

# A Brief Tutorial of the R Package **bigRR** for analytical chemistry

## Setup

To use the **bigRR** package, an R environment is required. Visit:

<http://www.r-project.org>

and install R for the operating system.

Start R and in the R console, type the following command to install the package:

```
install.packages('bigRR')
```

Something like the following should show:

```
trying URL ...
Content type ... length ... bytes (... Kb)
opened URL
=====
downloaded ... Kb
```

Now the package is installed in the R library. By default, the “install.packages” command will install the package from CRAN - the comprehensive R-archive network. The latest developer version of **bigRR** can be obtained from R-Forge, installed by:

```
install.packages('bigRR', repos = 'http://r-forge.r-project.org')
```

## Example

Type:

```
require(bigRR)
```

or:

```
library(bigRR)
```

to load the package (the depended package **hglm** and **DatABEL** are required to be installed as well), then two main functions in the **bigRR** package are ready to use - “bigRR” and “bigRR\_update”, where the former is an efficient tool for fitting high-dimensional ridge regression based on linear mixed models, and the latter is an “updater” for the former function call to produce the generalized ridge regression - HEM (heteroscedastic effects model). Note that when the dataset is extremely large that cannot even be loaded into memory, “hugeRR” and “hugeRR\_update” are the powerful alternative functions to use.

This document can be loaded by:

```
vignette('bigRR')
```

Run the following commands to load the example data:

```
data(chemometrics)
```

There are now two objects in the current working space:

```
ls()
```

```
[1] "analytes" "FTIR"
```

"analytes" is an 8-by-48 data frame of analytes' measurements from two different studies, each with 4 observations, measured via high-performance liquid chromatography (HPLC), and "FTIR" is a matrix containing 1055 wavelength absorbance signals of an FTIR (Fourier transform infrared spectroscopy) spectrum for all the 8 observations. We can check the first 5 observations of succinic acid and ethanol concentration and the first 5 wavelengths in "FTIR" as:

```
analytes[,c(1, 32, 34)]
```

	IR_Name	Succi	EtOH
1	EkoF1	1.900	2.600
2	EkoF2	1.800	2.800
3	EkoF3	1.800	2.800
4	EkoF4	2.000	3.100
5	D1	0.745	0.275
6	D2	0.878	0.497
7	D4	0.843	0.852
8	D7	0.834	0.515

```
FTIR[1:5,]
```

	Pin.number	Wavelength	Absorbance_EkoF1	Absorbance_EkoF2	Absorbance_EkoF3
1	PIN241	10.7553	0.0062	-0.0021	-0.0012
2	PIN242	10.7108	0.0020	0.0001	0.0015
3	PIN243	10.6667	-0.0010	-0.0077	-0.0009
4	PIN244	10.6230	-0.0039	-0.0111	-0.0091
5	PIN245	10.5797	-0.0088	-0.0113	-0.0169

	Absorbance_EkoF4	Absorbance_D1	Absorbance_D2	Absorbance_D4	Absorbance_D7
1	-0.0082	-0.0133	0.0035	0.0026	-0.0061
2	-0.0032	-0.0054	-0.0059	-0.0022	-0.0064
3	-0.0033	0.0004	-0.0111	-0.0049	-0.0109
4	-0.0092	-0.0015	-0.0104	-0.0054	-0.0072
5	-0.0169	-0.0018	-0.0073	-0.0040	-0.0042

where the row and column names of “FTIR” are simply the wavenumbers and the ID’s of the observations. Let’s take the ethanol content for example, as the first step, let’s fit a big ridge regression to model the data using the 4 observations from the first experiment:

```
RR.model <- bigRR(y = analytes[1:4,34], X = matrix(1, 4, 1), Z = t(FTIR[,3:6]))
```

The definition of a column of one’s in “X” is the design matrix for obtaining an intercept estimate in a linear mixed model. The ridge regression model is now fitted. In order to fit a HEM (heteroscedastic effects model) to the example data, just the simple updating function in the **bigRR** package is required:

```
HEM.model <- bigRR_update(RR.model, Z = t(FTIR[,3:6]))
```

The saved objects for the two steps are similar in structure:

```
summary(RR.model)
```

	Length	Class	Mode
phi	1	-none-	numeric
lambda	1	-none-	numeric
beta	1	-none-	numeric
hglm	24	hglm	list
u	1055	-none-	numeric
leverage	1055	-none-	numeric
GCV	1	-none-	numeric
Call	4	-none-	call
y	4	-none-	numeric
X	4	-none-	numeric

```
summary(HEM.model)
```

	Length	Class	Mode
phi	1	-none-	numeric
lambda	1	-none-	numeric
beta	1	-none-	numeric
hglm	24	hglm	list
u	1055	-none-	numeric
leverage	1055	-none-	numeric
GCV	1	-none-	numeric
Call	9	-none-	call
y	4	-none-	numeric
X	4	-none-	numeric

The output contains a sub-object “u” which are the estimated wavelength effects, which are the key values that we would use in predictions, and “beta” is the estimated intercept (or overall mean). If there is a new FTIR spectrum for a different sets of observations, loaded in a new matrix named “FTIR.new”, with observations as the rows and

wavelengths as columns, one can predict the ethanol content in the new sample using HEM as:

```
HEM.model$beta + FTIR.new %*% HEM.model$u
```

Details about the other values in the fitted objects can be seen in the R documentation for the package:

```
help("bigRR")
```

Let us compare the wavelength effects from the fitted ridge regression and HEM.

```
split.screen(c(1, 2))
split.screen(c(2, 1), screen = 1)
screen(3)
plot(1:length(HEM.model$u), HEM.model$u, type = 'l', col = 'red3', main = 'HEM',
     xlab = 'Wavelength index', ylab = 'Estimated effect size')
screen(4)
plot(1:length(RR.model$u), RR.model$u, type = 'l', col = 'blue3', main = 'RR',
     xlab = 'Wavelength index', ylab = 'Estimated effect size')
screen(2)
plot(RR.model$u, HEM.model$u, cex = .6, col = 'darkmagenta',
     xlab = 'Estimated effect size via RR', ylab = 'Estimated effect size via HEM')
```

The resulted figure is shown in **Figure T1**. We can see that HEM shows much stronger “shrinkage” effects for the minor wavelengths and highlights the major wavelength effects.

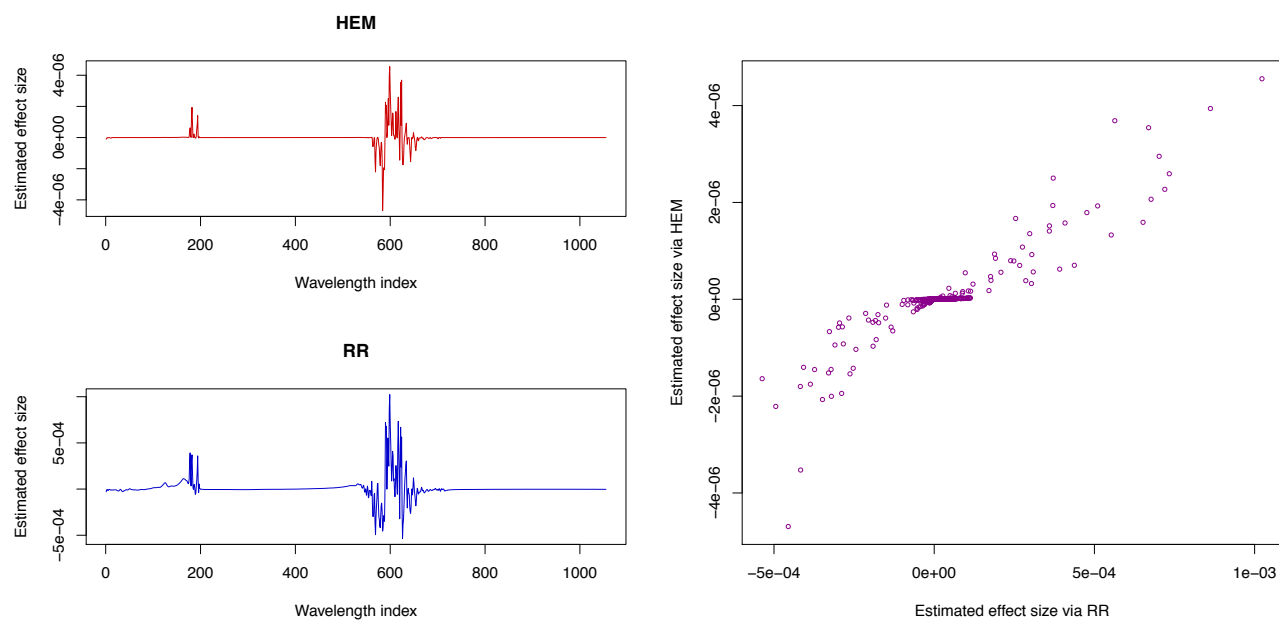
---

## Remarks

The package page on CRAN is <http://cran.r-project.org/web/packages/bigRR/>.

The developer versions of the package are maintained on the R-Forge project page: <https://r-forge.r-project.org/projects/bigrr/>.

The detailed description of the data analyzed in this tutorial can be found at: <https://docs.google.com/spreadsheet/ccc?key=0AmixEvB0Gwt6dHJJNnN6ZWg4N1N1UE5kdIhleWxXUWc&usp=sharing>



**Figure T1: Comparison of the estimated wavelength effects along the FTIR spectrum of the ethanol content in the first experiment. RR: ridge regression; HEM: heteroscedastic effects model.**