Advantages of Bayesian hierarchical modelling for constructing genetic linkage maps

Timothy Bilton[1,2], Matthew Schofield[2], Ken Dodds[1] and Mik Black[3]

[1] Invermay Agricultural Centre, AgResearch, Mosgiel, NZ
[2] Department of Mathematics and Statistics, University of Otago, Dunedin, NZ
[3] Department of Biochemistry, University of Otago, Dunedin, NZ

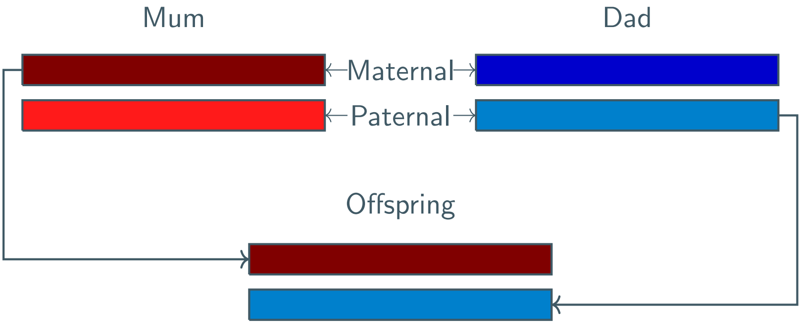
Acknowledgements

- Funding:
  - UO doctoral scholarship
  - UO Postgraduate Publishing Bursary
  - The Genomics for Production & Security in a Biological Economy programme
- AgResearch
- New Zealand eScience Infrastructure

Introduction

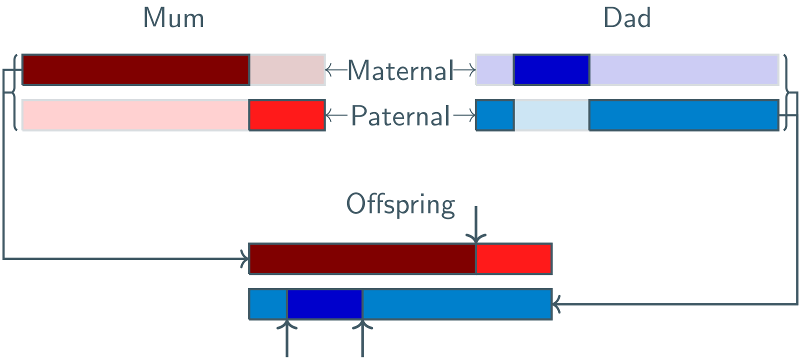
- Genetic maps give a 1-D representation of inheritance on a chromosome
  - Genetic markers (positions on a genome where variation is present)
  - Genetic distance between markers
- They form the basis of a number of genetic analyses, e.g.
  - Multipoint linkage analysis
  - Quantitative trait locus analysis
  - Estimation of historic population size

The genetic linkage map problem



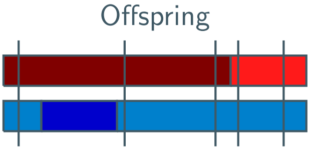
- Offspring inherits one chromosome from each parent

The genetic linkage map problem



- Offspring inherits one chromosome from each parent
- Meioses ⇒ genetic material inherited from both grandparents
- Change points are known as crossovers

The genetic linkage map problem



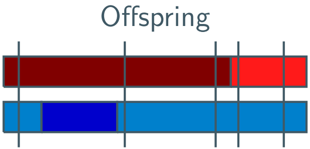
⇒

$$\begin{array}{r ccccc} \text{SNP} = & 1 & 2 & 3 & 4 & 5 \\ \boldsymbol{W}_{i\cdot 0} = & (\;0 & 0 & 0 & 1 & 1\;) \\ \boldsymbol{W}_{i\cdot 1} = & (\;1 & 1 & 1 & 1 & 1\;) \end{array}$$

- Introduce notation for genetic information at markers
  - W_{ijk} gives inheritance information by marker:
    - individual i $(i = 1, \ldots, N)$
    - marker j $(j = 1, \ldots, M)$
    - parent k $(k = 0$: mother, $k = 1$: father)

$$\boldsymbol{W}_{ij} = (\underbrace{W_{ij0}}_{mum}, \underbrace{W_{ij1}}_{dad})^T \qquad W_{ijk} = \begin{cases} 0 & \text{if maternally derived} \\ 1 & \text{if paternally derived} \end{cases}$$

Recombination



$$\Rightarrow \quad \begin{array}{rccccc} \text{SNP} = & 1 & 2 & 3 & 4 & 5 \\ \boldsymbol{W}_{i\cdot 0} = (& 0 & 0 & 0 & 1 & 1\,) \\ \boldsymbol{W}_{i\cdot 1} = (& 1 & 1 & 1 & 1 & 1\,) \end{array}$$

- Recombination: genetic material derived from different grandparents
- Occurs when odd # of crossovers
- Recombination fraction (ρ_j) between marker j and $j+1$:
    - Probability of a recombination, $\rho_j \in [0, 0.5]$

$$\hat{\rho}_j = \frac{1}{2N}\sum_{i=1}^{N}\left(|W_{ij+1\,0} - W_{ij0}| + |W_{ij+1\,1} - W_{ij1}|\right)$$

Genetic distance

- Genetic distance: # crossovers per chromosome between marker j and $j+1$.
  - Not a physical distance
  - Unit is Morgan (M): Average # of crossovers for 1 generation
    - Typically centimorgen (cM) is used (e.g., 1 cM = 0.01 M)
  - Genetic distance (δ_j) is a monotonic increasing function of ρ_j
    - Haldane mapping function

$$\delta_j = -0.5\log(1-2\rho_j)$$

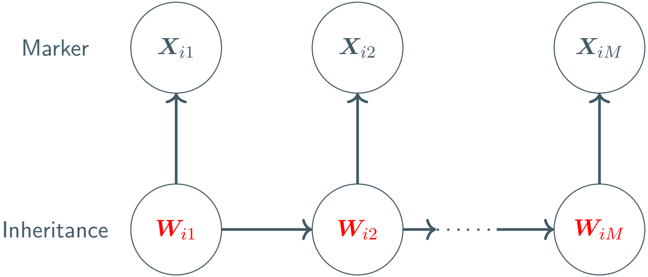
- Cumulative genetic distance (from marker j to h):

$$\Delta_{jh} = \sum_{m=j}^{h-1} \delta_m, \quad h > j$$

- Typically $50 < \Delta_{jh} < 150$

Genetic map: Modelling

- In practice, inheritance is unobserved (i.e., $\boldsymbol{W}_{ij}$ is latent)
- Marker genotypes are observed
  - Assuming biallelic SNPs (& diploids)
  - $X_{ij} = \#$ of major alleles in genotype $(X_{ij} = 0, 1, 2)$



Genetic maps for high-throughput sequencing (HTS)

- Marker data obtained using HTS technology
  - e.g., genotyping-by-sequencing, exome capture

- Low depth HTS data
  - Data consists of “reads”
    - Short sequence of DNA from a subset of the genome
    - Each read is derived from one of the parental chromosomes
    - Reads from one or both parents may not be observed
    - $\Rightarrow$ true marker information $\boldsymbol{X}$ is unobserved
  - Extend HMM to account for uncertainty in genotypes
    - Bilton et al. (2018) *Genetics*, 209:65-76
    - R package GUSMap (github.com/tpbilton/GUSMap)

High-throughput sequencing: HMM

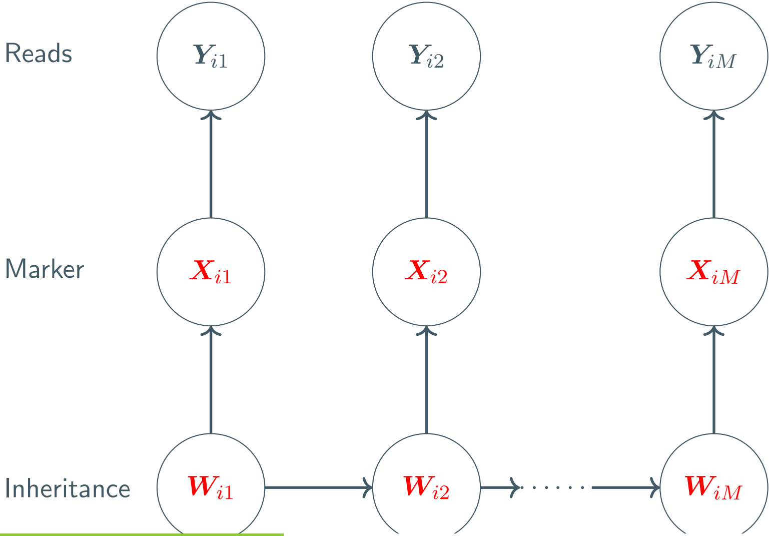
- Model reads conditional on the genotypes as:

$$Y_{ij}|(X_{ij}=x) \sim \text{Bin}\left(d_{ij}, p_{\varepsilon_j}\right)$$

$$p_{\varepsilon_j} = \begin{cases} \varepsilon_j & x=0 \\ 0.5 & x=1 \\ 1-\varepsilon_j & x=2 \end{cases}$$

- Y_{ij} is the # of reads of major allele
- d_{ij} = number of reads
- ε_j = probability of sequencing error
- i = individual & j = marker

High-throughput sequencing: model extension

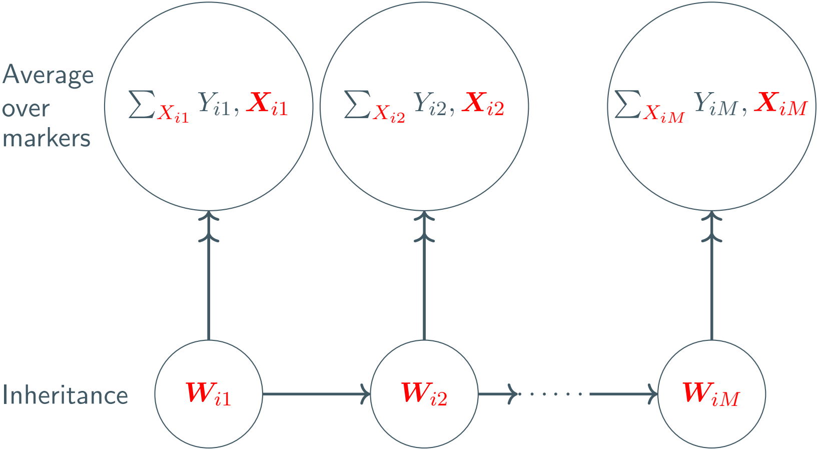


Reads $\boldsymbol{Y}_{i1}$ $\boldsymbol{Y}_{i2}$ $\boldsymbol{Y}_{iM}$

Marker $\boldsymbol{X}_{i1}$ $\boldsymbol{X}_{i2}$ $\boldsymbol{X}_{iM}$

Inheritance $\boldsymbol{W}_{i1}$ $\boldsymbol{W}_{i2}$ $\cdots\cdots$ $\boldsymbol{W}_{iM}$

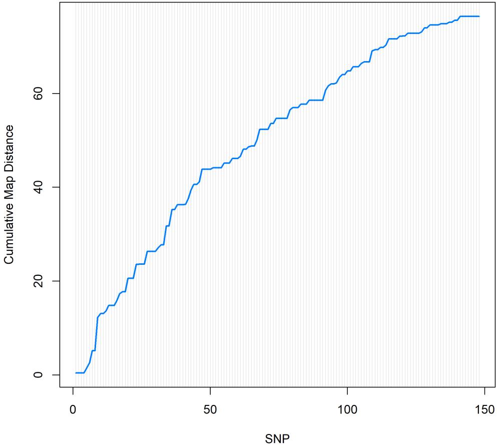
High-throughput sequencing: HMM


Average over markers
$\sum_{X_{i1}} Y_{i1}, \boldsymbol{X}_{i1}$
$\sum_{X_{i2}} Y_{i2}, \boldsymbol{X}_{i2}$
$\sum_{X_{iM}} Y_{iM}, \boldsymbol{X}_{iM}$
Inheritance
$\boldsymbol{W}_{i1}$
$\boldsymbol{W}_{i2}$
$\cdots\cdots$
$\boldsymbol{W}_{iM}$

Mānuka data

- *Leptospermum scoparium*
  - Native to NZ and South-East Australia
- Full-sib family of 177 plants
- Subset of SNPs located on chromosome 11
- SNPs filtered based on a range of criteria: 149 remaining
  - 95 are low depth (mean read depth: $\bar{d}_{\cdot j} < 6$)
  - 54 are high depth (80% of individuals had $d_{ij} \geq 20$)

Mānuka data



Mānuka plants:

- sequenced using GBS
- $N = 177$
- $M = 149$

Overall Map distance

- Usually between 50 and 150

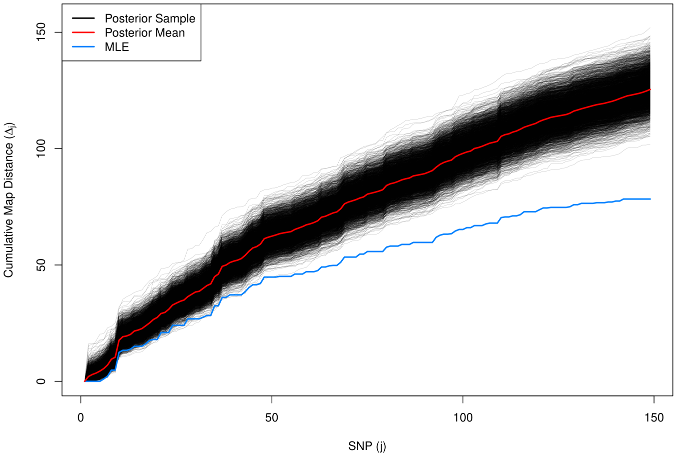
Why use Bayes?

Why considering using a Bayesian framework?

- Obtain uncertainty intervals
  - Many estimates on boundary $\rho_j = 0$
  - Various functions of parameters are of interest: Δ_{ij}
  - Makes quantifying uncertainty challenging in frequentist framework
- Enable more complex models to be fitted
  - Bayesian Hierarchical modelling
    - 'Borrow strengh' across parameters to improve estimates
    - Effectively applies shrinkage
    - Simplifies prior specification

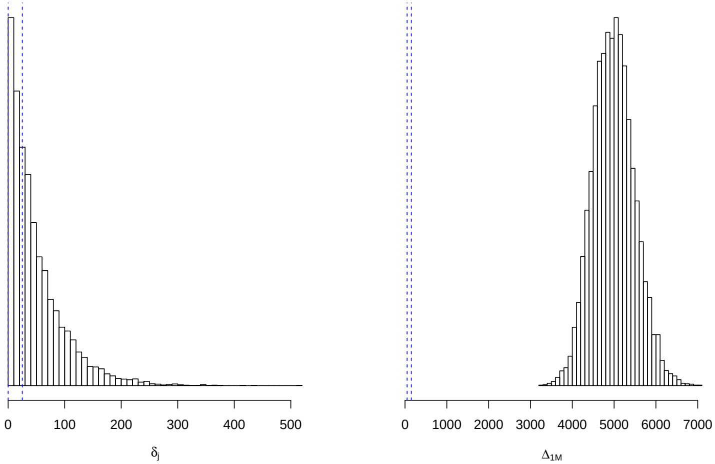
Bayes with uniform priors: Mānuka data

Independent uniform prior for ρ_j and ε_j



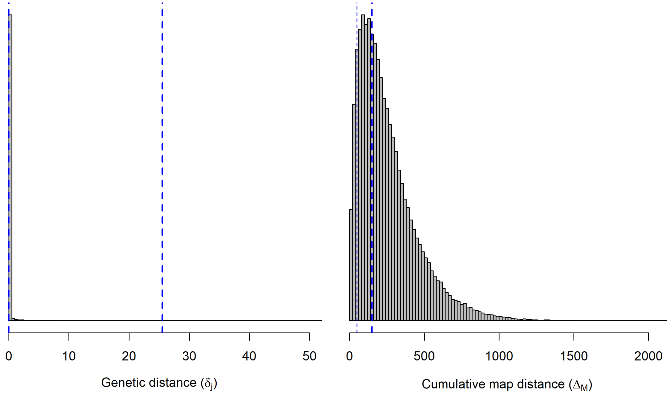
Bayes with uniform priors for ρ_j: Implied priors

- Implied priors for δ_j and Δ_{1M}



Bayes with gamma priors: Implied priors

- Gamma prior for δ_j
  - shape $=1.6384/(M-1)$, rate $= 0.0064$
  - Implied prior for Δ_{1M}: gamma with mode 100 and sd 200



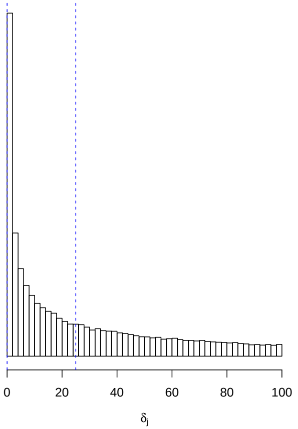
Bayesian hierarchical model: High level details

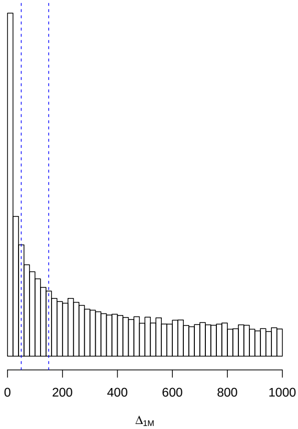
- Hierarchical components
  - cloglog$(2\rho_j) \sim N(\mu_\rho, \sigma^2_\rho), \quad j = 1, \ldots, M-1$
  - logit$(\epsilon_j) \sim N(\mu_\epsilon, \sigma^2_\epsilon), \quad j = 1, \ldots, M$
- Priors for means (on original scale and transform)
  - $f(\mu_\rho) \propto \left(1 - \exp(-e^{\mu_\rho})^2\right)^{a-1} \exp(\mu_\rho - e^{\mu_1}) \quad (a = 0.5)$
  - $f(\mu_\epsilon) \propto \exp(a\mu_\epsilon)(1 + \exp(\mu_\epsilon))^{-(a+b)} \quad (a = b = 0.5)$
- Priors for variance parameters:
  - $f(\sigma_\rho) =$ half-t$_3(0, 1)$
  - $f(\sigma_\epsilon) =$ half-t$_3(0, 1)$
- Use a non-centered parameterization
  - Improved MCMC convergence

agresearch
āta mātai, mātai whetū

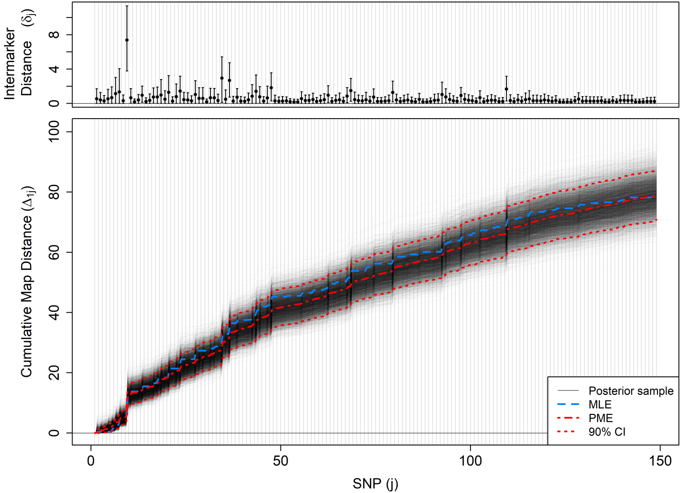
Bayesian hierarchical model: Priors

- Common model for ρ_j
  - 'Vague' priors for δ_j and Δ_{1M}

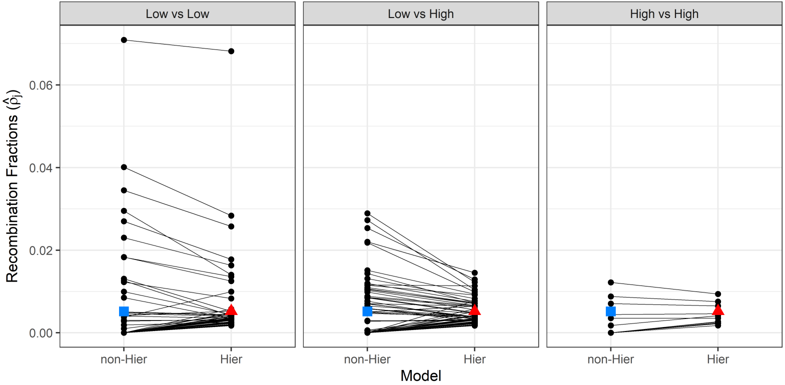




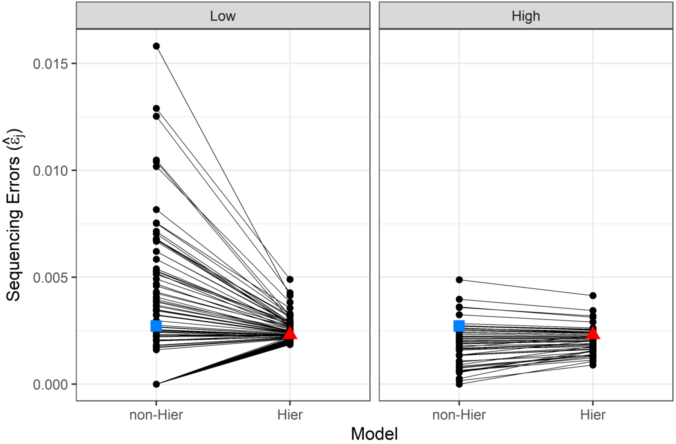
Mānuka data: Cumulative map distance



Mānuka data: Shrinkage plots (ρ)



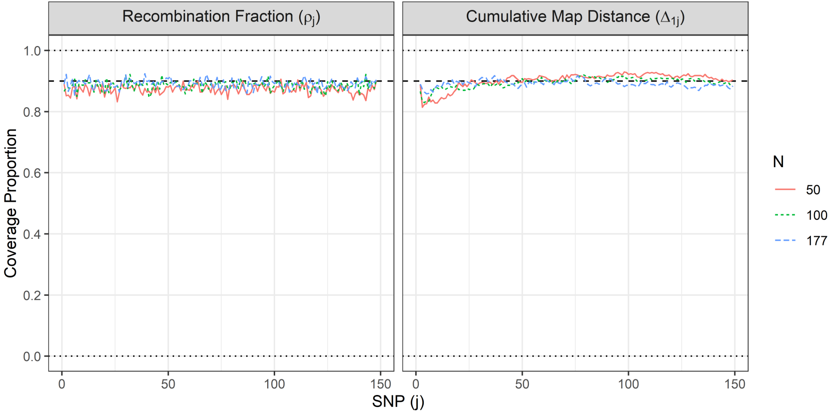
Mānuka data: Shrinkage plots (ϵ)



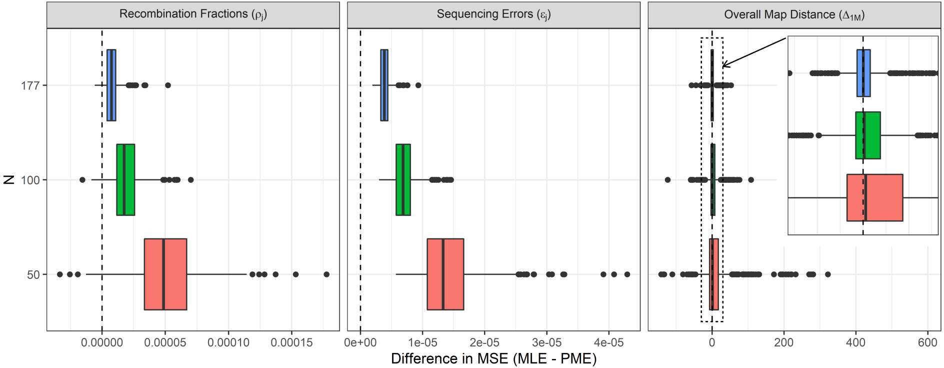
Simulations

- To examine properties of the hierarchical model in terms of:
  - Mean square error
  - Coverage
- Simulate data to resemble mānuka data
  - 500 simulated datasets
  - Δ_{1M} log normal with mean 80 and variance 5
  - $\delta_1, \dots, \delta_{M-1} = \Delta_{1M} \times z \;\; (z \sim Dir(0.25, \dots, 0.25))$
  - $\epsilon_j \sim Beta(3, 1497)$
  - $N = 50, 100, 177$ (randomly sample individuals)
  - Genotypes simulated using PedigreeSim (Voorrips & Maliepaard; 2012)
  - d_{ij} were those observed in the mānuka data
  - Y_{ij} were simulated based on the model given X_{ij}

Simulations: coverage



Simulations: mean square error



Summary

Bayesian modelling for genetic maps

- Prior specification very important
  - Consider implied priors
- Provides reliable uncertainty intervals
  - Parameter estimates on boundary
  - Parameters of interest that are function of other parameters

Bayesian hierarchical modelling

- Hierarchical model is more straightforward to fit with Bayes
- Simplifies prior specification
- Improves parameters estimates (i.e., MSE)