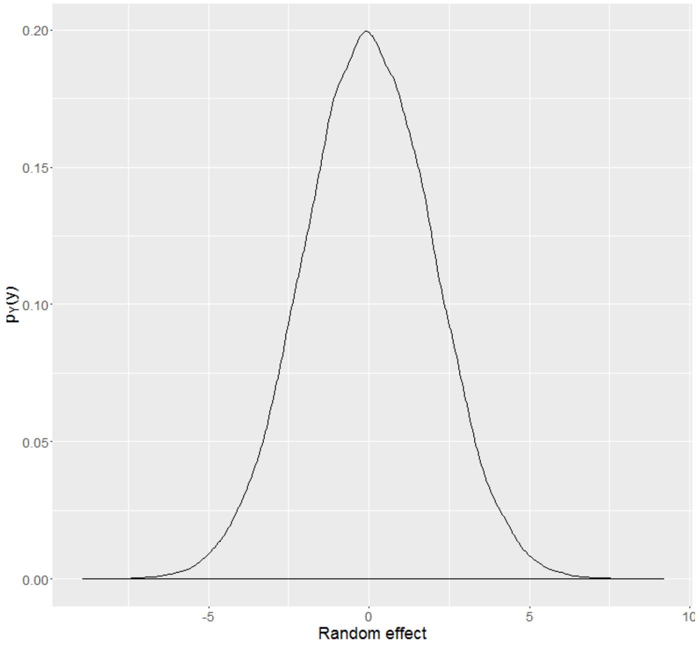
The use of mixture distributions in a Bayesian linear mixed effects model

Anirudh TOMER

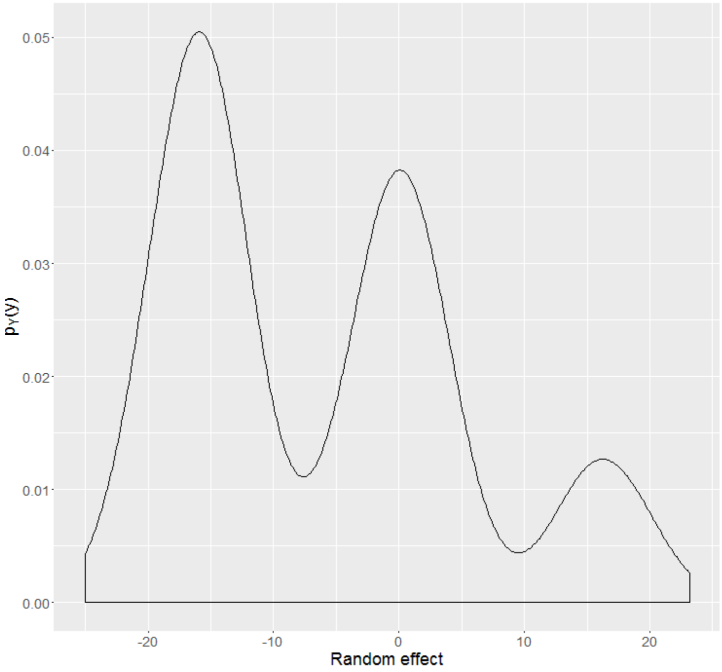
Promoter: Professor Emmanuel LESAFFRE

KU LEUVEN

Bayesian heterogeneity model



Vs.



What is the problem we are facing?

Estimation of mixture component density parameters θ_G

Criteria for choice of number of components (DIC, Marginal likelihood, Predictive methods)

Classification of observations into groups

DIC (deviance information criteria)

$DIC = \overline{D(\boldsymbol{\theta})} + p_D$

Where $p_D = \overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}})$ is the effective number of parameters in the model.

$D(\theta) = -2 \log p(\boldsymbol{y}|\boldsymbol{\theta}) + 2 \log f(\boldsymbol{y})$ is called the Bayesian deviance.

Various definitions of DIC proposed by Celeux et al., (2006) for missing data models.

Marginal/Complete/Conditional likelihood

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} f_N(\boldsymbol{y}_i; \boldsymbol{X}_i\beta + \boldsymbol{Z}_i \boldsymbol{b}_k^C, \boldsymbol{Z}_i G_k \boldsymbol{Z}_i^T + R_i)\eta_k$$

$$\begin{aligned} f(\boldsymbol{y}^F|\boldsymbol{\theta}) &= \prod_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{b}_i, S_i) f(\boldsymbol{b}_i|S_i) f(S_i) \\ &= \prod_{i=1}^{n} f_N(\boldsymbol{y}_i; \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i, R_i) f_N(\boldsymbol{b}_i; \boldsymbol{b}_{S_i}^C, G_{S_i}) f(S_i; \boldsymbol{\eta}) \end{aligned}$$

$$p(\boldsymbol{y}|\boldsymbol{\theta}^{\text{cond}}) = \prod_{i=1}^{n} f_N(\boldsymbol{y}_i; \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i, R_i)$$

$$\boldsymbol{\theta}^{\text{cond}} = (\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{S}).$$

X_i, Z_i

β, b_k^C

G_k, R_i

η_k, S_i, b_i

?

Marginal DIC definitions

$$\text{DIC}_1 = -4\text{E}_{\theta|y}(\log p(\boldsymbol{y}|\boldsymbol{\theta})) + 2\log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}}))$$

$$\text{DIC}_2 = -4\text{E}_{\theta|y}(\log p(\boldsymbol{y}|\boldsymbol{\theta})) + 2\log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_M))$$

$$\hat{\boldsymbol{\theta}}_M = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{y})$$

$$\text{DIC}_3 = -4\text{E}_{\theta|y}(\log p(\boldsymbol{y}|\boldsymbol{\theta})) + 2\log \hat{p}(\boldsymbol{y})$$

$$\hat{p}(\boldsymbol{y}) = \prod_{i=1}^{n} \hat{p}(\boldsymbol{y}_i) = \prod_{i=1}^{n} \frac{1}{m} \sum_{l=1}^{m} \sum_{k=1}^{K} f_N(\boldsymbol{y}_i; \boldsymbol{X}_i\beta^{(l)} + \boldsymbol{Z}_i\boldsymbol{b}_k^{C^{(l)}}, Z_i G_k^{(l)} \boldsymbol{Z}_i^T + R_i^{(l)})\eta_k^{(l)}$$

Complete and conditional DIC definitions

$$\mathsf{DIC}_4 = -4\mathsf{E}_{\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{S}|\boldsymbol{y}}(\log p(\boldsymbol{y}^F|\boldsymbol{\theta})) + 2\mathsf{E}_{\boldsymbol{b},\boldsymbol{S}|\boldsymbol{y}}(\log p(\boldsymbol{y}^F|\bar{\boldsymbol{\theta}}))$$

$$\mathsf{DIC}_5 = -4\mathsf{E}_{\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{S}|\boldsymbol{y}}(\log p(\boldsymbol{y}^F|\boldsymbol{\theta})) + 2\log p(\boldsymbol{y},\hat{\boldsymbol{b}}_M,\hat{\boldsymbol{S}}_M|\hat{\boldsymbol{\theta}}_M)$$

$$\mathsf{DIC}_6 = -4\mathsf{E}_{\boldsymbol{\theta}^{\mathsf{cond}}}(\log p(\boldsymbol{y}|\boldsymbol{\theta}^{\mathsf{cond}})) + 2\log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_M^{\mathsf{cond}}))$$

$$\boldsymbol{\theta}^{\mathsf{cond}} = (\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{S}).$$

$$\hat{\boldsymbol{\theta}}_M^{\mathsf{cond}} = \arg\max_{\boldsymbol{\theta}^{\mathsf{cond}}} p(\boldsymbol{\theta}^{\mathsf{cond}}|\boldsymbol{y}), \text{ and } \mathsf{E}_{\boldsymbol{\theta}^{\mathsf{cond}}}(\log p(\boldsymbol{y}|\boldsymbol{\theta}^{\mathsf{cond}}))$$

Marginal likelihood using Chib’s approximation

$$m(\boldsymbol{y}) = p(\boldsymbol{y}|M) = \frac{L(\boldsymbol{\theta}|\boldsymbol{y}, M)p(\boldsymbol{\theta}|M)}{p(\boldsymbol{\theta}|\boldsymbol{y}, M)}$$

$$\log \hat{m}(\boldsymbol{y}) = \log L(\boldsymbol{\theta}^*|\boldsymbol{y}) + \log p(\boldsymbol{\theta}^*) - \log p(\boldsymbol{\theta}^*|\boldsymbol{y})$$

$$\begin{aligned}\log p(\boldsymbol{\theta}^*|\boldsymbol{y}) = \sum_{k=1}^{K} \log p(G_k^*|\boldsymbol{y}) + \sum_{k=1}^{K} \log p(\boldsymbol{b}_k^{C^*}|G_k^*, \boldsymbol{y}) + \log p(\sigma^{2^*}|G_k^*, \boldsymbol{b}_k^{C^*}, \boldsymbol{y}) \\ + \log p(\beta^*|G_k^*, \boldsymbol{b}_k^{C^*}, \sigma^{2^*}, \boldsymbol{y}) + \log p(\boldsymbol{\eta}^*|G_k^*, \boldsymbol{b}_k^{C^*}, \sigma^{2^*}, \beta^*, \boldsymbol{y})\end{aligned}$$

$$\begin{aligned}\prod_{k=1}^{K} p(G_k^*|\boldsymbol{y}) &= \int \prod_{k=1}^{K} p(G_k^*|\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{S}, \boldsymbol{b}_k^C) p(\boldsymbol{b}_1^C, \boldsymbol{b}_2^C, ..., \boldsymbol{b}_K^C, \boldsymbol{b}, \boldsymbol{S}|\boldsymbol{y}) \, \mathrm{d}\boldsymbol{b}_1^C \, \mathrm{d}\boldsymbol{b}_2^C \ldots \mathrm{d}\boldsymbol{b}_K^C \, \mathrm{d}\boldsymbol{b} \, \mathrm{d}\boldsymbol{S} \\ &\approx \frac{1}{m} \sum_{l=1}^{m} \prod_{k=1}^{K} p(G_k^*|\boldsymbol{y}, \boldsymbol{b}^{(l)}, \boldsymbol{S}^{(l)}, \boldsymbol{b}_k^{C(l)}) \\ &\approx \frac{1}{m} \sum_{l=1}^{m} \prod_{k=1}^{K} f_{\mathcal{W}^{-1}}(G_k^*; n_k^{(l)} + n_0, \Psi + \sum_{i=1}^{n_k^{(l)}} (\boldsymbol{b}_i^{(l)} - \boldsymbol{b}_k^{C(l)})(\boldsymbol{b}_i^{(l)} - \boldsymbol{b}_k^{C(l)})^T)\end{aligned}$$

KU LEUVEN

Posterior predictive checks

$$T(\boldsymbol{r}) = \frac{1}{\sum_{i=1}^{n} m_i} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (r_{ij} - \bar{r}_{i.})^2$$

$$r_{ij} = \boldsymbol{z}_{ij}\tilde{\boldsymbol{b}}_i + \varepsilon_{ij} = y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta},$$

$$\bar{r}_{i.} = \frac{1}{m_i} \sum_{j=1}^{m_i} r_{ij}.$$

Motivation: Testing for overfitting.
Idea: Sample big values from empty components to obtain inflated test statistic.

KU LEUVEN

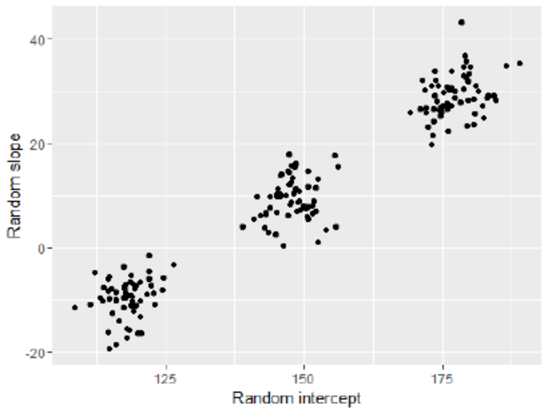
Data sets

Zebu cow weights

Predictors were gender, birth year, age, time of measurement.

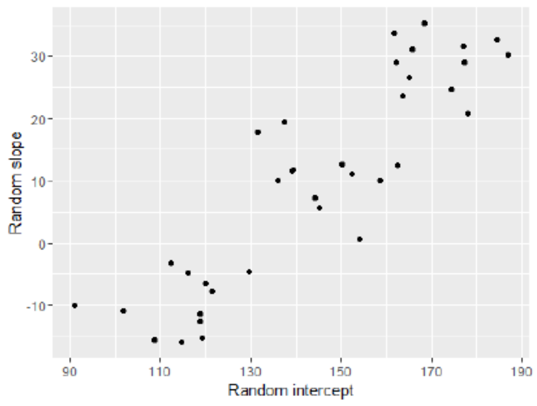
Various versions of the dataset which differed in number of mixture components for random effects, number of subjects, separation of mixture components and number of subjects per component.

DIC results



(a) Data set 2

# Components Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
	✓	✓	✓	✓	✓	✗
1	9966	9959	9965	12921	10531	7855
2	9865	9849	9864	12498	10458	7860
3	9664	9665	9663	11847	10244	7870



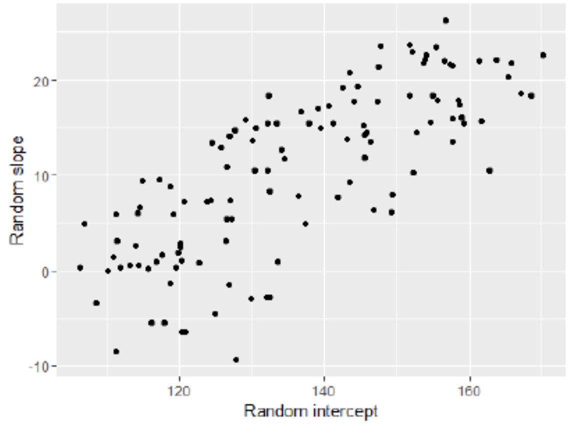
(b) Data set 3

# Components Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
	✓	✓	✓	✓	✗	✗
1	2013	2012	2012	2611	2118	1570
2	1989	1949	1987	2497	2013	1562
3	1942	1942	1940	2339	2039	1571

DIC4 is most discerning.

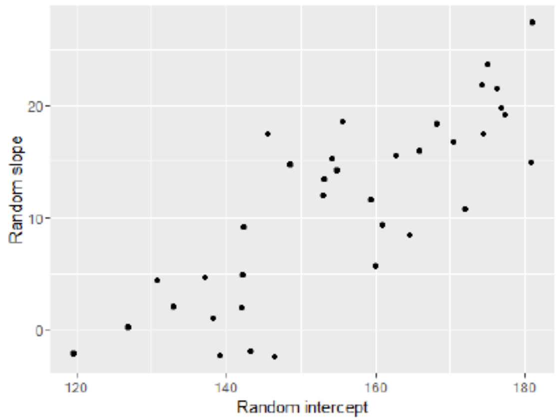
KU LEUVEN

DIC results



(a) Data set 4

# Components Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
1	6568	6566	6566	8454	6899	5197
2	6531	6523	6530	8263	6946	5253
3	6497	6492	6497	8017	6898	5263



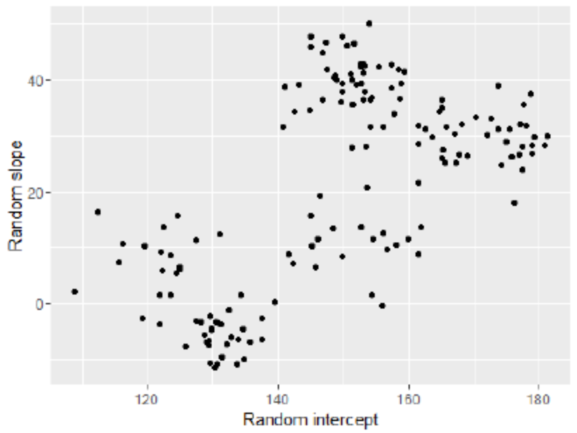
(b) Data set 5

# Components Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
1	1944	1943	1943	2500	1879	1364
2	1934	1936	1936	2438	2042	1526
3	1921	1920	1922	2350	2023	1537

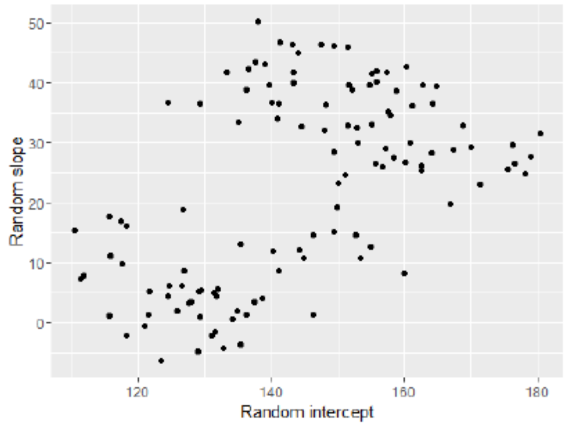
Less sample size, fused components→ DIC is not much discerning for anything other than DIC4

KU LEUVEN

DIC results



(a) Data set 6



(b) Data set 7

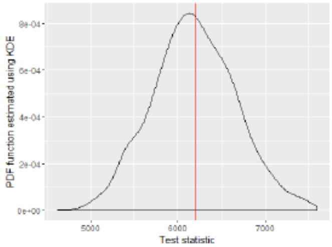
# Components Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
1	8982	8981	8980	11847	9251	6655
2	8829	8827	8827	11327	9293	6838
3	8745	8742	8744	11036	9251	6895
4	8669	8672	8677	10737	9208	6925
5	8649	8643	8648	10601	9165	6909

# Components Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
1	6708	6707	6706	8819	6977	5071
2	6606	6605	6604	8443	6946	5135
3	6539	6538	6537	8178	6944	5204
4	6506	6514	6521	8078	6915	5196
5	6505	6500	6508	7984	6896	5202

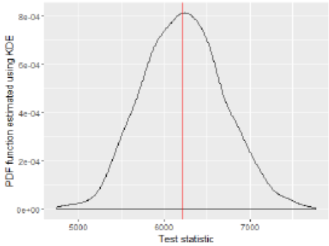
Marginal likelihood results ($\log \hat{m}(y)$)

Fitted	1 Comp	2 Comp	3 Comp	4 Comp	5 Comp
Data set 1	-2120				
Data set 2	-5019	-4989	-4937		
Data set 3	-1038	-1044	-1042		
Data set 4	-3317	-3318	-3322		
Data set 5	-1001	-986	-993		
Data set 6	-4545	-4492	-4477	-4467	-4473
Data set 7	-3397	-3379	-3373	-3380	-2749

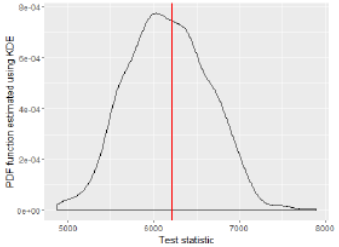
PPC results (data set 6)
Using Wishart prior for covariance matrix



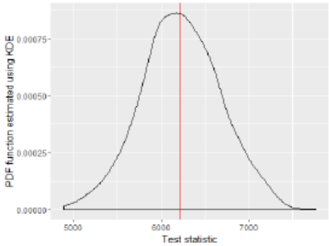
(a) # Components fitted = 5



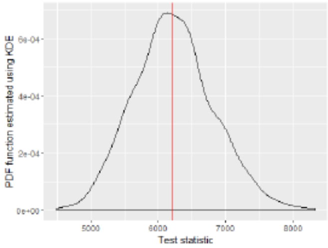
(b) # Components fitted = 4



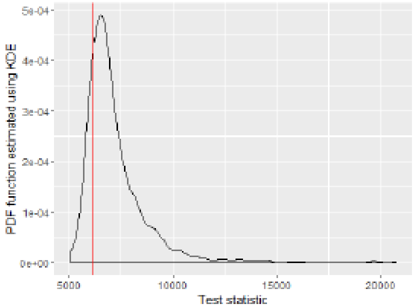
(c) # Components fitted = 3



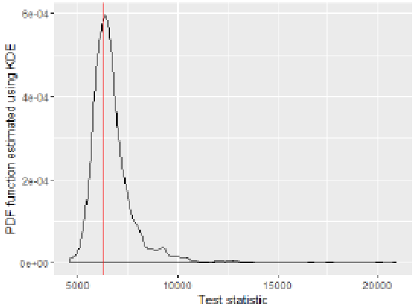
(d) # Components fitted = 2



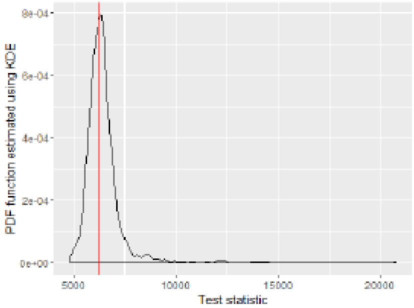
(e) # Components fitted = 1



(a) # Components fitted = 8



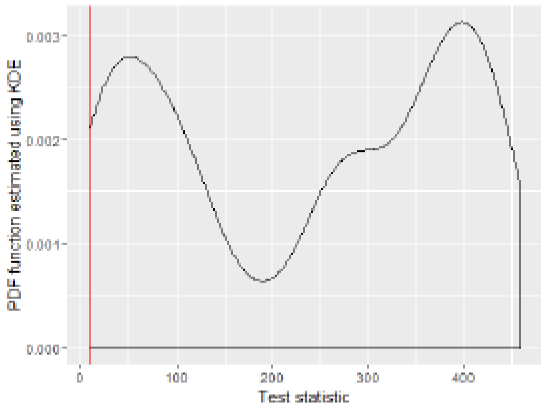
(b) # Components fitted = 7



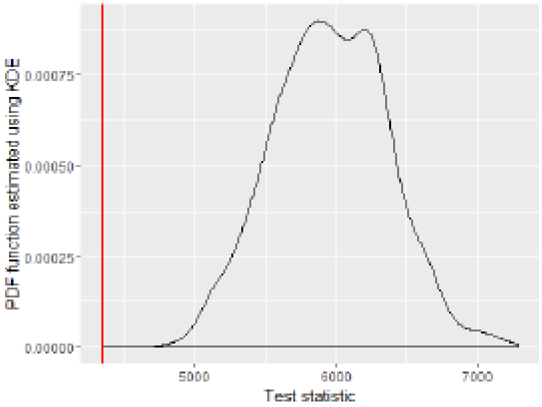
(c) # Components fitted = 6

KU LEUVEN

PPC results
Using U(-1,1) prior on ρ Gamma $(10^{-4}, 10^{-4})$ on precision



(a) 4 components fitted for data set 2. log scale is used.

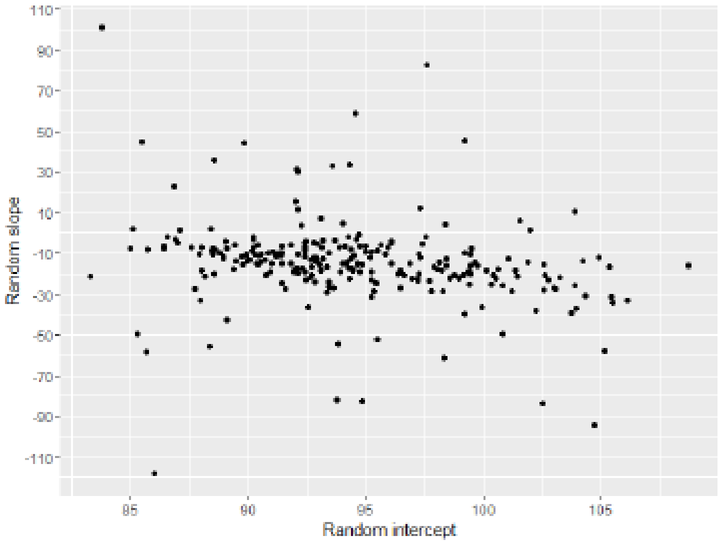


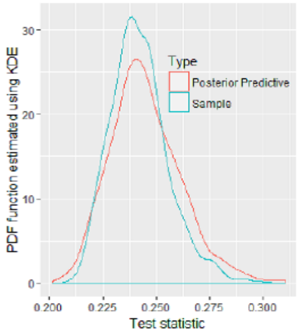
(b) 3 components fitted for data set 2

a) Overfitting has a big penalty
b) posterior variances of random effects were underestimating the sample data's variance covariance of the random effects. Bad model fit is detected with the test statistic

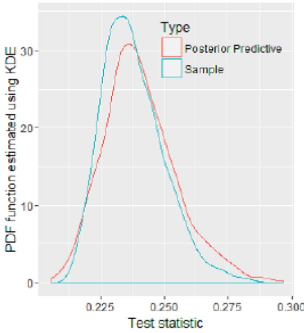
Blood donor data set analysis

# Components Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
1 comp	4817	4816	4818	7077	7532	4353
2 comp	4808	4805	4811	6956	7376	4306

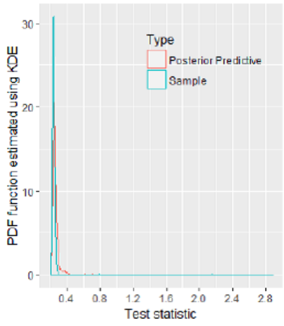




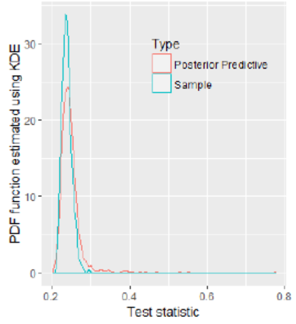
(a) #Components fitted = 1



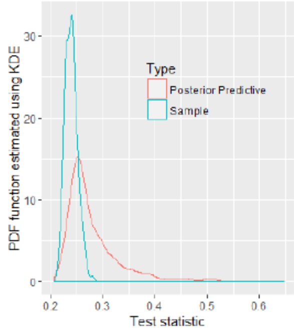
(b) #Components fitted = 2



(a) #Components fitted = 3



(b) #Components fitted = 4



(c) #Components fitted = 5

KU LEUVEN

Blood donor data set analysis

Component 1

- Baseline random component of Hb level is low
- Random slope is less negative relative to group 2.
- Correlation between random slope and random intercept is most likely 0.

Component 2

- Baseline random component of Hb level is higher.
- Average subject from group 2 has a more rapid decrease in Hb level levels with frequent donations.
- Higher the baseline Hb level level, more likely to have a rapid decrease in Hb level levels with donations

Other interesting results

Dirichlet prior with large parameter values. Incrementally increasing hyperparameter?

Chib’s approximation and chain length. Chib’s approximation when posterior’s do not belong to well known parametric families.

Wishart prior: underestimated posterior component variances if actual within subject variability > between subject variability.