Bayesian growth mixture models to distinguish hemoglobin value trajectories in blood donors

Kazem Nasserinejad[1] Joost van Rosmalen[1] Mireille Baart[2]
Katja van den Hurk[2] Dimitris Rizopoulos [1] Emmanuel Lesaffre [1,3]
Wim de Kort [2,4]

[1]Department of Biostatistics, Erasmus University Medical Center, Rotterdam, the Netherlands

[2]Sanquin Research, Department of Donor Studies, Amsterdam, the Netherlands

[3]L-Biostat, KU Leuven, Leuven, Belgium

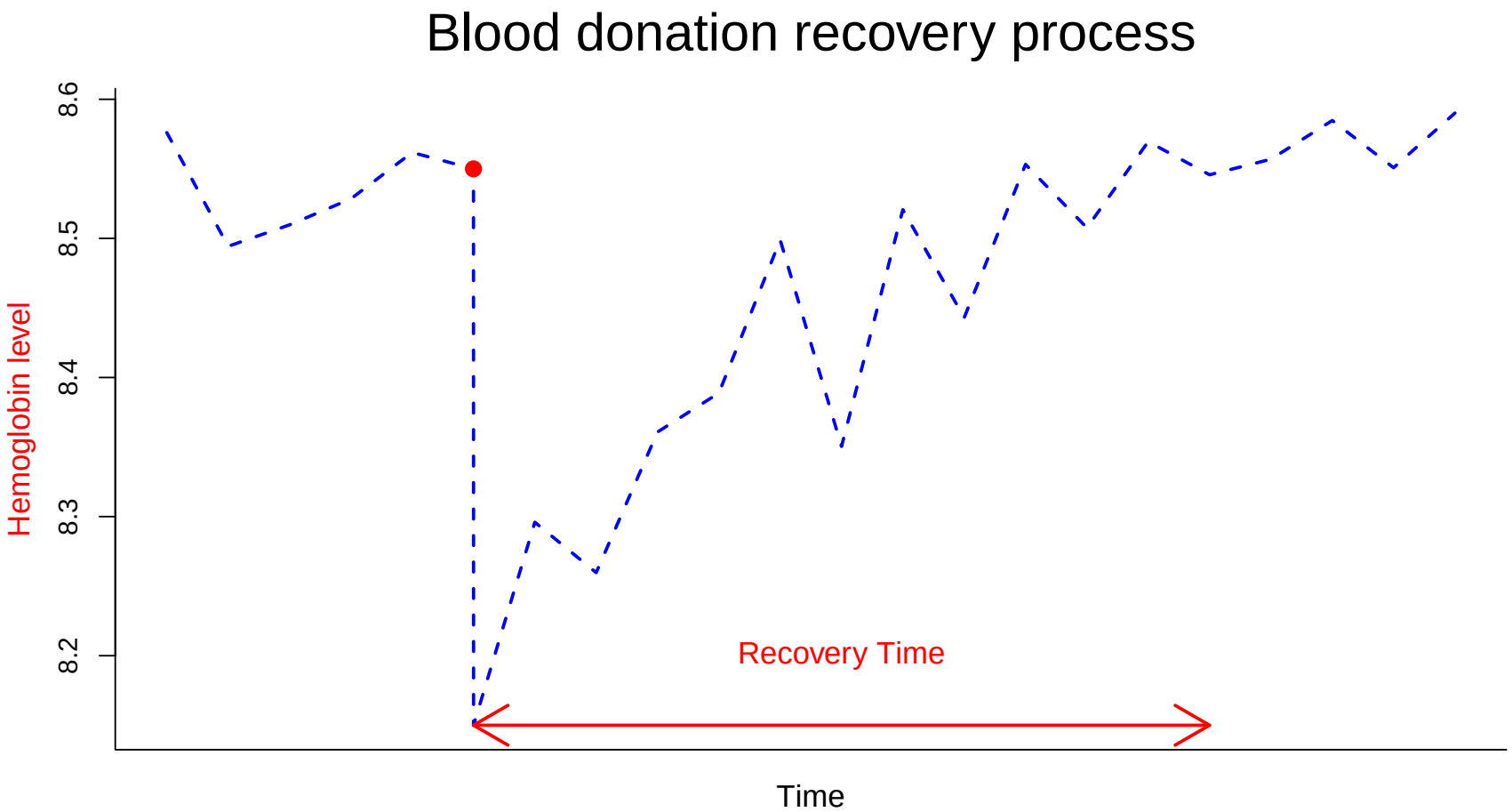
[4]Department of Public Health, Academic Medical Center, Amsterdam, the Netherlands

Bayes2014, June, 2014

Introduction

- Blood donation leads to a temporary reduction in the Hb level. This needs a period after donation for the Hb value to recover to its pre-donation level.
- Individual donors may differ in their recovery and minimum interval (56 days) may not be safe for each individual.
- We aim to classify the longitudinal Hb values measured in blood donors.

Blood donation recovery process



- Hb threshold for eligibility is 8.4(7.8) mmol/l for male(female)

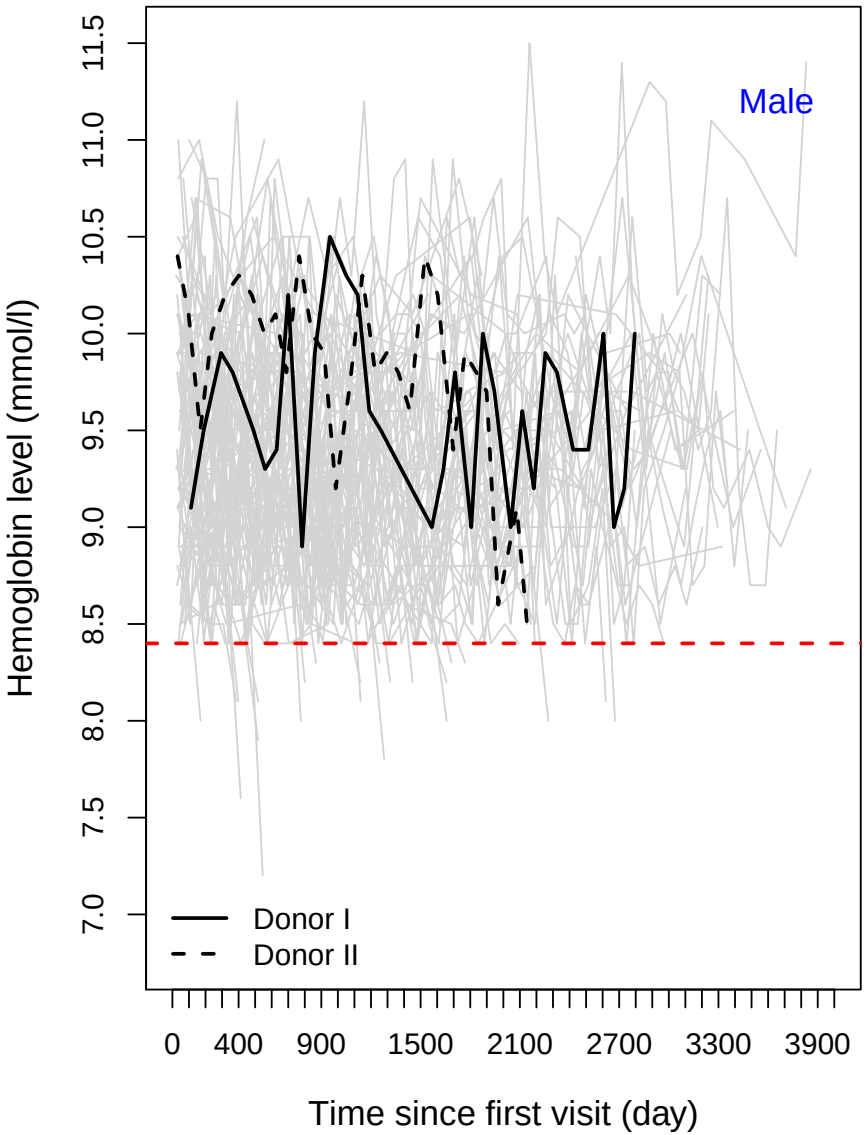
Data

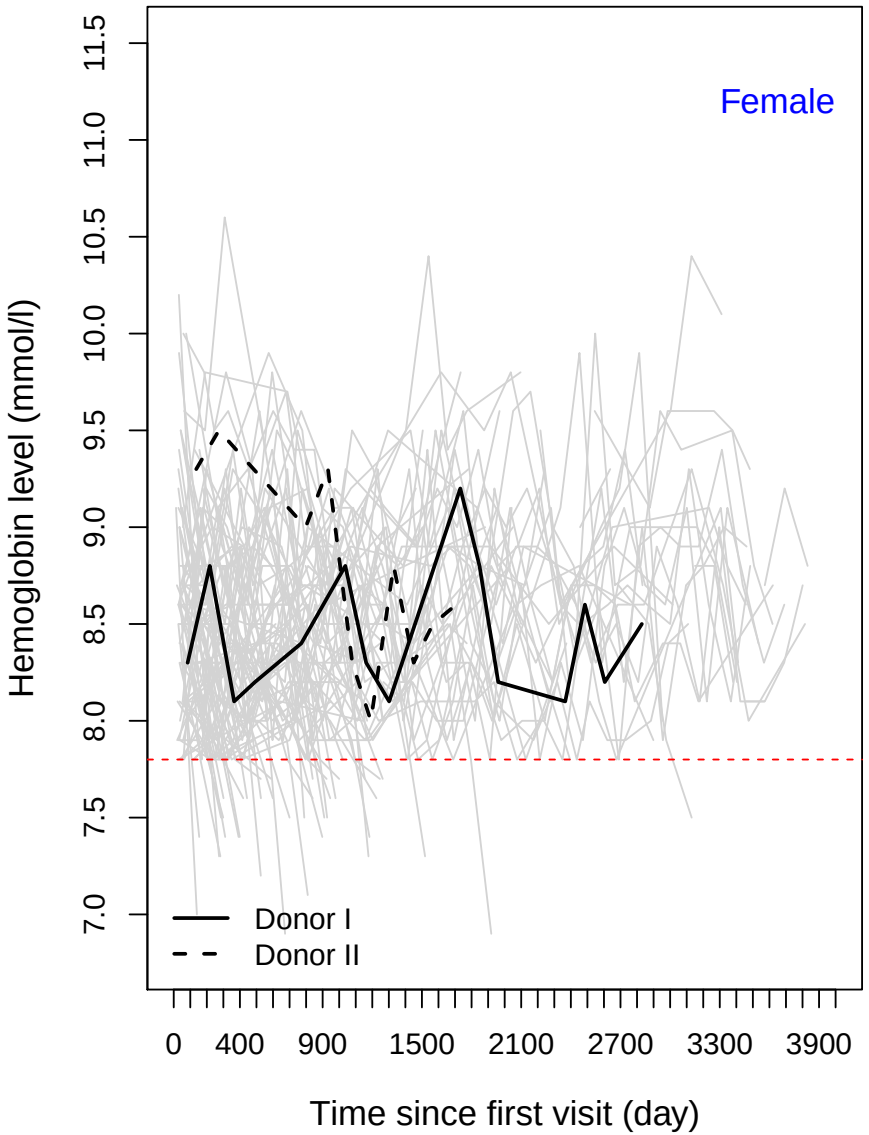
- A random sample of 2000 male and 2000 female donors collected by Sanquin Blood Supply (the Netherlands) (2005-2012).
  New-entrant donors, at least one donation and one visit afterward.

Covariates

- Season that donation takes place
- Age at 1st visit/donation
- Time since previous donation
- Number of donations in past two years
- Gender

Hemoglobin profiles for male and female donors





Descriptive statistics

	Male	**Female**
Donors deferred at least once due to low Hb	18.4%	32.3%
Donations in cold season	49.9%	49.5%
Age at screening visit (years)	34.8 (24.1, 45.8)	29.4 (21.6, 42.2)
Number of donations	9 (4, 16)	5 (2, 9)
Hb value at screening visit (mmol/l)	9.4 (9.0, 9.9)	8.4 (8.0, 8.8)
Inter-donation interval (days)	90 (74, 126)	138 (120, 188)

A statistical model

Capture the unobserved heterogeneity by (continuous latent classes) random effects.

- **Mixed-effects model**

$$Hb_{it} = \theta_1 + b_{i0} + \beta_1 Age_{i1} + \beta_2 Season_{it} + \beta_3 TSPD_{it}$$
$$+\beta_4 TSPD_{it}^2 + \beta_5 BFD_{it} + (\theta_2 + b_{i1}) NODY2_{it} + \epsilon_{it}$$

Non-informative priors!

Results of LMM (Male)

	Estimate	**CI**	σ^2
θ_1	9.589	9.556 9.624	3.1×10^{-4}
θ_2	-0.0545	-0.0596 -0.0496	6.3×10^{-6}

A statistical model

Capture the unobserved heterogeneity by latent class mixed model (growth mixture model).

- The density function of $\mathbf{y}$ in a Gaussian growth mixture model can be expressed as:

$$f(\mathbf{y}|\theta) = \sum_{k=1}^{K} w_k f_k(y|\theta_k),$$

- $f_k(y|\theta_k)$ (k=1,...,K) are density functions describe each class
- w_k is mixing distribution, $0 \leq w_k \leq 1$ and $\sum_{k=1}^{K} w_k = 1$.

A GMM with 2 latent classes

A growth mixture model with two classes where one class restricted to be stable regarding to "number of donations".

If person i belongs to class 1: (Stable class)

$$Hb_{it} = \theta_{11} + \mathbf{b_{i10}} + \beta_1 Age_{i1} + \beta_2 Season_{it} + \beta_3 TSPD_{it}$$
$$+\beta_4 TSPD_{it}^2 + \beta_5 BFD_{it} + \epsilon_{it}$$

If person i belongs to class 2:

$$Hb_{it} = \theta_{21} + \mathbf{b_{i20}} + \beta_1 Age_{i1} + \beta_2 Season_{it} + \beta_3 TSPD_{it}$$
$$+\beta_4 TSPD_{it}^2 + \beta_5 BFD_{it} + (\theta_{22} + \mathbf{b_{i21}}) NODY2_{it} + \epsilon_{it}$$

Data dependent priors! Elliott et al. 2005 Biostatistics.

Tends to force the posterior trajectories of the classes to be more equal!
Identifiable!

Two scenarios for the mixing distribution

Scenario 1:

Latent Class Model – no covariates, Dirichlet prior on mixing distribution.

$g[j] \sim dcat(w[j,])$ # Latent class indicator

$w[i, 1:k] \sim ddirch(a_1, a_2, ..., a_k)$

Fruhwirth-Schnatter 2006: $a_k > 1$ to ensure no empty class!

Rousseau et al., 2011: $a_k < d/2$ Ensure no over fitting!

Two scenarios for the mixing distribution

Scenario 2:
Latent Class Model – depends on some other covariates

$g[j] \sim dcat(w[j,])$ # Latent class indicator

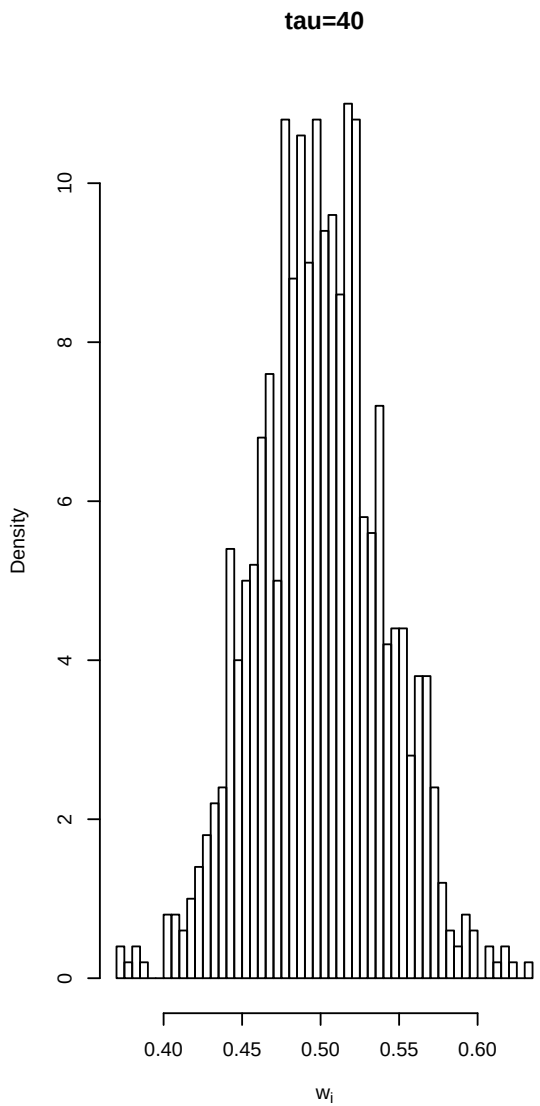
$w[j, class] < -phi[j, class]/sum(phi[j,])$

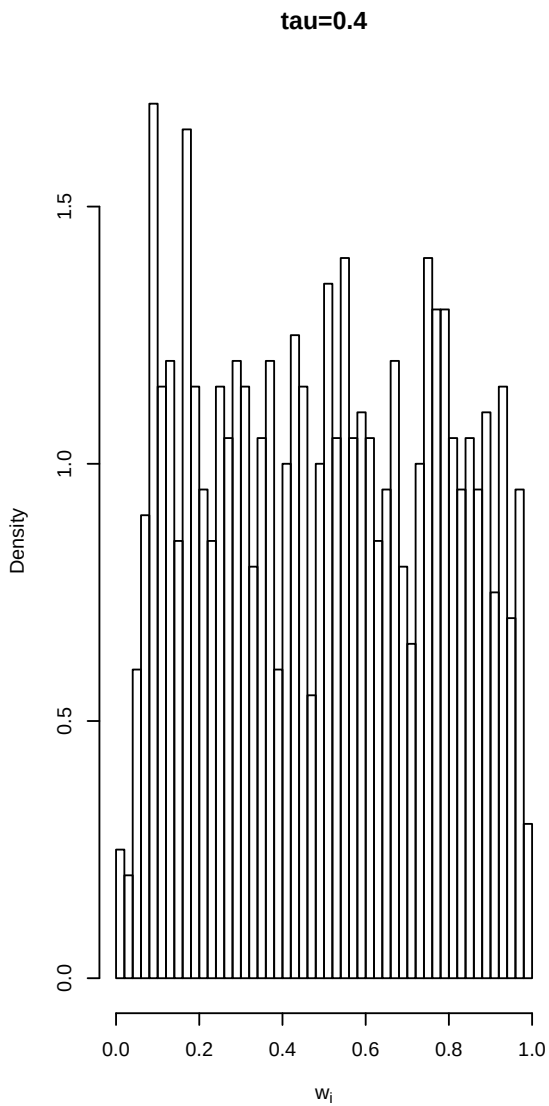
$log(phi[j, class])$ <-
$gamma[class, 1] + gamma[class, 2] * Age0[j] + gamma[class, 3] * Hb0[j]$

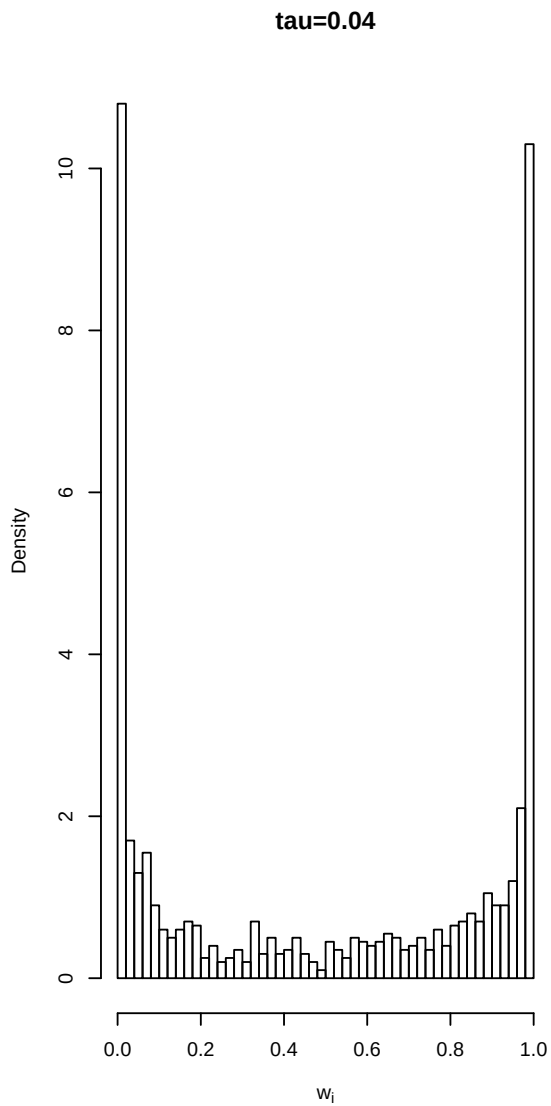
Informative or non-informative priors? $gamma[class, c] \sim dnorm(0, 4/9)$

After log transformation, this provides relatively flat prior!
Garrett and Zeger, 2000 Biometrics.

Priors distribution after log transformation!







Results of GMM with 2 latent classes

	Estimate	**CI**	σ^2	w_i
θ_{11}	9.037	8.998 9.072	3.5×10^{-4}	**42.5%**
θ_{21}	9.831	9.789 9.872	4.3×10^{-4}	**57.5%**
θ_{22}	-0.06849	-0.0749 -0.0622	1.1×10^{-5}	

Software

- **Jags** via Rjags package in R! http://mcmc-jags.sourceforge.net/
- **Stan** via Rstan package in R! http://mc-stan.org/index.html

- Stan, less iterations to be converged, e.g 800 - 1500 samples!

- Jags, more iterations to be converged, e.g 20,000 - 100,000 samples!

- Both computationally expensive and regarding to time almost the same!

Jags/Bugs syntax for growth mixture model

```
model {
for(j in 1:n) {
for(i in offset[j]:(offset[j+1]-1)){
Hb[i]~dnorm(mu[i],tau)
mu[i]<-beta[1]*Age1[j]+beta[2]*Season[i]+...+beta[5]*d[i]+
equals(g[j],1)*(theta11+b01[j])+     #  class I
equals(g[j],2)*(theta21+b02[j,1]+(theta22+b02[j,2])*NODY2[i]) # class  II
 }
g[j] ~ dcat(w[j,]) # Latent class indicator
for(class in 1:2)
{
w[j,class]<-phi[j,class]/sum(phi[j,])
log(phi[j,class])<-gamma[class,1]+ gamma[class,2]*Age0[j]+gamma[class,3]*Hb0[j]
}
}
```

Stan syntax for growth mixture model

```
transformed parameters {
vector[n] w1;    //mixing proportions
vector[n] w2;    //mixing proportions
vector[n] ppi1;  //
for(j in 1:n){
ppi1[j]<-exp(gamma[1,1]+ gamma[2,1]*Age0[j]+ gamma[3,1]*Hb0[j]);
                }
for(j in 1:n){
w1[j]<-ppi1[j]/(ppi1[j]+1);
w2[j]<-1/(ppi1[j]+1);
                }     }
model {
for(j in 1:n)
                 {
for(i in offset[j]:(offset[j+1]-1)){

 psi[1]<- log(w1[j])+normal_log(Hb[i],theta11+b1[j]+beta[1]*Age1[i]+..., sigma);// Class I
 psi[2]<- log(w2[j])+normal_log(Hb[i],theta21+b2[j,1]+beta[1]*Age1[i]+...+(theta22+b2[j,2])*DONY2[i], sigma);// Class II
 increment_log_prob(log_sum_exp(psi));
       }      }       }
```

A GMM with 3 latent classes

A growth mixture model with three classes where one class restricted to be stable regarding to "number of donation".

If person i belongs to class 1: (Stable class)

$$Hb_{it} = \theta_{11} + \mathbf{b_{i10}} + \beta_1 Age_{i1} + ... + \beta_5 BFD_{it} + \epsilon_{it}$$

If person i belongs to class 2:

$$Hb_{it} = \theta_{21} + \mathbf{b_{i20}} + \beta_1 Age_{i1} + ... + \beta_5 BFD_{it} + (\theta_{22} + \mathbf{b_{i21}}) NODY2_{it} + \epsilon_{it}$$

If person i belongs to class 3:

$$Hb_{it} = \theta_{31} + \mathbf{b_{i30}} + \beta_1 Age_{i1} + ... + \beta_5 BFD_{it} + (\theta_{32} + \mathbf{b_{i31}}) NODY2_{it} + \epsilon_{it}$$

Data dependent priors! Unidentifiable!

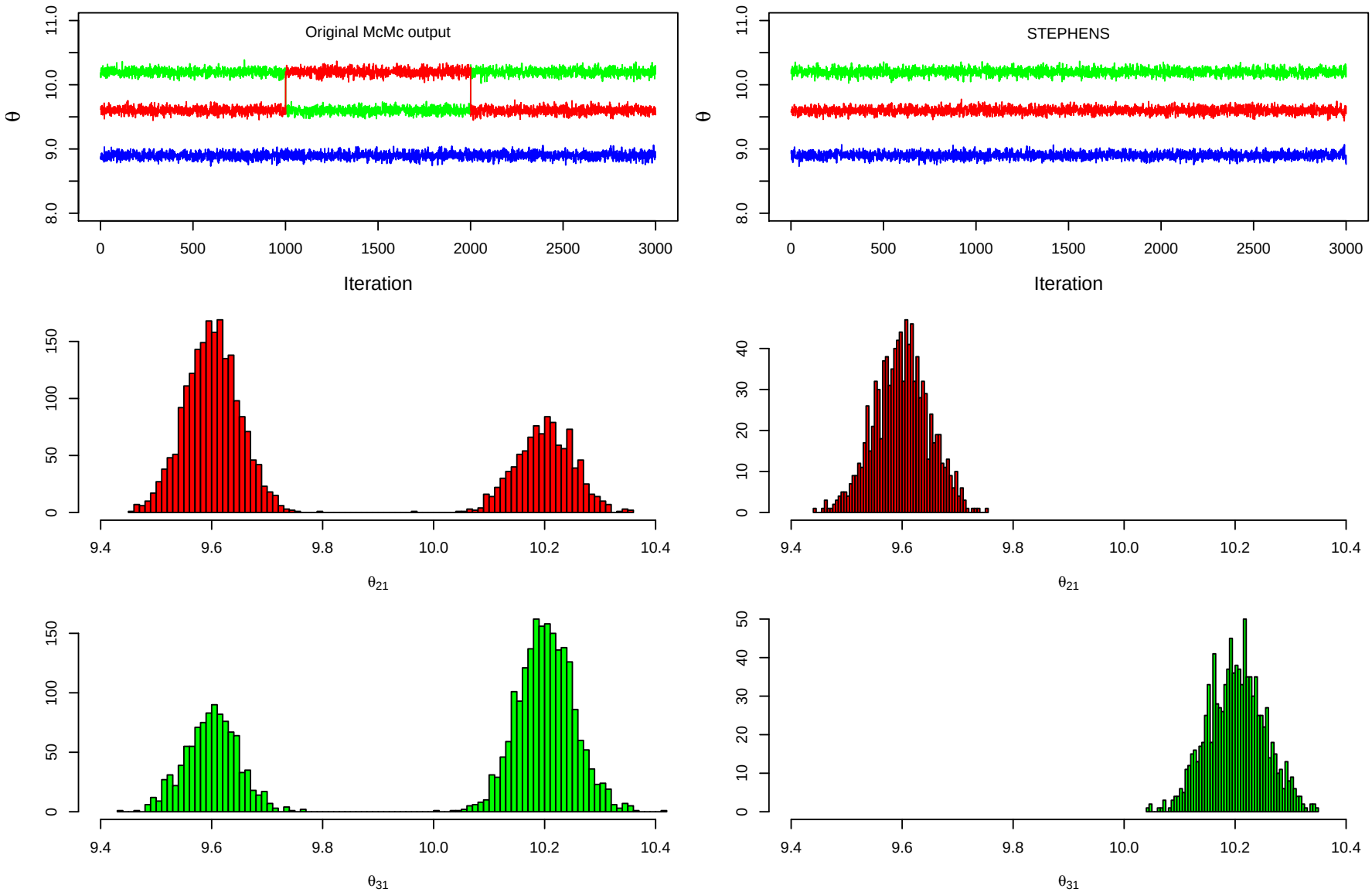
Label switching problem

$$f(y|\theta) = \sum_{k=1}^{K} w_k f_k(y|\theta_k) = w_1 f_1(y|\theta_1) + ... + w_k f_k(y|\theta_k)$$

K! permutations

- Simple solution: Imposing artificial identifiability constraints on the class parameters, to make them identifiable.
- Relabeling algorithm (on McMc output): Stephens 2000.
- More algorithms (on McMc output)!!!
- When the data set is large, label switching very seldom occurs! Pritchard et al., 2000; Guillot et al., 2005

Label switching problem example



Results of GMM with 3 latent classes

	Estimate	**CI**	σ^2	w_i
θ_{11}	8.96	8.91 9.02	7.6×10^{-4}	**22.5%**
θ_{21}	9.59	9.54 9.63	6.5×10^{-4}	**50.3%**
θ_{22}	-0.053	-0.063 -0.044	2.4×10^{-5}	
θ_{31}	10.21	10.14 10.28	1.3×10^{-3}	**27.2%**
θ_{32}	-0.085	-0.102 -0.069	6.7×10^{-5}	

Model comparison (How many classes?)

- WinBUGS currently does not compute DIC if the likelihood depends on any discrete parameters
- Jags currently needs more than one chain to compute DIC. (between chains label switching problems)
- Computing DIC manually outside the Bugs for large data set is almost impossible!
- Solution: Computing BIC manually outside the Bugs based on marginal posterior parameter!
- GMM with 4 latent classes selected.

Assess model fit

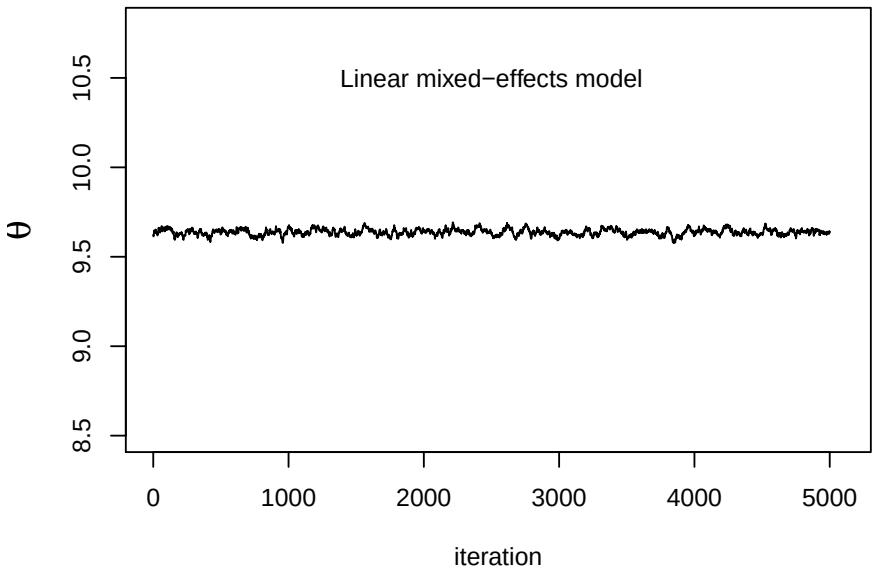
- Using PPC by computing a Bayesian P-value, χ^2 discrepancy measure to test both the distributional and latent class number assumption of the model.
- Posterior classification table.

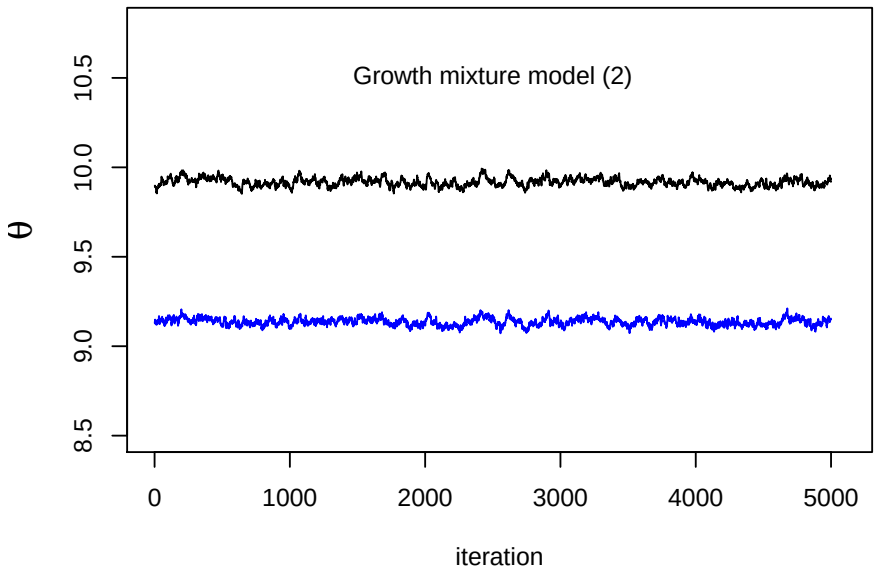
Assess model fit

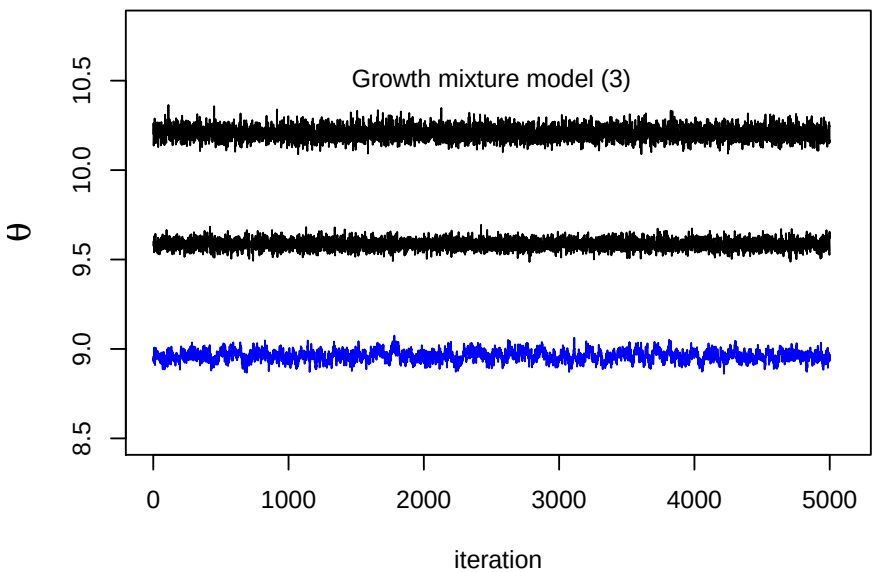
- Ex: Class in each iteration: 1 1 1 2 1 3 1 .....4 1 1
- Posterior classification table.

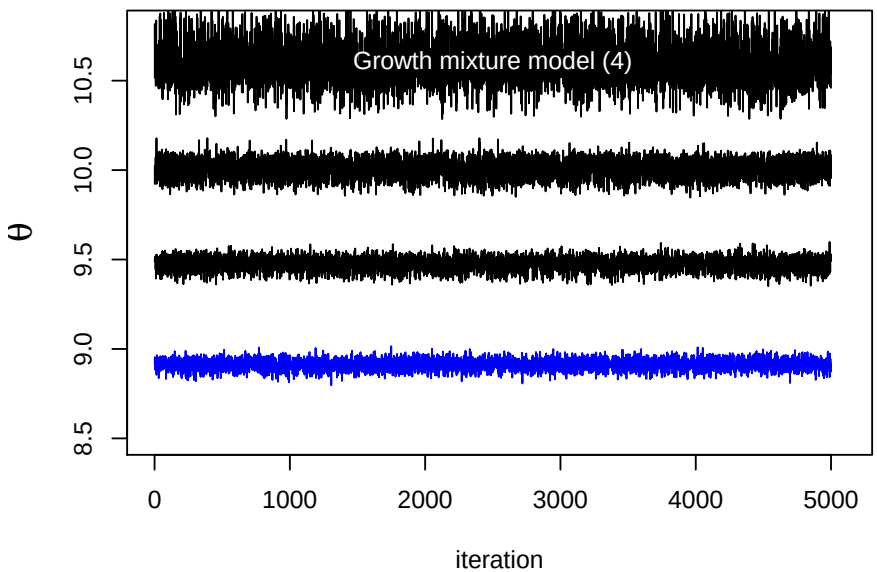
	Male				**Female**			
Latent Class	**Class I**	**Class II**	**Class III**	**Class IV**	**Class I**	**Class II**	**Class III**	**Class IV**
Class I	0.795	0.204	0.001	0.000	0.771	0.227	0.002	0.000
Class II	0.135	0.717	0.144	0.004	0.103	0.744	0.142	0.001
Class III	0.001	0.150	0.736	0.113	0.001	0.163	0.680	0.156
Class IV	0.000	0.001	0.191	0.808	0.000	0.008	0.121	0.780

Intercepts of different classes for male donors





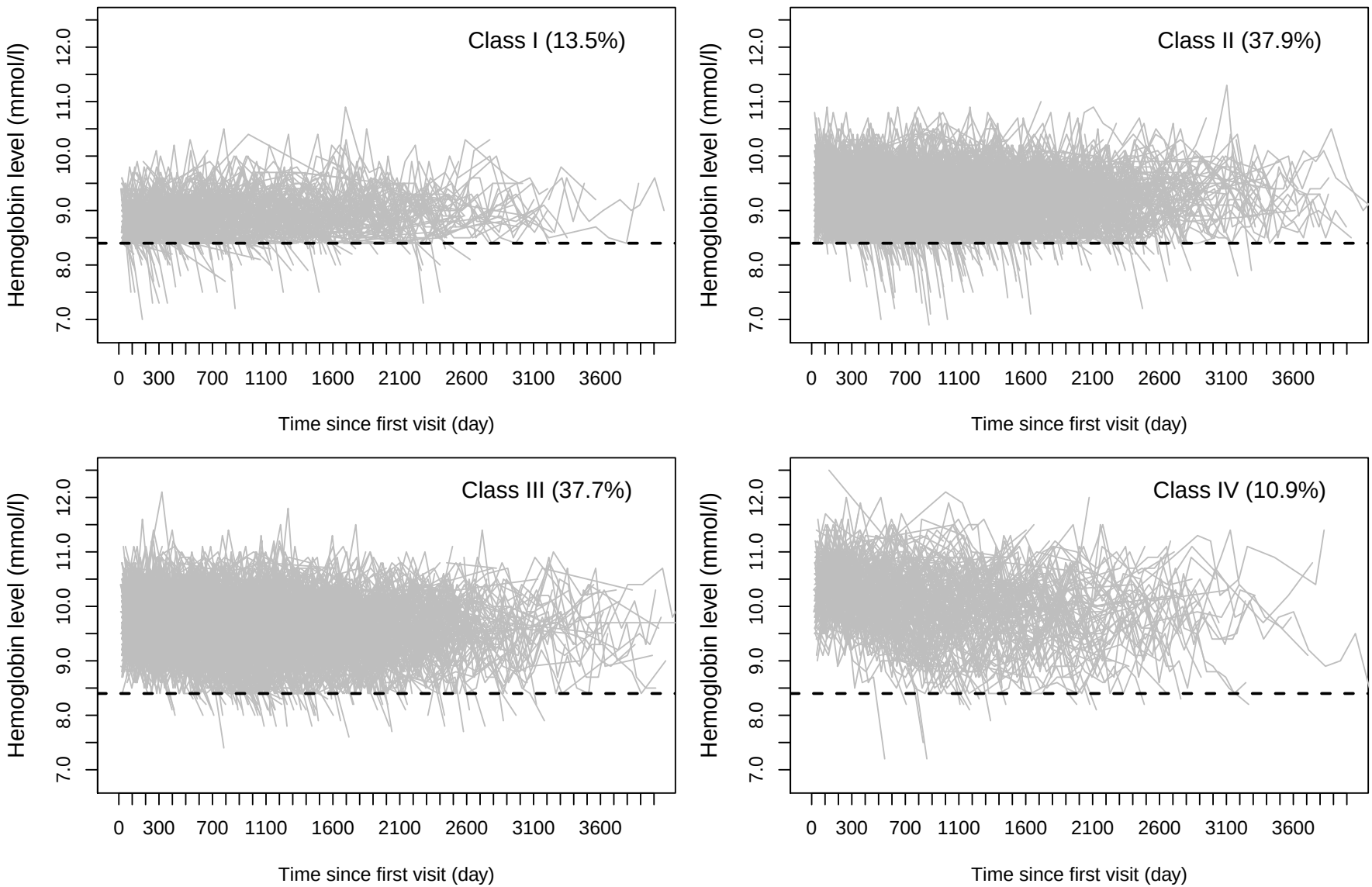




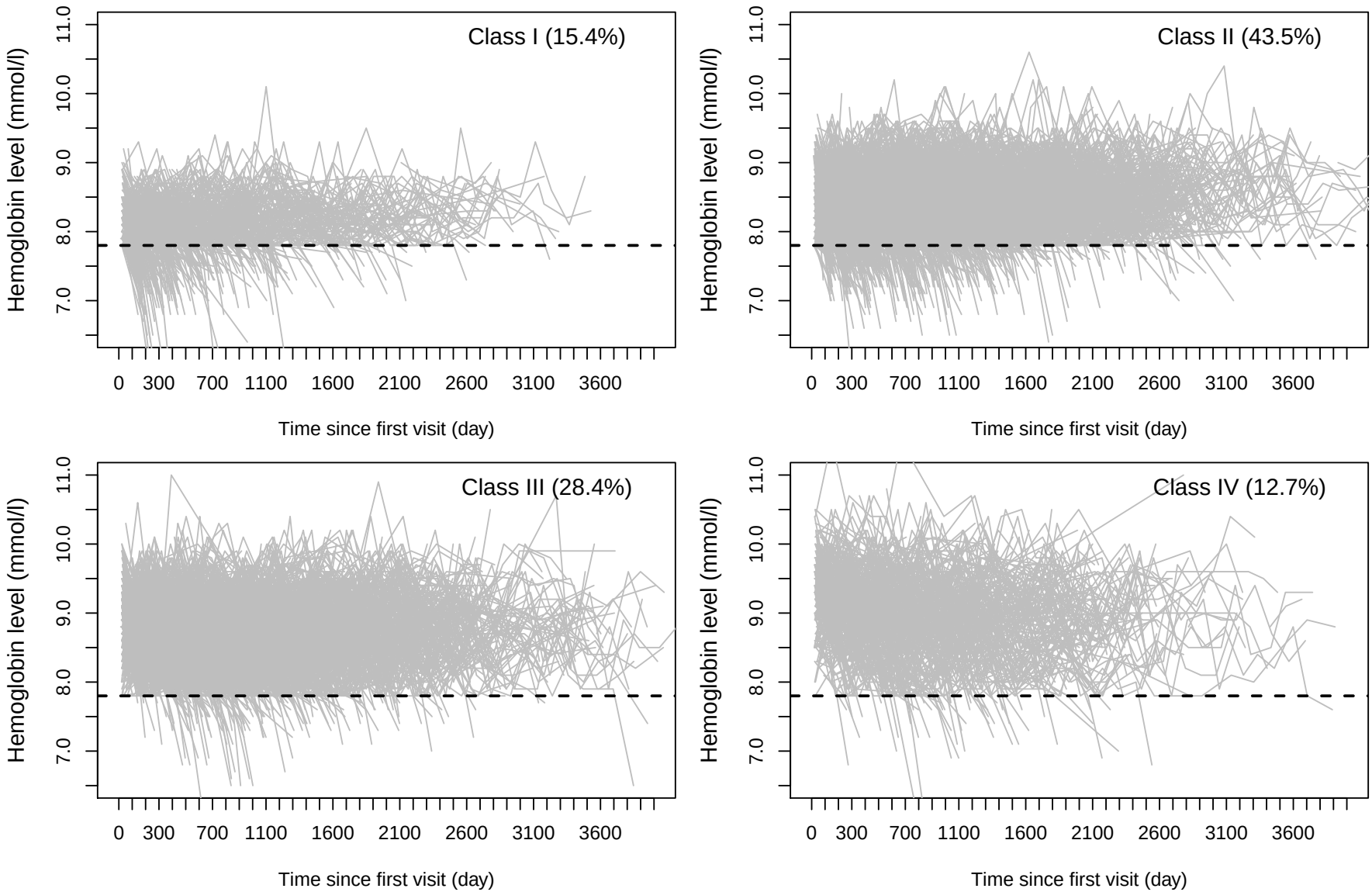
Parameter estimation (GMM IV) for male donors

	Male donors			Female donors		
Parameter	**Estimation**	**95% CI**		**Estimation**	**95% CI**	
$Intercept_{I}$	8.88	8.81	8.89	7.93	7.85	8.00
$Intercept_{II}$	9.34	9.27	9.44	8.30	8.24	8.36
$Intercept_{III}$	9.84	9.78	9.94	8.77	8.68	8.84
$Intercept_{IV}$	10.38	10.23	10.59	9.13	9.02	9.25
$NODY2_{II}$	-0.05	-0.06	-0.04	-0.02	-0.04	-0.01
$NODY2_{III}$	-0.06	-0.08	-0.05	-0.06	-0.10	-0.03
$NODY2_{IV}$	-0.09	-0.12	-0.07	-0.14	-0.17	-0.10
$Age_0(year)$	1.7×10^{-4}	-2.8×10^{-3}	3.1×10^{-3}	9.6×10^{-3}	6.4×10^{-3}	1.3×10^{-2}
$Season$	-7.7×10^{-2}	-8.9×10^{-2}	-6.5×10^{-2}	-5.2×10^{-2}	-6.2×10^{-2}	-4.2×10^{-2}
$TSPD(month)$	2.5×10^{-2}	1.9×10^{-2}	3.0×10^{-2}	1.9×10^{-2}	1.4×10^{-2}	2.3×10^{-2}
$TSPD^2(month)$	-1.0×10^{-4}	-1.3×10^{-4}	-7.3×10^{-5}	-4.5×10^{-5}	-6.2×10^{-5}	-2.8×10^{-5}
BFD	-8.7×10^{-4}	-3.2×10^{-2}	3.1×10^{-2}	8.3×10^{-2}	5.4×10^{-2}	11.1×10^{-2}

Hemoglobin profiles for different classes for male donors



Hemoglobin profiles for different classes for female donors

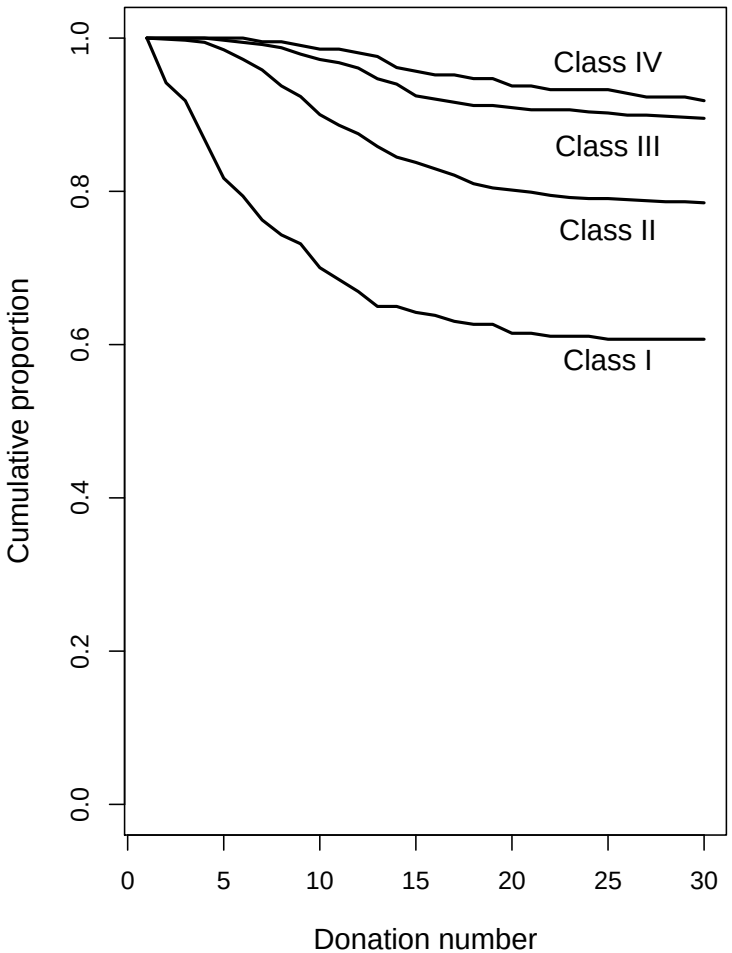


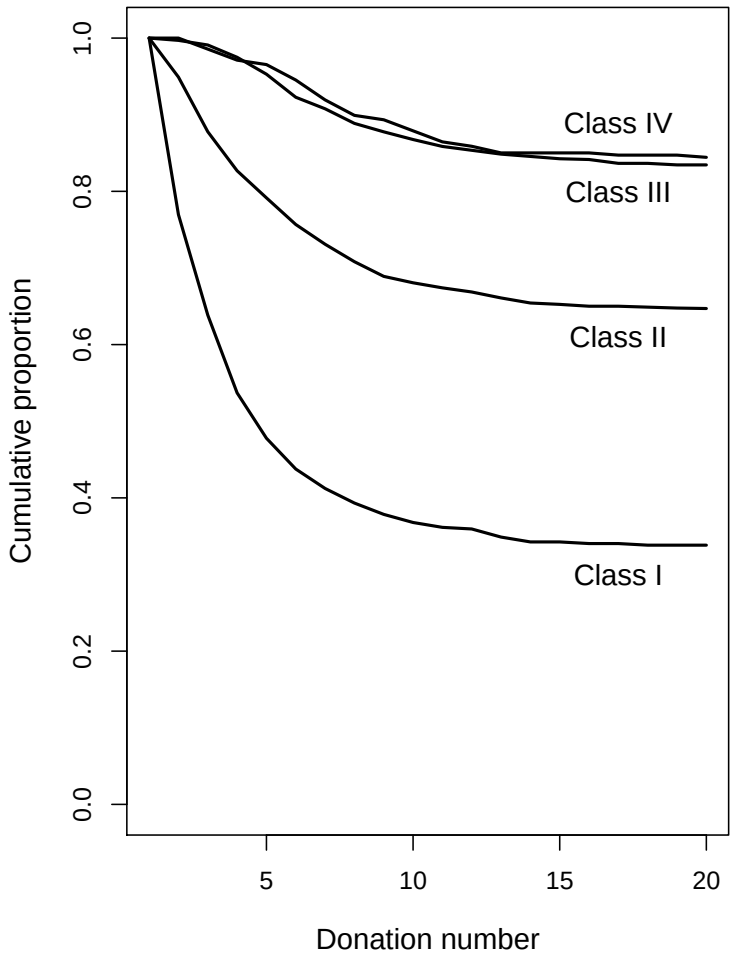
Latent classes description

Male donors	**Class I**	**Class II**	**Class III**	**Class IV**
Size of the class	13.5%	37.9%	37.7%	10.9%
Donors deferred at least once due to low Hb	39.3 %	21.6 %	10.6%	8.2 %
Number of donations	7(2, 13)	9(5, 16)	10(5, 17)	11(5, 18)
Hb value at screening visit (mmol/l)	8.7(8.5, 8.9)	9.2(9.0, 9.4)	9.7(9.4, 10.0)	10.4(10.1, 10.7)
Age at screening visit (year)	39(29, 51)	36(26, 48)	32(24, 43)	33(24, 44)

Female donors	**Class I**	**Class II**	**Class III**	**Class IV**
Size of the class	15.4%	43.5%	28.4%	12.7%
Donors deferred at least once due to low Hb	66.6%	35.4%	16.6%	15.8%
Number of donations	2(1, 5)	4(2, 9)	6(3, 11)	6(3, 9)
Hb value at screening visit (mmol/l)	7.8(7.6, 8.0)	8.2(8.0, 8.5)	8.7(8.5, 9.0)	9.3(9.1, 9.6)
Age at screening visit (year)	43.5(33, 52)	29(22, 41)	27(21, 40)	23(20, 31)

Donors deferred proportion Kaplan-Meier curves of the latent classes





Predicting latent class membership

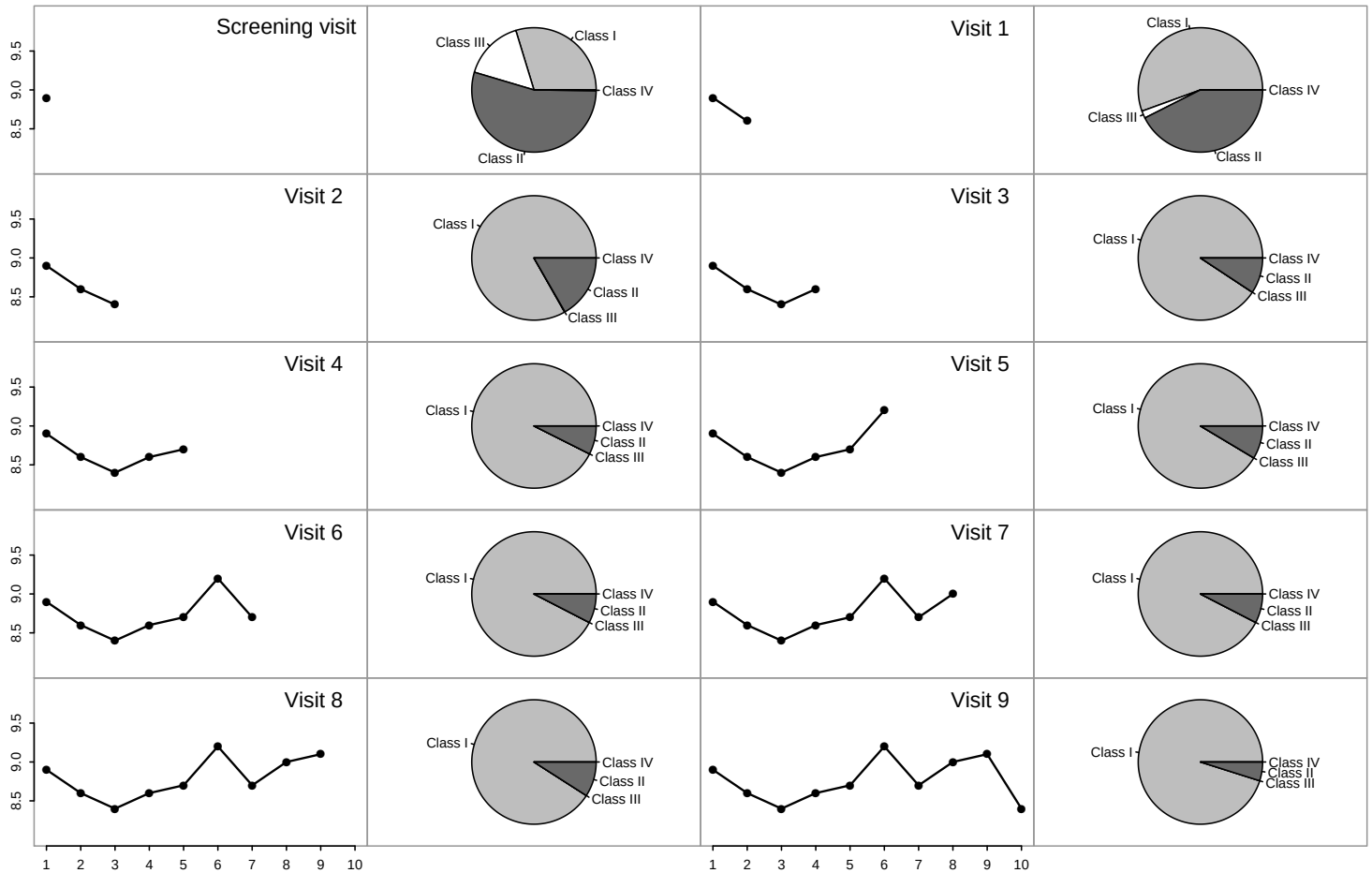
- The probability that individual i belongs to the kth latent class can be calculated:

$$P(c_i = k|Hb_i, Age_{0i}, Hb_{0i}, \theta_k) = \frac{w_{ik} f_k(Hb_i|\theta_k)}{\sum_{k=1}^{K} w_{ik} f_k(Hb_i|\theta_k)}$$

- This is a dynamic prediction for the latent class of a donor. In each visit it can be updated!

EX:Predicting latent class membership

For a male donor with Hb value of 8.9 mmol/l and age of 29 years at the screening visit.



Conclusion

- To capture the unobserved heterogeneity of Hb profiles, we implemented a Bayesian GMM. This model assumes that each donor belongs to one of several latent classes.
- Within each class, the Hb trajectory follows a linear mixed model.
- In addition we let the latent class membership depend on the age and hemoglobin value at first visit.

Conclusion

- Our fitted GMM suggests 4 different classes of Hb trajectories.
- This model gives some insight in the donation process and is a start to better predict for which donors care needs to be exercised not to produce a too low Hb level.

Future research

- Find a better way to determine the number of optimum latent classes:
  Reversible Jump McMC (Green 1995)
  Variational Bayes algorithm (Bruneau 2010)
- Robust Bayesian growth mixture model.
- Cost-effectiveness analysis.

Thank you!

K.nasserinejad@erasmusmc.nl