Introduction to Bayesian methods - II

Rita Almeida

2nd of February, 2016

Overview

Previous part:

- Bayes theorem
- Examples with discrete and continuous normal variables
- Posteriors, priors and likelihood
- Comparison with frequentist approaches

This part:

- Bayesian estimation
- Bayesian testing
- Bayesian model comparison
- Methods to calculate the posterior
- Bayesian linear and hierarchical linear models

Overview

- Bayesian estimation
- Bayesian testing
- Bayesian model comparison
- Methods to calculate the posterior
- Bayesian linear and hierarchical linear models

Bayesian estimation

The posterior represents all the information for θ.
Sometimes one wants to provide a single value - estimator.

- For $\delta(Y)$ to be a good estimator of θ the probability of $\delta(Y) - \theta$ being close to 0 must be high.
- Let $L(\theta, a)$ be the loss when the estimate is a and the true value is θ.
  - A common loss function is $L(\theta, a) = (\theta - a)^2$.
- The Bayes estimator $\delta^*(y)$ of θ is:

$$E[L(\theta, \delta^*(y))|y] = \min_{a \in \Omega} E[L(\theta, a)|y]$$

Bayesian estimation - example

Someone answers a test with 12 true-or-false equally difficult questions. The answers are independent. 9 correct answers. Estimate the probability of answering a question correctly.

- Likelihood: binomial $n = 12$, $y = 9$

$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

- Prior: uniform - Beta distribution with parameters $\alpha_0 = 1$, $\beta_0 = 1$
  - Mean of a Beta distribution with α, β is $\frac{\alpha}{\alpha+\beta}$

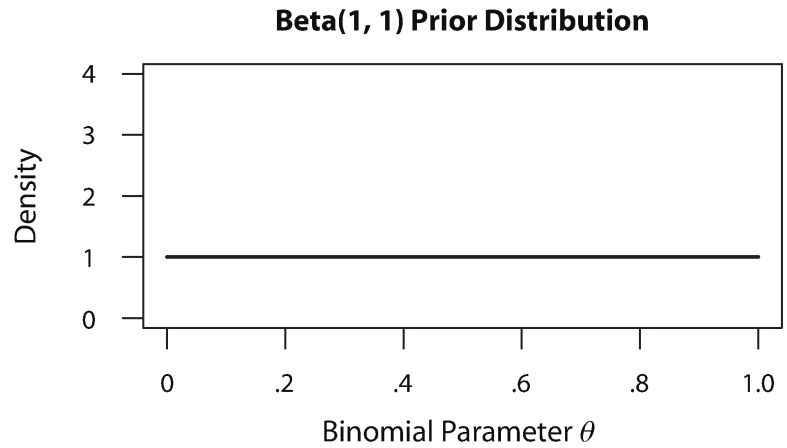


Figure adapted from Wagenmakers 2007

Bayesian estimation - example

- Posterior: Beta with parameters
  - $\alpha = \alpha_0 + y = 10$,
  - $\beta = \beta_0 + n - y = 4$ and $n = 12$

$$p(\theta|y = 9) = 13\binom{12}{9}\theta^9(1-\theta)^3$$

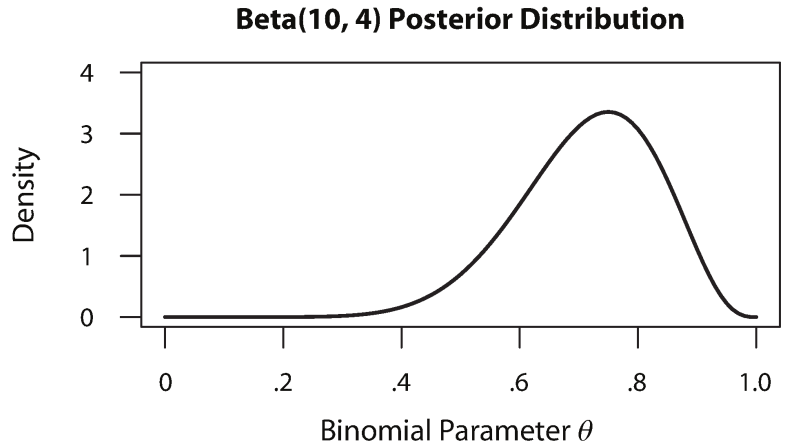


Figure adapted from Wagenmakers 2007

- Considering $L(\theta, a) = (\theta - a)^2$:

$$\delta^*(y) = \frac{\alpha_0 + y}{\alpha_0 + y + \beta_0 + n - y} = 0.71$$

  - The estimator is calculated from the posterior.

Bayesian estimation - example with normal

- Suppose we have a random sample $y_1, y_2, \ldots, y_n$ from a normal distribution with unknown mean θ and known variance σ^2.
- Considering a normal prior with mean μ and variance ν^2:

$$p(\theta) \quad \text{is} \quad N(\mu, \nu^2)$$

- The posterior becomes:

$$\theta|y_1, y_2, \ldots, y_n \sim N(\mu_1, \nu_1^2) \qquad \text{with} \qquad \mu_1 = w\ \mu + (1-w)\ \overline{y}_n$$

- Considering $L(\theta, a) = (\theta - a)^2$: $\qquad \delta^*(y) = \mu_1$

Overview

- Bayesian estimation
- Bayesian testing
- Bayesian model comparison
- Methods to calculate the posterior
- Bayesian linear and hierarchical linear models

Bayesian testing

- Hypotheses can be described as prior beliefs:
  $p(\mathrm{H}_0)$, $p(\mathrm{H}_1)$, $p(\mathrm{H}_2)$,...
- Posteriors after observing the data (2 hypotheses):
  $p(\mathrm{H}_0|y)$, $p(\mathrm{H}_1|y)$
- Posterior odds in favor of H_0 comparing to H_1

$$\underbrace{\frac{p(\mathrm{H}_0|y)}{p(\mathrm{H}_1|y)}}_{\text{posterior odds}} = \underbrace{\frac{p(y|\mathrm{H}_0)}{p(y|\mathrm{H}_1)}}_{\text{Bayes factor}} \underbrace{\frac{p(\mathrm{H}_0)}{p(\mathrm{H}_1)}}_{\text{prior odds}}$$

- If H_0 is complement of H_1:

$$p(\mathrm{H}_0|y) = \left(\frac{1}{B}\frac{p(\mathrm{H}_1)}{p(\mathrm{H}_0)} + 1\right)^{-1}$$

- If $p(\mathrm{H}_0) = p(\mathrm{H}_1)$ one can report Bayes factor.

Bayesian testing - example

Someone answers a test with 12 true-or-false equally difficult questions. The answers are independent. 9 correct answers. What is the probability of random gessing?

- $H_0 : \theta = 1/2$, θ is fixed
  $H_1 : \theta \neq 1/2$, $\theta \sim \text{Beta}(1, 1)$
- The Bayes factor is:

$$B = \frac{p(y|\theta = 1/2)}{p(y|\theta \sim \text{Beta}(1,1))} = \frac{\binom{12}{9}(\frac{1}{2})^{12}}{\int_0^1 p(y|\theta)p(\theta)d\theta} = 0.7$$

- The data is $1/0.7 = 1.4$ times more likely under H_1.
- If the priors are equal $p(H_0|y) = 0.41$.
- One determines $p(H_0|y)$ instead of $p(y|H_0)$!

Summary

Advantages of Bayesian methods:

- Focus on the data collected, not on the average if one would collect many sets of data.
- Logic / more intuitive
  - Frequentist hypotheses testing does not necessarily do what one would like:
  - Does not give the probability of the null hypothesis.
  - If the sample is large enough a small effect becomes significant.
- Principled way to take into consideration prior knowledge and incorporate new evidence.
- Unified flexible approach

Overview

- Bayesian estimation
- Bayesian testing
- Bayesian model comparison
- Methods to calculate the posterior
- Bayesian linear and hierarchical linear models

Bayesian model comparison

- In general, for a model m:

$$p(m|y) = \frac{p(y|m)p(m)}{p(y)} \qquad p(y|m) \text{ is the model evidence}$$

- Hypotheses can be seen as models.
- One can compute a Bayes factor (B_{ij}) for two models m_i and m_j:

$$B_{ij} = \frac{p(y|m_i)}{p(y|m_j)}$$

- With equal priors for models i and j, B_{ij} is enough to compare the models.
- If B_{ij} is large model i is more likely than model j.

Bayesian model comparison

- For a model m with parameters θ:

$$p(\theta|y,m) = \frac{p(y|\theta,m) \cdot p(\theta|m)}{p(y|m)}$$

- the model evidence is:

$$p(y|m) = \int p(y|\theta,m)p(\theta|m)d\theta$$

Overview

- Bayesian estimation
- Bayesian testing
- Bayesian model comparison
- Methods to calculate the posterior
- Bayesian linear and hierarchical linear models

Posterior distribution calculations

- Bayesian inference is based on the posterior distribution and functions of the posterior distribution.
- These calculations are often difficult.
- Methods used:
  - Monte Carlo/ sampling approximations
    - asymptotically exact
    - simple idea
    - computationally expensive
    - Markov Chain Monte Carlo (MCMC) sampling
  - Deterministic approximations
    - not exact
    - computationally efficient
    - Variational Bayes

Monte Carlo methods

Idea: instead of calculating the posterior analytically, approximate it (or a function of it) by sampling.

Example:

- $g(\theta)$ is a function of θ.
- The aim is to estimate $E[g(\theta)|y]$, but $p(\theta|y)$ is not possible to integrate analytically.
  - Example from estimation: $E[L(\theta, a)|y], \quad L(\theta, a) = (\theta - a)^2$.
- One generates a sample of size n of θ:
  - set of independent and identically distributed $\theta_1, \theta_2, \ldots, \theta_n$
- $E(g(\theta)|y)$ is estimated by:

$$\hat{g} = \frac{1}{n} \sum_{i=1}^{n} g(\theta_i)$$

Markov Chain Monte Carlo

Problem: How to generate a sample $\theta_1, \theta_2, \ldots, \theta_n$ from a posterior distribution $p(\theta|y)$?

- In general it is not simple to generate random numbers according to a given distribution.
- Markov Chain Monte Carlo (MCMC) methods provide a way to generate the samples.
  - MCMC methods are based on constructing a Markov chain on the space of θ whose steady distribution as the desired probability $p(\theta|y)$.
  - The Metropolis-Hastings algorithm and Gibbs sampling are MCMC methods.
  - Computationally expensive.

Variational Bayes

Idea: approximate the posterior distribution by a distribution that is easier to use.

- Variational Bayes (VB) is a particular approach to approximate a distribution.
- It is not exact.
- It is less computationally expensive - fast.
- Often difficult to derive.
- It has been very used in neuroimaging.

Overview

- Bayesian estimation
- Bayesian testing
- Bayesian model comparison
- Methods to calculate the posterior
- Bayesian linear and hierarchical linear models

General linear model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \beta_1 \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix} + \beta_2 \begin{bmatrix} x_{12} \\ \vdots \\ x_{n2} \end{bmatrix} + \ldots + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$Y = X\beta + \epsilon \qquad \epsilon \sim N(0, \sigma^2 I)$$

- One finds parameters $\hat{\beta}$'s so that $\hat{y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots$ is as close as possible to y.
- The ordinary least square estimator of β is:
  $\hat{\beta} = (X'X)^{-1}X'y$
- If $\epsilon \sim N(0, \Sigma)$ generalized least squares estimator is:
  $\hat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$

Bayesian analysis of the general linear model

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N(0, \Sigma)$$

- Likelihood: $p(y|\beta) \sim N(X\beta, \Sigma)$
- Assuming a normal prior: $p(\beta) \sim N(\beta_0, \Sigma_0)$
- The posterior is normal: $p(\beta|y) \sim N(\hat{\beta}^*, \Sigma_p)$

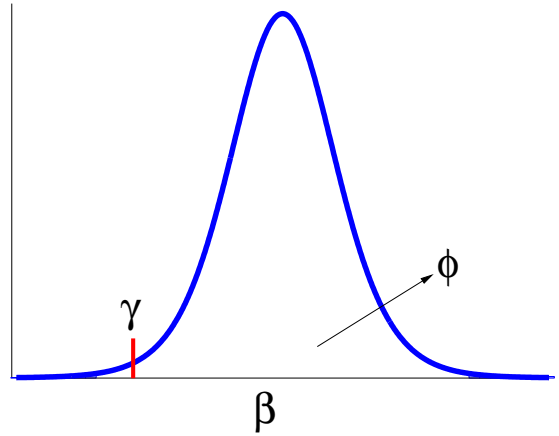
$$\hat{\beta}^* = (X'\Sigma^{-1}X + \Sigma_0^{-1})^{-1}(X'\Sigma^{-1}y + \Sigma_0^{-1}\beta_0)$$

- If Σ_0 is large one gets GLS: $\hat{\beta}^* \approx (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$

Bayesian analysis of the general linear model

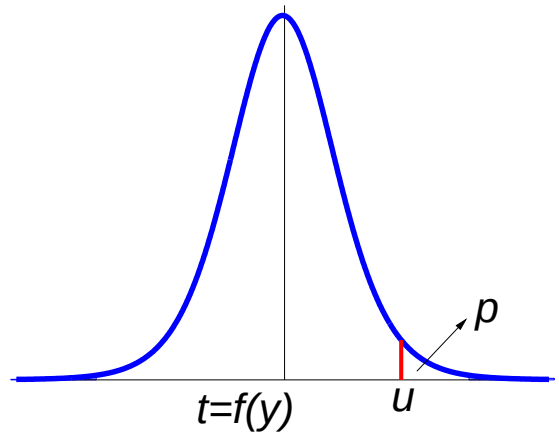
Bayesian approach: calculate probability ϕ that β exceeds some specific threshold γ given the data.

$$\phi = p(\beta > \gamma | y)$$



Frequentist approach: p values reflect how probable the data is given that there is no effect.

$$p = p(f(y) > u | \beta = 0)$$

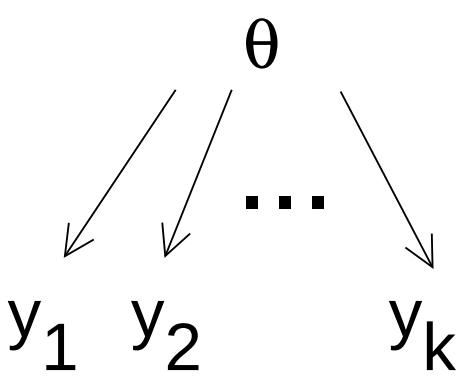


Multilevel/hierarchical models

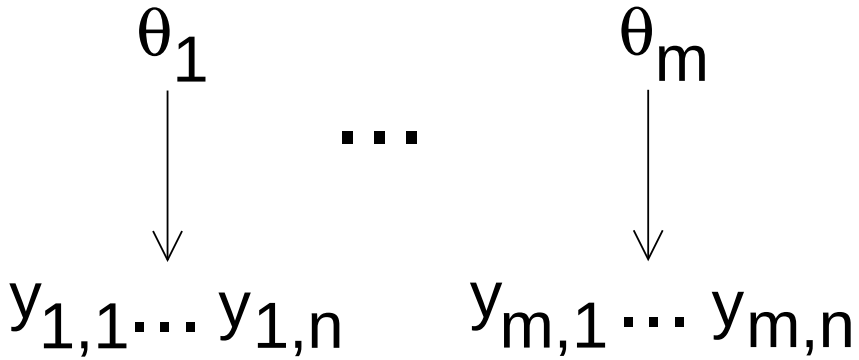
- Multilevel / hierarchical data: data with a clustered structure.
  - Example: repeated measures nested in subjects, subjects nested in medical doctor, doctor nested in hospital.

For repeated measures in subjects one could consider:

taking one measure per subject

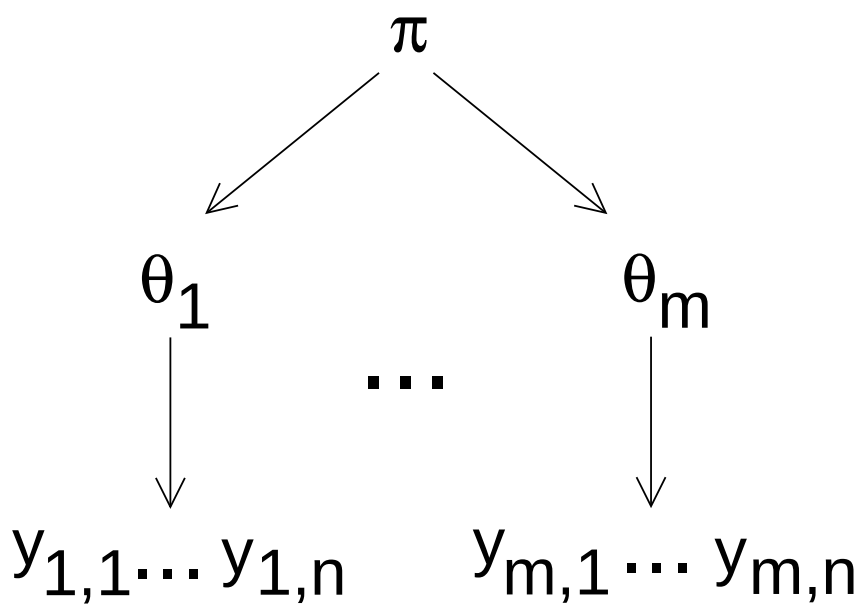


analyzing each subject separately



Multilevel/hierarchical models

- Multilevel / hierarchical models can be used to analyze such data.
  - Some parameters will model the data within a cluster and others across clusters.
  - Combine information across groups, but allow group specific characteristics.
  - Example: each subject can have its specific response, but there is also a common aspect.



Hierarchical Bayesian linear models

- Linear models are very commonly used.
- For clustered data linear models can be written as hierarchical models.
- Bayesian approach - hierarchical Bayesian linear models
  - using priors and estimating posterior distributions.

Hierarchical Bayesian linear models- example

Single-subject level:

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \Sigma)$$

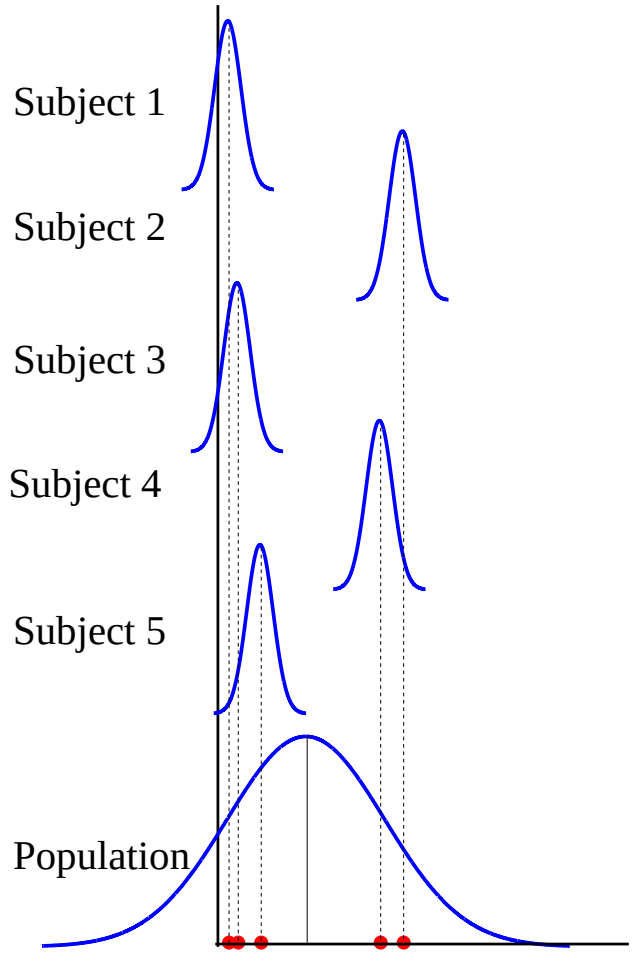
$$p(y|\beta) \sim N(X\beta, \Sigma)$$

Group level:

$$\beta = \beta_g + \eta, \quad \eta \sim N(0, \Sigma_g)$$

$$p(\beta|\beta_g) \sim N(\beta_g, \Sigma_g)$$

$$\beta_g \sim f$$

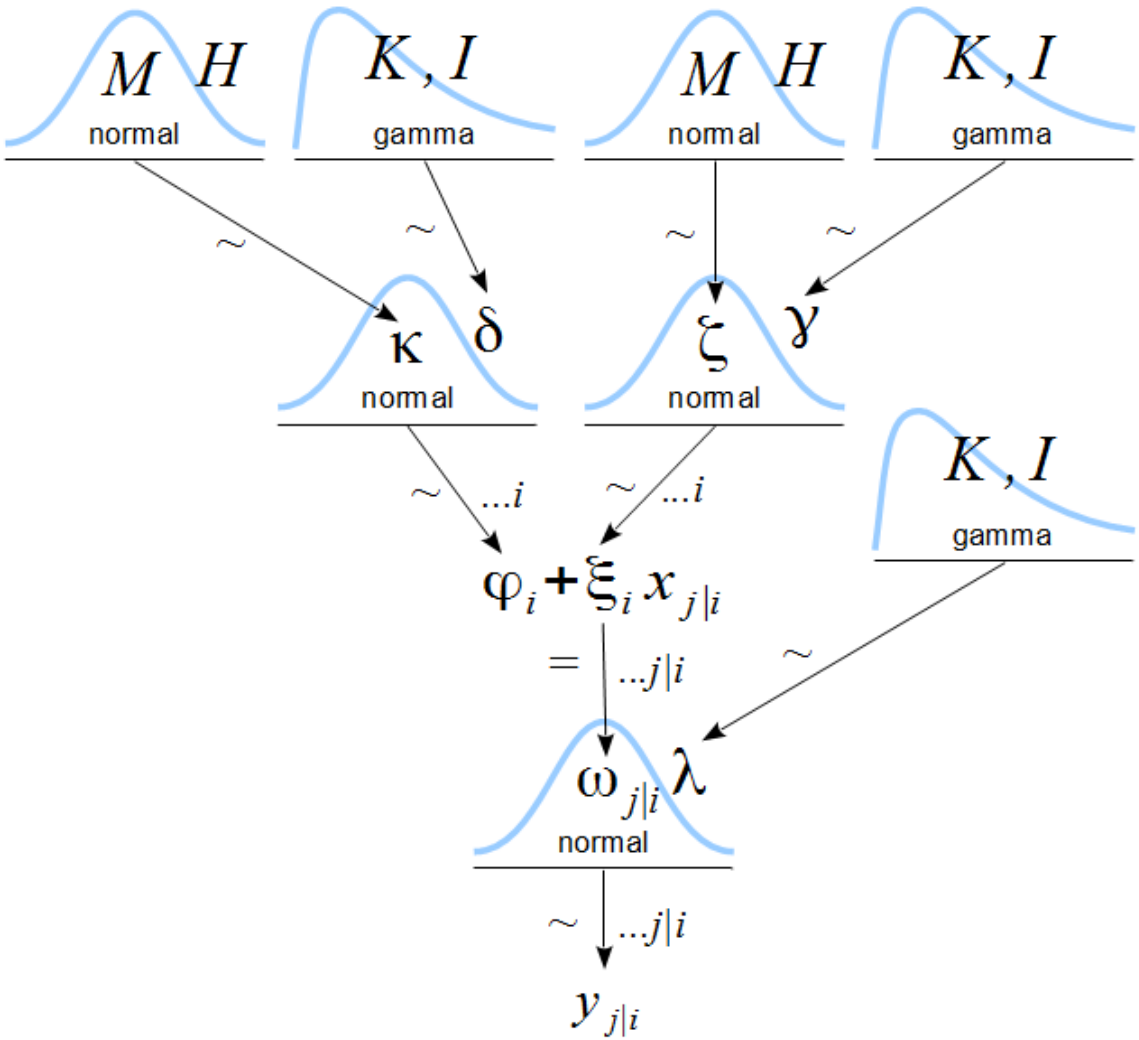


- Shrinkage: the cluster specific estimates get shrunk towards the common mean.
  - More uncertainty implies more shrinkage.
  - Example: unreliable subjects will count less.

Hierarchical Bayesian linear models - example

Example: Response measured across time for each subject.

- $y_{j|i}$ response for subject i at time $x_{j|i}$



$$y_{j|i} \sim N(\omega_{j|i}, \lambda)$$

$$\omega_{j|i} = \varphi_i + \xi_i x_{j|i}$$

φ_i individual intercept

ξ_i individual slope

$$\varphi_i \sim N(\kappa, \delta)$$

$$\xi_i \sim N(\zeta, \gamma)$$

$$\kappa, \zeta \sim N(M, H)$$

$$\lambda, \delta, \gamma \sim \Gamma(K, I)$$

Figure adapted from Kruschke 2014.

Empirical Bayes methods

If a prior distribution is known - full Bayesian approach.

If a prior distribution is not known:

- The data can be used to estimate:
  - the parameters of the prior distribution - hyperparameters or
  - the prior distribution.
- Empirical Bayes is used in inference with hierarchical Bayesian models.

References

- M. DeGroot and M. Schervish, *Probability and Statistics*, 2002.
- A. Gelman et al., *Bayesian Data Analysis*, 2014.
- J.K. Kruschke, *Doing Bayesian Data Analysis*, 2014.
- E.-J. Wagenmakers, A practical solution to the pervasive problems of p values, *Psychonomic Bulletin & Review*, 2007.