5-4

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(Rcpp)
```

1 Data Pre-processing

```
Ratings <- read.csv("archive/Ratings.csv")
Ratings <- Ratings[Ratings$Book.Rating != 0, ]
Ratings_by_users <- group_by(Ratings, User.ID)
user.Rating <- Ratings_by_users %>% summarise(
  num = length(Book.Rating),
  avg.rating = mean(Book.Rating)
)
num.Rating <- group_by(user.Rating, num) %>% summarise(
  users = length(avg.rating)
)
```

1.1 ISBN

```
N <- max(Ratings$User.ID)
book_idx <- unique(Ratings$ISBN)
M <- length(book_idx)
book_idx <- 1:M
names(book_idx) <- unique(Ratings$ISBN)
train <- read.csv("archive/train0.csv")
train$ISBN <- book_idx[train$ISBN]
test <- read.csv("archive/test0.csv")
test$ISBN <- book_idx[test$ISBN]
```

1.2 Romve Some Entries

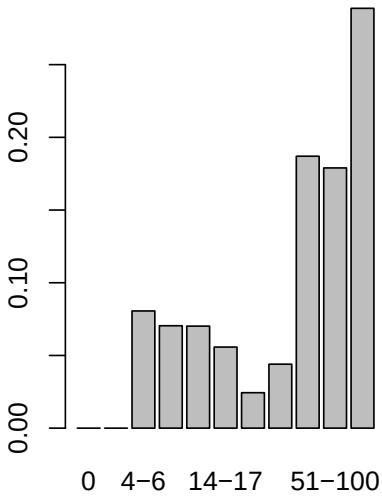
```
train.book.avg <- group_by(train, ISBN) %>% summarise(
  avg.rating = mean(Book.Rating)
)
test <- test[which(test$ISBN %in% train.book.avg$ISBN),]
```

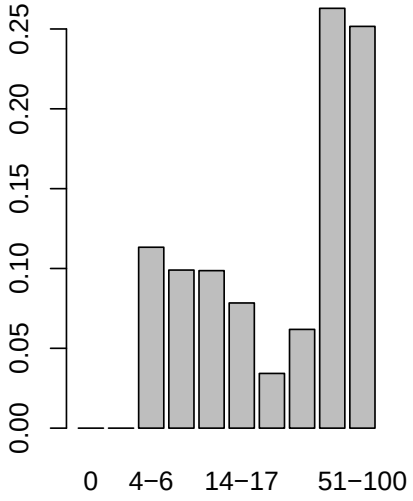
```
## There are 278854 users and 185973 books.
```

```
## There are 334954 examples in the training set and 64857 examples in the test set.
```

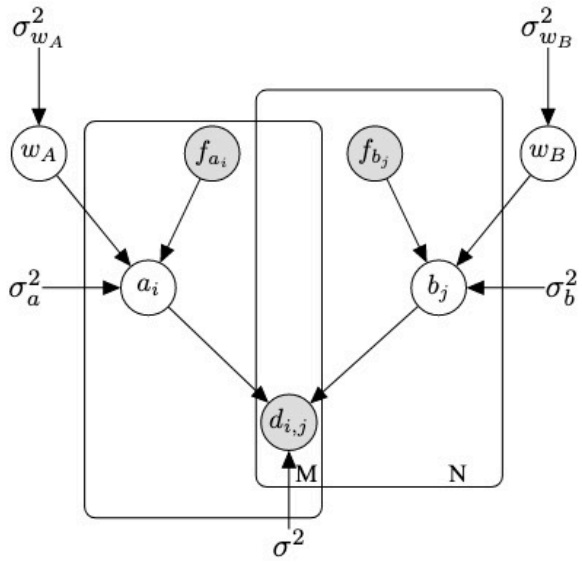
1.3 User Groups

```
par(mfrow = c(1, 2))
barplot(grp_num/sum(grp_num))
barplot(grp_num[-11]/sum(grp_num[-11]))
```


0.00
0.10
0.20
0
4−6
14−17
51−100


0.00
0.05
0.10
0.15
0.20
0.25
0
4−6
14−17
51−100

2 warm-start LPMF



$$W_A = [W_{A_1}, ...W_{A_k}], W_B = [W_{B_1}, ...W_{B_k}]$$

$$F_A = [F_{A_1}, ...F_{A_M}], F_B = [F_{B_1}, ...F_{B_N}]$$

$$p(W_A|\sigma^2_{W_A}) = \prod_{l=1}^{k} \mathcal{N}((w_A)_l|0, \sigma^2_{W_A}\mathbf{I}),\ p(W_B|\sigma^2_{W_B}) = \prod_{l=1}^{k} \mathcal{N}((w_B)_l|0, \sigma^2_{W_B}\mathbf{I})$$

$$p(A|F_A, W_A\sigma^2_a) = \prod_{i=1}^{M} \mathcal{N}(a_i|W_A^T f_{a_i}, \sigma^2_a\mathbf{I}),\ p(B|F_B, W_B\sigma^2_b) = \prod_{j=1}^{N} \mathcal{N}(b_j|W_B^T f_{b_j}, \sigma^2_b\mathbf{I})$$

$$p(D|A, B, \sigma^2) = \prod_{j=1}^{N}\prod_{i=1}^{M} \left[\mathcal{N}(d_{ij}|a_i^T b_j, \sigma^2)\right]^{I_{ij}}$$

$$
\begin{aligned}
log\, p(A,B|D,\sigma^2,\sigma_a^2,\sigma_b^2) = & -\frac{1}{2\sigma^2}\|P_\Omega(D-A^TB)\|_F^2 - \frac{1}{2\sigma_a^2}\|A-W_A^TF_A\|_F^2 \\
& - \frac{1}{2\sigma_b^2}\|B-W_B^TF_B\|_F^2 - \frac{1}{2\sigma_{W_A}^2}\|W_A\|_F^2 - \frac{1}{2\sigma_{W_B}^2}\|W_B^T\|_F^2 \\
& - \frac{1}{2}\left(\left(\sum_{i=1}^{M}\sum_{j=1}^{N} I_{ij}\right) log\,\sigma^2 + M\,k\,log\,\sigma_a^2 + N\,k\,log\,\sigma_b^2\right) \\
& - \frac{1}{2}k\,L\left(log\,\sigma_{W_A}^2 + log\,\sigma_{W_B}^2\right) + C
\end{aligned}
$$

$$
E = \frac{1}{2}\|P_\Omega(D-A^TB)\|_F^2 + \frac{1}{2}(\lambda_a\|A-W_A^TF_A\|_F^2 + \lambda_b\|B-W_B^TF_B\|_F^2 + \lambda_{W_A}\|W_A\|_F^2 + \lambda_{W_B}\|W_B\|_F^2)
$$

此处，$\lambda_a = \frac{\sigma^2}{\sigma_a^2}$，$\lambda_b = \frac{\sigma^2}{\sigma_b^2}, \lambda_{W_A} = \frac{\sigma^2}{\sigma_{W_A}^2}, \lambda_{W_B} = \frac{\sigma^2}{\sigma_{W_B}^2}$。

3 Users' Information

```
Users <- read.csv("archive/Users.csv")
Users <- Users[1:N, ]
summary(Users)
```

```
##     User.ID         Location              Age        
##  Min.   :     1   Length:278854      Min.   :  0.00  
##  1st Qu.: 69714   Class :character   1st Qu.: 24.00  
##  Median :139428   Mode  :character   Median : 32.00  
##  Mean   :139428                      Mean   : 34.75  
##  3rd Qu.:209141                      3rd Qu.: 44.00  
##  Max.   :278854                      Max.   :244.00  
##                                      NA's   :110759
```

```
sum(is.na(Users$Age))/length(Users$Age)
```

```
## [1] 0.3971935
```

```
length(unique(Users$Location))
```

```
## [1] 57338
```

3.1 pre-processing

```
Users[is.na(Users$Age), "Age"] <- 0
Users[which(Users$Age > 100), "Age"] <- 0
Users[which(Users$Age < 7), "Age"] <- 0
```

3.2 Check intuition

```
user.Rating <- merge(user.Rating, Users)
Rating_by_age <- group_by(user.Rating, Age) %>% summarise(
  user.num = length(num),
  avg.rating = sum(num*avg.rating)/sum(num)
)
head(Rating_by_age, 10)
```

```
## # A tibble: 10 x 3
##      Age user.num avg.rating
##    <dbl>    <int>      <dbl>
##  1     0    31455       7.35
##  2     7       10       7.84
##  3     8       25       7.81
##  4     9       28       7.08
##  5    10       33       8.49
##  6    11       53       7.76
##  7    12       62       8.07
##  8    13      241       7.69
##  9    14      504       7.78
## 10    15      568       7.93
```

3.3 Group by Age

```
FU <- matrix(0, nrow = N, ncol = 2)
FU[, 1] <- 1
FU[, 2] <- Users$Age
FU[which(Users$Age %in% 81:90), 2] <- 81
FU[which(Users$Age %in% 91:100), 2] <- 82
FU[which(Users$Age == 0), 2] <- 6
FU[, 2] <- FU[, 2] - 6
FU <- t(FU)
save(FU, file = "FU.Rda")
```

4 Books' Information

```
Books <- read.csv("archive/Books.csv")
Books <- Books[Books$ISBN %in% names(book_idx), ]
Books$Year.Of.Publication <- as.integer(Books$Year.Of.Publication)
```

```
## Warning: NAs introduced by coercion
```

```
Books$Year.Of.Publication[which(Books$Year.Of.Publication > 2004)] <- 0
Books$Year.Of.Publication[is.na(Books$Year.Of.Publication)] <- 0
summary(Books)
```

```
##      ISBN            Book.Title        Book.Author        Year.Of.Publication
##  Length:149836      Length:149836      Length:149836      Min.   :   0
##  Class :character   Class :character   Class :character   1st Qu.:1990
##  Mode  :character   Mode  :character   Mode  :character   Median :1996
##                                                           Mean   :1958
##                                                           3rd Qu.:2000
##                                                           Max.   :2004
##   Publisher          Image.URL.S        Image.URL.M        Image.URL.L
##  Length:149836      Length:149836      Length:149836      Length:149836
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
length(unique(Books$Book.Title))
```

```
## [1] 135567
```

```
length(unique(Books$Book.Author))
```

```
## [1] 62114
```

```
length(unique(Books$Publisher))
```

```
## [1] 11576
```

```
length(unique(Books$Year.Of.Publication))
```

```
## [1] 94
```

4.1 check intuition

```
book_rating <- group_by(Ratings, ISBN)%>% summarise(
  num = length(Book.Rating),
  avg.rating = mean(Book.Rating)
)
book_rating <- merge(book_rating, Books)
Rating_by_year <- group_by(book_rating, Year.Of.Publication) %>% summarise(
  user.num = length(num),
  avg.rating = sum(num*avg.rating)/sum(num)
)
Rating_by_year[50:60, ]
```

```
## # A tibble: 11 x 3
```

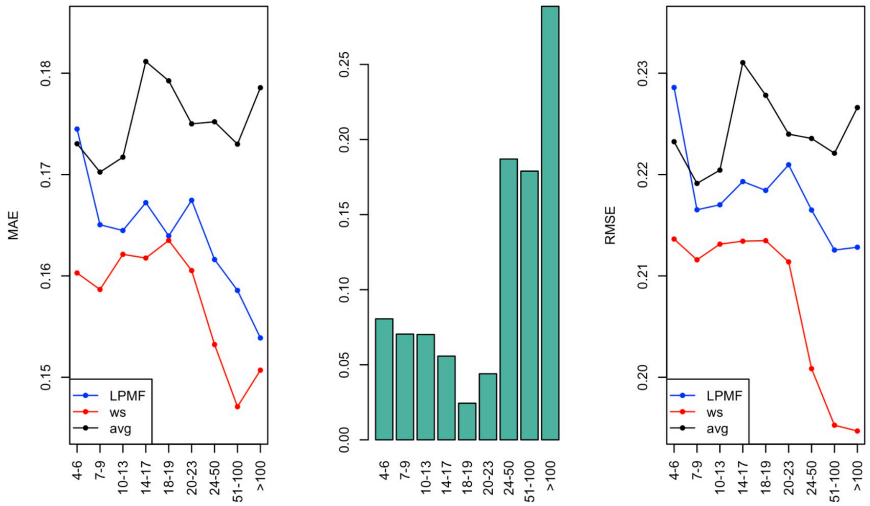
```
##    Year.Of.Publication user.num avg.rating
##                  <dbl>    <int>      <dbl>
##  1                1960       65       8.33
##  2                1961       76       8.24
##  3                1962       73       7.86
##  4                1963       70       8.21
##  5                1964       72       7.87
##  6                1965       94       7.76
##  7                1966       92       7.65
##  8                1967       89       7.85
##  9                1968      124       8.22
## 10                1969      178       7.51
## 11                1970      221       7.98
```

4.2 Group by Year

```
FV <- matrix(0, ncol = 2, nrow = M)
FV[, 1] <- 1
FV[book_idx[Books$ISBN], 2] <-  Books$Year.Of.Publication
FV[which(FV[, 2] < 1920 & FV[, 2] != 0), 2] <- 1
FV[which(FV[, 2] %in% 1920:1929), 2] <- 2
FV[which(FV[, 2] %in% 1930:1939), 2] <- 3
FV[which(FV[, 2] %in% 1940:1949), 2] <- 4
FV[which(FV[, 2] > 1949), 2] <- FV[which(FV[, 2] > 1949), 2] - 1945
FV <- t(FV)
save(FV, file = "FV.Rda")
```

5 Results

```
knitr::include_graphics("plot/warm start.png")
```


MAE
0.15
0.16
0.17
0.18
LPMF
ws
avg
4-6
7-9
10-13
14-17
18-19
20-23
24-50
51-100
>100
0.00
0.05
0.10
0.15
0.20
0.25
RMSE
0.20
0.21
0.22
0.23