We can therefore write the data model (likelihood) as:

$$Y|\beta_0, \beta_1, \sigma \sim p(\beta_0, \beta_1, \sigma)$$

Prior Distributions

In Bayesian statistics, we "encode" **prior assumptions** about each unknown parameter θ using **priors**.

→ Describes plausible values of θ
→ Quantifies uncertainty before seeing current data



We can encode this information using **probability distributions**:

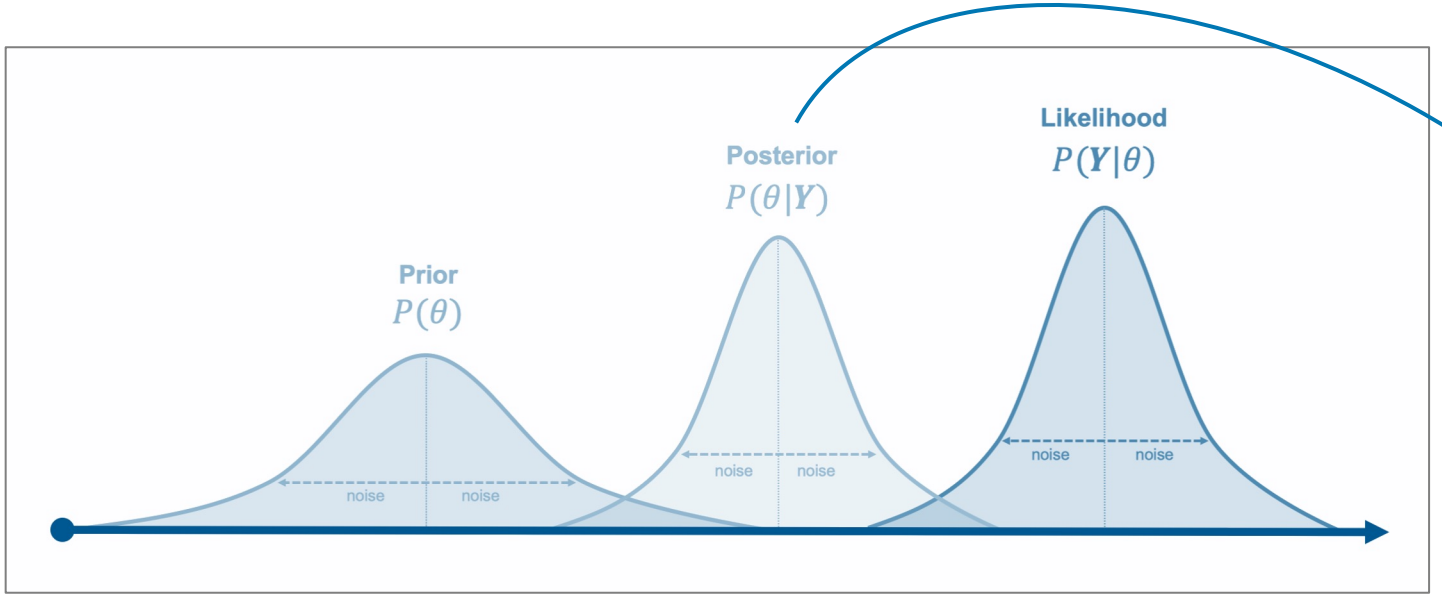
- **Wide distributions**: large uncertainty/little prior knowledge
- **Narrow distribution**: high certainty, high "weight" given to our prior knowledge

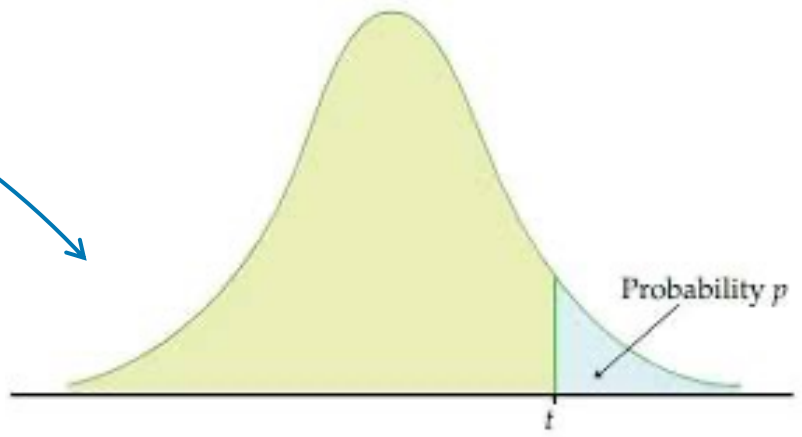
Bayesian Updating

Bayes' rule states that we can combine prior and likelihood to generate a **posterior distribution**:

→ $p(\theta \mid Y) \propto p(Y \mid \theta)\, p(\theta)$, where ∝ stands for "proportional to"

- By combining it with the data/likelihood, we "update" our priors, leading to the posterior.
- The posterior distribution can then be used to **make inferences** (e.g. estimate that θ is between two values)





e.g. Distribution of β_1

METAPSY

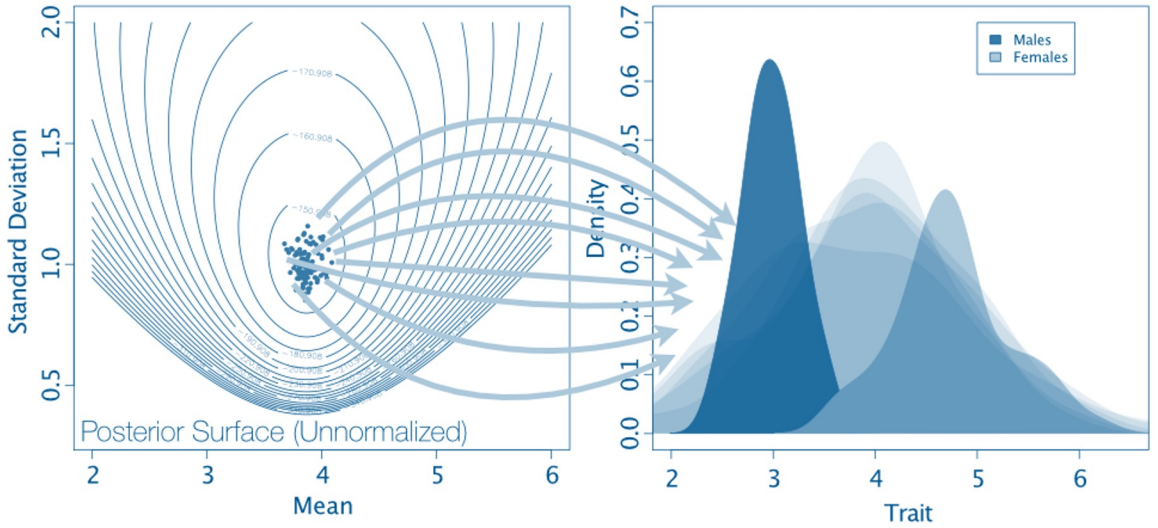
Prediction

Bayesian inference naturally leads to prediction when we **integrate over the posterior distributions**:

$$p(Y_{new} \mid y) = \int p(Y_{new} \mid \theta) p(\theta \mid y) d\theta$$

This is known as the **posterior predictive distribution**.

It accounts for **parameter uncertainty** automatically.



Höhna, Landis, Heath, Boussau, Lartillot, Moore, Huelsenbeck, Ronquist. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65:726-736.

Prediction

Linear regression example: we have posterior distributions for β_0 and β_1.

→ Predict value Y_{new}, given predictor value X_0.

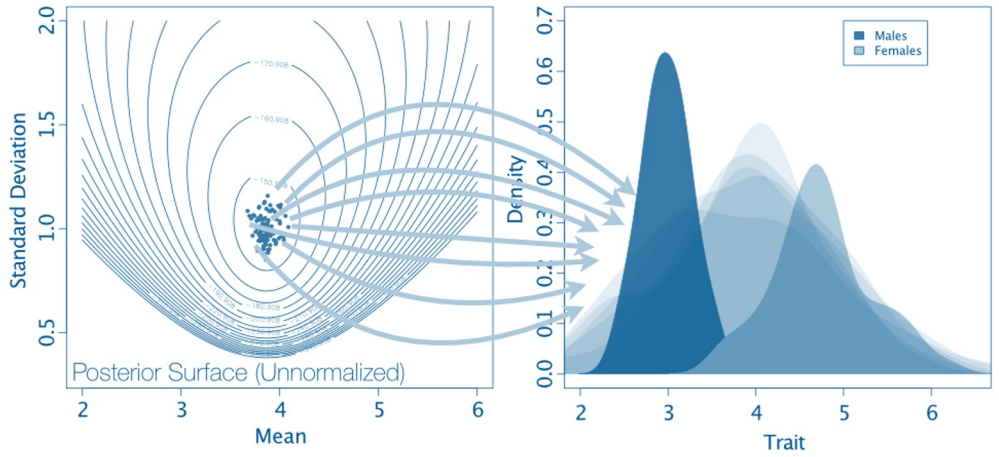
$$p(Y_{new}|X_o)$$

$$= \int p(Y_{new}|\beta_0, \beta_1, \sigma, X_0) p(\beta_0, \beta_1, \sigma|Y) d\beta_0 d\beta_1 d\sigma$$

We integrate over parameter uncertainty in both β_0 and β_1.

This leads **not to a single predicted value**, but an **entire posterior distribution** for the predicted value.

→ Easily allows to assess certainty in our prediction.



Höhna, Landis, Heath, Boussau, Lartillot, Moore, Huelsenbeck, Ronquist. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65:726-736.

METAPSY

Frequentist vs. Bayesian Inference

Frequentist vs. Bayesian Inference

“Conventional” statistics using p-values and confidence intervals is based on an alternative or **frequentist interpretation of probability**.

Frequentist statistics only focuses on the **likelihood of the model**, which tries to maximise; for example using optimization algorithms or matrix algebra (OLS).
→ There is no “updating” of prior beliefs, and no posterior distribution.

Ronald A. Fisher

Maximum Likelihood Estimation (MLE): $\hat{\theta} = \text{argmax}_{\theta}\ p(Y|\theta)$

Linear regression (OLS): $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$

In frequentist statistics/MLE:

- Parameters are fixed (not random)
- No integration over parameters
- We solve an optimization problem instead

Frequentist vs. Bayesian Inference

- For Bayesian inference with posterior distributions, intervals, tail probabilities, posterior predictions, etc., we need to **solve integrals**.

- For some models and prior distributions, such **integrals can be computed easily.**

- **Conjugate prior distributions:** prior distribution family that that lead to the same distribution in the posterior → **facilitates computation with closed formulas**

Frequentist vs. Bayesian Inference

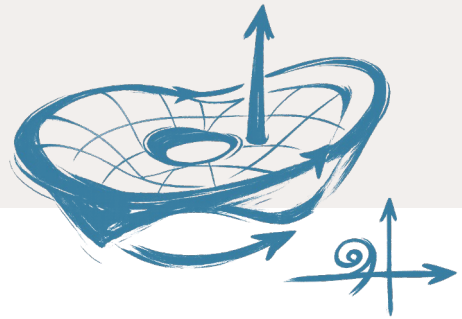
TABLE 3.1
Univariate conjugate prior distributions for various one-parameter likelihoods from a sample of size n. Also given are the corresponding posterior parameters and the predictive distribution for a single new observation $\tilde{y}$[†]. See Appendix C and/or Bernardo and Smith (1994), pp. 427–435, for definitions of distributions.

Sampling distribution	Conjugate prior	Posterior parameters	Predictive distribution
$y\|\theta \sim \text{Binomial}(\theta, n)$ including Bernoulli $(n = 1)$	$\theta \sim \text{Beta}(a, b)$	$a_n = a + y,$ $b_n = b + n - y$	Beta-Binomial(a_n, b_n, n)
$y\|\mu \sim \prod_{i=1}^n \text{Normal}(\mu, \sigma^2)$	$\mu \sim \text{Normal}(\gamma, \omega^2 = \frac{\sigma^2}{n_0})$	$\gamma_n = \frac{n_0\gamma + n\bar{y}}{n_0+n},$ $\omega_n^2 = \frac{\sigma^2}{n_0+n}$	Normal$(\gamma_n, \omega_n^2 + \sigma^2)$[‡]
$y\|\sigma^2 \sim \prod_{i=1}^n \text{Normal}(\mu, \sigma^2)$	$\sigma^{-2} \sim \text{Gamma}(a, b)$	$a_n = a + \frac{n}{2},$ $b_n = b + \frac{1}{2}\sum_i (y_i - \mu)^2$	Student-$t(\mu, \frac{b_n}{a_n}, 2a_n)$[§]
$y\|\theta \sim \prod_{i=1}^n \text{Poisson}(\theta)$	$\theta \sim \text{Gamma}(a, b)$	$a_n = a + n\bar{y},$ $b_n = b + n$	NegBin$(\frac{b_n}{b_n+1}, a_n)$
$y\|\theta \sim \prod_{i=1}^n \text{Gamma}(\alpha, \theta)$ including Exponential $(\alpha = 1)$	$\theta \sim \text{Gamma}(a, b)$	$a_n = a + n\alpha,$ $b_n = b + n\bar{y}$	Gamma-Gamma(a_n, b_n, α)
$y\|\theta \sim \prod_{i=1}^n \text{Uniform}(0, \theta)$	$\theta \sim \text{Pareto}(a, b)$	$a_n = a + n,$ $b_n = \max\{b, y\}$	$\begin{cases} \frac{a_n}{a_n+1}\text{Uniform}(0, b_n), & \tilde{y} \le b_n \\ \frac{1}{a_n+1}\text{Pareto}(a_n, b_n), & \tilde{y} > b_n \end{cases}$
$y\|\theta \sim \text{NegBin}(\theta, r)$ including Geometric $(r = 1)$	$\theta \sim \text{Beta}(a, b)$	$a_n = a + r,$ $b_n = b + y$	Negative-Binomial-Beta(a_n, b_n, r_p)

BUGS Book, p. 46

METAPSY

Frequentist vs. Bayesian Inference



Alas, in realistic/more complex models:

High-dimensional integrals with many parameters

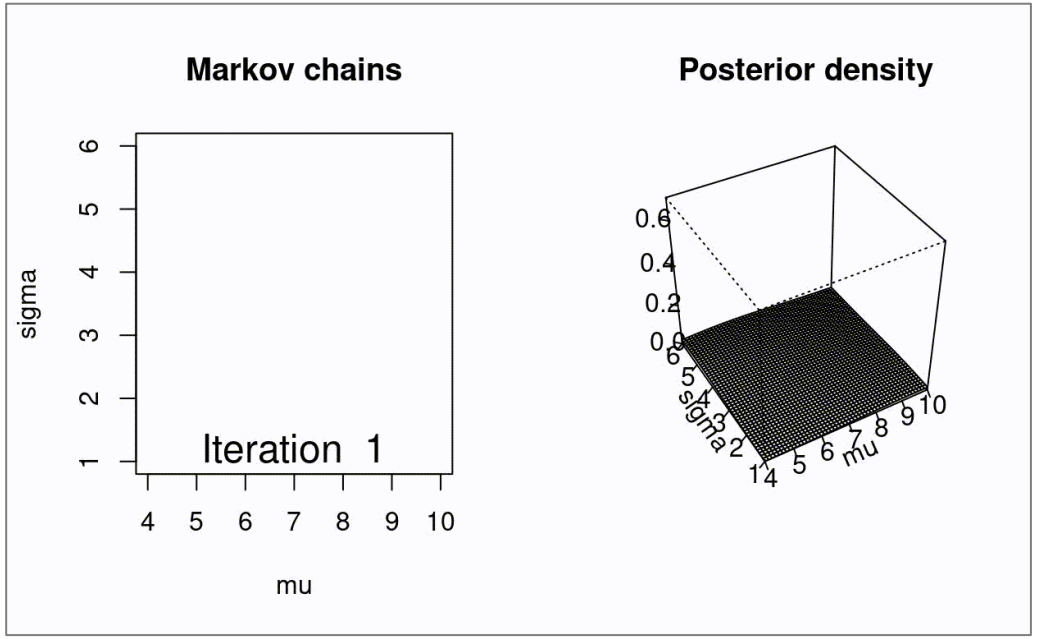
No conjugate prior(s)

No closed-form posterior

→ **Major ”bottleneck” for application of Bayesian statistics over centuries!**

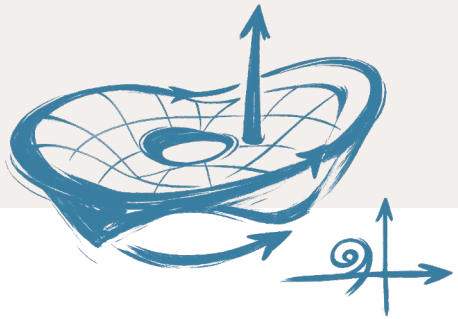
Solution: **Markov Chain Monte Carlo Sampling**

- Possible with high computing power (starting from the late 1980s)
- With enough samples, the resulting values will approximate the mathematically correct shape.



METAPSY

Markov Chain Monte Carlo

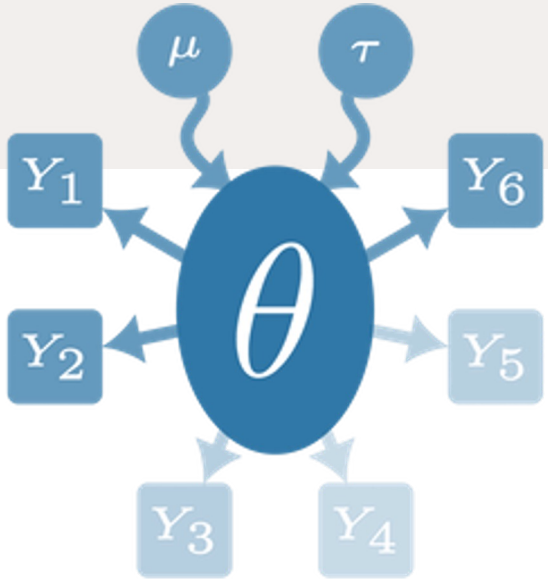


With MCMC sampling, Bayesian modelling becomes very attractive (e.g. in comparison to MLE)

- **Very high flexibility**: define the sampling model/likelihood and priors, obtain results via sampling
- Easy to **tweak parameters** of a model by "normal" users, without having to derive complex formulas for the likelihood (as in MLE)
- Particularly helpful for **hierarchical models** in which information is pooled (e.g., meta-analysis)
- Easy to **quantify uncertainty** or exact probabilities, since we obtain the full posterior distributions.

BUGS and JAGS

BUGS



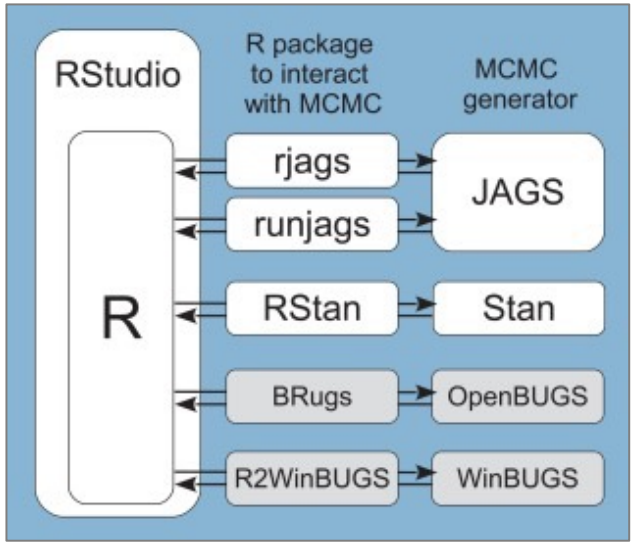
- In the late 1980s: **computational breakthroughs** in MCMC through Gibbs sampling (Gelfand & Smith, 1990)
- At the same time: **BUGS project** at Cambridge (1989)

What is BUGS?

Bayesian inference **U**sing **G**ibbs **S**ampling

Declarative model language for specifying full joint distributions via **directed acyclical graphs (DAGs)**:

- Model specification (model language)
- Inference engine (MCMC algorithms)



METAPSY

How Does Gibbs Sampling Work?

For Bayesian inference, we want to sample from the posterior distribution.

→ Yet, usually, this posterior cannot be integrated analytically, so it is not available

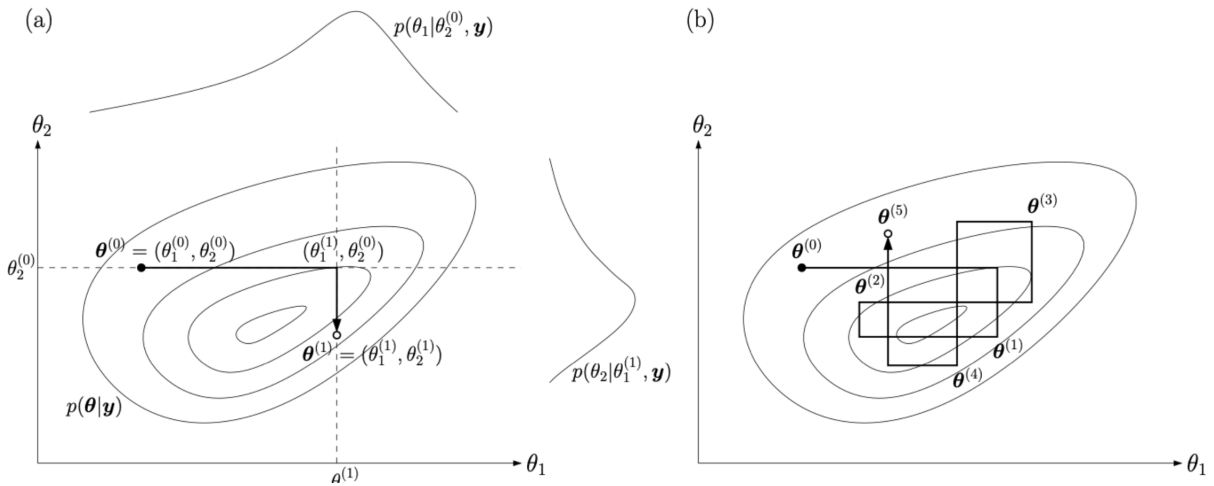
Gibbs sampling

Instead of sampling from the full joint posterior

$$p(\theta_1, \theta_2, \ldots, \theta_k \mid Y),$$

We sample from an easier, conditional distribution of one parameter θ_j given all other parameters $\boldsymbol{\theta}_{-j}$, and the data:

$$p(\theta_j|\boldsymbol{\theta}_{-\boldsymbol{j}}, Y)$$



METAPSY

How Does Gibbs Sampling Work?

Example: Two Parameters θ_1 and θ_2

Step 1: Choose starting values
Pick initial values: $\theta_1^{(0)}, \theta_2^{(0)}$

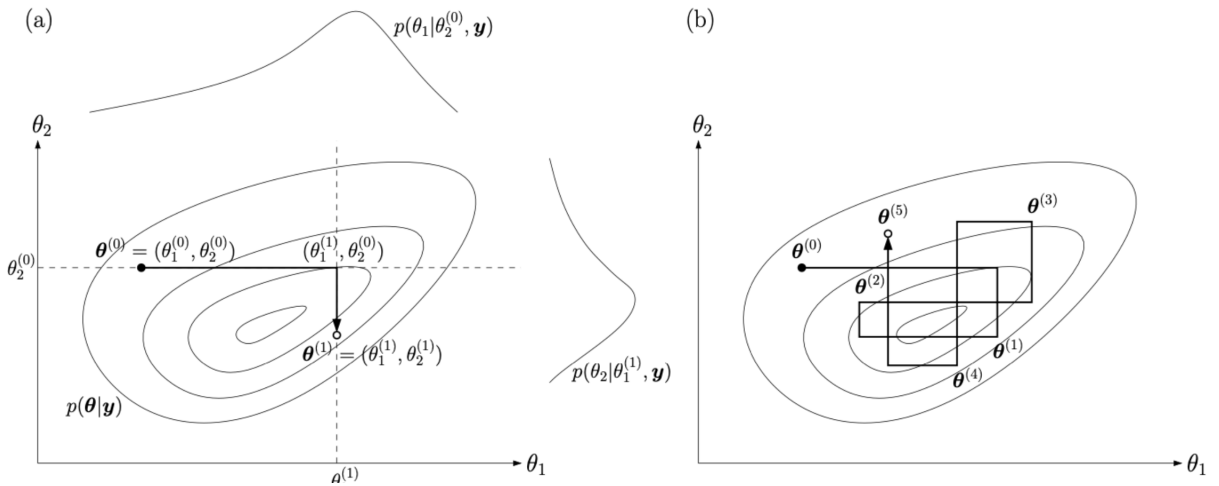
Step 2: Update first parameter
Draw: $\theta_1^{(1)} \sim p\left(\theta_1, \mid \theta_2^{(0)}, Y\right)$

Step 3: Update second parameter
Draw: $\theta_2^{(1)} \sim p\left(\theta_2, \mid \theta_1^{(1)}, Y\right)$

Step 4: Repeat many times
Continue cycling:

$$\theta_1^{(t)} \sim p\left(\theta_1, \mid \theta_2^{(t-1)}, Y\right)$$
$$\theta_2^{(t)} \sim p\left(\theta_2, \mid \theta_1^{(t)}, Y\right)$$

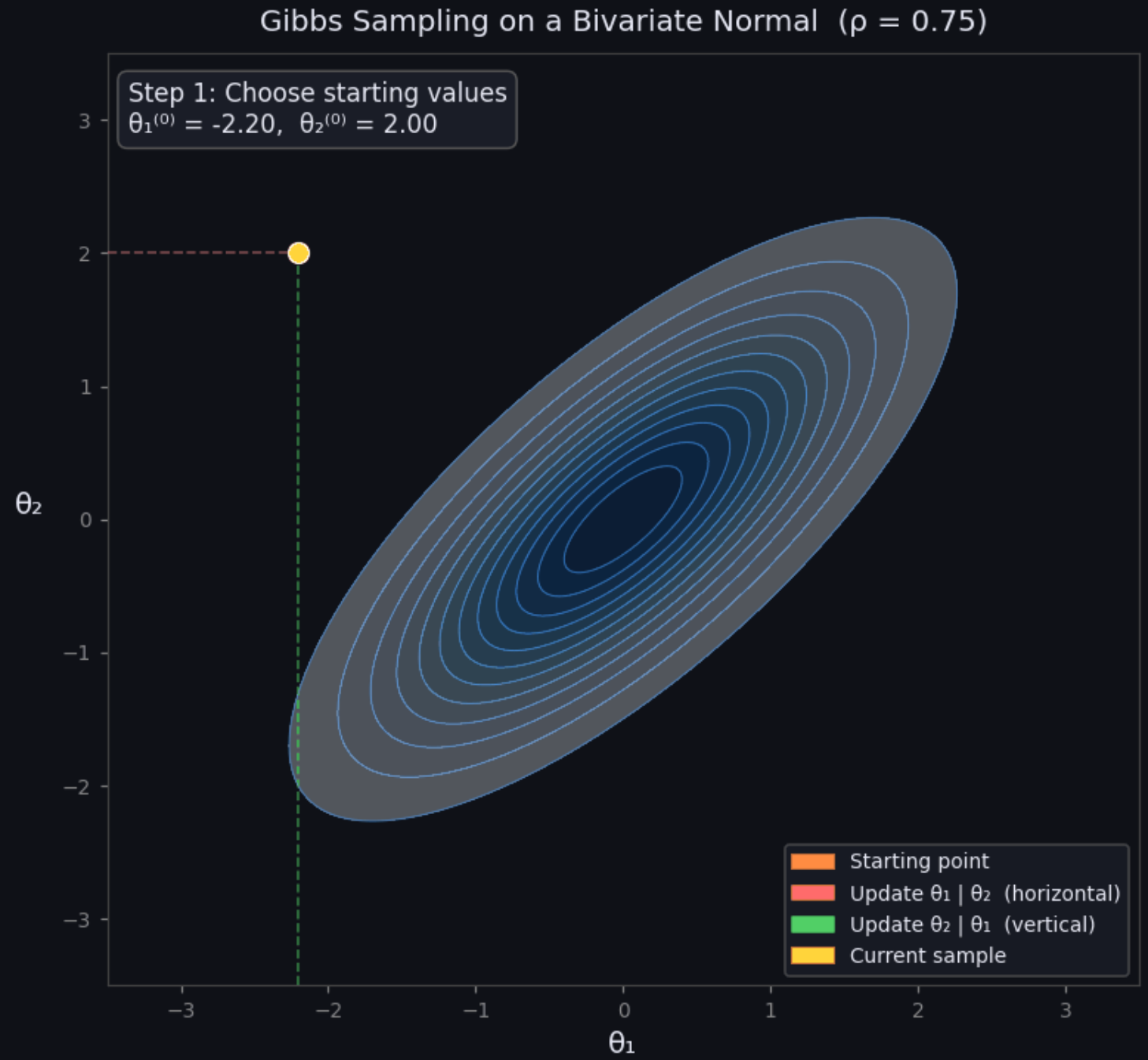
.

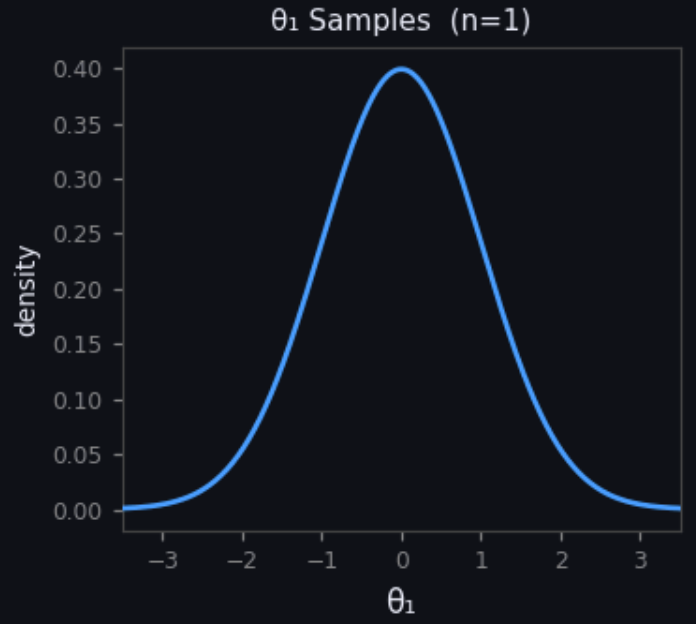


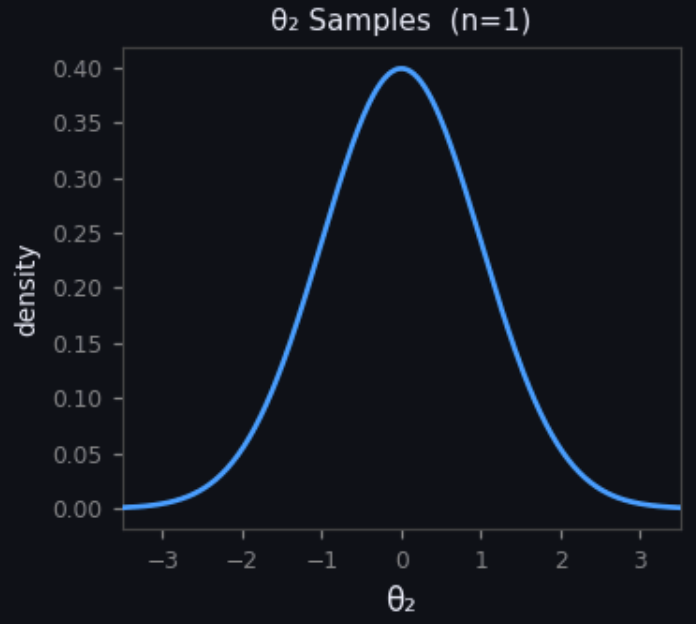
After many iterations, the joint draws **explore the full posterior distribution**:

$$\left(\theta_1^{(t)}, \theta_2^{(t)}\right) \sim p(\theta_1, \theta_2 \mid Y)$$

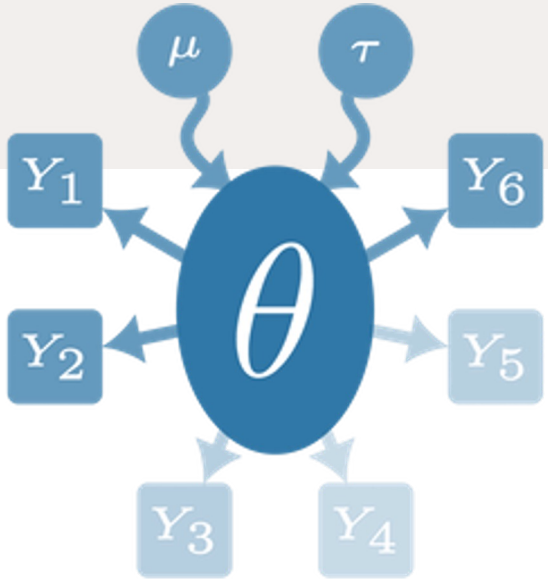
METAPSY


Gibbs Sampling on a Bivariate Normal (ρ = 0.75)
Step 1: Choose starting values
θ₁⁽⁰⁾ = -2.20, θ₂⁽⁰⁾ = 2.00
θ₂
θ₁
Starting point
Update θ₁ | θ₂ (horizontal)
Update θ₂ | θ₁ (vertical)
Current sample


θ₁ Samples (n=1)
density
θ₁


θ₂ Samples (n=1)
density
θ₂

What is JAGS?



- Stands for **J**ust **A**nother **G**ibbs **S**ampler
- Free, open-source Bayesian modeling software
- You write your model (likelihood + priors) in the **BUGS language**
- JAGS then handles the MCMC automatically
- Runs on Windows, Mac, Linux; connects to R via `runjags`

What is Stan?

- A similar but newer language for Bayesian modelling
- Stan uses **no-U-turn (NUTS)** instead of Gibbs sampling
- Stan converges faster on complex/correlated models, but has a steeper learning curve
- R interface: `rstan`, `brms`

→ Not covered in this course

JAGS: Components & Structure

Model block

- Defines the **probabilistic structure of the model**.
- Loops: used for the likelihood/model → lets data "flow" into the model
- Prior block: defined for all parameters θ we are uncertain about.

Code components

- **Stochastic relations (~):** derived using sampling from a distribution
- **Deterministic relations (<−):** algebraic/direct functions of other variables

  Constants / data: observed values in our dataset (e.g. x, y, N)

```
model {

  for (i in 1:N) {
    y[i] ~ dnorm(mu[i], tau)  # likelihood
    mu[i] <- beta0 + beta1 * x[i]
  }

  alpha ~ dnorm(0, 0.001)      # prior
  beta  ~ dnorm(0, 0.001)      # prior
  tau   ~ dgamma(0.001, 0.001) # prior
}
```

Example JAGS Code for a Linear Regression

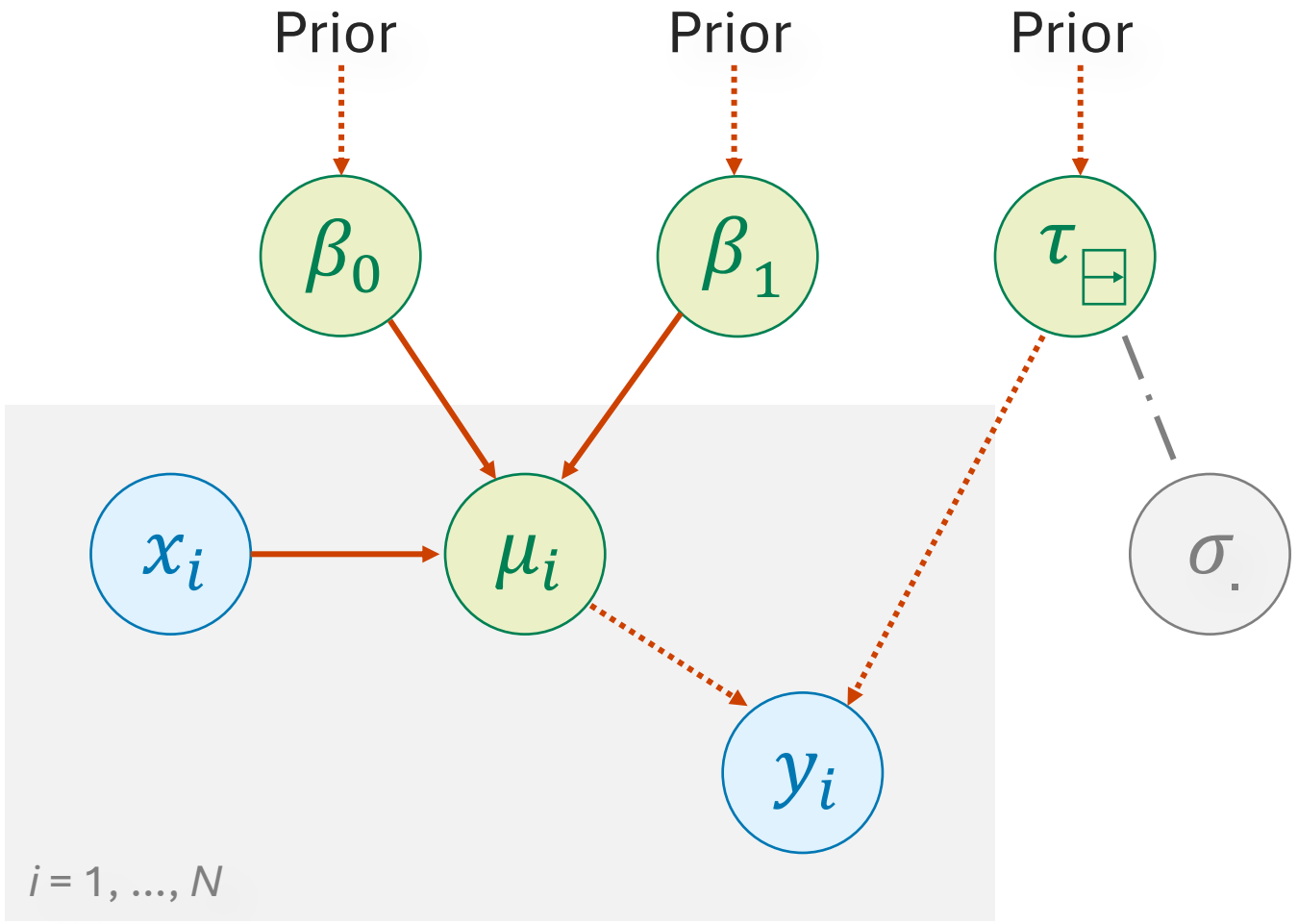
METAPSY

JAGS: Components & Structure

Model code in JAGS can be described as **Directed Acyclic Graphs** (DAGs), ...

... in which **nodes** (parameters, quantities) are connected with **edges** (deterministic or probabilistic relationships).

This can help to identify parameters with no **"parent nodes"** → require priors



METAPSY

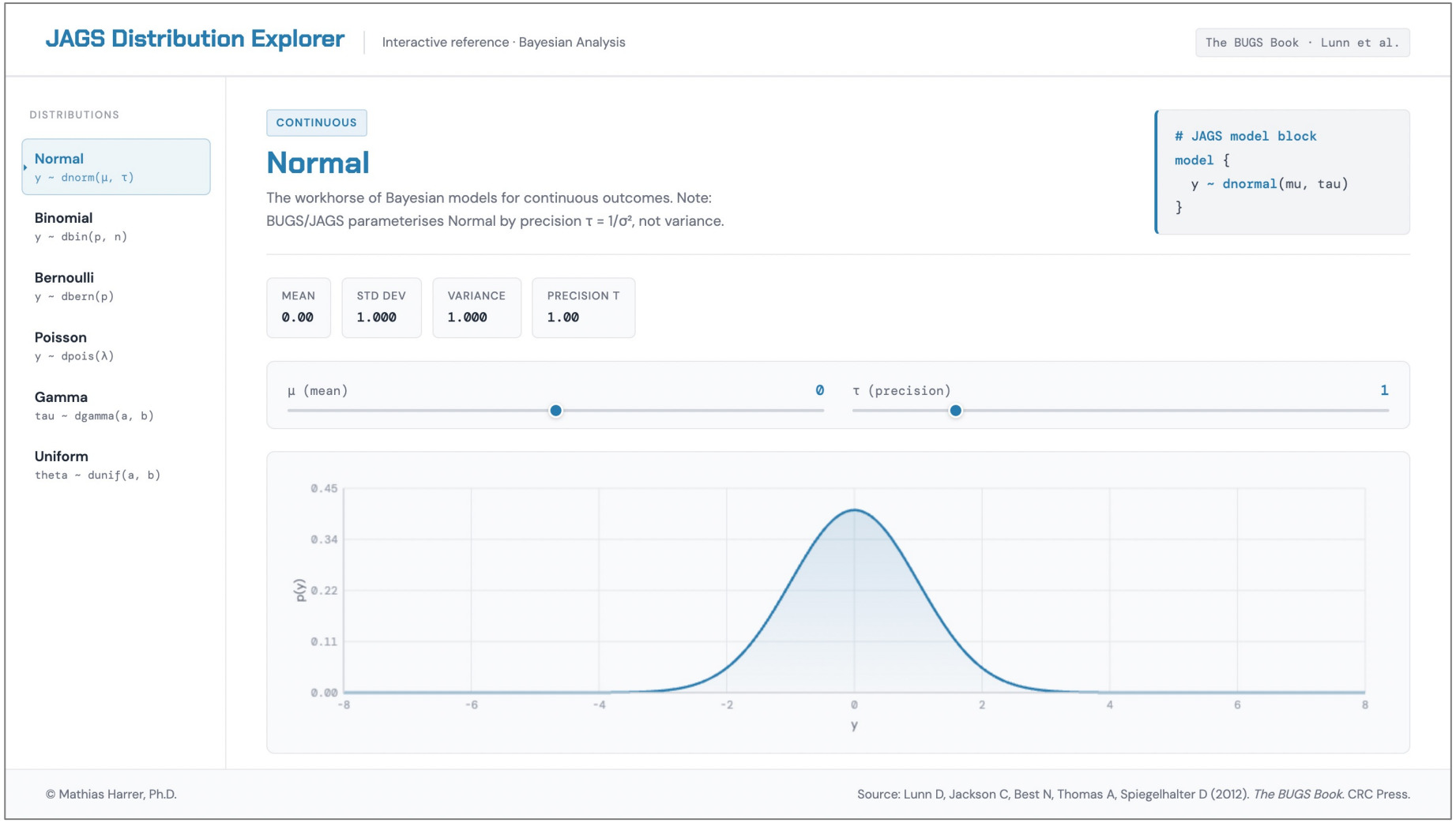
JAGS: Components & Structure

Distribution	JAGS syntax	Typical use
Normal	`y ~ dnorm(mu, tau)`	Continuous outcomes
Binomial	`y ~ dbin(p, n)`	Binary / counts of successes
Bernoulli	`y ~ dbern(p)`	Binary outcomes
Poisson	`y ~ dpois(lambda)`	Count data
Gamma	`tau ~ dgamma(a, b)`	Prior for precision
Uniform	`theta ~ dunif(a, b)`	Weak / bounded prior

The Precision Parameter (τ) in JAGS

In JAGS, the normal distribution uses **<u>precision</u> instead of variance**: $Y \sim Normal(\mu, \tau)$
Thus, $\tau = (\sigma^2)^{-1}$; Low precision (τ) means high variance → wide distribution
You can add a direct conversion in the code: `tau <- 1/pow(sigma, 2)`

Explore JAGS Distributions



METAPSY

JAGS in Practice

A First Example

Linear Regression

Model formula/likelihood

For $i = 1, \ldots, N$:

$$y_i \sim \mathcal{N}(\mu_i, \tau)$$
$$\mu_i = \beta_0 + \beta_1 x_i$$

$$\beta_0 \sim \mathcal{N}(0, 0.0001)$$
$$\beta_1 \sim \mathcal{N}(0, 0.0001)$$
$$\sigma \sim U(0, 100)$$
$$\tau = 1/\sigma^2$$

Model code

01-regression.R

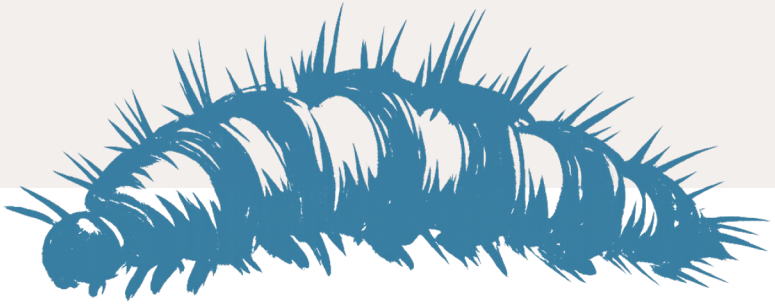
Vague Priors

Residual SD

```
model {

  # Likelihood
  for (i in 1:n) {
    y[i]  ~ dnorm(mu[i], tau)
    mu[i] <- beta0 + beta1 * x[i]
  }

  # Priors
  beta0 ~ dnorm(0, 0.0001)
  beta1 ~ dnorm(0, 0.0001)
  sigma ~ dunif(0, 100)

  # Derived quantity
  tau <- 1 / pow(sigma,2)

}
```

Model Diagnostics

Model Diagnostics



Check convergence of the sampler before interpreting posterior estimates.

Potential Scale Reduction Factor

(PSRF; $\hat{R}$, Gelman–Rubin diagnostic)

- Compares **within-chain** and **between-chain** variance across multiple chains.
- **Ideal:** $\hat{R} \approx 1.00$; acceptable if $\hat{R} < 1.01$)

If $\hat{R}$ is too large:

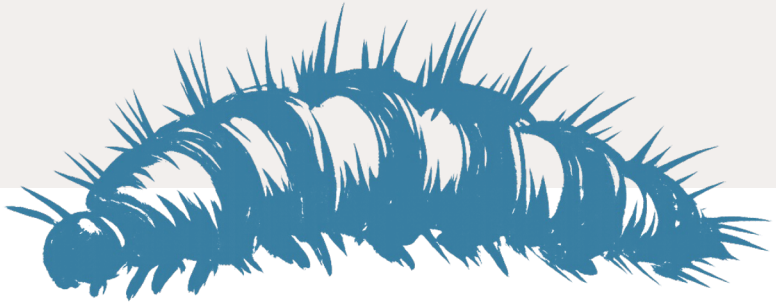
- Run the sampler for **more iterations** or improve the model specification.

Inspect trace plots:

- Chains should **mix well** and reach **stationarity**.
- A well-behaved chain resembles a **"fat hairy caterpillar"** (stable mean, good mixing).

METAPSY

Model Diagnostics



Effective Sample Size (ESS)

- MCMC draws are **not independent** because successive samples are **autocorrelated**.
- ESS estimates how many **independent samples** the chain is equivalent to.
- If N samples are drawn but autocorrelation is high, the **effective sample size is much smaller**.

$$\text{ESS} = \frac{\text{Total number of MCMC samples}}{1 + 2 \sum_{k=1}^{\infty} \rho_k}$$

ρ_k: autocorrelation at lag k

Large ESS → reliable posterior estimate

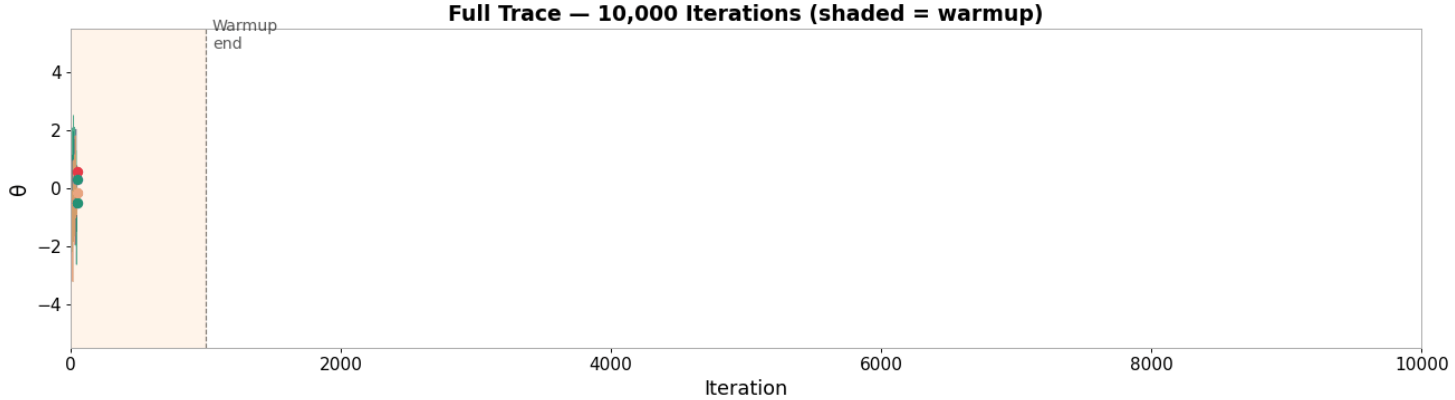
Small ESS → high autocorrelation, inefficient sampling

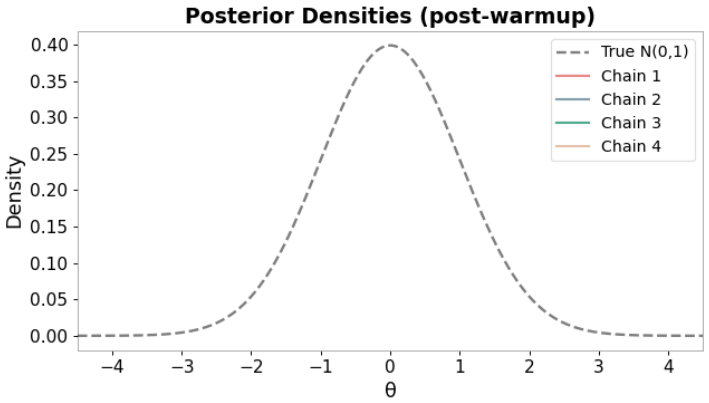
What to do if ESS is small

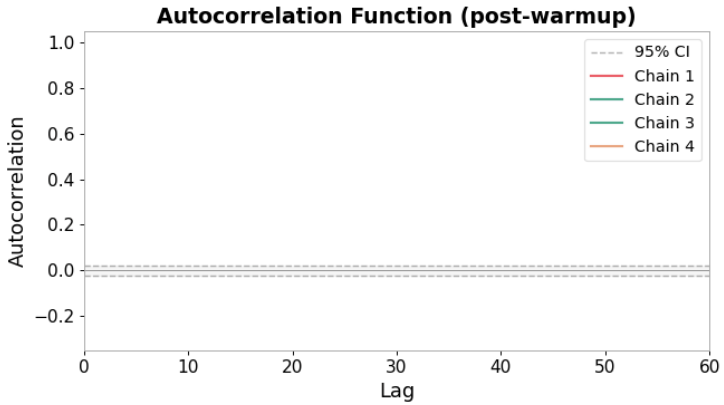
- Run **more iterations**
- Check **autocorrelation plots**

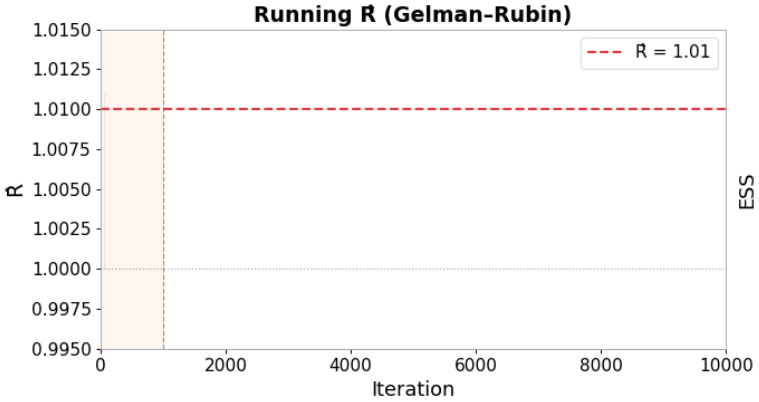
Rule of thumb: Aim for **ESS ≥ 1000** for key parameters

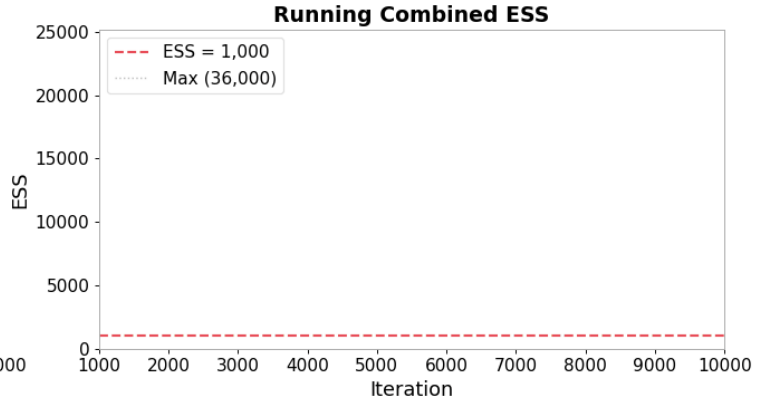
METAPSY

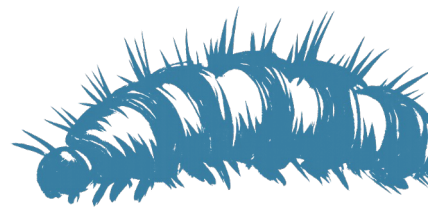

Full Trace — 10,000 Iterations (shaded = warmup)
Warmup end
θ
Iteration


Posterior Densities (post-warmup)
True N(0,1)
Chain 1
Chain 2
Chain 3
Chain 4
Density
θ


Autocorrelation Function (post-warmup)
95% CI
Chain 1
Chain 2
Chain 3
Chain 4
Autocorrelation
Lag


Running R̂ (Gelman–Rubin)
R̂ = 1.01
R̂
Iteration


Running Combined ESS
ESS = 1,000
Max (36,000)
ESS
Iteration



METAPSY

Model Diagnostics

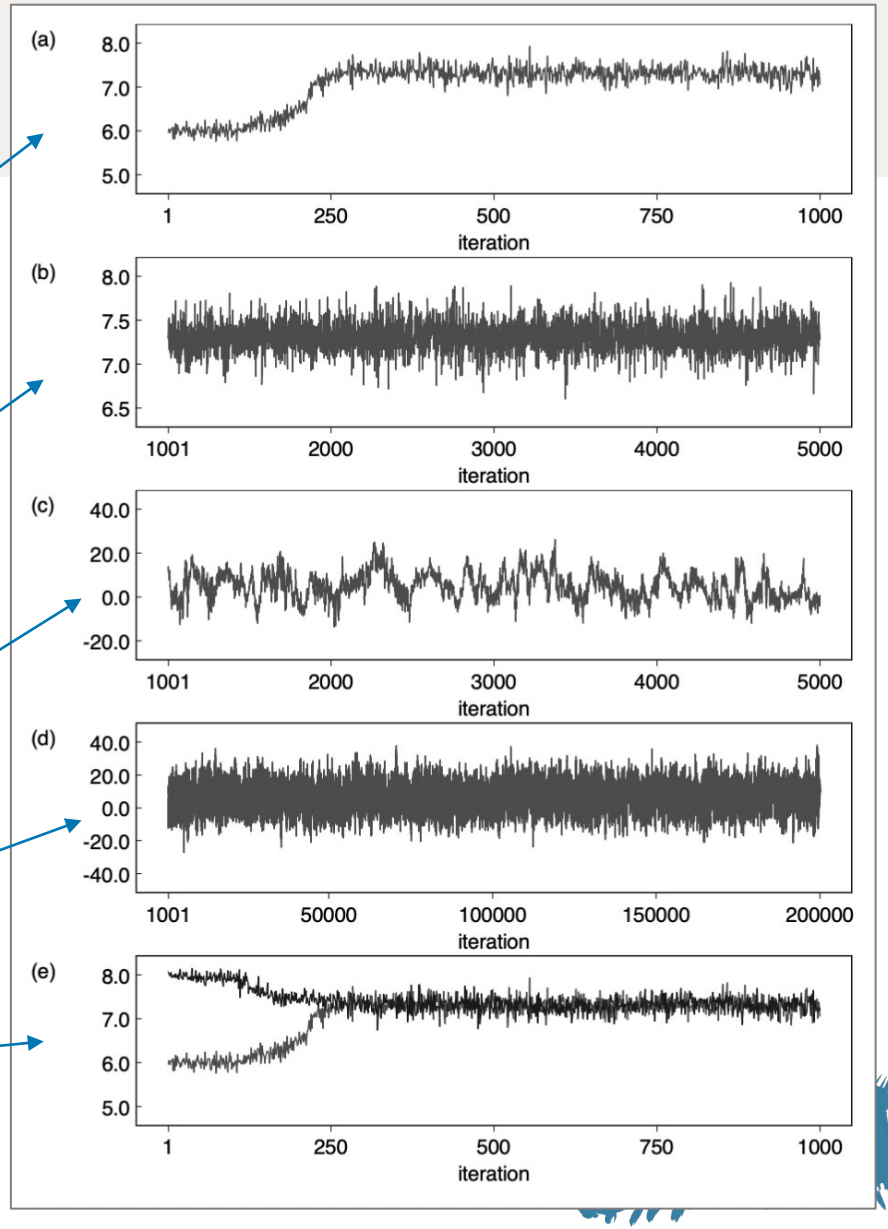
a) After convergence, a Markov chain appears as a **random scatter around a stable mean** (stationary distribution).

b) A well-behaved chain resembles a **“fat hairy caterpillar”**, indicating good mixing and sufficient information for inference.

c) A **“snake-like” chain** suggests **high autocorrelation** — more samples are needed.

d) Running the chain for **many more iterations** can reduce autocorrelation and produce the desired pattern.

e) Multiple chains with different starting values should **overlap and mix**, indicating convergence to the same distribution.



METAPSY

Meta-Analysis

Meta-Analysis

The standard meta-analysis model is based on a multilevel or hierarchical structure with two levels:

Level 1: The observed treatment effect in study *k* differs from the "true" effect due sampling error:

$$y_i|\theta_k, \sigma_k \sim \mathcal{N}(\theta_k, \sigma_k)$$

Level 2: However, even the true effect of *k* is only part of a "universe" of true effect sizes the treatment can have across contexts:

$$\theta_k \mid \mu, \tau \sim \mathcal{N}(\mu, \tau^2)$$

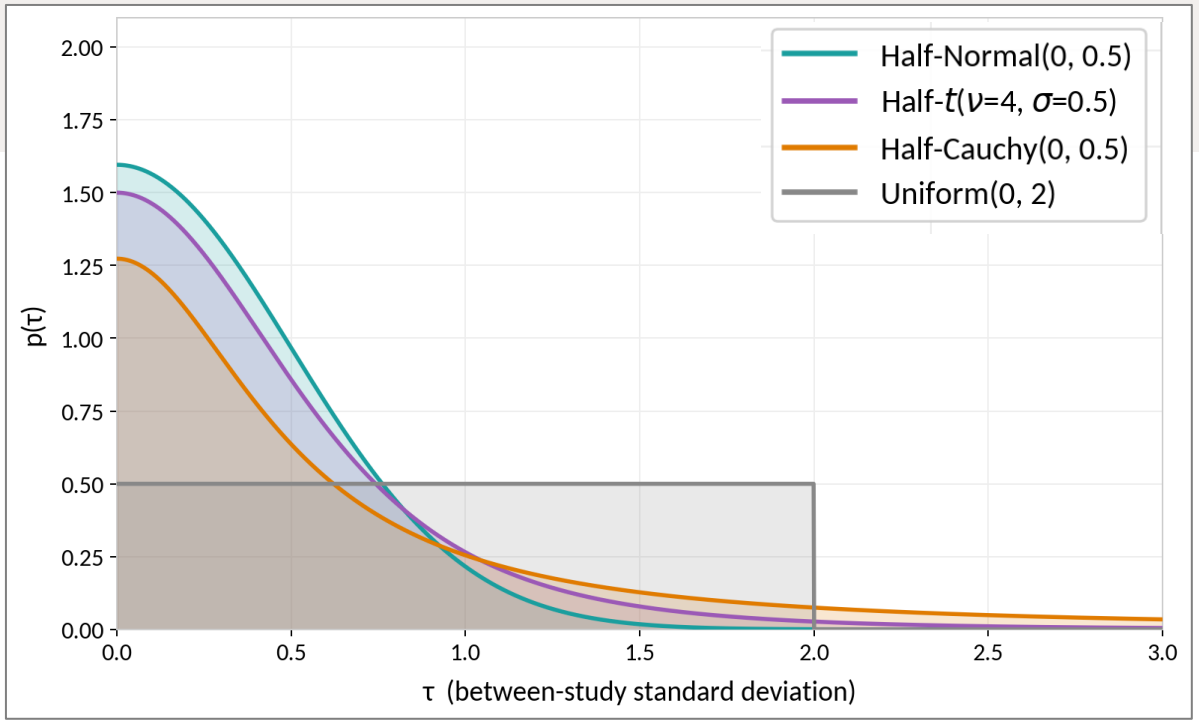
→ Since both formulas sample from a normal distribution, we sometimes call this a **"normal-normal" hierarchical model**.

```
model {

  # Level 1
  for (j in 1:J) {
    yi[j] ~ dnorm(theta[j], prec.yi[j])
    prec.yi[j] <- 1 / vi[j]
  }

  # Level 2
  for (j in 1:J) {
    theta[j] ~ dnorm(mu, prec.tau)
  }

  tau ~ dnorm(0, pow(0.5, -2)) T(0, )
  mu  ~ dnorm(0, pow(10, -2))
  prec.tau <- 1 / (tau * tau)
}
```

Prior Distributions for τ

Why weakly informative priors?

- With **few studies**, noninformative/improper priors (e.g., uniform) can produce improper or extremely heavy-tailed posteriors.
- Conjugate inverse-gamma priors are problematic too: often place too much mass on large τ values.
- Weakly informative priors **stabilize estimation** while still letting the data dominate.



Common prior choices for τ

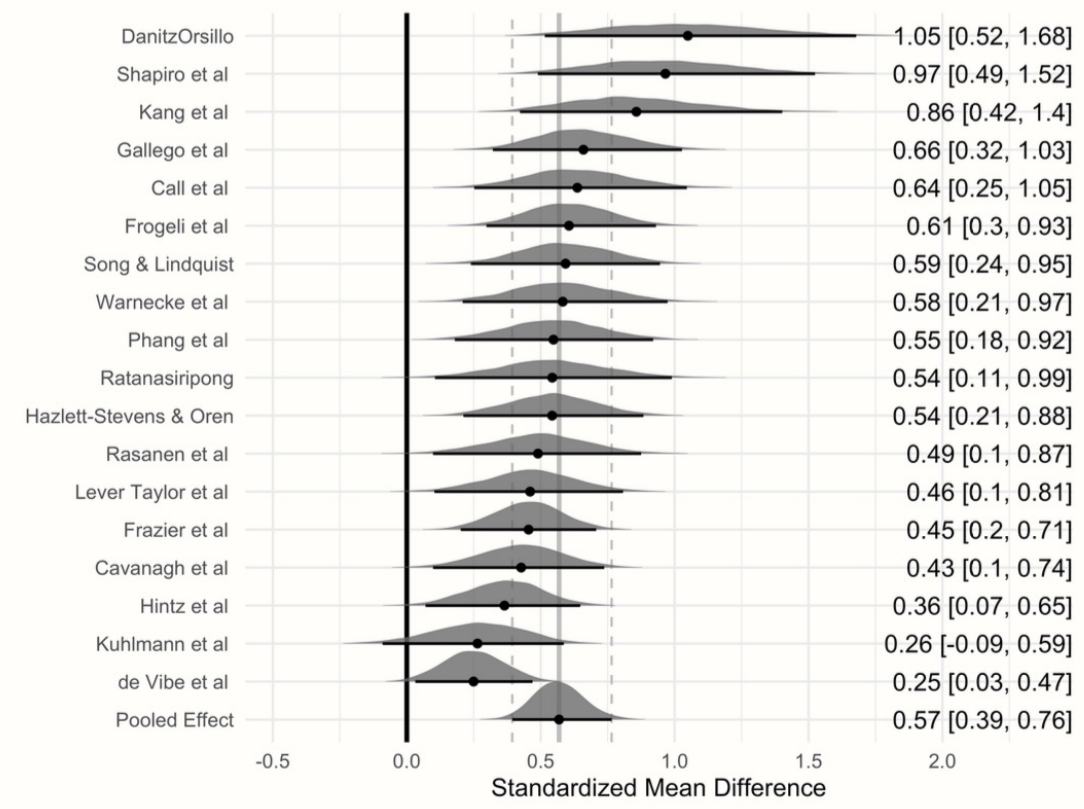
- **Half-normal** → simple, numerically stable, quickly decaying tail
- **Half-Cauchy** → heavy tails; robust when τ range is highly uncertain.
- **Half-Student-t(v)** → flexible family bridging half-normal (large v) and half-Cauchy (v=1).

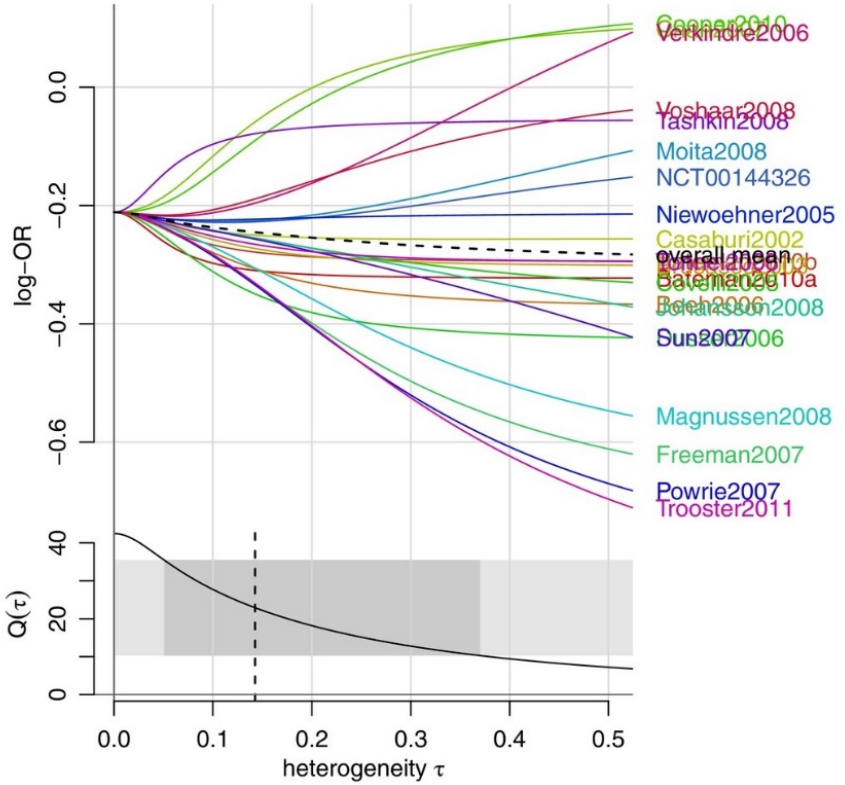
Practical default (Röver, 2021, RSM)

- **Half-Normal(0, 0.5):** places ≈95% prior mass on τ < 1 (reasonable–high heterogeneity for log-OR/log-RR, SMDs).
- Still allows large τ values (~5%), preserving robustness.

METAPSY

Shrinkage & BLUPs





The `theta[j]` are the best linear unbiased predictors (BLUPs): precision-weighted shrinkage estimates that blend each study's observed effect y_j toward the grand mean μ, with noisier studies shrunk more strongly → **the smaller τ, the stronger the shrinkage overall**

METAPSY

Binary Outcomes

Handling Binary Outcome Data in Meta-Analysis

Approach 1: Keep Normal–Normal Model

- Convert binary outcomes into an effect size (e.g., **log-risk ratio**, **log-odds ratio**, or **risk difference**)
- Assume estimated effects follow a **normal distribution** with known sampling variance

Binary Outcomes

Handling Binary Outcome Data in Meta-Analysis

Approach 2: Binomial Likelihood Model

Model the **raw event counts directly** using a binomial likelihood:

$$y_k^0 \sim Bin\left(n_k^0, p_k^0\right) \qquad y_k^1 \sim Bin\left(n_k^1, p_k^1\right)$$

$$\ln\left(\frac{p_k^0}{1-p_k^0}\right) = \alpha_k \qquad \ln\left(\frac{p_k^1}{1-p_k^1}\right) = \alpha_k + \delta_k$$

$$\delta_k \sim \mathcal{N}(\mu, \tau^2);\ e^{\mu} = exp(\mu) = OR$$

→ Avoids normal approximation and fully respects the **discrete data structure**

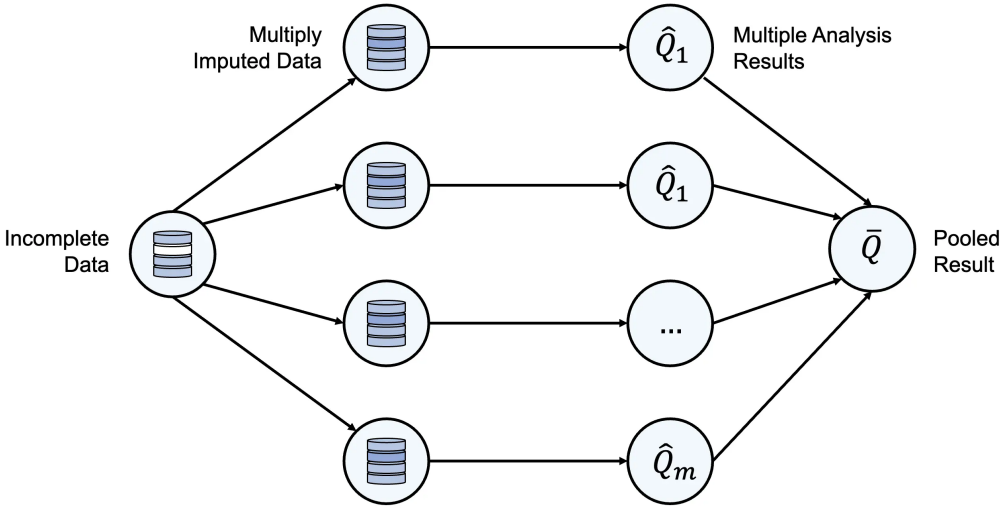
```
model {
  for (k in 1:K) {
    y1[k] ~ dbin(p1[j], n1[k])
    y0[k] ~ dbin(p0[j], n0[k])
    logit(p1[k]) <- alpha[k] + delta[k]
    logit(p0[k]) <- alpha[k]
  }
  for (k in 1:K) {
    delta[k] ~ dnorm(mu,      prec.tau)
    alpha[k]   ~ dnorm(mu.base, prec.base)
  }
  prec.tau  <- 1 / (tau * tau)
  prec.base <- 1 / (tau.base * tau.base)
  tau      ~ dnorm(0, pow(0.5, -2)) T(0, )
  tau.base ~ dnorm(0, pow(1.0, -2)) T(0, )
  mu       ~ dnorm(0, pow(10,  -2))
  mu.base  ~ dnorm(0, pow(10,  -2))
}
```

Multiple Imputation

Multiple Imputation with MICE

Multiple Imputation (MI) replaces each missing value with several plausible values drawn from a predictive distribution.

1. Generate m complete datasets by imputing missing values (typically $m \approx 50–100$).
2. Analyse each dataset separately using the intended analysis model.
3. Pool estimates using **Rubin's Rules**.



Advantages: retains all observations and produces valid uncertainty estimates under MAR (missing at random).
→ Most commonly used approach: **Multivariate Imputation by Chained Equations (MICE).**

METAPSY

The Core Idea Behind MICE

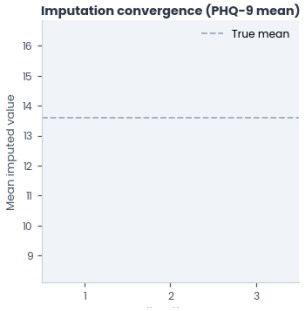
MICE imputes missing values **variable-by-variable using conditional models** („fully conditional specification“; FCS)

Algorithm:

1. Start with initial placeholder values for missing entries.
2. Select one incomplete variable and fit a regression model using other variables as predictors.
3. Draw new imputations from the predictive distribution.
4. Repeat for all variables in cycles (“chained equations”)

→ Similar to Gibbs Sampling, but **no strict joint distributional models**: we can choose e.g. logistic regression for one variable, random forest for the other, etc.

PHQ-9	Treatment (Complete)	Sex	Age	Antidepressants
?	1	No	20	No
19	0	?	54	?
?	1	No	?	No
10	1	No	38	?
7	1	?	26	Yes
?	1	Yes	?	No
6	1	Yes	35	Yes
25	1	Yes	21	?
18	1	Yes	42	Yes
22	1	?	31	Yes
10	0	?	?	?
10	0	Yes	43	Yes
?	1	No	19	Yes
20	1	Yes	?	Yes
3	1	No	45	Yes
7	0	Yes	64	Yes
23	1	No	24	Yes
2	0	Yes	61	No



METAPSY

MICE for Multilevel Data

- Many datasets are **clustered** (e.g., patients nested within studies in IPD meta-analysis).
- Standard MICE assumes **independent observations** and ignores clustering → underestimated uncertainty and unrealistic imputations.

Solution: multilevel (two-level) imputation models

Each imputation model includes random effects to account for cluster structure.

- **Benefits:** preserves between-cluster heterogeneity and ensures compatibility with mixed-effects analysis models.
- **Disdvantage:** often more complex to run in practice, convergence issues, dealing with system-missing data (variable missing in the entire cluster/study).

Rubin’s Rules

We fit the analysis model separately in each multiply imputed dataset.
Combine results using **Rubin’s Rules:**

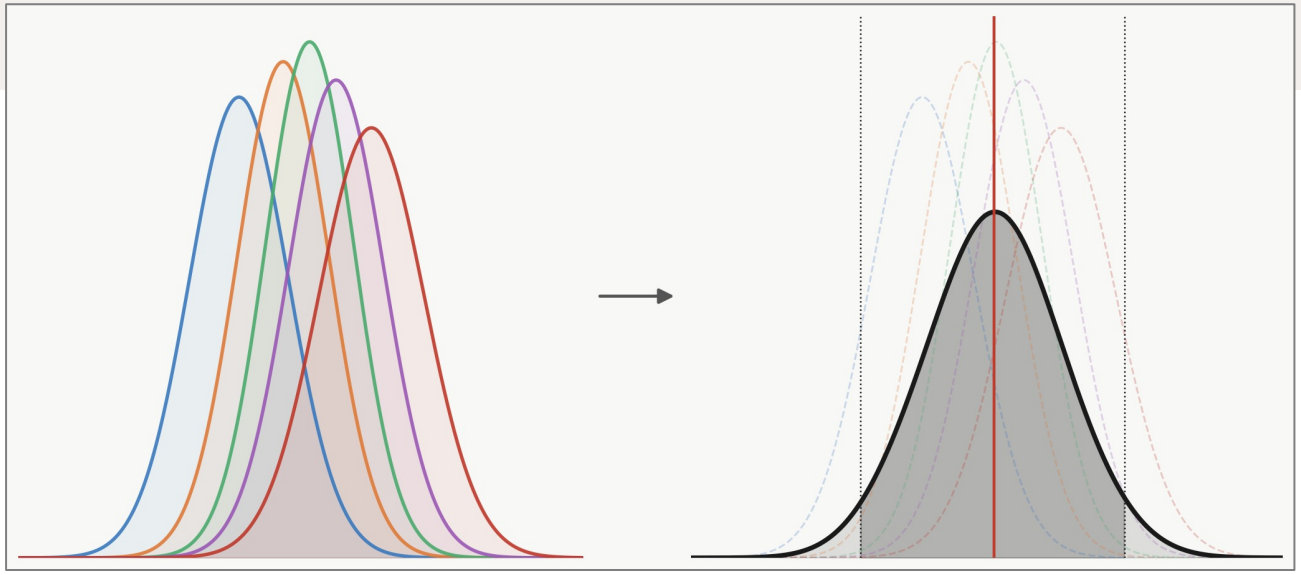
Pooled estimate: mean of parameter estimates across datasets.
Variance/confidence intervals: combines within-imputation and between-imputation variance.

Donald B. Rubin

$$\mathrm{Var}[Q|Y_{\text{obs}}] = \underbrace{\mathbb{E}\left[\mathrm{Var}[Q|Y_{\text{obs}}, Y_{\text{mis}}] \,\middle|\, Y_{\text{obs}}\right]}_{\substack{\text{mean of the variances across MI sets} \\ \rightarrow \textbf{Within-Variance } (\bar{U})}} + \underbrace{\mathrm{Var}\left[\mathbb{E}[Q|Y_{\text{obs}}, Y_{\text{mis}}] \,\middle|\, Y_{\text{obs}}\right]}_{\substack{\text{variance of the means across MI sets} \\ \rightarrow \textbf{Between-Variance } (B)}}$$

This accounts for uncertainty introduced by missing data.
Important: never merge imputed datasets before analysis.

Posterior Draw Mixing



Rubin's Rules were derived for **frequentist estimates** (point estimate & standard error).

When the analysis model is Bayesian and produces full posterior distributions, a more direct approach can be used (Zhou & Reiter 2014):

- Run the Bayesian model separately in each of the m imputed datasets
- Obtain posterior draws $p\left(\theta \mid Y^{(s)}\right)$ from each dataset
- Combine draws across datasets (equal number from each), and calculate results (median, credibility interval, …)

Key advantages:

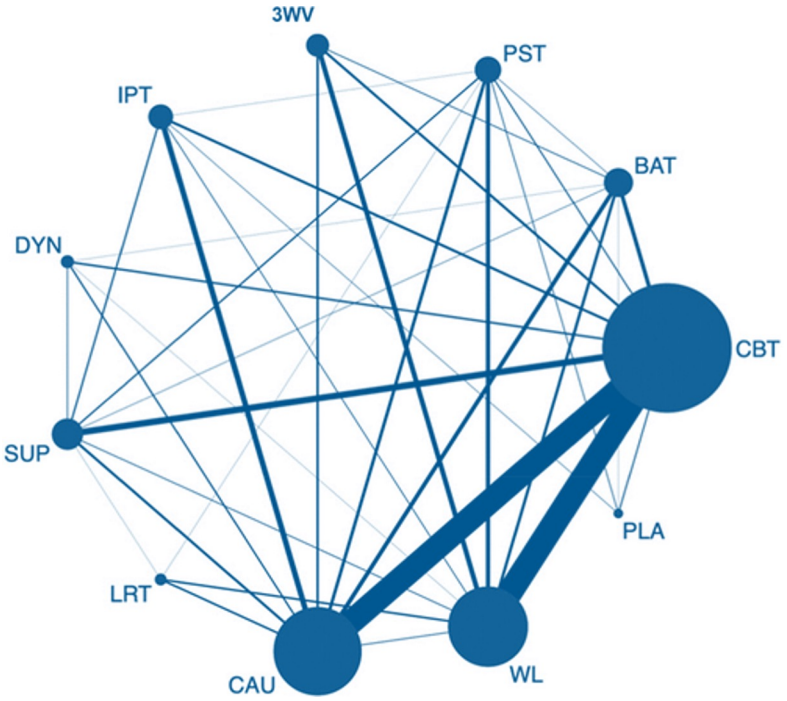
- Produces the full posterior distribution, not just mean and variance
- Preserves skewness, heavy tails, and nonlinear parameter relationships

METAPSY

Network Meta-Analysis

(IPD) Network Meta-Analysis

- Combines **evidence across many treatments** in a single statistical model
- Integrates **direct and indirect comparisons** to estimate all treatment contrasts
- Models **patients nested within studies** in a hierarchical framework
- Uses **individual patient data (IPD)** to study effect modifiers and heterogeneity
- Enables **treatment ranking and personalized treatment effect prediction**



→ **NMA involves complex hierarchical model**

→ **JAGS offers a straightforward and flexible implementation**

METAPSY

Network Meta-Analysis (Aggregate Data)

Likelihood with observed effects (y) and S.E.

$$\bar{y}_{i,k} \mid \mu_i, \delta_{i,k} \sim \mathcal{N}\left(\mu_i + \delta_{i,k},\ \widehat{\mathrm{se}}_{i,k}^{2}\right)$$

Random effects

$$\delta_{i,2} \mid d, \tau \sim \mathcal{N}\left(d_{t_{i,2}} - d_{t_{i,1}},\ \tau^2\right)$$

Random effects (adjusted P-matrix for multi-arm trial)

$$\boldsymbol{\delta}_i \mid d, \tau \sim \mathcal{N}_{n_a-1}\left(\boldsymbol{\theta}_i,\ \mathbf{P}_{n_a}^{-1}\right), \qquad \theta_{i,k} = d_{t_{i,k+1}} - d_{t_{i,1}}$$

$$P_{jl} = \tau^{-2} \times \begin{cases} \dfrac{2(n_a - 1)}{n_a} & j = l \\ -\dfrac{2}{n_a} & j \neq l \end{cases}$$

```
model {
  d[1] <- 0
  for (t in 2:Nt) { d[t] ~ dnorm(0, 0.001) }

  tau      ~ dunif(0, 5)
  prec.tau <- 1 / (tau * tau)

  # --- 2-arm studies ---
  for (i in 1:N2) {
    mu.2[i] ~ dnorm(0, 0.0001)
    y.2[i,1] ~ dnorm(mu.2[i],                prec.2[i,1])
    y.2[i,2] ~ dnorm(mu.2[i] + delta.2[i],   prec.2[i,2])
    delta.2[i] ~ dnorm(d[t.2[i,2]] - d[t.2[i,1]], prec.tau)
  }

  # --- 3-arm studies (multivariate RE) ---
  for (i in 1:N3) {
    ...
    xi.3[i,1:2] ~ dmnorm(theta.3[i,1:2], prec.la.3[1:2,1:2])
    for (kk in 1:2) { theta.3[i,kk] <- d[t.3[i,kk+1]] - d[t.3[i,1]] }
  }

  # Lu-Ades precision matrix (3-arm)
  prec.la.3[1,1] <- 1.3333 * prec.tau;  prec.la.3[1,2] <- -0.6667 * prec.tau
  prec.la.3[2,1] <- -0.6667 * prec.tau; prec.la.3[2,2] <-  1.3333 * prec.tau
}
```

Priors

$$d_1 = 0, \quad d_t \sim \mathcal{N}(0, 10^3)\ (t \geq 2), \quad \mu_i \sim \mathcal{N}(0, 10^4), \quad \tau \sim \mathrm{Uniform}(0, 5)$$

Network Meta-Analysis (IPD)

Stage 1: Individual-level model with moderators γ

$$y_{ij} \sim \mathcal{N}(\mu_{ij},\ \sigma_i^2)$$
$$\mu_{ij} = \alpha_i + \delta_{i,k(j)} + \boldsymbol{\beta}_i^\top \mathbf{x}_{ij}^c + \boldsymbol{\gamma}_{i,k(j)}^\top \mathbf{x}_{ij}^c$$
$$\delta_{i,1} = 0, \quad \boldsymbol{\gamma}_{i,1} = \mathbf{0} \quad \text{(within-study reference arm)}$$

Small ridge penality
(facilitates computation)

Data summaries for stage 2

$$\hat{\boldsymbol{\theta}}_i = \big(\underbrace{\delta_{i,2},\dots,\delta_{i,n_a}}_{\text{trt effects}},\ \underbrace{\gamma_{i,2,1},\dots,\gamma_{i,n_a,1},\ \gamma_{i,2,2},\dots,\gamma_{i,n_a,P}}_{\text{interactions, by covariate}}\big)^\top \qquad \hat{\boldsymbol{\Sigma}}_i = \mathrm{Cov}\big(\text{concat draws}\big) + 10^{-6}\mathbf{I}$$

Stage 2: Multivariate NMA

$$\hat{\boldsymbol{\theta}}_i \ \sim\ \mathcal{N}_{D_i}\big(\boldsymbol{M}_i,\ \hat{\boldsymbol{\Sigma}}_i\big)$$

$$\text{Treatment-effect block (RE):}\ \ \boldsymbol{\delta}_i \sim \mathcal{N}_{n_a-1}\big(d_{t_{i,2:n_a}} - d_{t_{i,1}},\ \mathbf{P}_{n_a}^{-1}\big)$$
$$\text{Interaction block (FE):}\ \ M_{i,\,\mathrm{mod}(k,p)} = \gamma^*_{t_{i,k+1},p} - \gamma^*_{t_{i,1},p}$$

Moderators for each treatment

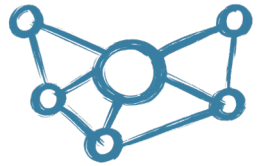
Personalized prediction for some patient

$$\Delta_t(\mathbf{x}^*) = \big(d_t + \boldsymbol{\gamma}_t^{*\top}\mathbf{x}^{*c}\big) - \big(d_1 + \boldsymbol{\gamma}_1^{*\top}\mathbf{x}^{*c}\big)$$

Priors on Level 2

$$d_1 = 0, \quad d_t \sim \mathcal{N}(0,1)\ (t \geq 2), \quad \gamma^*_{1,p} = 0, \quad \gamma^*_{t,p} \sim \mathcal{N}(0, 0.1)\ (t \geq 2), \quad \tau \sim \mathrm{Uniform}(0,2)$$

```
## Level 1 Model
model {
  for (j in 1:N) {
    y[j] ~ dnorm(mu[j], prec.y)
    mu[j] <- intercept
           + trt.eff[trt[j]]
           + inprod(beta[1:Nprog],        prog.mat[j, 1:Nprog])
           + inprod(tmod[trt[j], 1:Nprog], prog.mat[j, 1:Nprog])
  }
  prec.y    <- 1 / (sigma * sigma);  sigma     ~ dunif(0, 5)
  intercept  ~ dnorm(0, 0.01)
  trt.eff[1] <- 0
  for (k in 2:na) { trt.eff[k] ~ dnorm(0, 0.01) }
  for (p in 1:Nprog) {
    beta[p]    ~ dnorm(0, 0.01);  tmod[1, p] <- 0
    for (k in 2:na) { tmod[k, p] ~ dnorm(0, 0.01) }
  }
}

## Level 2 Model
model {
  d[1] <- 0
  for (t in 2:Nt) { d[t] ~ dnorm(0, 1) }

  for (p in 1:Nprog) { tmod.global[1,p] <- 0 }
  for (t in 2:Nt) { for (p in 1:Nprog) { tmod.global[t,p] ~ dnorm(0, 0.1) } }

  tau  ~ dunif(0, 2);  prec <- 1/(tau*tau)

  # --- 2-arm studies ---
  for (i in 1:N2) {
    y2[i, 1:D2] ~ dmnorm(Mu2[i, 1:D2], invV2[i, 1:D2, 1:D2])
    Mu2[i, 1] ~ dnorm(d[t2[i,2]] - d[t2[i,1]], prec)   # RE treatment
    for (p in 1:Nprog) {                                # FE interactions
      Mu2[i, 1*p + 1] <- tmod.global[t2[i,2], p] - tmod.global[t2[i,1], p]
    }
  }
  # (3-arm groups analogous, with multivariate RE for treatment positions)
}
```



METAPSY

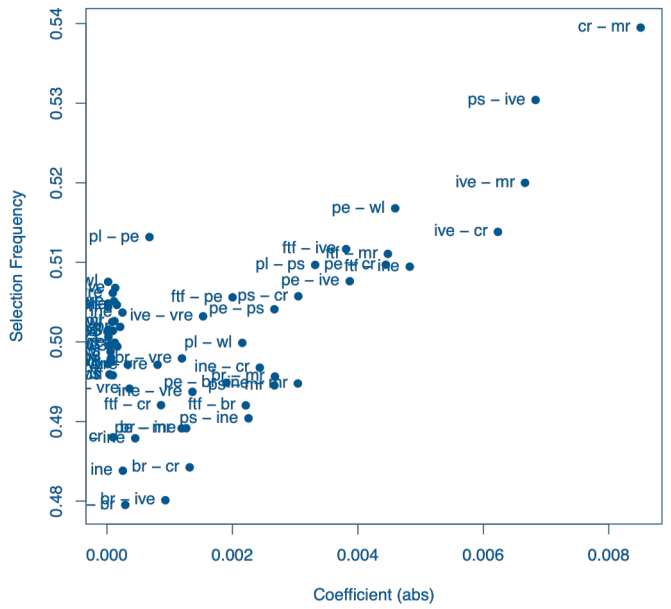
Other Models

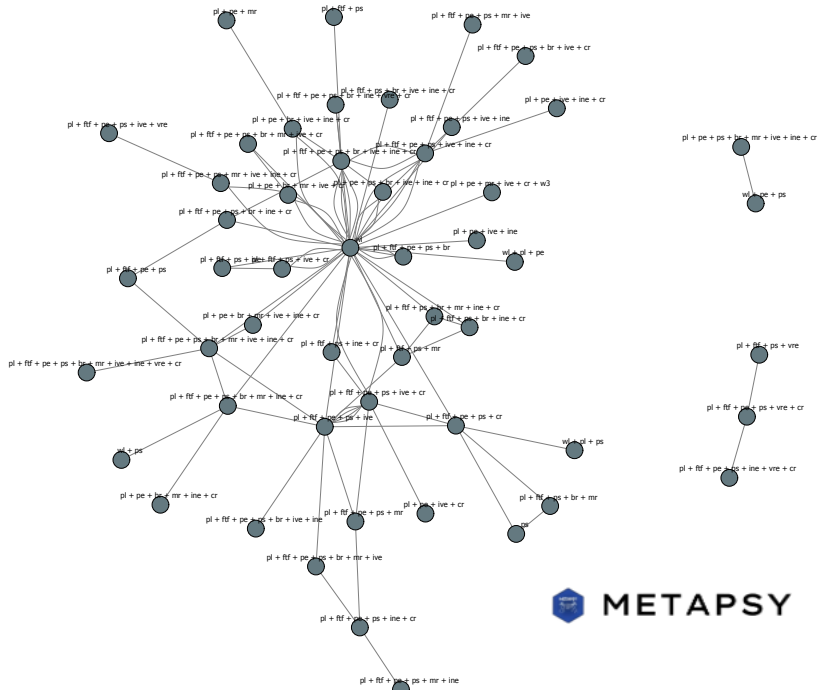
Other Models

Other example models included in the scripts folder:

- IPD-MA with moderators
- IPD-MA with moderators (binomial likelihood)
- IPD-cNMA with component interactions
- IPD-cNMA with component interactions (binomial likelihood)

	Depression severity (iMD of PHQ–9 scores), median (95% CrI)
Age	0·19 (–0·09 to 0·47)
Baseline depression, PHQ–9 scores	2·59 (2·32 to 2·85)
Gender*	–0·03 (–0·28 to 0·18)
Relationship†	–0·12 (–0·33 to 0·12)
Waiting component	0·42 (–0·75 to 1·53)
Non-specific treatment effects	–1·41 (–2·52 to –0·30)
Psychoeducation about depression	0·02 (–0·86 to 0·93)
Cognitive restructuring	0·30 (–0·87 to 1·41)
Behavioural activation	–1·83 (–2·90 to –0·80)
Interpersonal skills training	–0·54 (–1·59 to 0·52)
Problem solving	–0·64 (–1·41 to 0·09)
Relaxation	1·20 (0·17 to 2·27)
Third-wave components	–0·53 (–1·55 to 0·49)
Behaviour therapy for insomnia	–1·82 (–3·92 to 0·26)
Relapse prevention	0·35 (–0·69 to 1·32)
Homework required	0·31 (–0·69 to 1·35)
Initial face-to-face contact	0·85 (–1·80 to 3·41)
Automated encouragement to proceed with iCBT	–0·26 (–1·13 to 0·60)
Human encouragement to proceed with iCBT	–0·29 (–1·17 to 0·58)
Therapeutic guidance for iCBT	0·01 (–0·88 to 0·89)





METAPSY

VU VRIJE UNIVERSITEIT AMSTERDAM

WHO Collaborating Centre for Research and Dissemination of Psychological Interventions

Thank You!