

Analysis of eGFR trajectories from Hong Kong Diabetes Registry

Guozhi Jiang

May 16, 2015

Contents

1	Description of data	1
1.1	Data overview	1
1.2	Outcome	5
2	Analysis	10
2.1	3 classes	10
2.2	4 classes	17

Chapter 1

Description of data

1.1 Data overview

```
> rm(list=ls())
> start <- Sys.time()
> set.seed(1983)
> library(lcmm)
> library(Epi)
> library(lme4)
> library(nlme)
> print(sessionInfo(), l=F)

R version 3.1.0 (2014-04-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)

attached base packages:
[1] splines      stats      graphics  grDevices  utils      datasets  methods
[8] base

other attached packages:
[1] nlme_3.1-117      lme4_1.1-7        Rcpp_0.11.5      Matrix_1.1-3
[5] Epi_1.1.67        lcmm_1.7.2        survival_2.37-7

loaded via a namespace (and not attached):
[1] compiler_3.1.0  grid_3.1.0        lattice_0.20-29  MASS_7.3-31
[5] minqa_1.2.4     nloptr_1.0.4      tools_3.1.0
```

The data are comprised by two parts: the baseline data and the follow-up eGFR data. The baseline data, which has been extracted from the Hong Kong Diabetes Registry(HKDR), including the information of clinical assessments and laboratory investigations at enrollment, and the well-defined complication outcomes censored to 30th, June 2014. As we focus on the ESRD outcome in this analysis, we only select those Chinese patients with no history of ESRD which is defined according to the ICD-9 codes and eGFR <15 . Therefore, we obtain a cohort consisted of 1218 ESRD events and 8336 event-free patients at censoring date. All dates are transformed into continuous variables. The follow-up data include all the creatinine records from enrollment to 2014, and the eGFR is corresponding calculated using the Chinese-modified MDRD formula.

```
> base_dat <- read.table("../data/ESRD1_Prosp2014_CH-T2D_1218vs8336.csv", header=TRUE, sep=",")
> follow_dat <- read.table("../data/eGFR_19940714_20140630.csv", header=TRUE, sep=",")
```

```
> base_dat <- transform(base_dat, doin=cal.yr(date), dob=cal.yr(DOB, "%d/%m/%Y"),
+                       dox=cal.yr(ESRD1_DATE))
> base_dat <- transform(base_dat, dodm = pmin(YEAR_DIA + runif(length(YEAR_DIA)), doin))
> dob_na <- which(is.na(base_dat$dob))
> base_dat$dob[dob_na] <- (floor(base_dat$doin[dob_na]) - base_dat$AGE[dob_na]) +
+   runif(length(dob_na))
> base_subdat <- subset(base_dat, select=c("Obs_id", "doin", "dob", "dodm", "dox",
+   "SEX", "ESRD1_END"))
> dim(base_subdat)
```

```
[1] 9554    7
```

```
> names(base_subdat)
```

```
[1] "Obs_id"    "doin"      "dob"       "dodm"      "dox"       "SEX"
[7] "ESRD1_END"
```

```
> head(base_subdat)
```

	Obs_id	doin	dob	dodm	dox	SEX	ESRD1_END
1	1	2002.805	1940.598	2002.493	2014.493	0	0
2	2	1996.757	1939.194	1995.153	2014.493	1	0
3	3	1996.585	1935.172	1983.217	2014.493	1	0
4	4	2001.203	1927.048	1980.747	2004.648	1	0
5	5	1997.381	1924.527	1993.908	2014.493	1	0
6	6	1999.221	1922.345	1991.221	2010.564	0	0

```
> table(base_subdat$ESRD1_END)
```

```
 0    1
8336 1218
```

```
> follow_dat <- transform(follow_dat, dolab=cal.yr(test_date))
> follow_dat <- subset(follow_dat, select=-c(test_date))
> dim(follow_dat)
```

```
[1] 391551    4
```

```
> head(follow_dat)
```

	Obs_id	F_eGFR	creatinine	dolab
1	1	80.6474	84	2002.632
2	1	97.9131	71	2002.643
3	1	73.5245	91	2002.663
4	1	91.8750	75	2002.767
5	1	72.5693	92	2002.805
6	1	81.4311	83	2003.914

We merge the baseline data and follow-up eGFR data according to the id of subject:

```
> merged_dat <- merge(base_subdat, follow_dat, by=intersect("Obs_id", "Obs_id"), sort=F)
> dim(merged_dat)
```

```
[1] 366122    10
```

```
> head(merged_dat)
```

	Obs_id	doin	dob	dodm	dox	SEX	ESRD1_END	F_eGFR	creatinine
1	1	2002.805	1940.598	2002.493	2014.493	0	0	102.8468	66
2	1	2002.805	1940.598	2002.493	2014.493	0	0	91.8750	75
3	1	2002.805	1940.598	2002.493	2014.493	0	0	96.8863	71
4	1	2002.805	1940.598	2002.493	2014.493	0	0	80.6474	84
5	1	2002.805	1940.598	2002.493	2014.493	0	0	87.7935	77
6	1	2002.805	1940.598	2002.493	2014.493	0	0	103.2903	66
	dolab								
1	2014.359								
2	2002.767								
3	2005.951								
4	2002.632								
5	2007.558								
6	2012.812								

```
> addmargins(table(table(merged_dat$Obs_id)))
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
350	65	69	74	94	141	182	179	158	174	160	150	155	172	158	176
17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
172	222	214	200	220	160	174	196	178	164	177	159	159	164	153	119
33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
130	129	132	125	112	105	117	126	123	118	113	99	88	122	76	93
49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
93	98	73	75	69	62	54	56	68	59	62	55	54	48	43	34
65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
32	38	36	43	38	36	27	34	26	29	40	30	28	28	30	21
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
25	22	22	23	25	29	23	23	25	24	19	16	21	16	17	14
97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112
10	18	22	12	14	10	12	17	11	13	16	13	12	10	8	12
113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128
9	8	13	8	8	10	11	12	10	6	7	4	5	5	4	5
129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144
6	10	6	8	2	9	9	1	9	5	6	3	2	2	4	3
145	146	147	148	149	150	151	153	154	155	156	158	159	160	161	162
1	2	7	4	7	7	4	2	5	4	2	7	4	2	2	1
163	164	165	166	167	169	170	171	172	173	175	176	177	178	179	181
1	5	2	1	4	2	3	2	1	6	3	4	3	4	4	1
182	183	186	187	191	193	194	195	196	197	198	199	200	204	206	207
2	2	1	3	2	1	1	3	1	2	1	2	2	1	1	1
208	209	210	211	212	213	216	218	221	223	224	225	228	231	232	238
1	1	1	1	1	1	2	1	1	1	1	1	2	1	1	1
240	241	243	249	250	251	252	254	258	267	282	284	288	293	301	308
1	1	1	1	1	2	2	1	1	3	1	1	1	1	1	1
319	353	431	725	Sum											
1	1	1	1	9554											

We only select those records between enrollment and event/censoring dates, that said, those eGFR records before enrollment or after event/censoring dates need to be removed. Moreover, we also calculate the follow-up age(F_AGE), duration of diabetes(F_DMAGE), and the backward time gap between event/censoring date and measurement date of eGFR (BW_TIME).

```
> sub_merged <- subset(merged_dat, (1:nrow(merged_dat)) %in%
+   intersect(which(dolab >= doin), which(dolab <= dox)))
> sub_merged <- transform(sub_merged, F_AGE=dolab-dob, F_DMAGE=dolab-dodm, BW_TIME=dolab-dox)
> dim(sub_merged)
```

```
[1] 282607    13
```

```
> head(sub_merged)
```

	Obs_id	doin	dob	dodm	dox	SEX	ESRD1_END	F_eGFR	creatinine
1	1	2002.805	1940.598	2002.493	2014.493	0	0	102.8468	66
3	1	2002.805	1940.598	2002.493	2014.493	0	0	96.8863	71
5	1	2002.805	1940.598	2002.493	2014.493	0	0	87.7935	77
6	1	2002.805	1940.598	2002.493	2014.493	0	0	103.2903	66
7	1	2002.805	1940.598	2002.493	2014.493	0	0	81.2613	83
8	1	2002.805	1940.598	2002.493	2014.493	0	0	81.4311	83

	dolab	F_AGE	F_DMAGE	BW_TIME
1	2014.359	73.76044	11.865461	-0.1341547
3	2005.951	65.35250	3.457521	-8.5420945
5	2007.558	66.95962	5.064640	-6.9349760
6	2012.812	72.21355	10.318575	-1.6810404
7	2004.568	63.96988	2.074906	-9.9247091
8	2003.914	63.31554	1.420560	-10.5790554

```
> str(sub_merged)
```

```
'data.frame':      282607 obs. of  13 variables:
 $ Obs_id      : int   1 1 1 1 1 1 1 1 1 1 ...
 $ doin        : num   2003 2003 2003 2003 2003 ...
 $ dob         : num   1941 1941 1941 1941 1941 ...
 $ dodm        : num   2002 2002 2002 2002 2002 ...
 $ dox         : num   2014 2014 2014 2014 2014 ...
 $ SEX         : int    0 0 0 0 0 0 0 0 0 0 ...
 $ ESRD1_END   : int    0 0 0 0 0 0 0 0 0 0 ...
 $ F_eGFR      : num   102.8 96.9 87.8 103.3 81.3 ...
 $ creatinine  : num    66 71 77 66 83 83 92 74 73 69 ...
 $ dolab       : num   2014 2006 2008 2013 2005 ...
 $ F_AGE       : num    73.8 65.4 67 72.2 64 ...
 $ F_DMAGE     : num    11.87 3.46 5.06 10.32 2.07 ...
 $ BW_TIME     : num   -0.134 -8.542 -6.935 -1.681 -9.925 ...
```

```
> range(sub_merged$BW_TIME)
```

```
[1] -21.03491  0.00000
```

1.2 Outcome

We extract a complete-case cohort by removing those records with missing values. To have an overview of the outcome variables, we first plot the observed creatinine and eGFR values (Figure 1.1). From the figure, we can see there are some abnormal records, which may be due to measurement or typo errors.

```
> nomiss_dat <- sub_merged[complete.cases(sub_merged), ]
> dim(nomiss_dat)

[1] 279002    13

> with(nomiss_dat, table((F_eGFR>300) + (F_eGFR>1000)))

      0      1      2
278701  265   36

> with(nomiss_dat[sample(1:nrow(nomiss_dat), 5000), ],
+       plot(dolab, F_eGFR, pch=16, cex=0.3, xlab="Date of measurement", ylab="eGFR"))

> with(nomiss_dat[sample(1:nrow(nomiss_dat), 5000), ],
+       plot(dolab, creatinine, pch=16, cex=0.3, xlab="Date of measurement", ylab="Creatinine"))
```



Figure 1.1: *Distribution of the raw data of eGFR and creatinine.*

We then remove those records with eGFR ≥ 300 , which are considered to be errors. The updated distributions were shown in Figure 1.2.

```
> sub_nomiss <- subset(nomiss_dat, F_eGFR<300)
> dim(sub_nomiss)
```

```
[1] 278701      13

> with(sub_nomiss[sample(1:nrow(sub_nomiss), 2000), ],
+       plot(BW_TIME, F_eGFR, pch=16, cex=0.3,
+            xlab="Time before ESRD", ylab="eGFR"))
> abline(h=15, col="red")

> with(sub_nomiss[sample(1:nrow(sub_nomiss), 2000), ],
+       plot(BW_TIME, creatinine, pch=16, cex=0.3, #ylim=c(0,400),
+            xlab="Time before ESRD", ylab="Creatinine"))
```

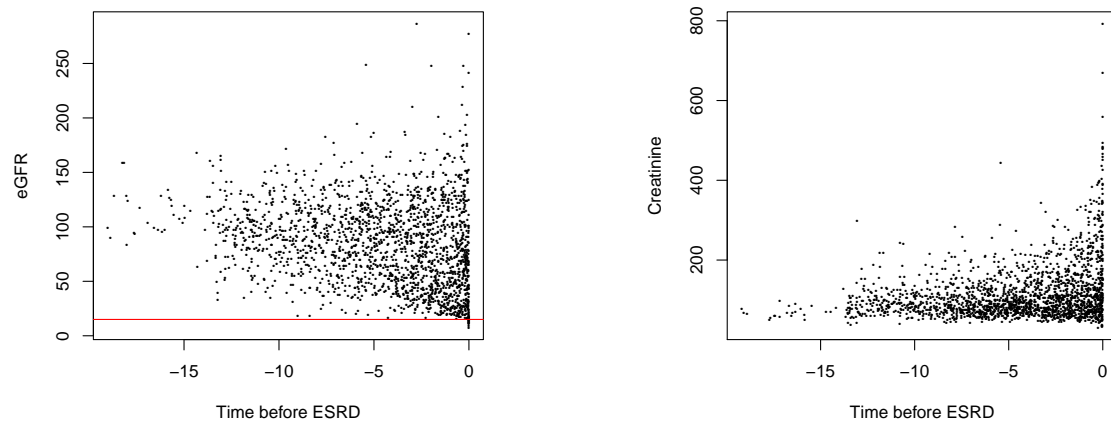


Figure 1.2: *Distribution of eGFR and creatinine with 2000 samples after removing $eGFR \geq 300$. The red line represents $eGFR = 15$.*

As those subjects with only one eGFR records contribut little to the trajectory, we further remove such patients from the cohort, and build the models on events and non-events separately. The summaries of the number of measurement for event and non-event subjects are shown in Figure 1.3.

```
> num_test <- table(sub_nomiss$Obs_id)
> id_keep <- names(which(num_test>1)) #removed those patients with only one measurement
> sub_nomiss <- subset(sub_nomiss, Obs_id %in% id_keep)
> sub_nomiss <- transform(sub_nomiss, log_eGFR=log(sub_nomiss$F_eGFR))
> event_dat <- subset(sub_nomiss, ESRD1_END==1) #event subjects
> dim(event_dat)

[1] 41412      14

> length(unique(event_dat$Obs_id))

[1] 1167
```

```
> (num_event <- table(table(event_dat$Obs_id)))
```

```

  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
51 21 23 18 17 19 23 18 21 21 22 17 21 23 20 18 18 23 14 16
22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
23 17 20 18 15 17 17 32 17 13 11 24 18 13 23 14 20 24 15 16
42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61
14 16 10 14 11 11 10 10 8 6 11 10 7 8 7 7 7 7 7 8
62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
10 11 4 8 10 5 5 8 4 2 4 2 3 3 4 8 2 3 3 1
82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 98 99 100 102 103
3 4 5 5 1 3 3 2 1 4 4 2 3 1 2 1 2 1 1 1
104 105 106 108 109 111 112 113 114 115 117 120 122 125 126 128 129 131 132 133
4 1 2 1 1 2 1 2 1 3 4 1 1 1 1 1 1 1 1 1
136 142 146 147 148 153 173
1 1 1 1 1 1 1
```

```
> nevent_dat <- subset(sub_nomiss, ESRD1_END==0) #non-event subjects
> dim(nevent_dat)
```

```
[1] 236907 14
```

```
> length(unique(nevent_dat$Obs_id))
```

```
[1] 7852
```

```
> (num_nevent <- table(table(nevent_dat$Obs_id)))
```

```

  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
77 70 81 119 175 214 205 181 213 172 193 183 175 186 210 203 219 193 219 200
22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
173 169 164 186 146 162 169 128 141 119 112 126 101 109 95 91 77 94 86 100
42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61
86 90 88 84 79 57 62 60 60 53 54 57 37 34 38 42 40 38 38 33
62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
42 27 20 22 24 22 22 19 19 22 22 18 19 13 16 19 13 8 8 19
82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 101 102
10 10 9 6 14 11 10 10 6 7 8 5 12 8 7 7 9 5 2 4
103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 123
5 4 2 3 2 10 8 2 2 6 5 1 1 2 2 2 6 4 1 4
124 126 127 128 129 130 131 132 134 137 139 142 143 144 145 147 148 151 155 156
2 4 2 3 1 1 2 1 3 2 1 3 1 1 1 3 1 2 1 1
157 158 159 161 163 166 167 168 176 183 185 187 190 194 199 203 207 220 221 288
1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1
295 621
1 1
```

```
> par(mar=c(5,4,1,2))
> plot(as.numeric(names(num_event)), num_event, type="h", lwd=3, xaxs="i", xlim=c(0,140),
+       xlab="No. of measurements", ylab="No. of subjects", main="ESRD subjects", yaxt="n")
> axis(side=2)
```

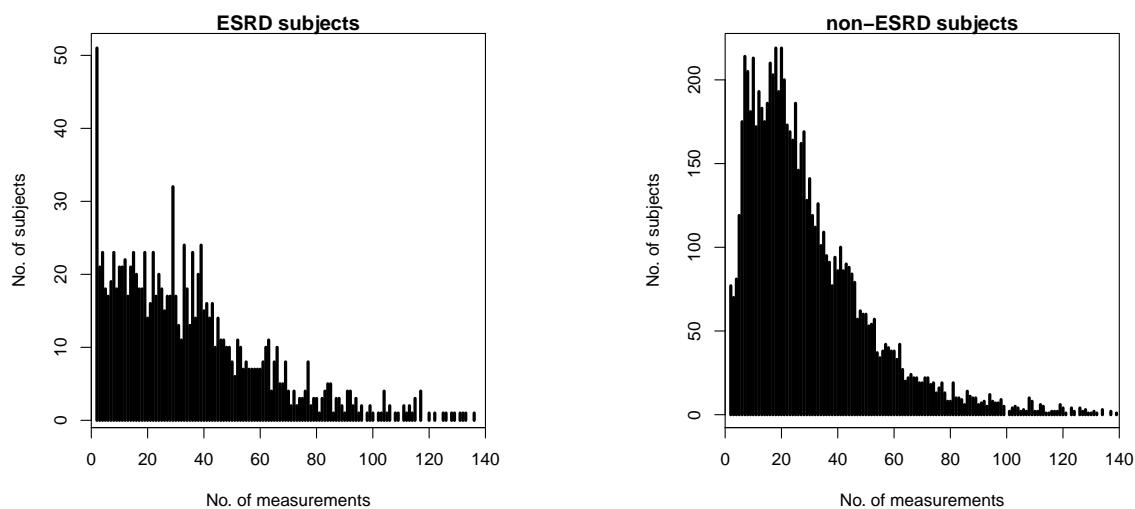


Figure 1.3: Number of eGFR records for event and non-event subjects.

```
> par(mar=c(5,4,1,2))
> plot(as.numeric(names(num_nevent)), num_nevent, type="h", lwd=3, xaxs="i", xlim=c(0,140),
+       xlab="No. of measurements", ylab="No. of subjects", main="non-ESRD subjects", yaxt="n")
> axis(side=2)
```

We check the patient with the largest number of eGFR measurements (Figure 1.4).

```
> (max_num <- max(as.numeric(names(num_nevent))))

[1] 621

> (id <- names(which(table(nevent_dat$Obs_id) == max_num)))

[1] "5673"

> subdat <- subset(nevent_dat, Obs_id %in% id)
> (range(subdat$dolab))

[1] 2002.641 2014.479

> with(subdat, plot(dolab, F_eGFR, xlab="Date of measurement", ylab="eGFR"))

> dest_eGFR <- density(event_dat$F_eGFR)
> plot(dest_eGFR, xlim=c(0,150), xlab="eGFR", lwd=3, yaxs="i",
+       ylab="Density", main="ESRD subjects", bty="n")
> abline(v=quantile(event_dat$F_eGFR, probs=c(50, 75, 90)/100),
+        col="red")

> dest_log_eGFR <- density(log(event_dat$F_eGFR))
> plot(dest_log_eGFR, lwd=3, xlab="Ln(eGFR)",
+       ylab="Density", main="ESRD subjects")
```

As shown in Figure 1.5, the original distribution of eGFR is skewed, whilst it's close to be normal after log-transformed.

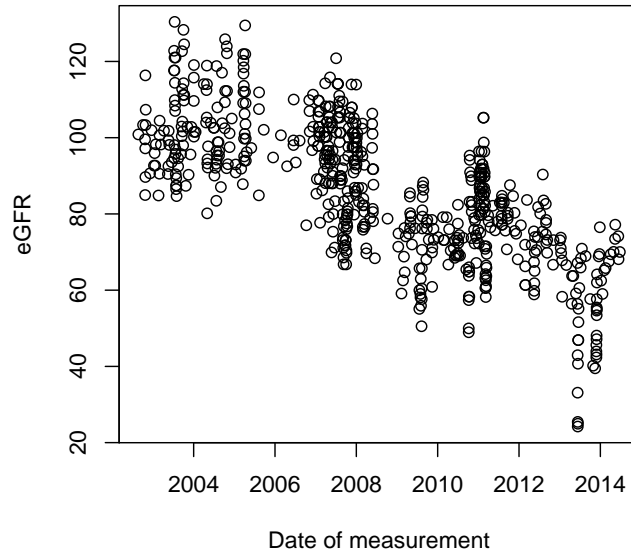


Figure 1.4: *The observed trajectory for the patient with the largest number of eGFR measurements.*

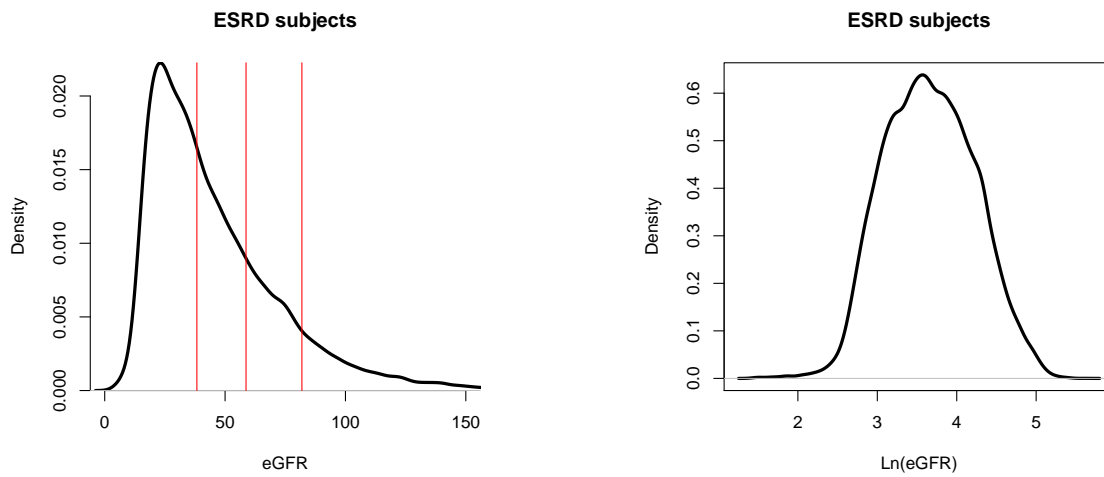


Figure 1.5: *Distribution of eGFR and $\log(eGFR)$ for ESRD subjects. Red lines represents the 50, 75 and 90 percent of subjects, respectively*

Chapter 2

Analysis

2.1 3 classes

We fit 3 different eGFR trajectories for the event data using the "hlme" function, and fit the non-event data using the ordinary linear mixed model.

```
> (kn <- seq(-12, 0, , 5))

[1] -12 -9 -6 -3 0

> MM <- Ns(event_dat$BW_TIME, knots=kn)
> dim(MM)

[1] 41412 4

> MM <- detrend(MM, event_dat$BW_TIME)
> dim(MM)

[1] 41412 3

> head(MM)

      1      2      3
[1,] 0.04689555 0.38709327 -0.10183513
[2,] -0.02867914 0.38150137 -0.07236472
[3,] -0.06988576 0.30174461 -0.04175845
[4,] -0.05881167 -0.13637325 0.03674734
[5,] -0.07504976 0.02845691 0.01176823
[6,] -0.34770860 -0.36093378 0.33529565

> (colnames(MM) <- paste("x", colnames(MM), sep=""))

[1] "x1" "x2" "x3"
```

```

> event_dat <- cbind(event_dat, MM)
> dim(event_dat)

[1] 41412    17

> #Fit the event model using the "hlme" function.
> event_model <- hlme(log_eGFR ~ BW_TIME + x1 + x2 + x3 + F_AGE + SEX + F_DMAGE,
+                     mixture = ~ BW_TIME + x1 + x2 + x3,
+                     random = ~ BW_TIME,
+                     subject = "Obs_id", ng=3, data=event_dat)

Be patient, hlme is running ...
The program took 2021.78 seconds

> event_model

Heterogenous linear mixed model
  fitted by maximum likelihood method

hlme(fixed = log_eGFR ~ BW_TIME + x1 + x2 + x3 + F_AGE + SEX +
      F_DMAGE, mixture = ~BW_TIME + x1 + x2 + x3, random = ~BW_TIME,
      subject = "Obs_id", ng = 3, data = event_dat)

Statistical Model:
  Dataset: event_dat
  Number of subjects: 1167
  Number of observations: 41412
  Number of latent classes: 3
  Number of parameters: 24

Iteration process:
  Convergence criteria satisfied
  Number of iterations: 16
  Convergence criteria: parameters= 6.5e-11
                      : likelihood= 2.6e-09
                      : second derivatives= 7.8e-16

Goodness-of-fit statistics:
  maximum log-likelihood: -4507.4
  AIC: 9062.8
  BIC: 9184.29

> postprob(event_model)

Posterior classification:
  class1 class2 class3
N 303.00 593.00 271.00
% 25.96 50.81 23.22

Posterior classification table:
--> mean of posterior probabilities in each class
  prob1 prob2 prob3
class1 0.8997 0.0996 0.0007
class2 0.0912 0.8463 0.0625

```

```

class3 0.0086 0.1246 0.8669

Posterior probabilities above a threshold (%):
      class1 class2 class3
prob>0.7  85.15  76.05  80.44
prob>0.8  78.55  65.60  70.85
prob>0.9  68.32  55.99  60.15

> str(event_model$pprob)

'data.frame':      1167 obs. of  5 variables:
 $ Obs_id: num  16 21 22 36 37 42 44 54 56 71 ...
 $ class : int  2 1 3 1 2 1 1 2 3 2 ...
 $ prob1 : num  5.33e-03 9.96e-01 2.51e-08 9.35e-01 1.71e-03 ...
 $ prob2 : num  0.942052 0.003643 0.000482 0.065047 0.998282 ...
 $ prob3 : num  5.26e-02 7.25e-31 1.00 5.41e-11 3.14e-06 ...

> ng <- event_model$ng

> #Fit the non-event model using the "lme" function.
> #nevent_model <- hlme(log_eGFR ~ BW_TIME + F_AGE + SEX + F_DMAGE,
> #                      random = ~ BW_TIME,
> #                      subject = "Obs_id", ng = 1, data = nevent_dat,
> #                      B = c(5.421, 0.0143, -0.0181, -0.0125, -0.0028,
> #                          0.1923, 0.0213, 0.0017, 0.1432))
> #nevent_model <- lme(F_eGFR ~ BW_TIME + F_AGE + SEX + F_DMAGE, random = ~BW_TIME | Obs_id,
> #                      data = nevent_dat)
> nevent_model <- lmer(log_eGFR ~ BW_TIME + F_AGE + SEX + F_DMAGE + (BW_TIME | Obs_id), nevent_dat)
> nevent_model

Linear mixed model fit by REML ['lmerMod']
Formula: log_eGFR ~ BW_TIME + F_AGE + SEX + F_DMAGE + (BW_TIME | Obs_id)
Data: nevent_dat
REML criterion at convergence: -75830.92
Random effects:
 Groups   Name                Std.Dev. Corr
Obs_id    (Intercept)  0.44283
          BW_TIME      0.03612  0.76
Residual                    0.18575
Number of obs: 236907, groups:  Obs_id, 7852
Fixed Effects:
(Intercept)      BW_TIME          F_AGE          SEX          F_DMAGE
  4.803249    -0.004661    -0.003337    -0.006691    -0.007128

```

Now we can further investigate the posterior probabilities of subjects in each class in event model (Figure 2.1 and 2.6).

```

> clr <- c("black", "red", "blue", "brown")
> par(mfrow=c(1,ng), mar=c(3,0,1,1), oma=c(0,3,0,0),
+     las=1, bty="n", mgp=c(3,1,0)/1.6)
> ppr <- event_model$pprob
> num <- table(ppr$class)
> for(i in 1:ng)
+ {
+   hist(ppr[ppr$class==i,i+2], breaks=0:50/50,

```

```

+       col=clrs[i], border=clrs[i], ylim=c(0,60),
+       main="", xlab="", yaxt="n", yaxs="i")
+   if(i==1) axis(side=2)
+   text(0.4,30,paste("Class",i," n=",num[i]),
+       font=2,col=clrs[i], cex=1)
+ }

```

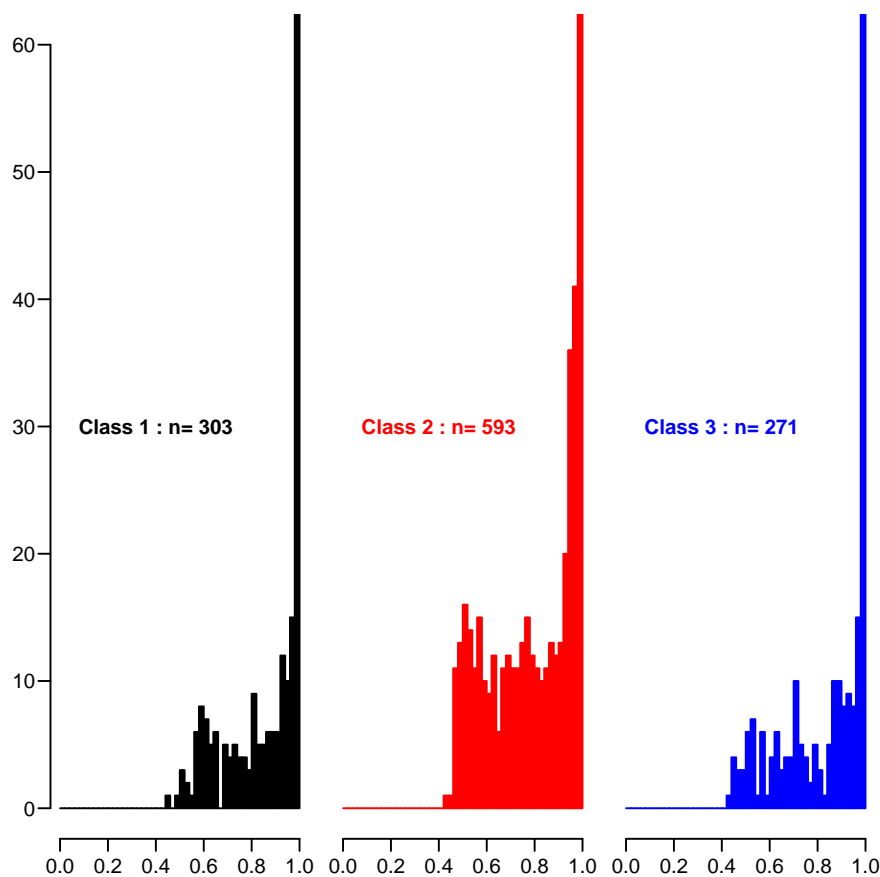


Figure 2.1: *Posterior probabilities of 3 classes for the event model.*

```

> pairs(ppr[, 2+1:ng], pch=16, col=clrs[ppr$class],
+       cex=0.4, gap=0)

```

We construct a data set for prediction to plot the estimated trajectories. Here the median age 65, median DM duration 15 and sex as male at time point 0 are used(Figure 2.3).

```

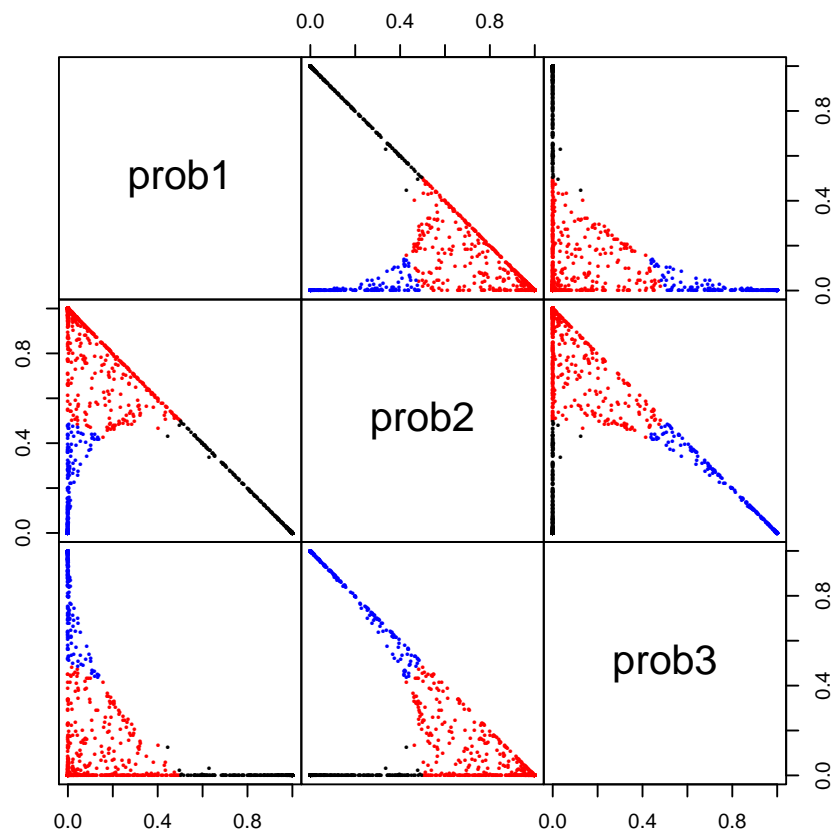
> (adjust_var <- event_model$Xnames[-c(1:5)])

```

```

[1] "F_AGE" "SEX" "F_DMAGE"

```

Figure 2.2: *Pairwise posterior probabilities from the event model.*

```
> length(table(event_dat$BW_TIME))

[1] 6964

> x <- sort(unique(event_dat$BW_TIME))
> length(x)

[1] 6964

> wh <- match(x, event_dat$BW_TIME)[1:69*100]
> length(wh)

[1] 69
```

```

> plotdata <- data.frame(1, event_dat$BW_TIME[wh],MM[wh,],
+                         F_AGE = 65 + event_dat$BW_TIME[wh],
+                         SEX = 1,
+                         F_DMAGE = 15 + event_dat$BW_TIME[wh])
> names(plotdata) <- event_model$Xnames
> head(plotdata)

  intercept  BW_TIME      x1      x2      x3  F_AGE SEX
1      1 -14.22040 -0.6919963 -0.001011389 -0.621725804 50.77960 1
2      1 -12.62149 -0.6178121 -0.138170718 -0.250384157 52.37851 1
3      1 -11.96715 -0.5874523 -0.194302593 -0.098414192 53.03285 1
4      1 -11.56468 -0.5682702 -0.228569177 -0.005717254 53.43532 1
5      1 -11.20602 -0.5490492 -0.258027568  0.073654734 53.79398 1
6      1 -10.87201 -0.5277820 -0.283754898  0.142452077 54.12799 1

  F_DMAGE
1 0.779603
2 2.378508
3 3.032854
4 3.435318
5 3.793977
6 4.127995

> range(plotdata$BW_TIME)

[1] -14.22039699 -0.08761123

> range(plotdata$F_DMAGE)

[1] 0.779603 14.912389

> x <- seq(-14, 0, by=0.5)
> nevent_plotdata <- data.frame(Obs_id=1, BW_TIME=x,
+                               F_AGE=65+x, SEX=1, F_DMAGE=15+x)
> head(nevent_plotdata)

  Obs_id BW_TIME F_AGE SEX F_DMAGE
1      1  -14.0  51.0  1    1.0
2      1  -13.5  51.5  1    1.5
3      1  -13.0  52.0  1    2.0
4      1  -12.5  52.5  1    2.5
5      1  -12.0  53.0  1    3.0
6      1  -11.5  53.5  1    3.5

> pred_event <- exp(predictY(event_model, plotdata, var.time="BW_TIME", draws=TRUE)$pred)
> pred_nevent <- exp(predict(nevent_model, nevent_plotdata, re.form=NA))
> boot_pred <- bootMer(nevent_model, FUN=function(x) exp(predict(x, nevent_plotdata, re.form=NA)),
+                     nsim=1000)
> lci_nevent <- apply(boot_pred$t, 2, quantile, 0.025) #95% CI for the prediction of non-event model
> uci_nevent <- apply(boot_pred$t, 2, quantile, 0.975)
> ylim <- range(pred_event)
> lwd_main <- 4
> lwd_ci <- 1
> for (i in 1:ng)

```

```

+ {
+   plot(y = pred_event[, i], x = plotdata$BW_TIME, type = "l", col = clr[i],
+        ylim = ylim, lwd = lwd_main, xlab = "Time before events(years)", ylab = "eGFR")
+   points(y = pred_event[, i + ng], x = plotdata$BW_TIME, type = "l", lty = "dashed",
+          col = clr[i], lwd = lwd_ci)
+   points(y = pred_event[, i + 2*ng], x = plotdata$BW_TIME, type = "l", lty = "dashed",
+          col = clr[i], lwd = lwd_ci)
+   par(new = TRUE)
+ }
> #curve for non-event subject
> points(y = pred_nevent, x = nevent_plotdata$BW_TIME, type = "l", col = clr[ng+1], lwd = lwd_main)
> points(y = lci_nevent, x = nevent_plotdata$BW_TIME, type = "l", lty = "dashed", col = clr[ng+1], lwd = lwd_ci)
> points(y = uci_nevent, x = nevent_plotdata$BW_TIME, type = "l", lty = "dashed", col = clr[ng+1], lwd = lwd_ci)
> abline(h = 15, lty = "dashed")
> legend("topright", legend = c(paste("class", 1:ng), "non-ESRD"), col=clr, lty=1, lwd=lwd_main, cex=0.8)
> par(new = FALSE)

```

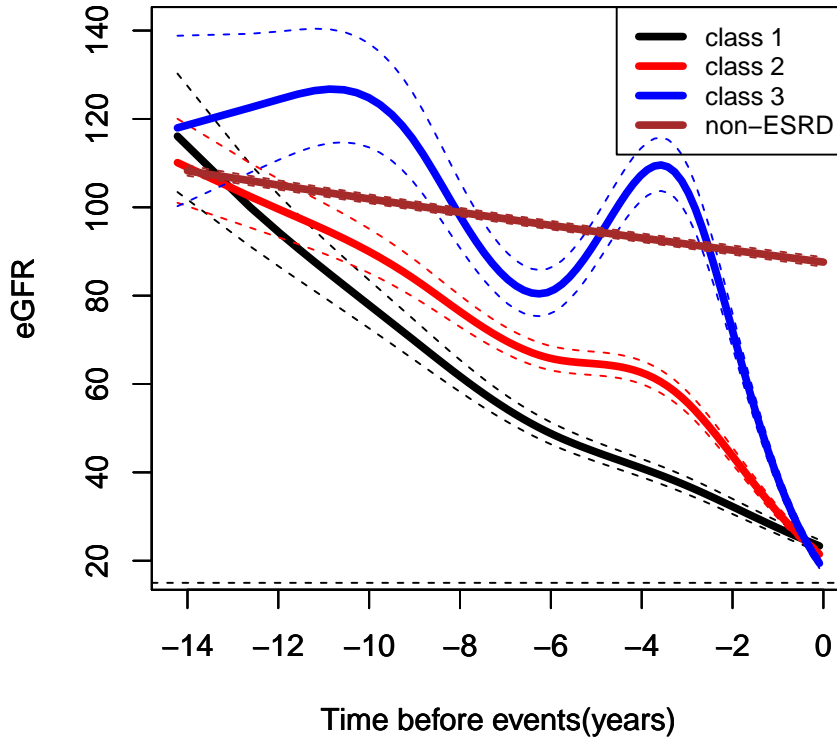


Figure 2.3: Mean trajectories of eGFR with three latent classes of subjects developing ESRD. The horizontal dashed line represents eGFR is equal to 15

We further plot the observed trajectories of 3 subjects with top posterior probability of each classes for event data (Figure 2.4).

```
> obj_num <- 3;                                     #selected number of subjects for each class
> par(mfrow=c(ng, obj_num), mar=c(0,0,0,0), oma=c(5,5,3,3),
+     bty="o", las=1)
> for (i in 1:ng)
+ {
+   sort_ppr <- ppr[order(ppr[i+2], decreasing=T), ];
+   top_ppr <- round(sort_ppr[1:obj_num, i+2], 2)      #Top ppr of class i
+   top_id <- sort_ppr[1:obj_num, 1]                  #Top ids of class i
+   lapply(1:obj_num, FUN=function(x) {
+     subdat <- subset(event_dat, Obs_id %in% top_id[x])
+     with(subdat, plot(BW_TIME, F_eGFR, xlim=c(-14,0), ylim=c(0,170),
+                      yaxt="n", xaxt="n"))
+     if(i==ng && (x==1 || x==obj_num)) axis(side=1, labels=T)
+     if(i==2 && x==1) axis(side=2)
+     if(i==1 && x==2) axis(side=3)
+     if((i==1 || i==3) && x==obj_num) axis(side=4)
+     leg_pos <- "bottomleft";
+     if(i==1) leg_pos <- "topleft";
+     legend(leg_pos, legend = c(paste("class:", i), paste("id:", top_id[x]),
+                                paste("ppr:", top_ppr[x])), cex=0.8)
+   })
+ }
> mtext("Time before ESRD (years)", side=1, outer=T, line=3);
> mtext("eGFR", side=2, outer=T, line=3, las=3);
```

2.2 4 classes

We then try to fit 4 trajectories for the event data.

```
> model_4C <- hlme(log_eGFR ~ BW_TIME + x1 + x2 + x3 + F_AGE + SEX + F_DMAGE,
+                  mixture = ~ BW_TIME + x1 + x2 + x3,
+                  random = ~ BW_TIME,
+                  subject = "Obs_id", ng=4, data=event_dat)
```

```
Be patient, hlme is running ...
The program took 5261.96 seconds
```

```
> model_4C
```

```
Heterogenous linear mixed model
fitted by maximum likelihood method
```

```
hlme(fixed = log_eGFR ~ BW_TIME + x1 + x2 + x3 + F_AGE + SEX +
      F_DMAGE, mixture = ~BW_TIME + x1 + x2 + x3, random = ~BW_TIME,
      subject = "Obs_id", ng = 4, data = event_dat)
```

```
Statistical Model:
Dataset: event_dat
Number of subjects: 1167
Number of observations: 41412
Number of latent classes: 4
Number of parameters: 30
```

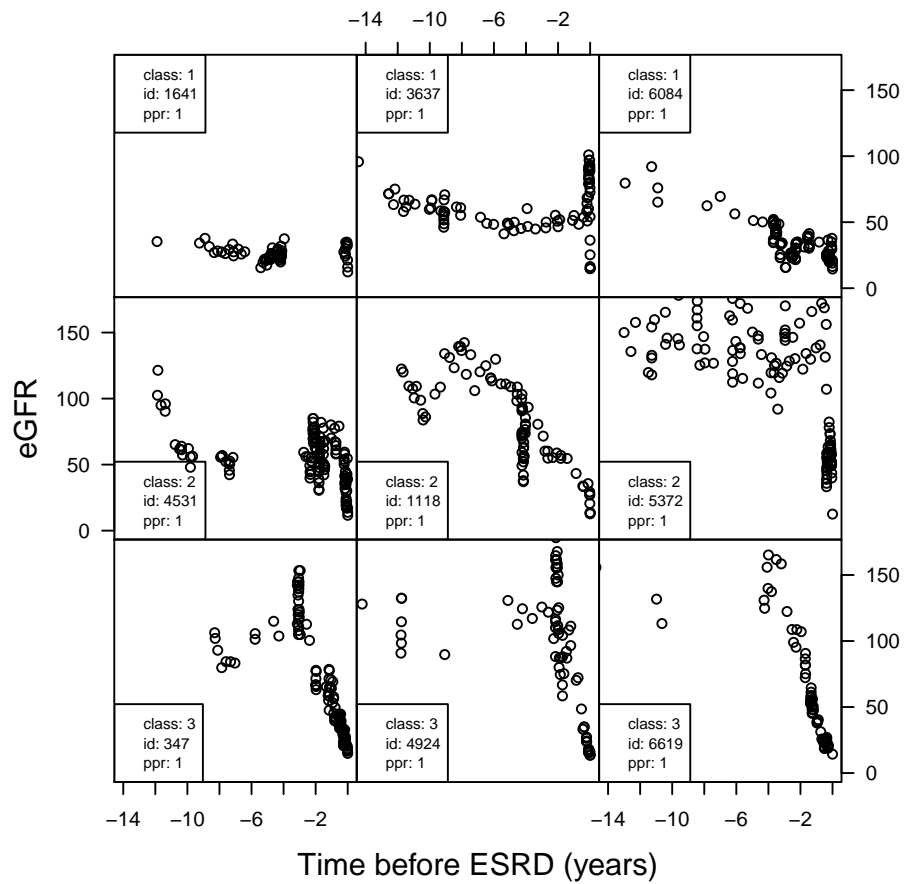


Figure 2.4: Observed eGFR trajectories 3 subjects with top posterior probability(ppr) of each classes.

```
Iteration process:
  Convergence criteria satisfied
  Number of iterations: 29
  Convergence criteria: parameters= 3.4e-08
                        : likelihood= 1.6e-06
                        : second derivatives= 4.1e-13
```

```
Goodness-of-fit statistics:
  maximum log-likelihood: -4265.84
  AIC: 8591.69
  BIC: 8743.55
```

```
> postprob(model_4C)
```

```
Posterior classification:
  class1 class2 class3 class4
N 273.00 304.00 436.00 154.0
```

```

% 23.39 26.05 37.36 13.2

Posterior classification table:
--> mean of posterior probabilities in each class
      prob1 prob2 prob3 prob4
class1 0.8572 0.1229 0.0195 0.0004
class2 0.0972 0.7891 0.1110 0.0027
class3 0.0420 0.1330 0.7514 0.0736
class4 0.0044 0.0194 0.1151 0.8612

Posterior probabilities above a threshold (%):
      class1 class2 class3 class4
prob>0.7  77.29  65.79  58.03  79.22
prob>0.8  67.77  52.30  45.18  71.43
prob>0.9  59.71  40.46  35.55  59.74

> str(model_4C$pprob)

'data.frame':      1167 obs. of  6 variables:
 $ Obs_id: num  16 21 22 36 37 42 44 54 56 71 ...
 $ class : int   3 1 4 2 2 1 1 3 3 3 ...
 $ prob1 : num  3.12e-03 7.50e-01 4.97e-09 4.06e-01 3.54e-05 ...
 $ prob2 : num  2.39e-01 2.50e-01 1.81e-06 5.94e-01 9.15e-01 ...
 $ prob3 : num  7.55e-01 8.15e-10 4.89e-03 2.42e-04 8.47e-02 ...
 $ prob4 : num  2.58e-03 5.46e-45 9.95e-01 1.04e-15 1.78e-11 ...

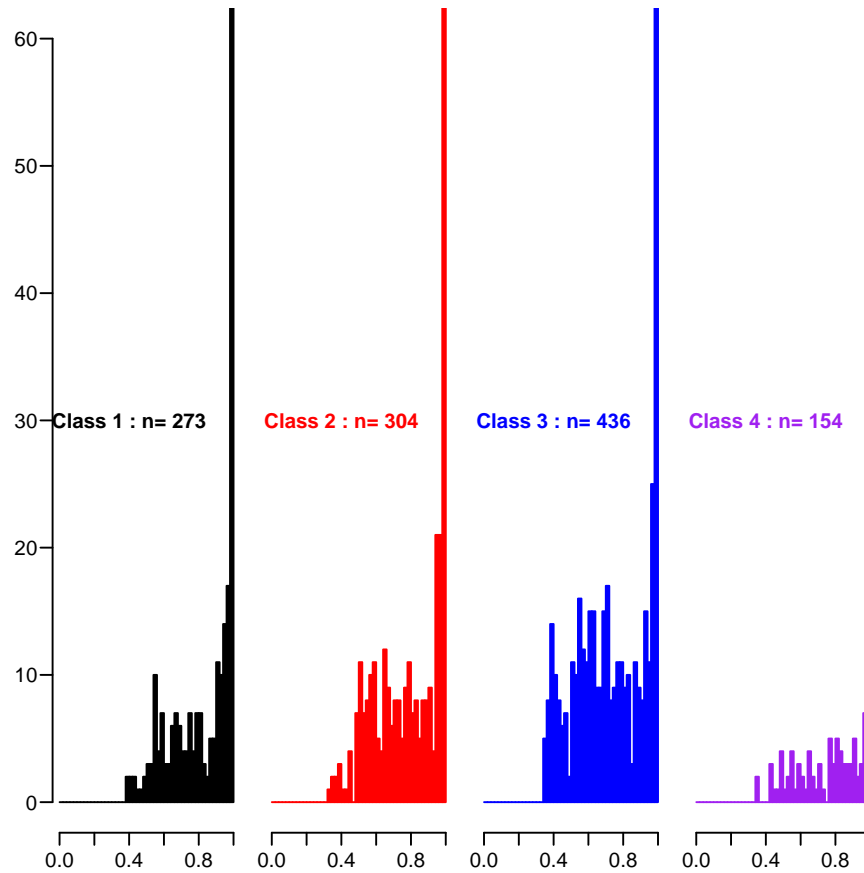
> ng <- model_4C$ng

> clr = c("black", "red", "blue", "purple", "brown")
> par(mfrow=c(1,ng), mar=c(3,0,1,1), oma=c(0,3,0,0),
+     las=1, bty="n", mgp=c(3,1,0)/1.6)
> ppr_4C <- model_4C$pprob
> num <- table(ppr_4C$class)
> for(i in 1:ng)
+ {
+   hist(ppr_4C[ppr_4C$class==i,i+2], breaks=0:50/50,
+        col=clr[i], border=clr[i], ylim=c(0,60),
+        main="", xlab="", yaxt="n", yaxis="i")
+   if(i==1) axis(side=2)
+   text(0.4,30,paste("Class",i,": n=",num[i]),
+        font=2,col=clr[i], cex=1)
+ }

> pairs(ppr_4C[, 2+1:ng], pch=16, col=clr[ppr_4C$class],
+       cex=0.4, gap=0)

> pred_4C <- exp(predictY(model_4C, plotdata, var.time="BW_TIME", draws=TRUE)$pred)
> ylim <- range(pred_4C)
> lwd_main <- 4
> lwd_ci <- 1
> for (i in 1:ng)
+ {
+   plot(y = pred_4C[, i], x = plotdata$BW_TIME, type = "l", col = clr[i],
+        ylim = ylim, lwd = lwd_main, xlab = "Time before events(years)", ylab = "eGFR")
+   points(y = pred_4C[, i + ng], x = plotdata$BW_TIME, type = "l", lty = "dashed",

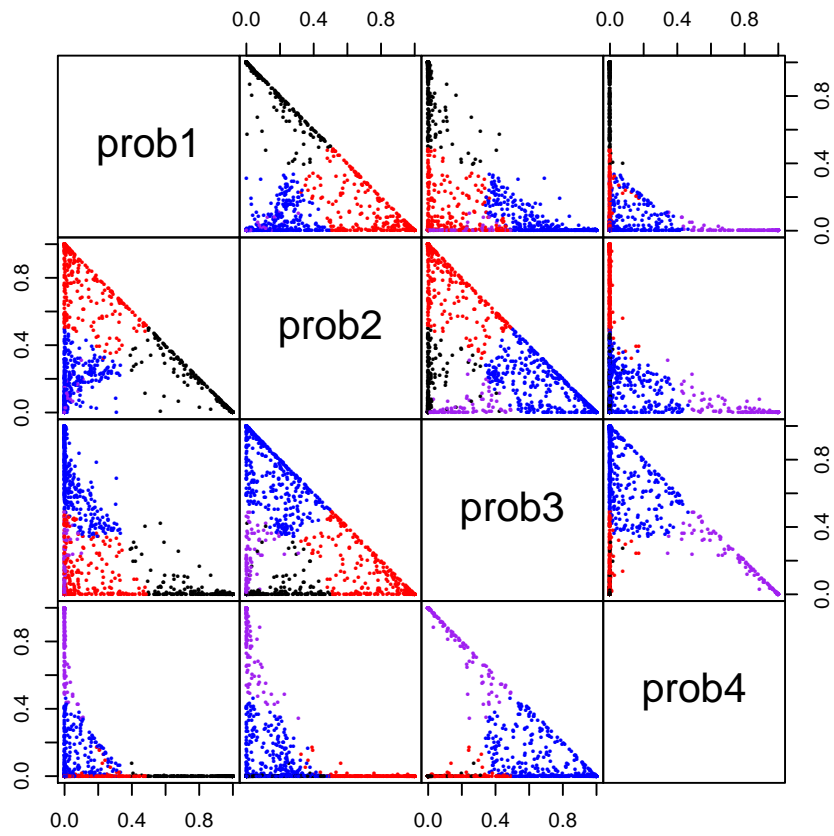
```

Figure 2.5: *Posterior probabilities of 4 classes for the event model.*

```

+         col = clrs[i], lwd = lwd_ci)
+   points(y = pred_4C[, i + 2*ng], x = plotdata$BW_TIME, type = "l", lty = "dashed",
+         col = clrs[i], lwd = lwd_ci)
+   par(new = TRUE)
+ }
> #curve for non-event subject
> points(y = pred_nevent, x = nevent_plotdata$BW_TIME, type = "l", col = clrs[ng+1], lwd = lwd_main)
> points(y = lci_nevent, x = nevent_plotdata$BW_TIME, type = "l", lty = "dashed", col = clrs[ng+1], lwd = lwd_ci)
> points(y = uci_nevent, x = nevent_plotdata$BW_TIME, type = "l", lty = "dashed", col = clrs[ng+1], lwd = lwd_ci)
> abline(h = 15, lty = "dashed")
> legend("topright", legend = c(paste("class", 1:ng), "non-ESRD"), col=clrs, lty=1, lwd=lwd_main, cex=0.8)
> par(new = FALSE)

```

Figure 2.6: *Pairwise posterior probabilities from the event model.*

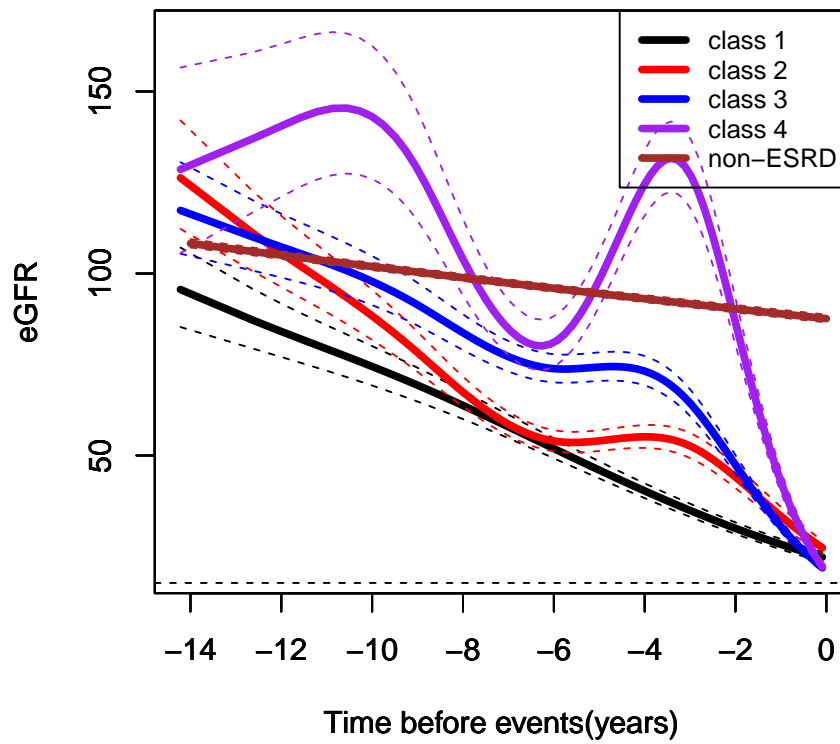


Figure 2.7: Mean trajectories of eGFR with four latent classes of subjects developing ESRD. The horizontal dashed line represents eGFR is equal to 15