Analysis of RNAseq data[1] using ShrinkBayes

Mark van de Wiel
mark.vdwiel@vumc.nl

Department of Epidemiology & Biostatistics
VUmc, Amsterdam

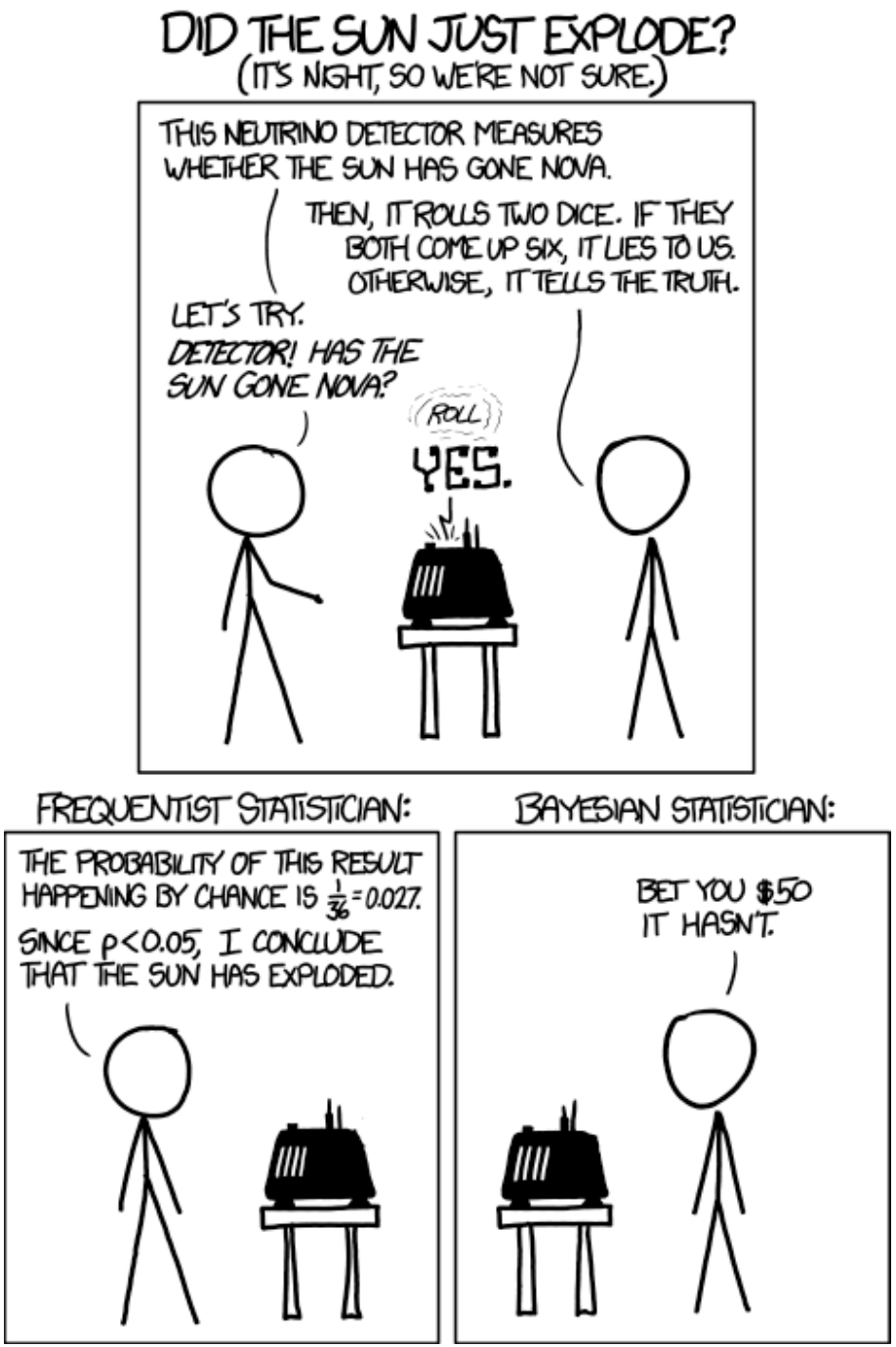
VU medisch centrum

[1]and other high-dimensional data

Contents

Overview

Some Bayesian statistics



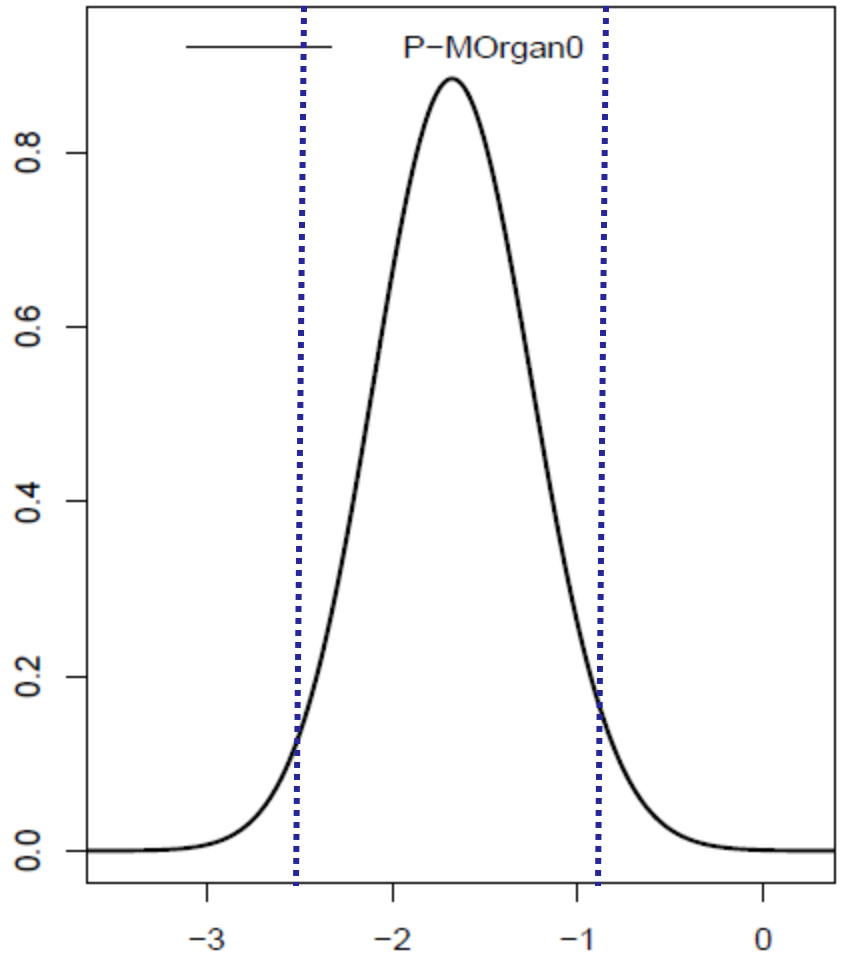
Source: http://xkcd.com/1132/

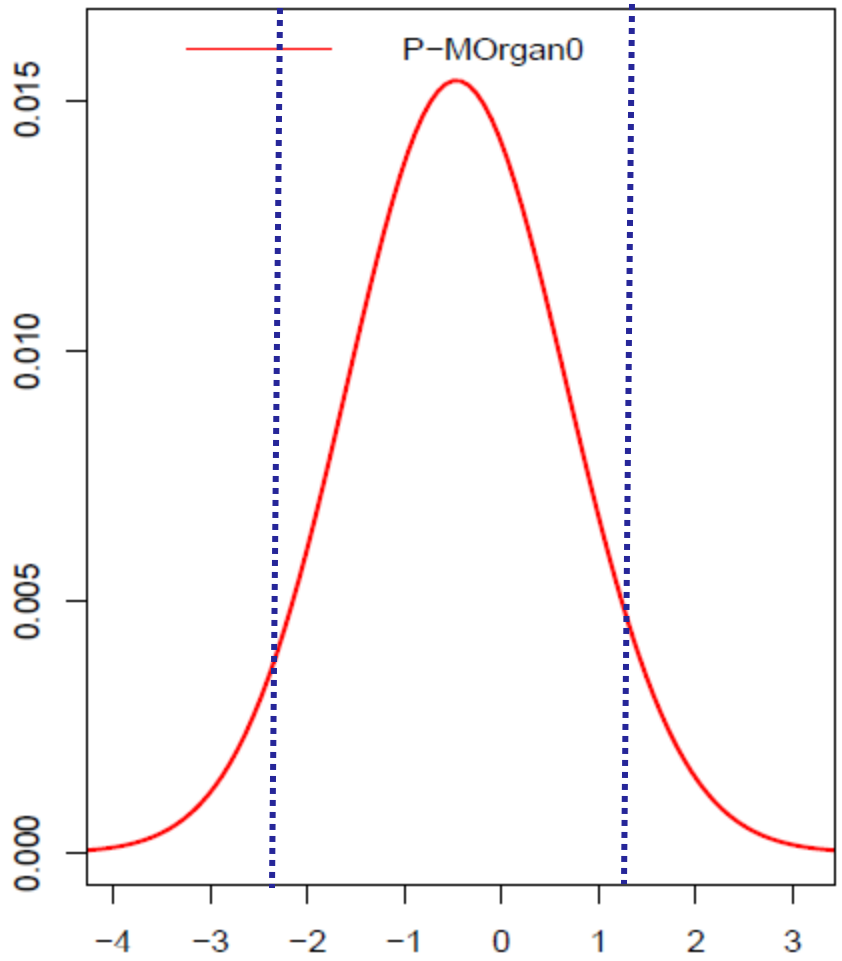
Bayes in a nutshell…

- Data: **Y**
- Parameter of interest: β
- Known: likelihood π(**Y**| β) and prior π(β)
- ?Posterior: π(β | **Y**)
- Bayes' rule: $\pi(\beta \mid \mathbf{Y}) = \frac{\pi(\mathbf{Y} \mid \beta)\pi(\beta)}{\pi(\mathbf{Y})} = \frac{\pi(\mathbf{Y} \mid \beta)\pi(\beta)}{\int_{\beta} \pi(\mathbf{Y} \mid \beta)\pi(\beta)d\beta}$
- 95% credibility interval: CI= $[q_{0.025}(\pi(\beta \mid \mathbf{Y}), q_{0.975}(\pi(\beta \mid \mathbf{Y})]$
- Simple inference: β ∈ CI?

Bayesian inference, single test

Parameter of interest: gene expression difference between two conditions: primary and metastasis. Posteriors:





Bayesian inference

Conjugate priors

Prior $\pi(\beta)$ is conjugate to likelihood $\pi(\mathbf{Y}\ \beta)$ when the posterior has an analytical form.

1. Normal-Normal
2. Normal-Normal-Inverse Gamma (for variance σ^2)
3. Binomial-Beta (=Beta-binomial distribution)
4. Poission-Gamma (= Negative Binomial)
5. .....

Mixtures of conjugates are also conjugate when the mixture components are known

Bayesian inference, conjugate priors, example

Suppose $Y_j \sim N(\mu, \sigma^2), \mu \sim N(\mu_0, \sigma_0^2)$

Assume all parameters known, except μ.

$$f(\mathbf{Y}|\mu) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \mu)^2\right]$$

$$\propto \exp\left[-\frac{1}{2}\left(\frac{\mu - \bar{y}}{\sigma/\sqrt{n}}\right)^2\right],$$

because $\sum_{j=1}^{n}(y_j - \mu)^2 = \sum_{j=1}^{n}(y_j - \bar{y})^2 + \sum_{j=1}^{n}(\mu - \bar{y})^2$

Bayesian inference, conjugate priors, example

Then, $\pi(\mu|\mathbf{Y}) \propto f(\mathbf{Y}|\mu)\pi(\mu) \propto \exp\left\{-\frac{1}{2}\left[\left(\frac{\mu-\mu_0}{\sigma_0^2}\right) + \left(\frac{\mu-\bar{y}}{\sigma/\sqrt{n}}\right)^2\right]\right\}$

Observe that term in exponent[1] can be written as $[(\mu - m)/s]^2$, so

$$\pi(\mu|\mathbf{Y}) = N(m, s),$$

with

$$m = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{1/\sigma_0^2 + n/\sigma^2}$$

$$s^2 = \frac{1}{1/\sigma_0^2 + n/\sigma^2}$$

[1]Ignoring constant -1/2

Bayesian inference, approximations

In many practical cases, exact formulas for the posteriors are not available, hence approximations are necessary:

1. Markov Chain Monte Carlo
   Very popular. Computationally intensive. Needs to be developed for each new problem. Very general though

2. Variational Bayes approximations.
   Fast. Approximate posteriors by analytical formulas assuming knowledge of conditional posteriors. Require specific development for each problem

3. Integrated Nested Laplace Integration (INLA)

INLA

Bayesian inference, GLM setting

Often: multiple unknown parameters, many regression parameters (β's) and unknown analytical solution.

$$Y_j \sim F(\mu_j, \boldsymbol{\theta}), \text{ e.g. } F() = N(\mu_j, \sigma^2), \text{ so } \boldsymbol{\theta} = \sigma^2$$

$$g(\mu_j) = \beta_0 + \beta_1 X_{1j} \quad (+\sum_{k>2} \beta_k X_{kj})$$

$$\pi(\beta_1|\mathbf{Y}) \propto f(\mathbf{Y}|\beta_1)\pi(\beta_1)$$

$$= \pi(\beta_1) \int f(\mathbf{Y}|\beta_1, \boldsymbol{\beta}_{(-1)}, \boldsymbol{\theta})\pi(\boldsymbol{\beta}_{(-1)})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}d\boldsymbol{\beta}_{(-1)}$$

Often: many β's, so difficult integral

INLA

- Approximations for *marginal* posteriors based on Laplace approximations
- Fast, accurate alternative for MCMC
    - Applicable to a wide variety of models, including GLM.
- Can also be used for count data
- Implementation in R: www.r-inla.org

INLA

Laplace approximation (Rue et al., 2009, INLA) : Often reasonable to use Gaussian priors for β_k.

Then, *conditional* on $\boldsymbol{\theta}$, $\pi(\beta_1|\boldsymbol{\theta}, \mathbf{Y}) \approx \pi_G(\beta_1|\boldsymbol{\theta}, \mathbf{Y})$.
Furthermore,

$$\pi(\beta_1|\mathbf{Y}) = \int \pi(\beta_1|\mathbf{Y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\theta}$$

For any $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\theta}|\mathbf{Y}) \propto \pi(\boldsymbol{\theta}, \mathbf{Y}) = \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{Y})}{\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{Y})/\pi(\boldsymbol{\theta}, \mathbf{Y})} = \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{Y})}{\pi(\boldsymbol{\beta}|\boldsymbol{\theta}, \mathbf{Y})}$$

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{Y}) \propto \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{Y})}{\tilde{\pi}_G(\boldsymbol{\beta}|\boldsymbol{\theta}, \mathbf{Y})}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\text{mode}}}$$

Length $\boldsymbol{\theta}$ is usually small: computations reduce to low-dimensional integral plus Gaussian approximations

Why go Bayesian?

Tossing of 10.000 coins, a simple example

- Setting: 10.000 coins, each flipped only three times

- Suppose 8000 fair coins: p(head) = 0.5; 1000 for which p(head) = 0.35 and 1000 for which p(head) = 0.65

- Some computations: $X \sim Bin(n,p)$

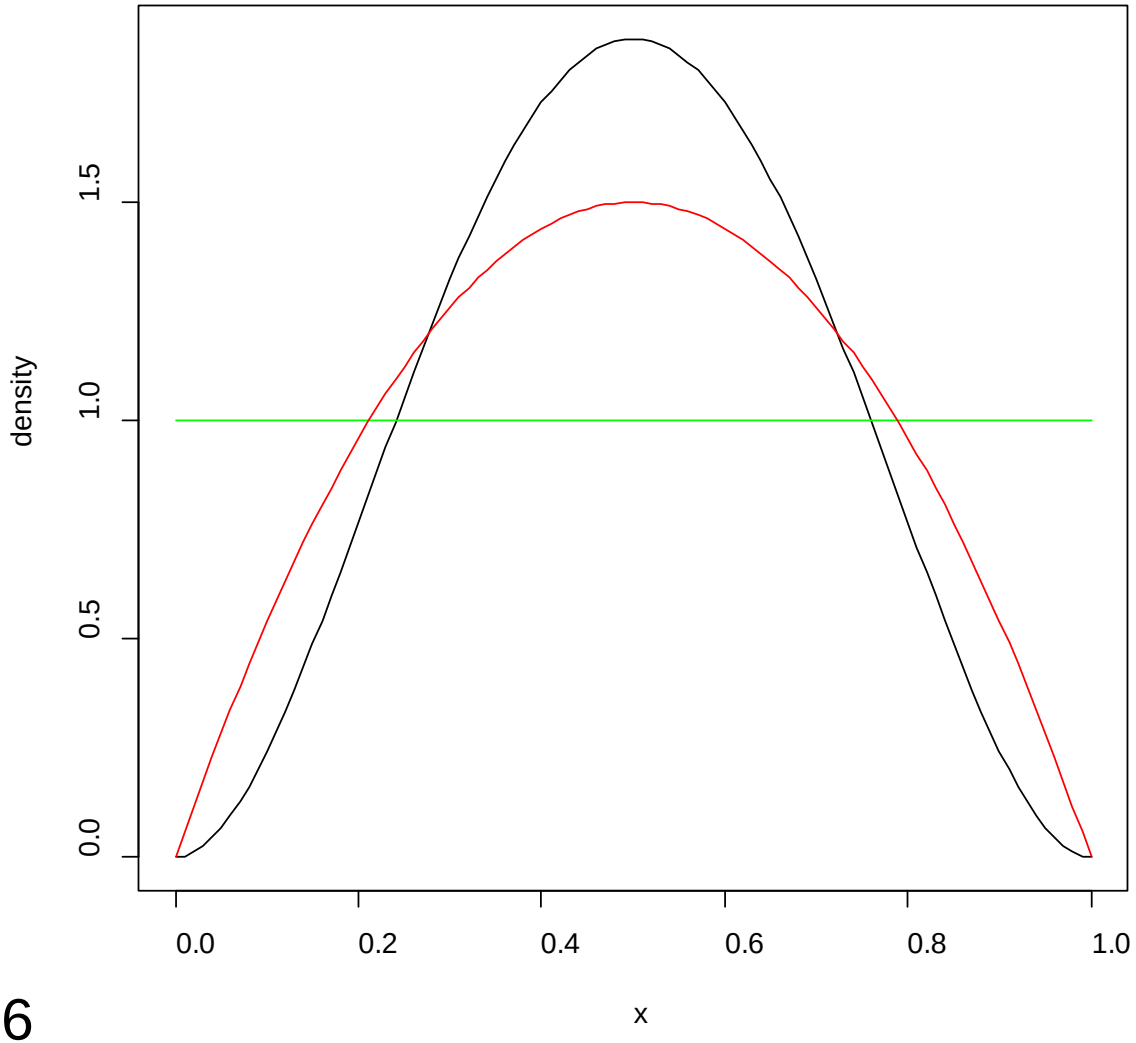
$$p \sim Beta(\alpha,\beta)$$

$$\text{Bayesian estimate} : E[p \mid k] = \frac{k+\alpha}{n+\alpha+\beta}$$

$$\text{Frequentist estimate} : \frac{k}{n}$$

Why go Bayesian?

Beta-distribution



Black: beta(3,3)
Red: beta(2,2)
Green: beta(1,1)
= U[0,1]

Why go Bayesian?

Estimates

$$\text{Bayesian estimate}: E[p \mid k] = B(k,n,\alpha,\beta) = \frac{k+\alpha}{n+\alpha+\beta}$$

$$\text{Frequentist estimate}: F(k,n) = \frac{k}{n}$$

Suppose k=0.

F(1,3) = 0. B(1,3,1,1) = 1/5; B(1,3,3,3) = 3/9

Suppose k=1.

F(1,3) = 1/3. B(1,3,1,1) = *2*/5; B(1,3,3,3) = *4*/9

In particular, the informative prior renders better estimates for the majority of coins (not for all)

Why go Bayesian?

Estimation also improves in terms of precision

Possibly more important: Bayesian estimates using an informative prior tend to be more precise (less variance)

Show this for the beta-binomial coin tossing experiment. [exer]

Why go Bayesian?

Informative prior renders better estimates

Nice, but we don’t know it. True, but we can estimate it!
Empirical Bayes: estimate the prior from data. Many different forms. Simplest one:

Equate the theoretical moments of the data to the empirical moments.

Why go Bayesian?

Empirical Bayes

Equate theoretical moments of the data to empirical ones.

$X \sim \text{Bin}(n,p), \quad p \sim \text{Beta}(\alpha,\beta)$

$$EX = E[E[X|p]] = n\alpha/(\alpha+\beta) = \overline{X} = \left(\sum_{i=1}^{m} X_i / m \right)$$

$$VX = V[E[X|p]] + E[V[X|p]] = n^2 V[p] - n(E[p^2] - E[p])$$

$$= (n^2 - n)V[p] - n(E[p])^2 + nE[p]$$

$$= (n^2 - n)\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta-1)] - n(\alpha/(\alpha+\beta))^2 + n\alpha/(\alpha+\beta)$$

$$= \hat{V} = 1/(m-1)\sum_{i=1}^{m} (X_i - \overline{X})^2$$

ShrinkBayes: motivating example

Motivating example

- 25 libraries, 70,000 tags (tag clusters)
- 5 brain regions
- 7 donors

- Main interest: which tags differ between brain regions (Groups)

Individual Brain Region	1	2	3	4	5	6	7
1		△	△	□		△	△
2	□	□	△	□	□		
3	□	□	△	△	□		
4	□		△	□		△	△
5	□	□	△	□	□		

Motivating example

Matrix Y with entries Y_{ij} : i denotes a feature, j a sample.

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	...	$j=24$	$j=25$
$i=1$	0	0	1	3	0	...	6	23
$i=2$	1	8	11	2	6	...	0	4
$i=3$	7	4	0	0	0	...	0	0
$i=4$	324	19	7	34	63	...	99	1451
$i=5$	0	0	0	5	0	...	0	1
$i=6$	.	.	.	.	.	...	.	.
.	.	.	.	.	.	...	.	.
.	.	.	.	.	.	...	.	.
.	.	.	.	.	.	...	.	.
$i=7*10^4$	65	43	176	3	12	...	5	9

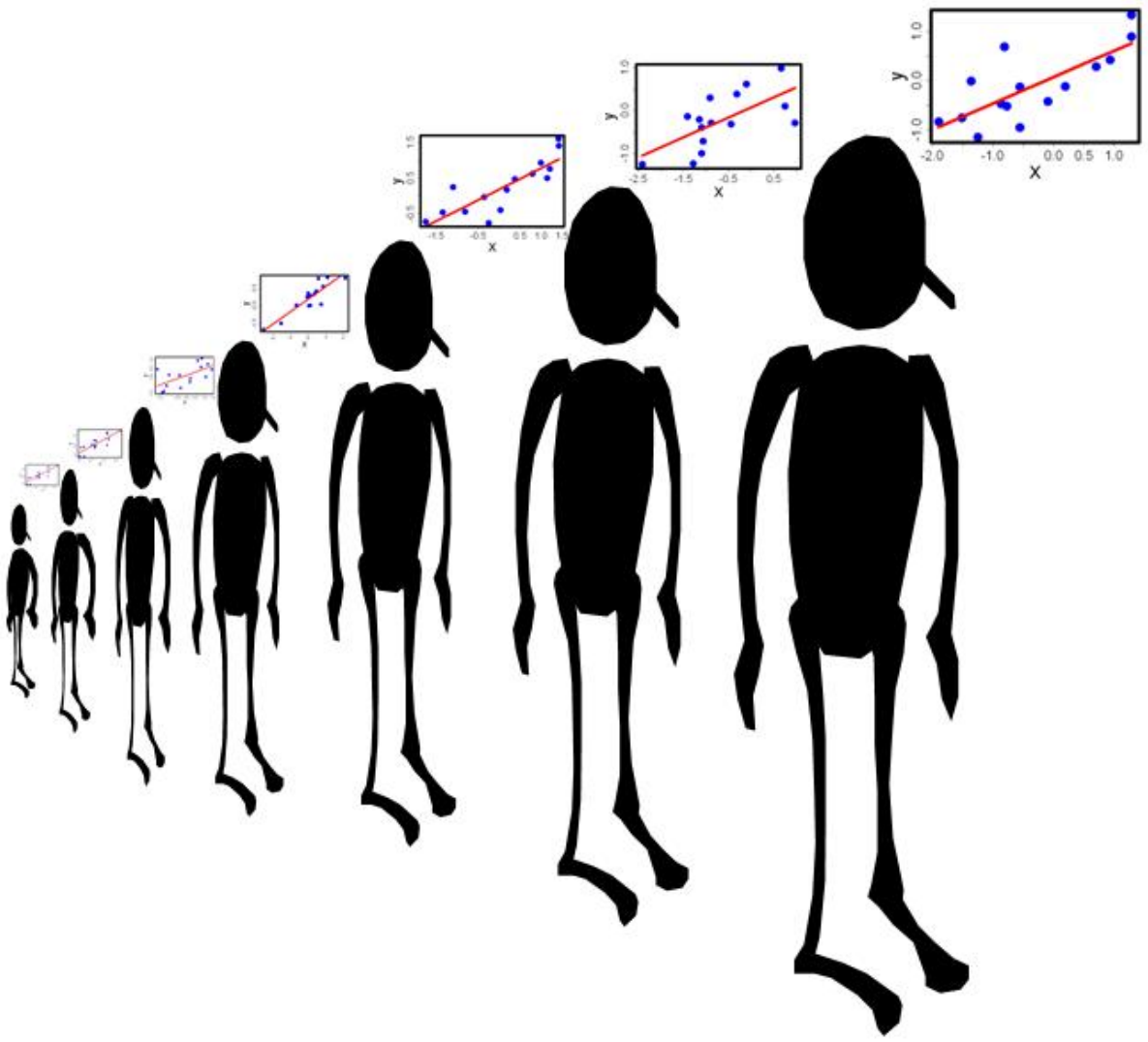
Motivating example

Matrix X with entries X_{jk} : j a sample, k a specific covariate. Dependence on feature i is allowed.

X for CAGE experiment (ANOVA-coding):

	$k = 1$: Group	$k = 2$: Ind.	$k = 3$: Batch
$j = 1$	1	2	1
$j = 2$	1	3	1
$j = 3$	1	4	2
$j = 4$	1	6	1
$j = 5$	1	7	1
$j = 6$	2	2	2
.	.	.	.
.	.	.	.
.	.	.	.
$j = 25$	5	3	1

Random effects



25 Source: http://anythingbutrbitrary.blogspot.fr/2012/06/random-regression-coefficients-using.html

ShrinkBayes: model

Model

MODEL

$Y_{ij} \sim ZI\text{-}NB(p_0, \mu_{ij}, \varphi_i) = p_{0i}\delta(0) + (1\text{-}p_{0i})NB(\mu_{ij}, \varphi_i)$

$\mu_{ij} = g^{-1}(\mathbf{X_j\beta_i})$

$\beta_{ik} \sim N(0,(\tau_k)^2)$ [or $\beta_{ik} \sim N(0,(\tau_{ik})^2)$; $(\tau_{ik})^{-2} \sim \Gamma(\alpha,\beta)$]

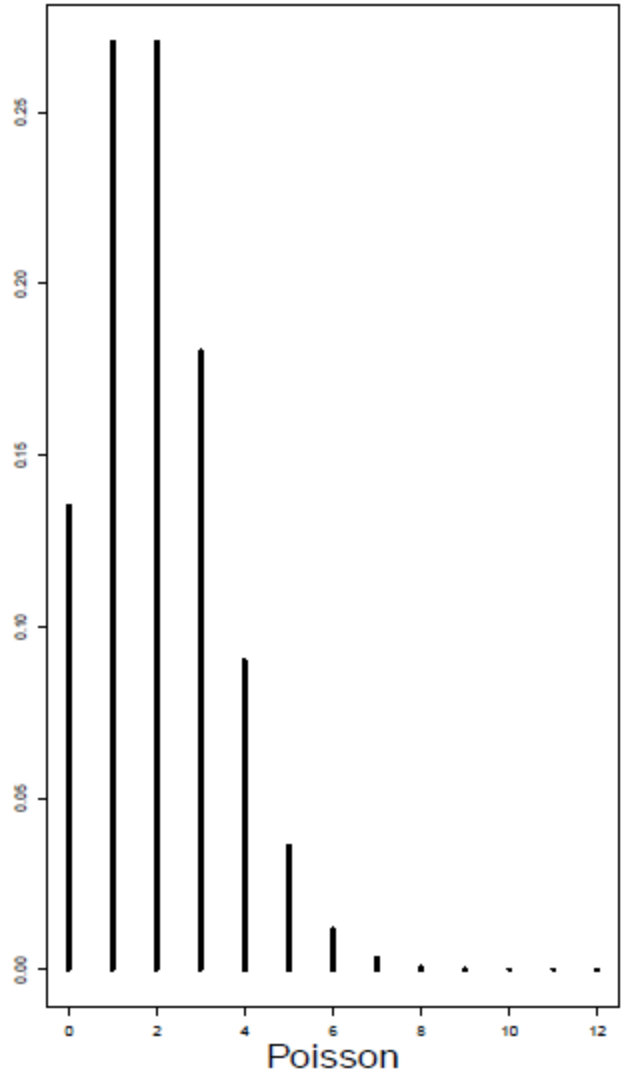
$\theta_{il} \sim F_l$

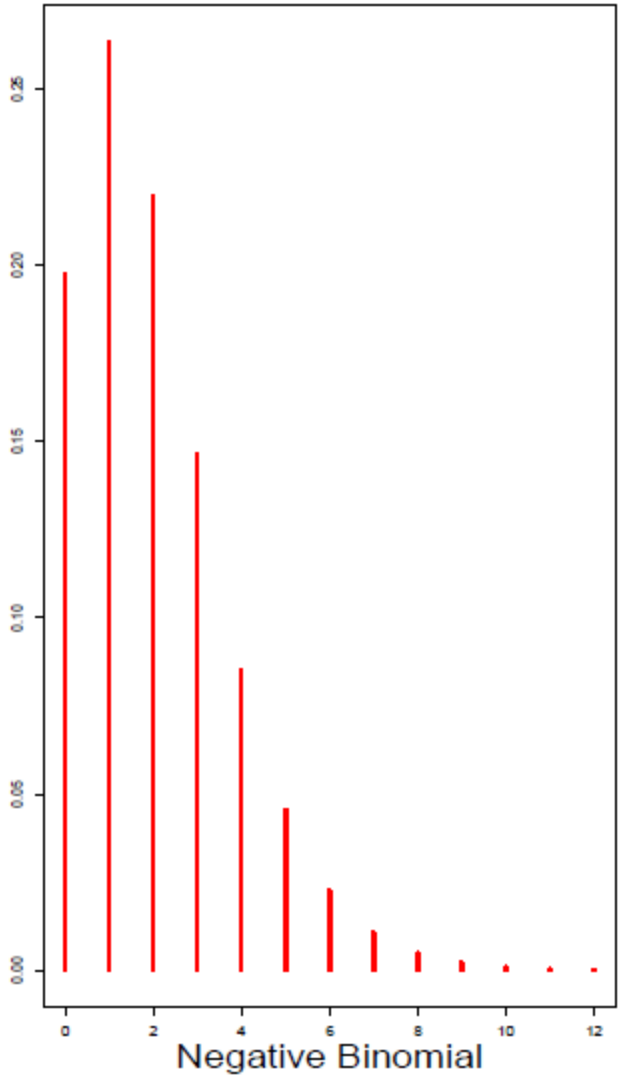
g : link function, e.g. g(x)=log(x)

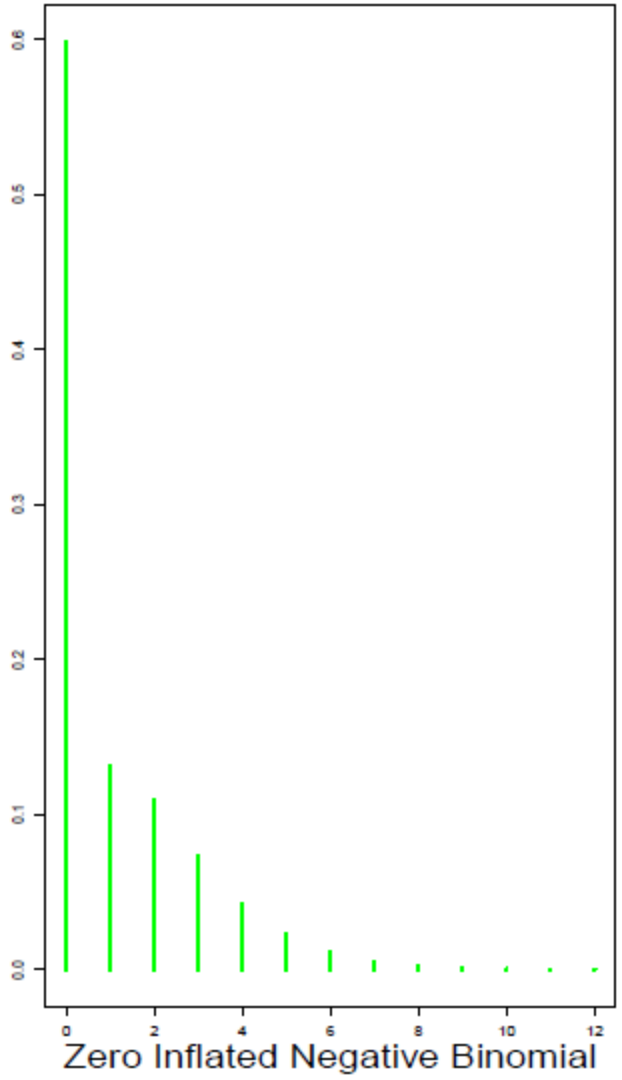
θ_{il} : hyper-parameter (not linked to mean): φ_i, p_{0i}, etc.

Model

Poisson,Negative Binomial, Zero-Inflated Negative Binomial





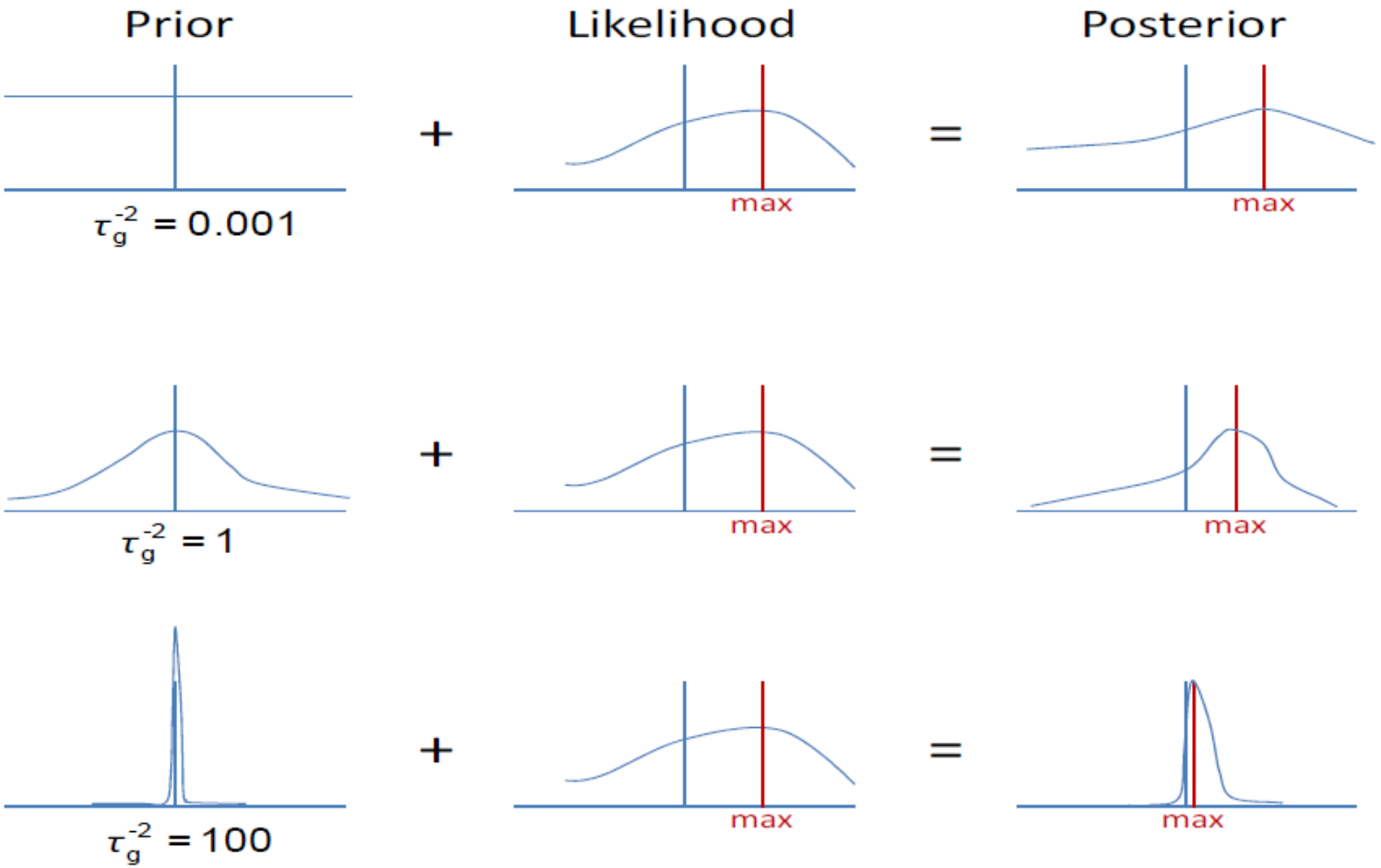


ShrinkBayes: shrinkage

ShrinkBayes

How does Bayesian shrinkage work?

Bayesian shrinkage by using informative priors, $\beta_{i,\text{group}} =^{d} N(0, \tau_g^2)$



ShrinkBayes

Why is shrinkage useful?

- Dispersion
  - enables borrowing information across features
  - leads to more stable estimates

- Parameters of interest (e.g. $\beta_{i,group}$)
  - accommodates multiplicity correction
  - corrects for 'selection bias' caused by regression to the mean

- Nuisance parameters: (e.g. $\beta_{i,batch}$)
  - automatic 'model selection' property: prior shrinks to null when unimportant

ShrinkBayes: shrinkage

Back to the example....

$Y_{ij} \sim p_{0i}\delta(0) + (1-p_{0i})NB(\mu_{ij}, \varphi_i)$, $\varphi_i = \log(v_i)$

$\mu_{ij} = \exp(\beta_{i0} + \beta_{i,group}X_g + \beta_{i,batch}X_b + \beta_{i,indj}X_{indj})$

Priors[1]

$$\log(\nu_i) =^d N(\mu, \tau^2)$$

$$\beta_{i,\text{group}} =^d N(0, \tau_g^2)$$

$$\beta_{i,\text{batch}} =^d N(0, \tau_b^2)$$

and $\tau_{i,\text{Ind}}^{-2} =^d \Gamma(\kappa_1, \kappa_2)$ for random effect $\beta_{i,\text{Ind}_j} =^d N(0, \tau_{i,\text{Ind}}^2)$

[1]Vague priors for p_{0i} and β_{i0}

ShrinkBayes: shrinkage

Priors for the example

$$\log(\nu_i) \quad =^d \quad N(\mu, \tau^2)$$

$$\beta_{i,\text{group}} \quad =^d \quad N(0, \tau_g^2)$$

$$\beta_{i,\text{batch}} \quad =^d \quad N(0, \tau_b^2)$$

and $\tau_{i,\text{Ind}}^{-2} =^d \Gamma(\kappa_1, \kappa_2)$ for random effect $\beta_{i,\text{Ind}_j} =^d N(0, \tau_{i,\text{Ind}}^2)$

How do we estimate the parameters of the priors?

$$A = \{(\kappa_1, \kappa_2), \tau_g, \tau_b, (\mu, \tau)\}$$

ShrinkBayes

Bayesian Empirical Bayes

Let $\pi_{\boldsymbol{\alpha}}(\theta)$ denote the common prior of parameters $\theta_i, i = 1, \ldots, p$.

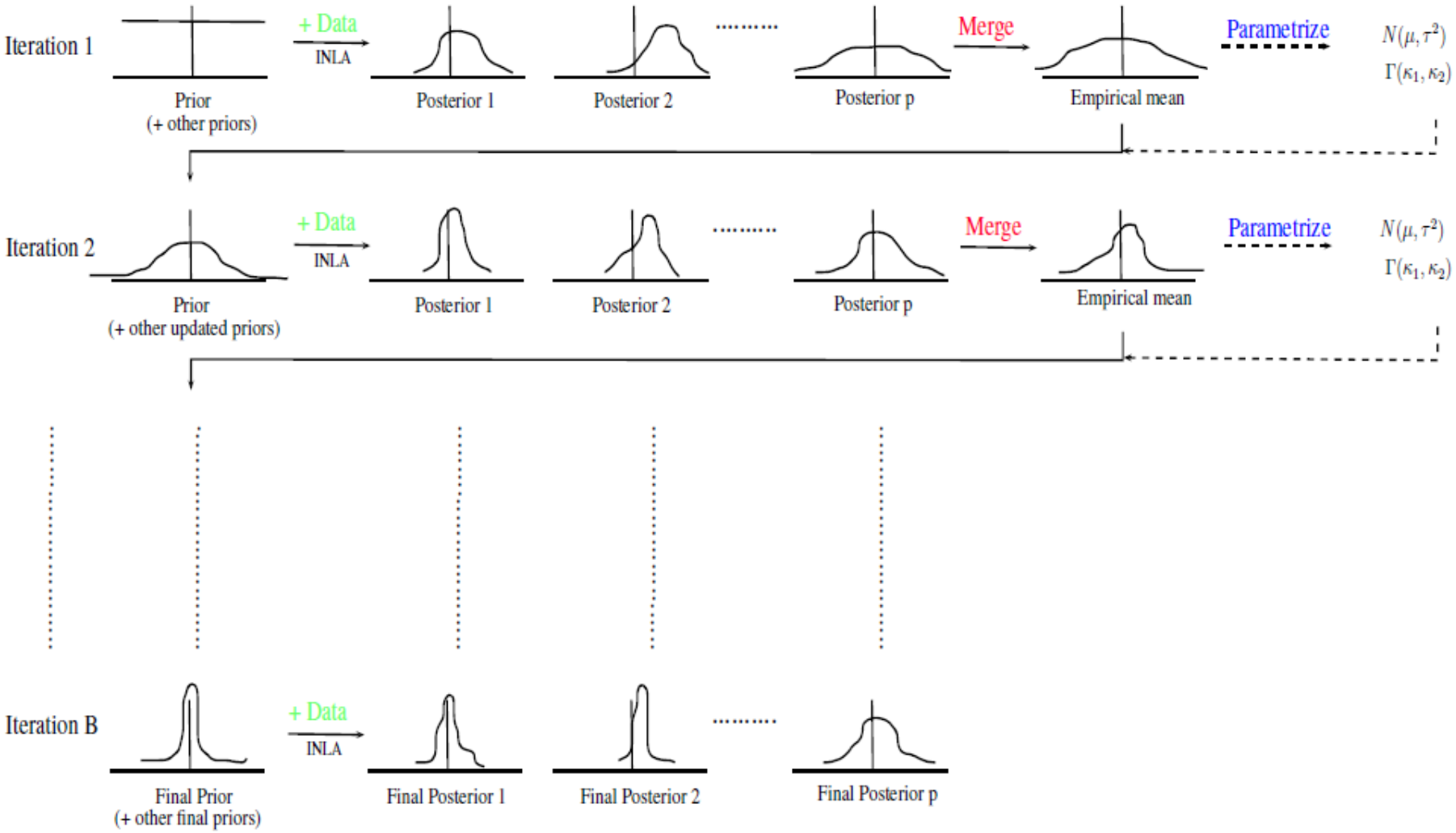
Empirical Bayes: $\mathbf{Y}_i, i = 1, \ldots, p$, are samples from density $f_A(y)$. Then,

$$\begin{aligned}\pi_{\boldsymbol{\alpha}}(\theta) &= \int \pi_A(\theta|y) f_A(y) d\mu(y) \approx \pi_A^{\text{Emp}}(\theta) \\ &= \frac{1}{p}\sum_{i=1}^{p} \pi_A(\theta|\mathbf{Y}_i) = \frac{1}{p}\sum_{i=1}^{p} \pi_{\{\boldsymbol{\alpha}\}\cup A-}(\theta|\mathbf{Y}_i).\end{aligned}$$

Idea: I) *Initialize* $\boldsymbol{\alpha}$ (and hence $\pi_{\boldsymbol{\alpha}}(\theta)$); II) *Estimate* posteriors; III) *Update* $\pi_{\boldsymbol{\alpha}}(\theta)$: fitting parametric shape to sample of $\pi_A^{\text{Emp}}(\theta)$; IV) *Iterate*.

ShrinkBayes

Iterative estimation of the prior parameters

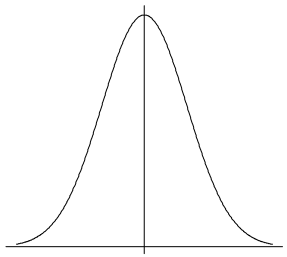
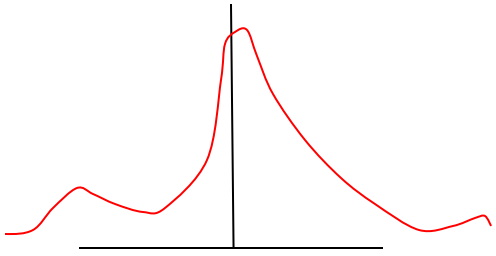


ShrinkBayes

Nonparametric prior

- Shape of prior may be important, in particular for central parameter of interest (e.g. $\theta_i = \beta_{i,group}$)

- How to get

from  to 

$\theta_i \sim N(0,\tau^2)$ $\theta_i \sim F_{np}$

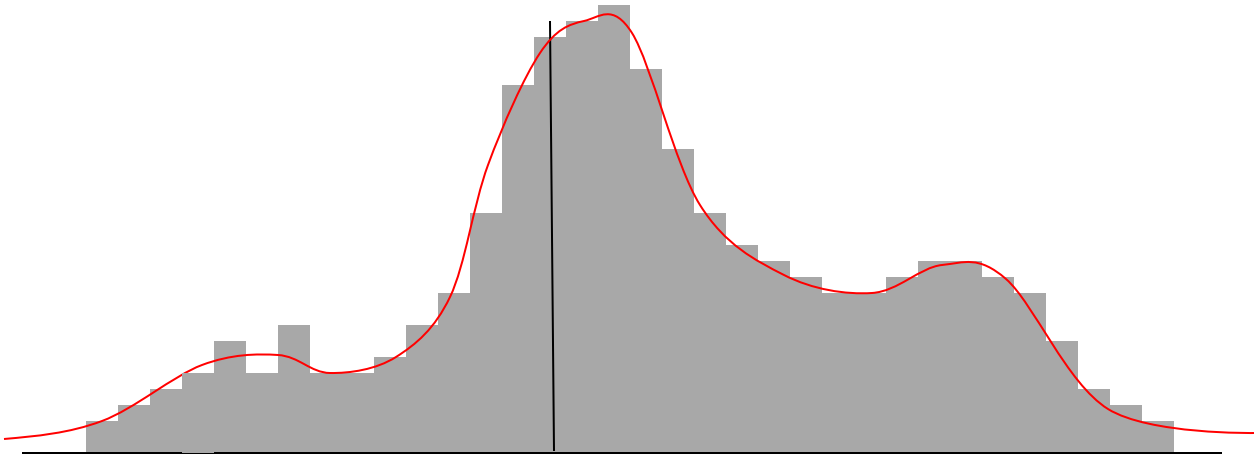
?

ShrinkBayes

Nonparametric prior

- Empirical Bayes fitting is easy, remember: $\pi(\theta) \approx \sum_{i=1}^{p} \pi(\theta \mid \mathbf{Y}_i)$

- So, we simply sample from this mixture (given an initial prior) and smoothen it (e.g. using ‘density’ in R):



ShrinkBayes

Nonparametric prior

- Nice, but INLA does not accept non–parametric priors!

- Solution:
  1. Use iterative procedure to find a good parametric shrinkage prior: $f_p(\theta)$
  2. Obtain posteriors under $f_p(\theta)$
  3. Apply the same iterative procedure to obtain a non-parametric estimate f_{np}, computing posteriors by:

$$\pi_{np}(\theta \mid \mathbf{Y}_i) = C\pi_p(\theta \mid \mathbf{Y}_i)\frac{f_{np}(\theta)}{f_p(\theta)} \quad \text{[exer]}$$

$$\int \pi_{np}(\theta \mid \mathbf{Y}_i)d\theta = 1 \square \qquad C = \left(1 \Big/ \int \pi_p(\theta \mid \mathbf{Y}_i)\frac{f_{np}(\theta)}{F_f(\theta)}d\theta \right)$$

ShrinkBayes: inference

ShrinkBayes: inference

Hypothesis testing in a Bayesian setting

- In a Bayesian setting it is very natural to test interval H_0: $|\theta_i| \leq \delta$

- Using $\delta \neq 0$ avoids detecting small, biologically non-relevant effects

- Testing H_0: $|\theta_i| \leq \delta$ trivial from posteriors:

$$\text{lfdr} = P(H_0 \mid \mathbf{Y}_i) = P(|\theta_i| \leq \delta \mid \mathbf{Y}_i) = \int_{-\delta}^{\delta} P(\theta_i \mid \mathbf{Y}_i) d\theta_i$$

...and BFDR(t) =E[lfdr|lfdr ≤ t] computed from lfdr as before[1]

[1]More on False Discovery Rate (FDR) in later lectures

ShrinkBayes: inference

Bayesian hypothesis testing: multiple comparisons

- Multiple comparisons: $\theta_{ikl} = \theta_{ik} - \theta_{il}$ for all (k,l). E.g. ($\beta_{i,group1} - \beta_{i,group2}$, $\beta_{i,group1} - \beta_{i,group3}$, $\beta_{i,group2} - \beta_{i,group3}$, etc.)

- H_0: $|\theta_{ikl}| \leq \delta$ for all (k,l).

$$
\begin{aligned}
\text{lfdr} &= P(H_0 \mid \mathbf{Y}_i) = P(\max_{(k,l)} |\theta_{ikl}| \leq \delta \mid \mathbf{Y}_i) = 1 - P(\max_{(k,l)} |\theta_{ikl}| > \delta \mid \mathbf{Y}_i) \\
&\leq 1 - \max_{(k,l)} P(|\theta_{ikl}| > \delta \mid \mathbf{Y}_i) = \min_{(k,l)} [1 - P|\theta_{ikl}| > \delta \mid \mathbf{Y}_i)] \\
&= \min_{(k,l)} P(|\theta_{ikl}| \leq \delta \mid \mathbf{Y}_i) = \min_{(k,l)} \text{lfdr}_{kl}
\end{aligned}
$$

- So: if minimum lfdr < α then certainly global lfdr < α

- Also works for BFDR

ShrinkBayes: inference

Point null-hypothesis in a Bayesian setting

- Recall that for testing H_0: $\theta_i = 0$ (e.g. $\theta_i = \beta_{i,group}$) prior needs to contain mass on 0!

E.g: $$\pi(\theta) = p_0\delta(0) + (1 - p_0)N(0, \tau^2)$$

Then[1]:

$$\pi_0(\mathbf{Y}) = \frac{p_0\pi(\mathbf{Y}\mid 0)}{p_0\pi(\mathbf{Y}\mid 0) + (1-p_0)\Pi(\mathbf{Y})} = \frac{p_0\pi(\mathbf{Y}\mid 0)}{p_0\pi(\mathbf{Y}\mid 0) + (1-p_0)\int \pi'(\theta)\pi(\mathbf{Y}\mid\theta)d\theta}$$

[1]dropping index i

ShrinkBayes: inference

Point null-hypothesis in a Bayesian setting

$$\pi_0(\mathbf{Y}) = \frac{p_0\pi(\mathbf{Y}\mid 0)}{p_0\pi(\mathbf{Y}\mid 0) + (1-p_0)\Pi(\mathbf{Y})} = \frac{p_0\pi(\mathbf{Y}\mid 0)}{p_0\pi(\mathbf{Y}\mid 0) + (1-p_0)\int \pi'(\theta)\pi(\mathbf{Y}\mid\theta)d\theta}$$

$\pi(\mathbf{Y}|0)$ and $\Pi(\mathbf{Y})$ are *marginal likelihoods* which are conveniently provided by INLA (where $\pi(\mathbf{Y}|0)$ is obtained after fitting the null-model: i.e. the model not containing β_{ik}

$$\mathrm{BFDR}(t) = E\left[\pi_0(\mathbf{Y}) \mid \pi_0(\mathbf{Y}) \le t\right] \hat{=} \frac{\sum_{i=1}^{p} \pi_0(\mathbf{Y}) I_{\{\pi_0(\mathbf{Y}) \le t\}}}{\sum_{i=1}^{p} I_{\{\pi_0(\mathbf{Y}) \le t\}}}$$

ShrinkBayes: inference

Posterior

Posterior is [exer]:

$$P(\theta = 0 \mid \mathbf{Y}) = \frac{p_0 \pi(\mathbf{Y} \mid 0)}{p_0 \pi(\mathbf{Y} \mid 0) + (1 - p_0)\pi(\mathbf{Y})}$$

$$f(\theta \mid \mathbf{Y}) = \frac{(1 - p_0) f'(\theta \mid \mathbf{Y})}{p_0 \pi(\mathbf{Y} \mid 0) / \pi(\mathbf{Y}) + (1 - p_0)}, \text{ for } \theta \neq 0$$

Where f'(θ|**Y**) is the posterior obtained by INLA under the Guassian prior $N(0,\tau^2)$

ShrinkBayes: Software

ShrinkBayes

Demo

See the R-Vignette ShrinkBayesVignette.pdf for many examples

ShrinkBayes: Plusses and minuses

ShrinkBayes

Plusses and minuses, comparison with edgeR

+ Better (in terms of power) for small sample sizes

+ Can deal with pairs, random effects, excess of zeros

+ Automatic model selection

+ More reproducible results

- More difficult to use

- More time-consuming

- Bayesian concept, less familiar to most biologists