PCBs covariance matrix investigation

Xuelong Wang

2019-08-22

Contents

1 Motivation

The sample covariates matrix of PCBs is unstructured and dense, so it is not straightforward to estimate a large, non-spare covariance matrix directly. One possible solution is to borrow the information from histrocial data or data from other studies with different responses but similar covariates. The report is to investigate the covariance matrix across all the available PCBs data from NHANES from 1999 - 2014. We are trying to see if there is common structure among all different surveies or within certain subgroups.

1.1 Decorrelation using histrocial data

1.1.1 Decorrelation steps

Let X be the covariates and the $\hat{\Sigma}_X = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X})$. Assume that there is a matrix A that we could used for decorrelation.

$$Z = XA \Rightarrow \hat{\Sigma}_Z = \frac{1}{n-1}(Z - \bar{X})^T(Z - \bar{Z}) = A^T\hat{\Sigma}_X A$$

2 PCBs data summary and covariance-correlation structure

2.1 1999-2004

There are 3 surveys were conducted during the 6 years. However, based on the data I collected, the types of PCBs are not extact same accross the 3 surveys. Followings are some facts.

The types of PCBs measured for each survey is

```
$`1999`
 [1] "PCB028" "PCB052" "PCB066" "PCB074" "PCB099" "PCB101" "PCB105"
 [8] "PCB118" "PCB128" "PCB138" "PCB146" "PCB153" "PCB156" "PCB157"
[15] "PCB167" "PCB170" "PCB172" "PCB177" "PCB178" "PCB180" "PCB183"
[22] "PCB187"

$`2001`
 [1] "PCB052" "PCB066" "PCB074" "PCB087" "PCB099" "PCB101" "PCB105"
```

```
 [8] "PCB110" "PCB118" "PCB128" "PCB138" "PCB146" "PCB149" "PCB151"
[15] "PCB153" "PCB156" "PCB157" "PCB167" "PCB170" "PCB172" "PCB177"
[22] "PCB178" "PCB180" "PCB183" "PCB187" "PCB189" "PCB194" "PCB195"
[29] "PCB196" "PCB199" "PCB206"

$`2003`
 [1] "PCB028" "PCB066" "PCB074" "PCB105" "PCB118" "PCB156" "PCB157"
 [8] "PCB167" "PCB189" "PCB044" "PCB049" "PCB052" "PCB087" "PCB099"
[15] "PCB101" "PCB110" "PCB128" "PCB138" "PCB146" "PCB149" "PCB151"
[22] "PCB153" "PCB170" "PCB172" "PCB177" "PCB178" "PCB180" "PCB183"
[29] "PCB187" "PCB194" "PCB195" "PCB196" "PCB199" "PCB206" "PCB209"
```

The common PCBs that were measured by each survey is

```
 [1] "PCB052" "PCB066" "PCB074" "PCB099" "PCB101" "PCB105" "PCB118"
 [8] "PCB128" "PCB138" "PCB146" "PCB153" "PCB156" "PCB157" "PCB167"
[15] "PCB170" "PCB172" "PCB177" "PCB178" "PCB180" "PCB183" "PCB187"
```

I will use those 21 PCBs to calculate their covariance matrix.

The total number of PCBs data from 1999-2004 is 7106. After I remove all the missing data, what I get is total 4873 observations.

Note that I only work on the PCBs measurement without adjustment, but I used the under the limit adjustment.

2.1.1 time

Followings are 3 heat-maps of correlation-matrix across the 3 surveys. In general, they have high correlations among those PCBs. But they do show some common pattern.

The number of observation across each survey is

```
   SDDSRVYR    N
1:        1 1647
2:        2 1435
3:        3 1791
```

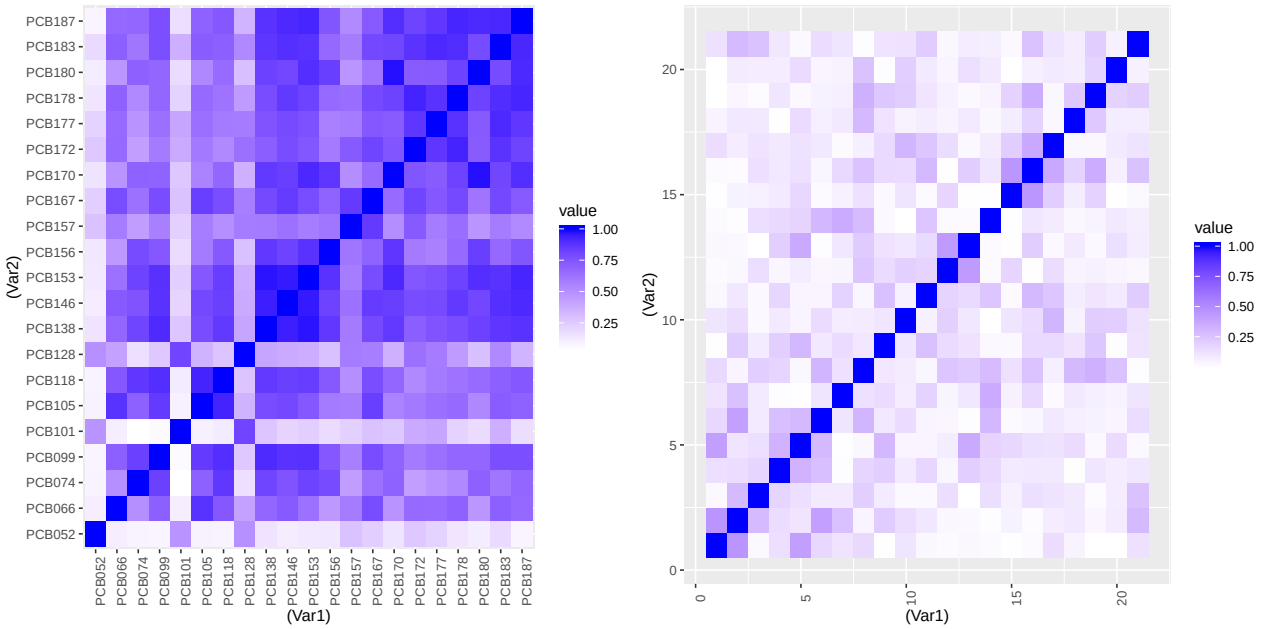


Figure 1: 1999-2000

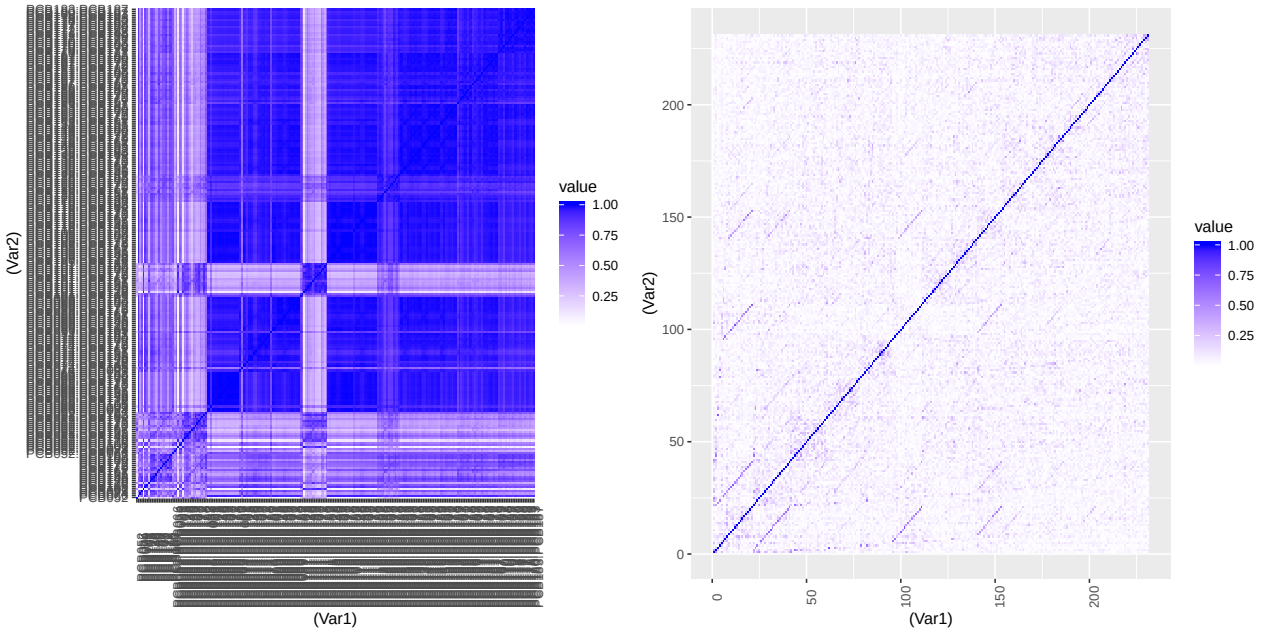


Figure 2: Combined main and interaction 1999-2000

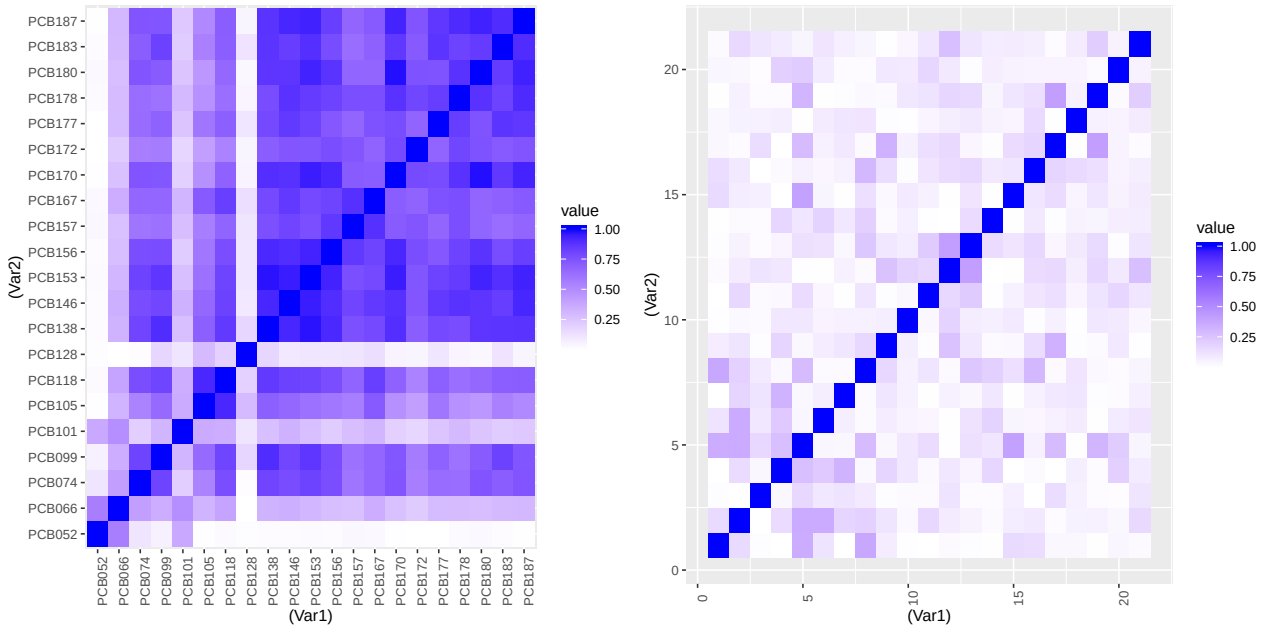


Figure 3: 2001-2002

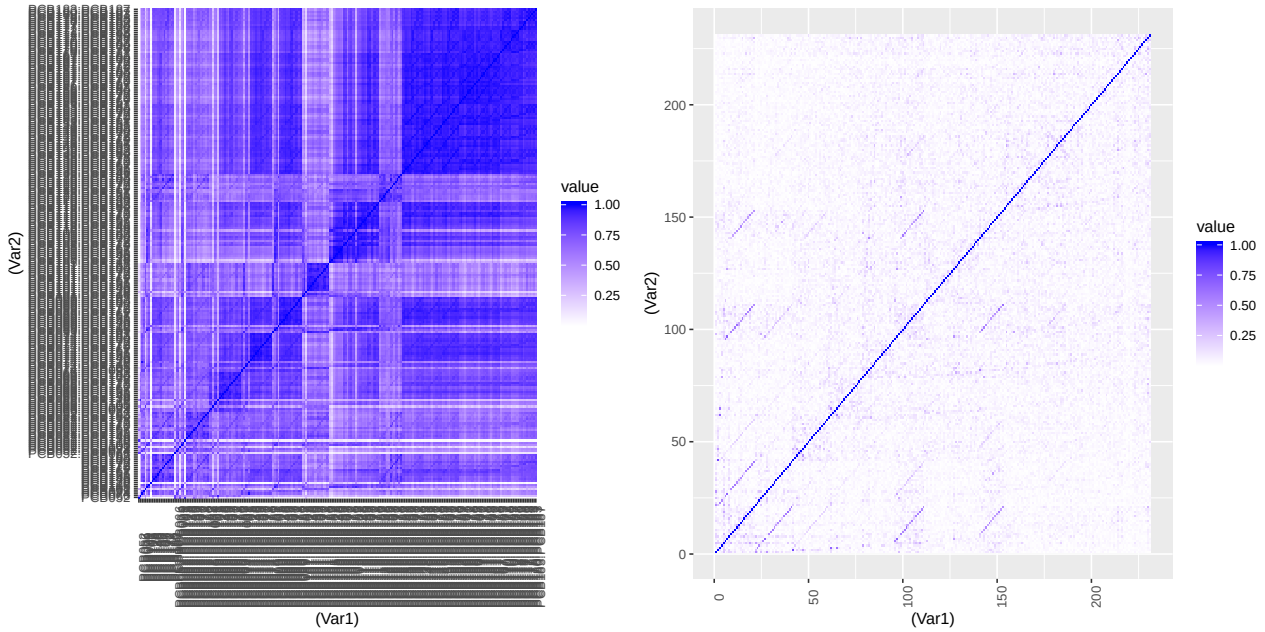


Figure 4: Combined main and interaction 2001-2002

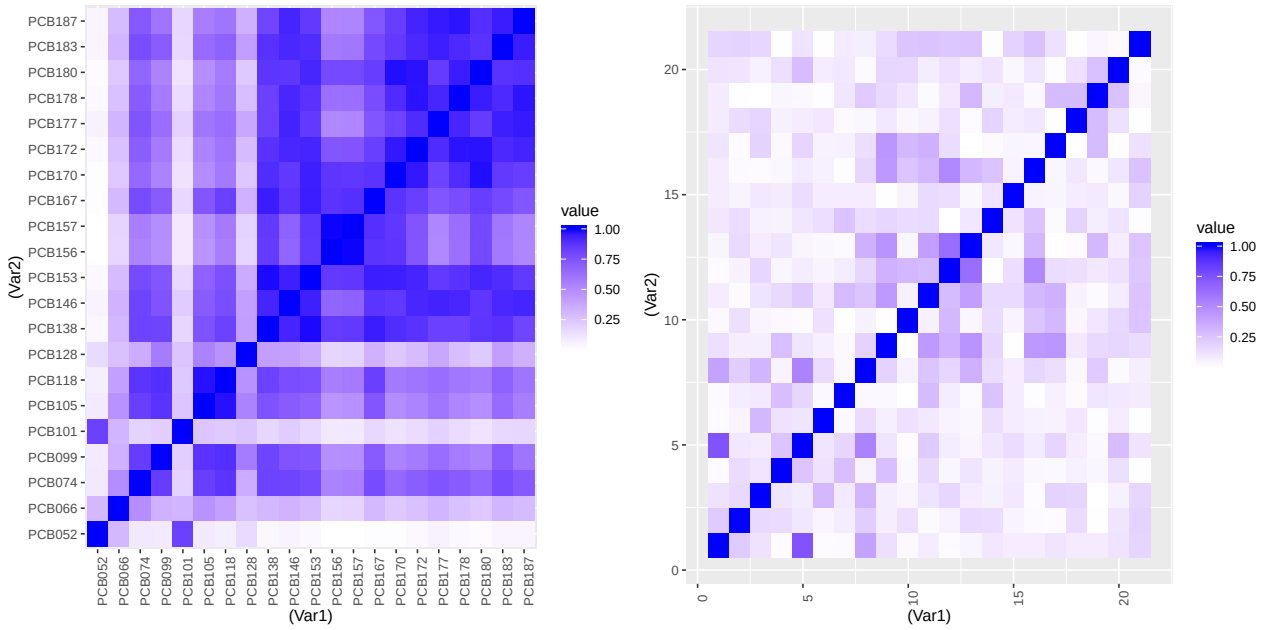


Figure 5: 2003-2004

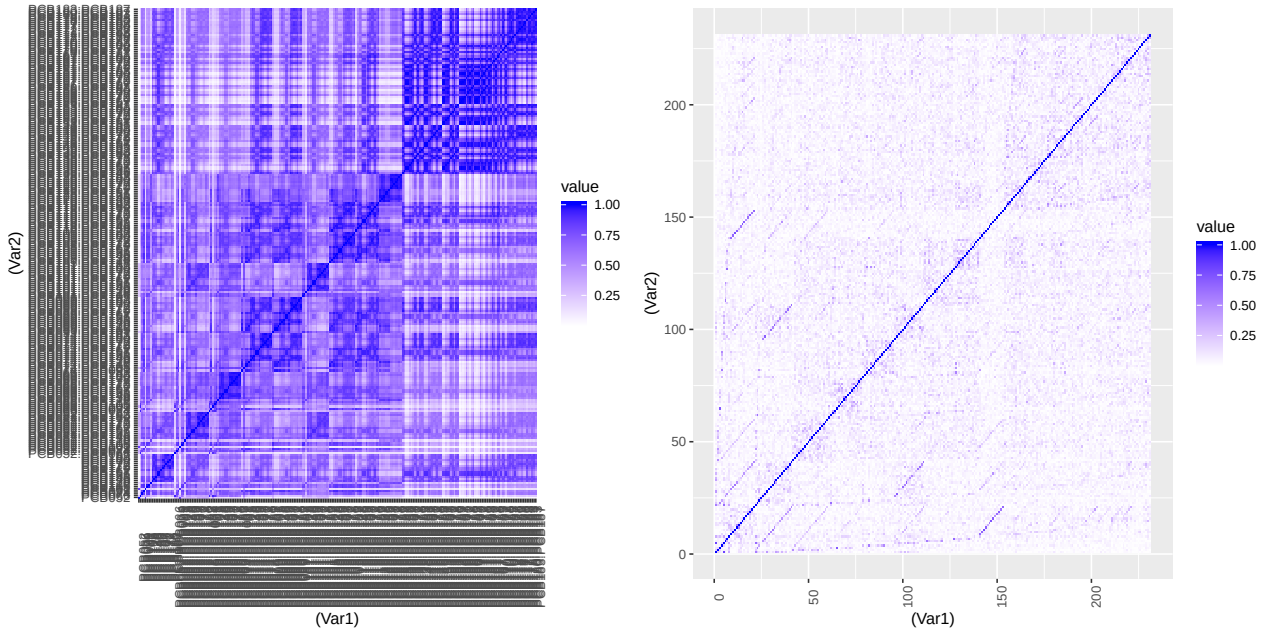


Figure 6: Combined main and interaction 1993-2004

2.1.2 gender

Followings are the heat-maps of correlation-matrix for different gender

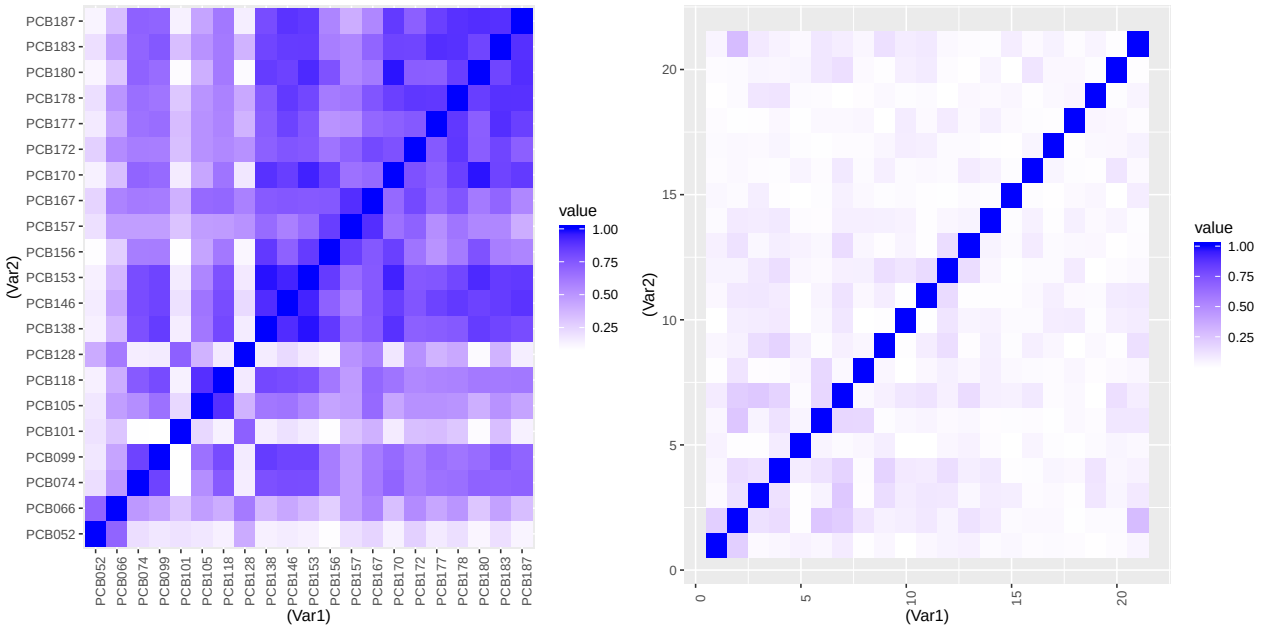


Figure 7: Male

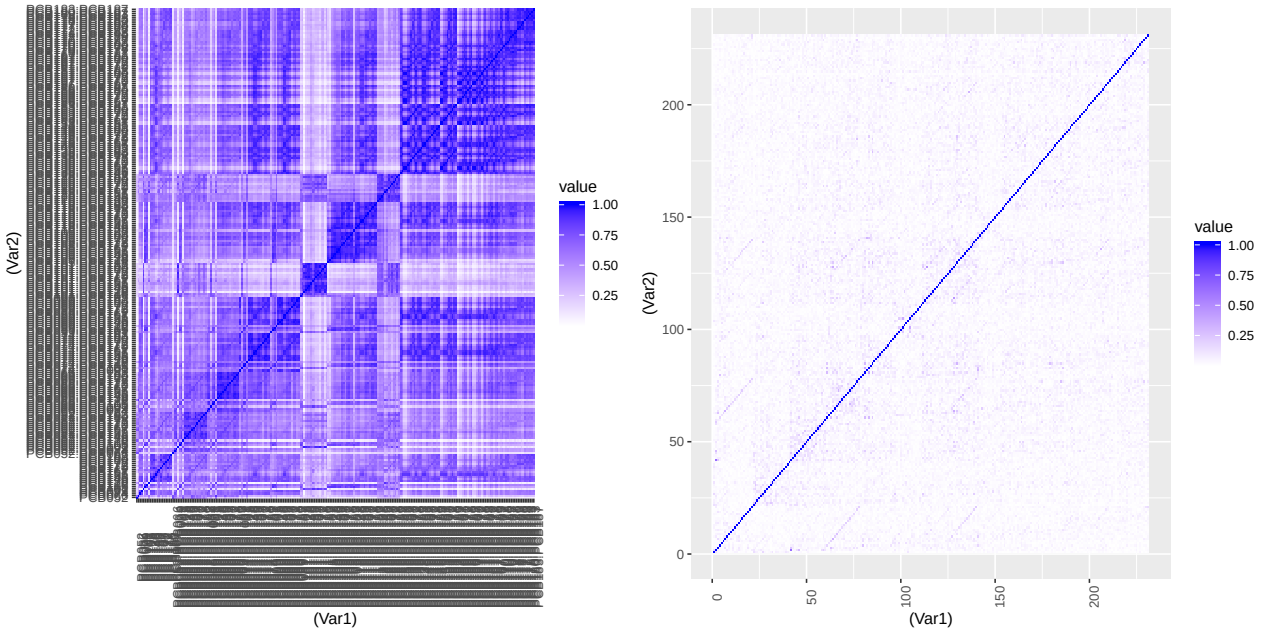


Figure 8: Combined main and interaction Male

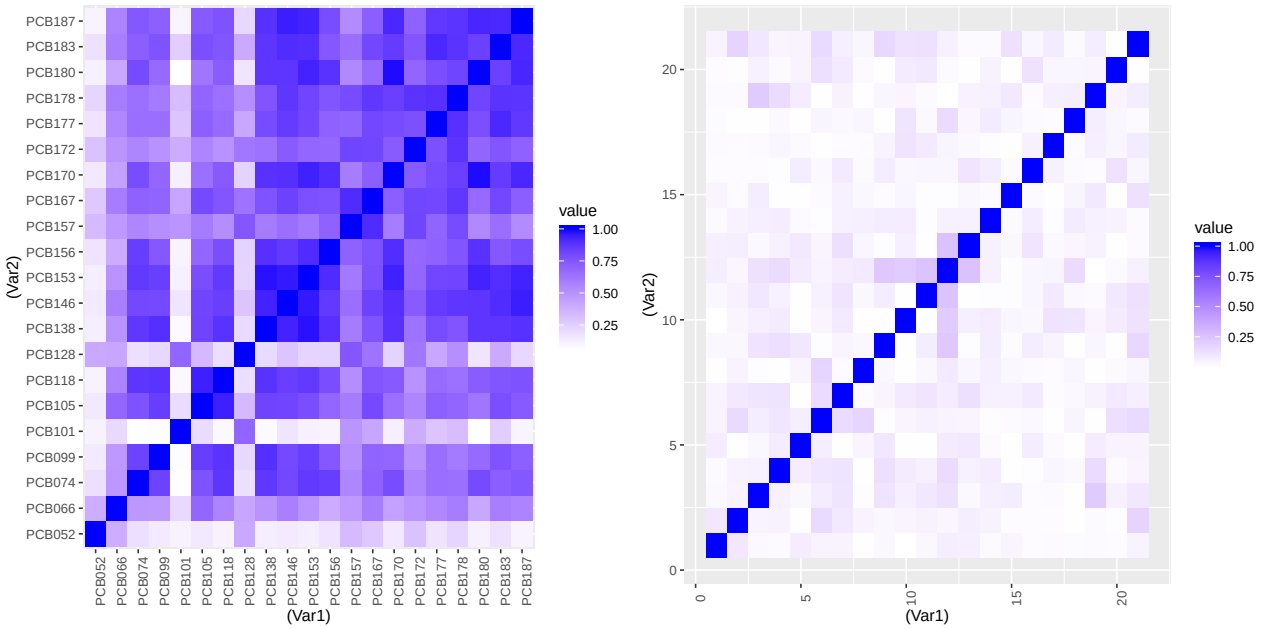


Figure 9: Female

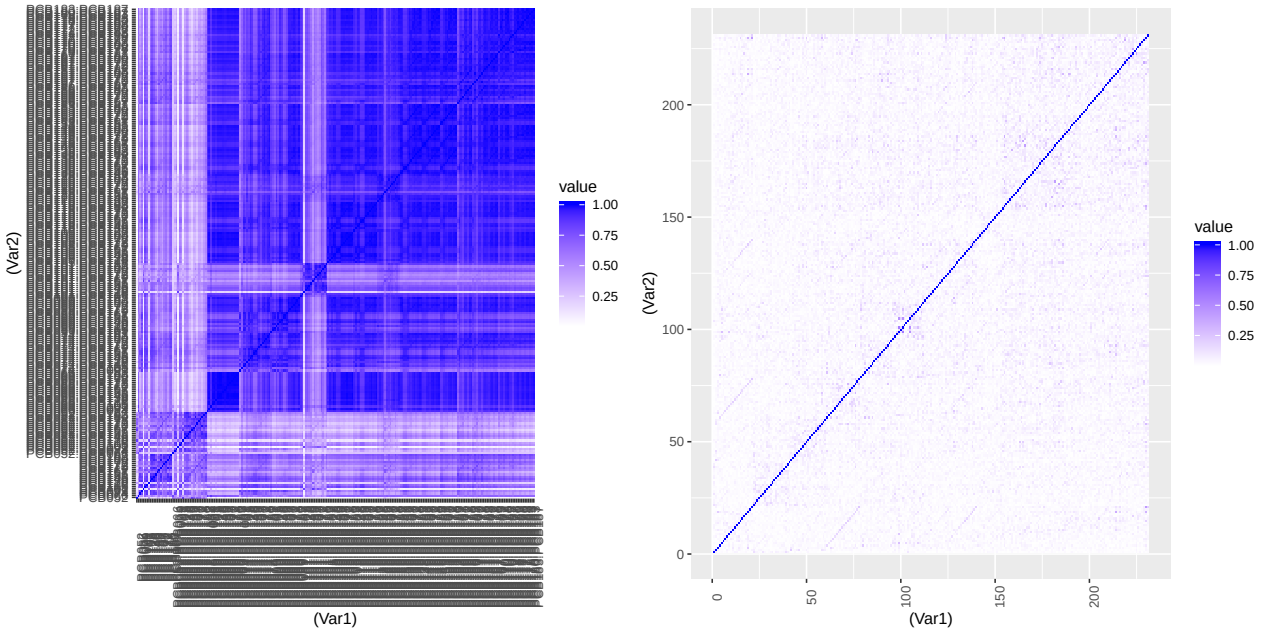


Figure 10: Combined main and interaction Female

2.2 2005-2014

2.2.1 the pooled-sample

NHANES adapted a pooled sampling method to collect the PCBs data since 2005. It seems that they collect all the blood samples from all the subjects but they only chose a sub-sample of them to measure the PCBs value. The basic idea is following:

1. Divide the whole subjects into 32 demographic groups

2. For all the subjects in each demographic group, split them into pools with sample size around 8. Note that the number of pools are proportion to the total number of subjects in the demographic group.

3. Draw a random sample from each pools, so that those sub-samples keeps same pattern in term of demographic groups ratio.

The follwoing is a summary table of sub-samples of 2005-2006 subjects' for getting PCBs measurements.

Table 2. Number of subjects per demographic group in the NHANES 2005–2006 one-third subsample, number of individual serum samples available, number of usable samples, and number of pools formed from usable samples.

Race/ Ethnicity	Gender	Number of subjects in the one-third subsample/ Number of Samples Available/Number of Usable Samples (Number of Pools)			
		12–19 years (RIDAGGRP=1)[1]	20–39 years (RIDAGGRP=2)	40–59 years (RIDAGGRP=3)	60+ years (RIDAGGRP=4)
Non-Hispanic White (ETHNICTY=1)[1]	Male (RIAGENDR=1)[1]	81/76/72 (9)	110/108/96 (12)	114/110/96 (12)	141/137/120 (15)
	Female (RIAGENDR=2)	94/85/80 (10)	143/136/128 (16)	116/111/104 (13)	149/146/136 (17)
Non-Hispanics Black (ETHNICTY=2)	Male (RIAGENDR=1)	129/114/104 (13)	60/57/48 (6)	56/51/40 (5)	55/52/40 (5)
	Female (RIAGENDR=2)	132/117/112 (14)	74/66/56 (7)	65/62/56 (7)	55/48/40 (5)
Mexican American (ETHNICTY=3)	Male (RIAGENDR=1)	106/96/88 (11)	87/84/72 (9)	44/43/32 (4)	38/38/32 (4)
	Female (RIAGENDR=2)	143/133/128 (16)	88/84/72 (9)	50/50/48 (6)	38/37/24 (3)
Other (ETHNICTY=4)	Male (RIAGENDR=1)	20/19/16 (2)	27/26/24 (3)	24/23/23[2] (3)	9/8/8 (1)
	Female (RIAGENDR=2)	31/26/24 (3)	38/34/32 (4)	18/18/16 (2)	10/6/6[3] (1)

1 Value of this categorical variable in the data set.
2 With only 23 usable samples, two 8 sample pools and one 7 sample pool were created.
3 With only 6 usable samples, one 6 sample pool was created.

The types of PCBs measured for each survey is

```
$`2005`
 [1] "PCB028" "PCB044" "PCB049" "PCB052" "PCB066" "PCB074" "PCB087"
 [8] "PCB099" "PCB101" "PCB105" "PCB110" "PCB114" "PCB118" "PCB123"
[15] "PCB128" "PCB138" "PCB146" "PCB149" "PCB151" "PCB153" "PCB156"
[22] "PCB157" "PCB167" "PCB170" "PCB172" "PCB177" "PCB178" "PCB180"
[29] "PCB183" "PCB187" "PCB189" "PCB194" "PCB195" "PCB196" "PCB199"
```

```
[36] "PCB206" "PCB209"

$`2007`
 [1] "PCB028" "PCB044" "PCB049" "PCB052" "PCB066" "PCB074" "PCB087"
 [8] "PCB099" "PCB101" "PCB105" "PCB110" "PCB114" "PCB118" "PCB123"
[15] "PCB128" "PCB138" "PCB146" "PCB149" "PCB151" "PCB153" "PCB156"
[22] "PCB157" "PCB167" "PCB170" "PCB172" "PCB177" "PCB178" "PCB180"
[29] "PCB183" "PCB187" "PCB189" "PCB194" "PCB195" "PCB196" "PCB199"
[36] "PCB206" "PCB209"

$`2009`
 [1] "PCB028" "PCB066" "PCB074" "PCB099" "PCB105" "PCB114" "PCB118"
 [8] "PCB138" "PCB146" "PCB153" "PCB156" "PCB157" "PCB167" "PCB170"
[15] "PCB178" "PCB180" "PCB183" "PCB187" "PCB189" "PCB194" "PCB196"
[22] "PCB199" "PCB206" "PCB209"

$`2011`
 [1] "PCB028" "PCB066" "PCB074" "PCB099" "PCB105" "PCB114" "PCB118"
 [8] "PCB138" "PCB146" "PCB153" "PCB156" "PCB157" "PCB167" "PCB170"
[15] "PCB178" "PCB180" "PCB183" "PCB187" "PCB189" "PCB194" "PCB196"
[22] "PCB199" "PCB206" "PCB209"

$`2013`
 [1] "PCB028" "PCB066" "PCB074" "PCB099" "PCB105" "PCB114" "PCB118"
 [8] "PCB138" "PCB146" "PCB153" "PCB156" "PCB157" "PCB167" "PCB170"
[15] "PCB178" "PCB180" "PCB183" "PCB187" "PCB189" "PCB194" "PCB196"
[22] "PCB199" "PCB206" "PCB209"
```

The common PCBs that were measured by each survey is

```
 [1] "PCB028" "PCB066" "PCB074" "PCB099" "PCB105" "PCB114" "PCB118"
 [8] "PCB138" "PCB146" "PCB153" "PCB156" "PCB157" "PCB167" "PCB170"
[15] "PCB178" "PCB180" "PCB183" "PCB187" "PCB189" "PCB194" "PCB196"
[22] "PCB199" "PCB206" "PCB209"
```

I will use those 24 PCBs to calculate their covariance matrix.

The total number of PCBs data from 2005-2014 is 1347. After I remove all the missing data, what I get is total 1228 observations.

Note that I only work on the PCBs measurement without adjustment, but I used the under the limit adjustment.

2.2.2 time

Followings are 5 heat-maps of correlation-matrix across the 3 surveys. In general, they have high correlations among those PCBs. But they do show some common pattern.

The number of observation across each survey is

```
   SDDSRVYR   N
1:        4 247
2:        5 264
3:        6 215
4:        7 248
5:        8 254
```

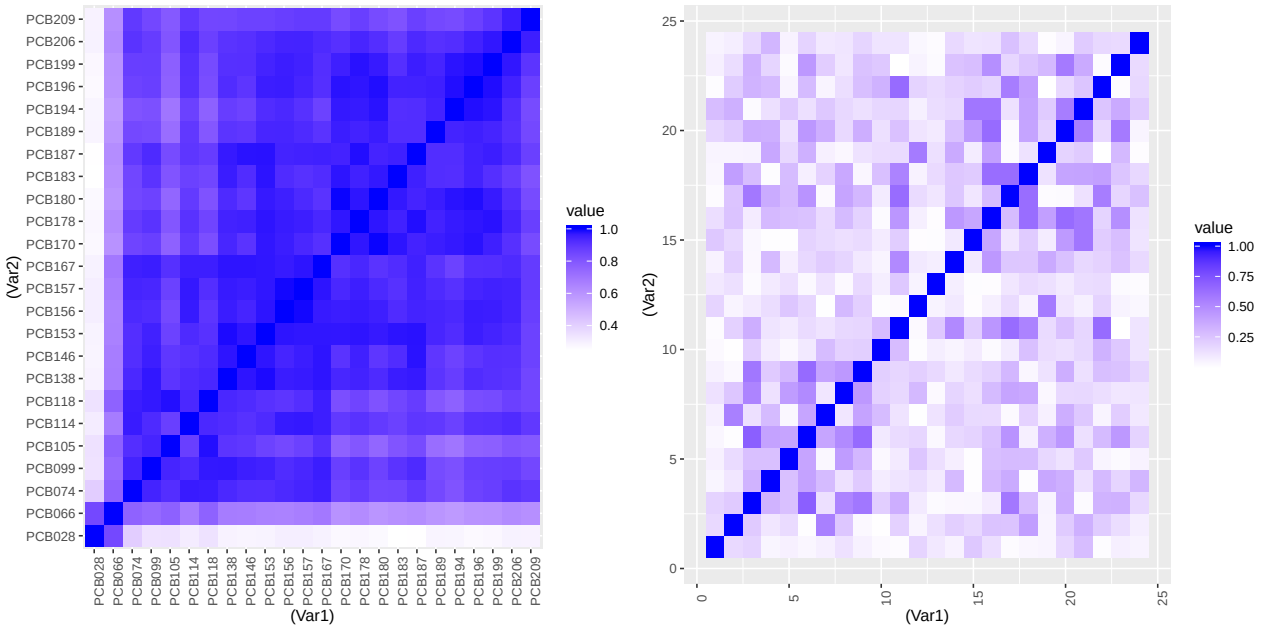


Figure 11: 2005-2006

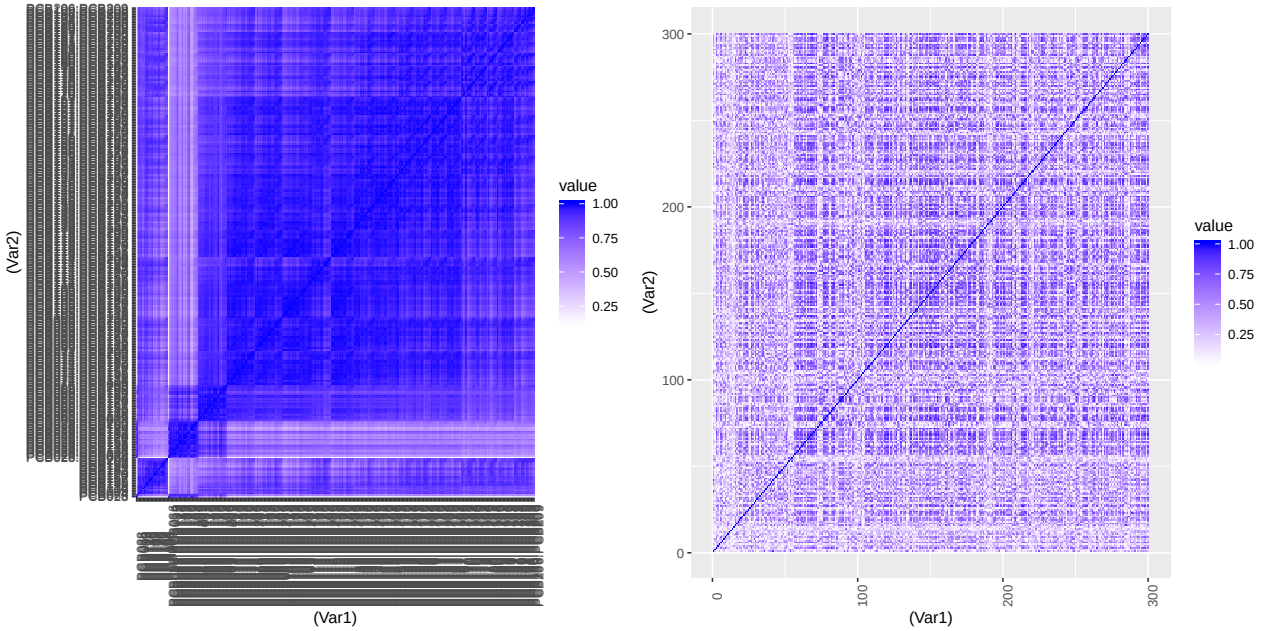


Figure 12: Combined main and interaction 2005-2006

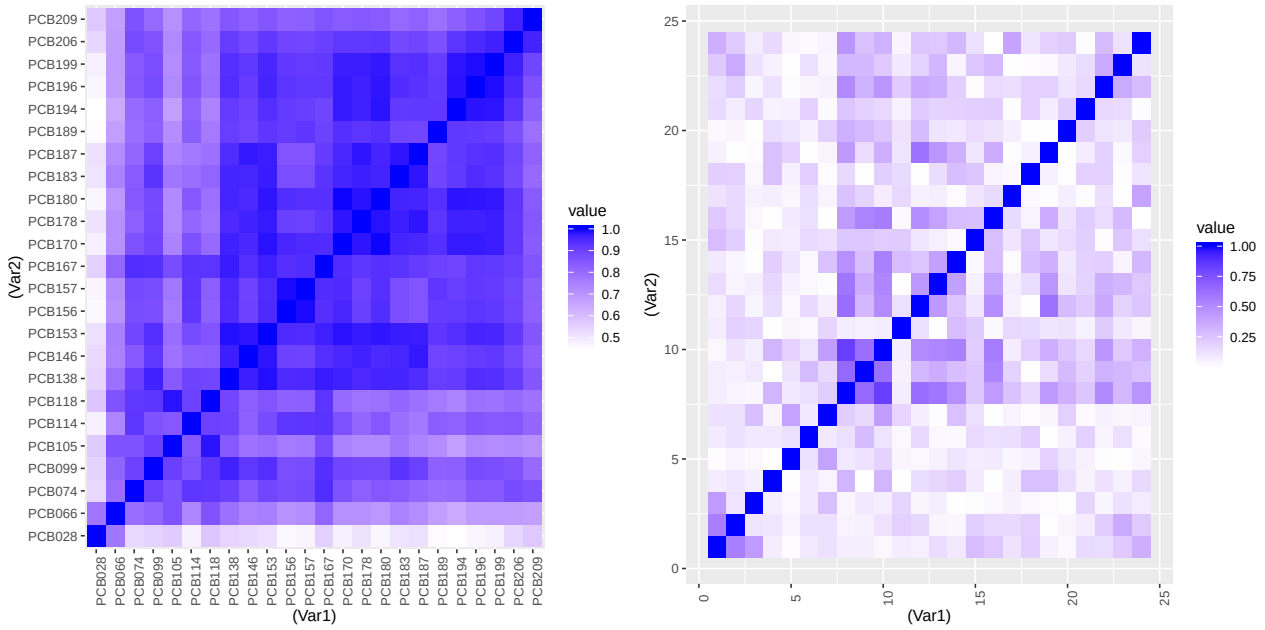


Figure 13: 2007-2008

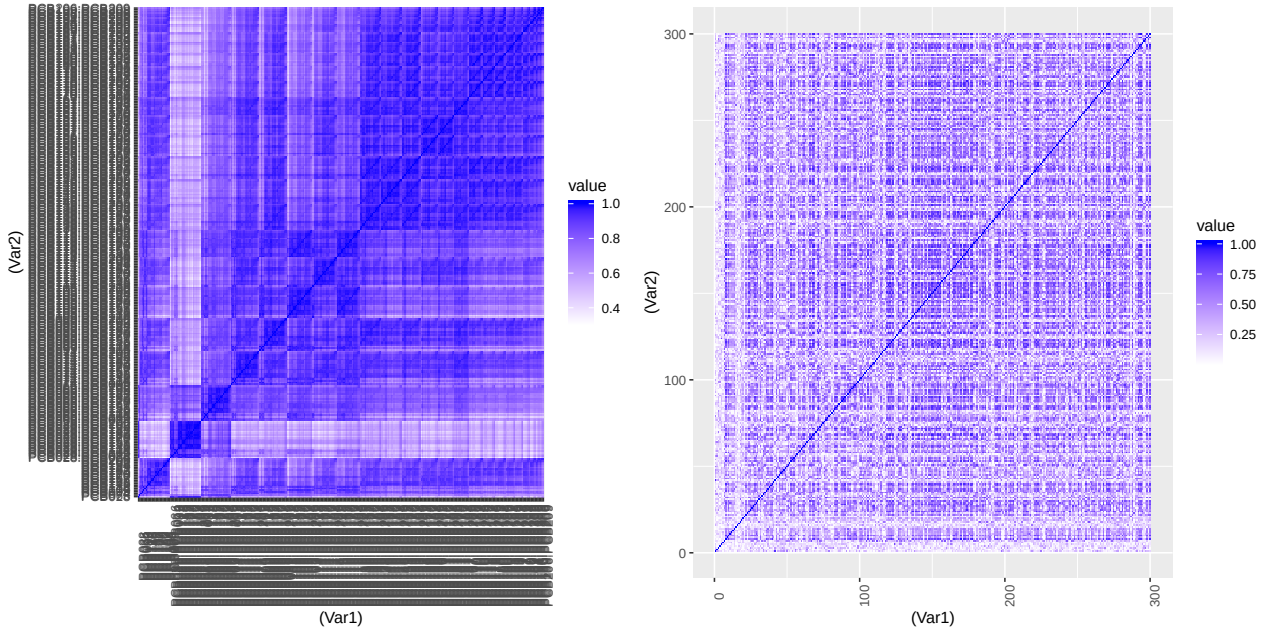


Figure 14: Combined main and interaction 2007-2008

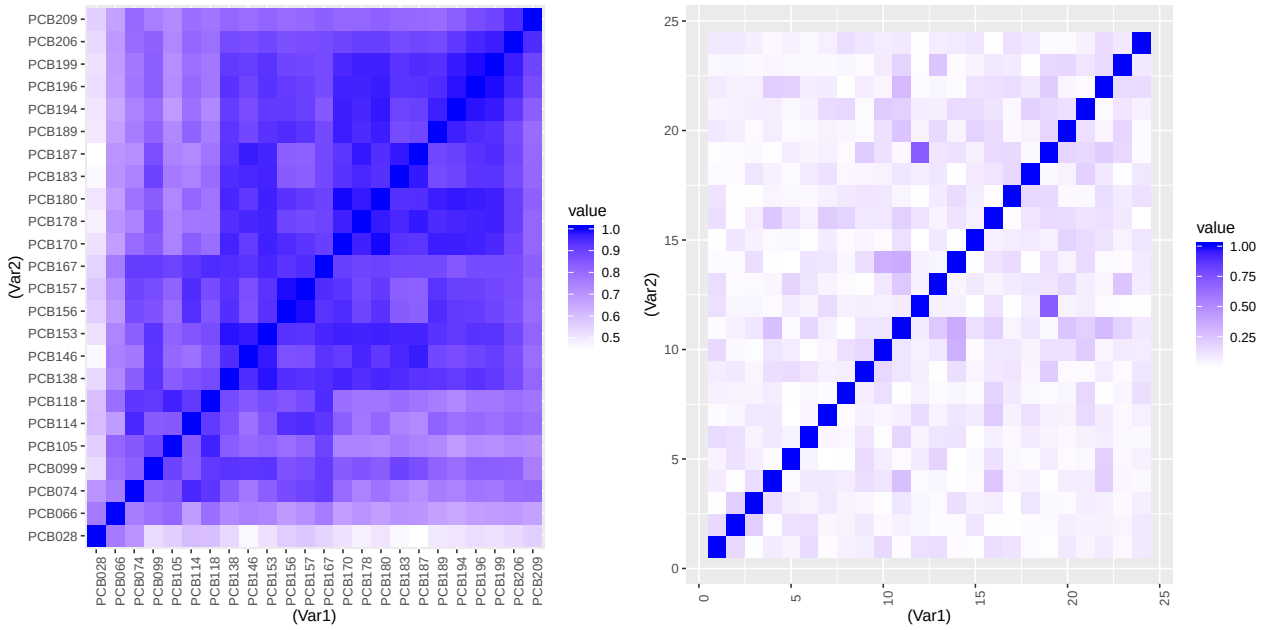


Figure 15: 2009-2010

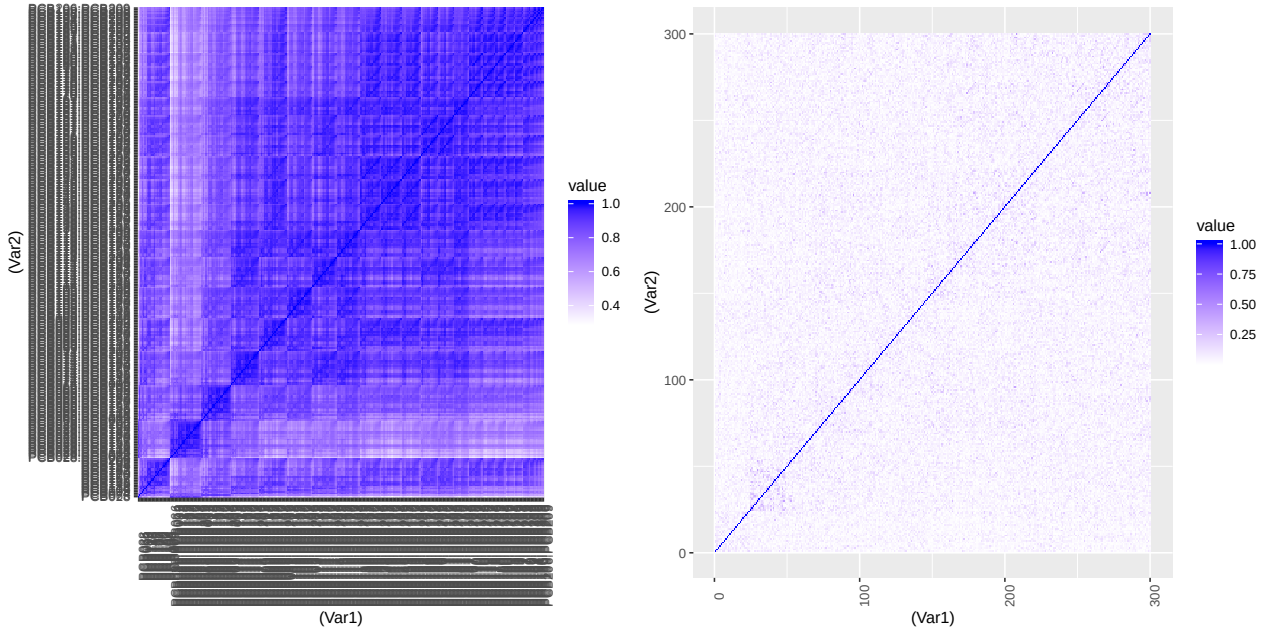


Figure 16: Combined main and interaction 2009-2010

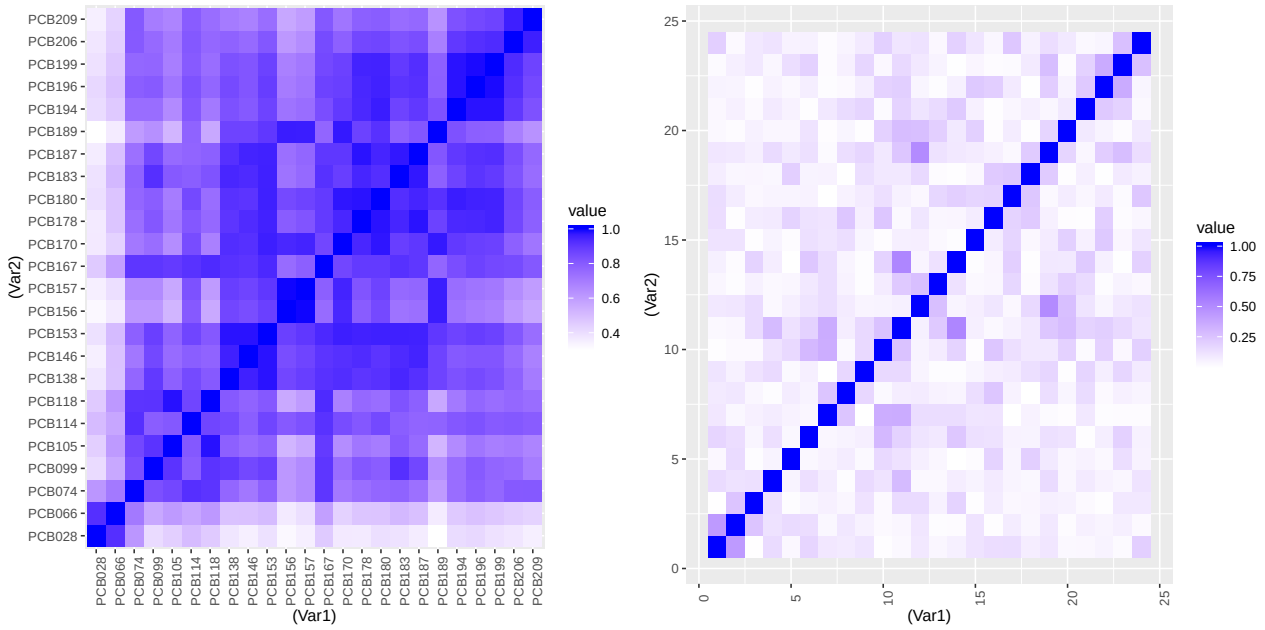


Figure 17: 2011-2012

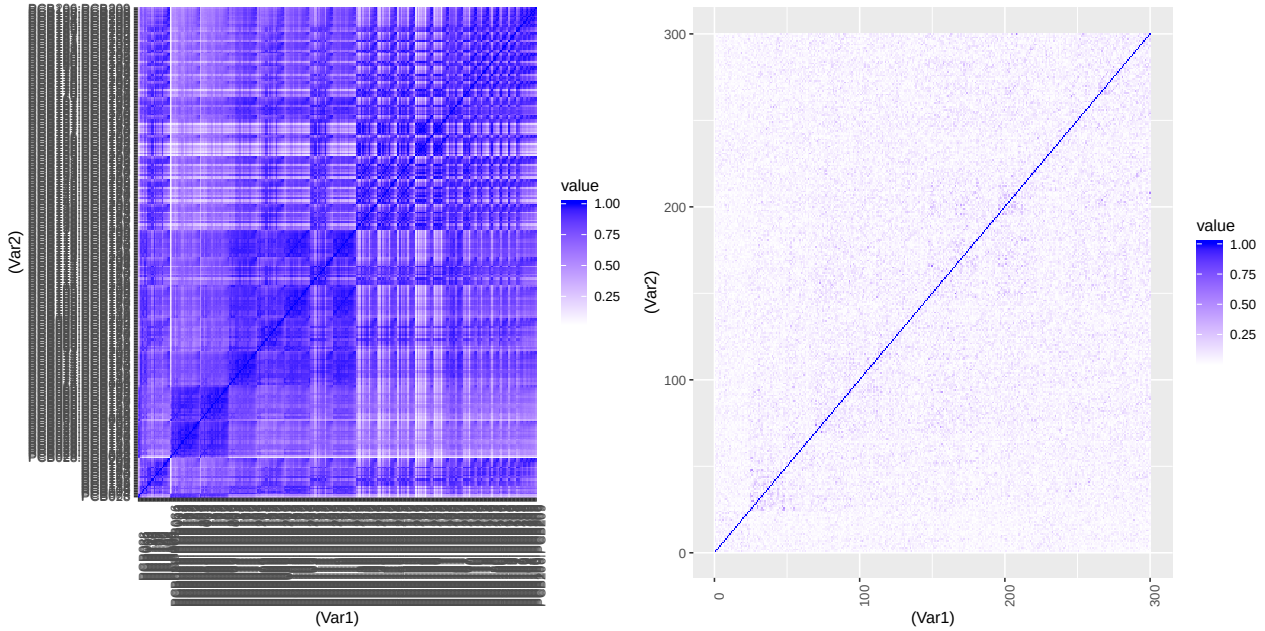


Figure 18: Combined main and interaction 2011-2012

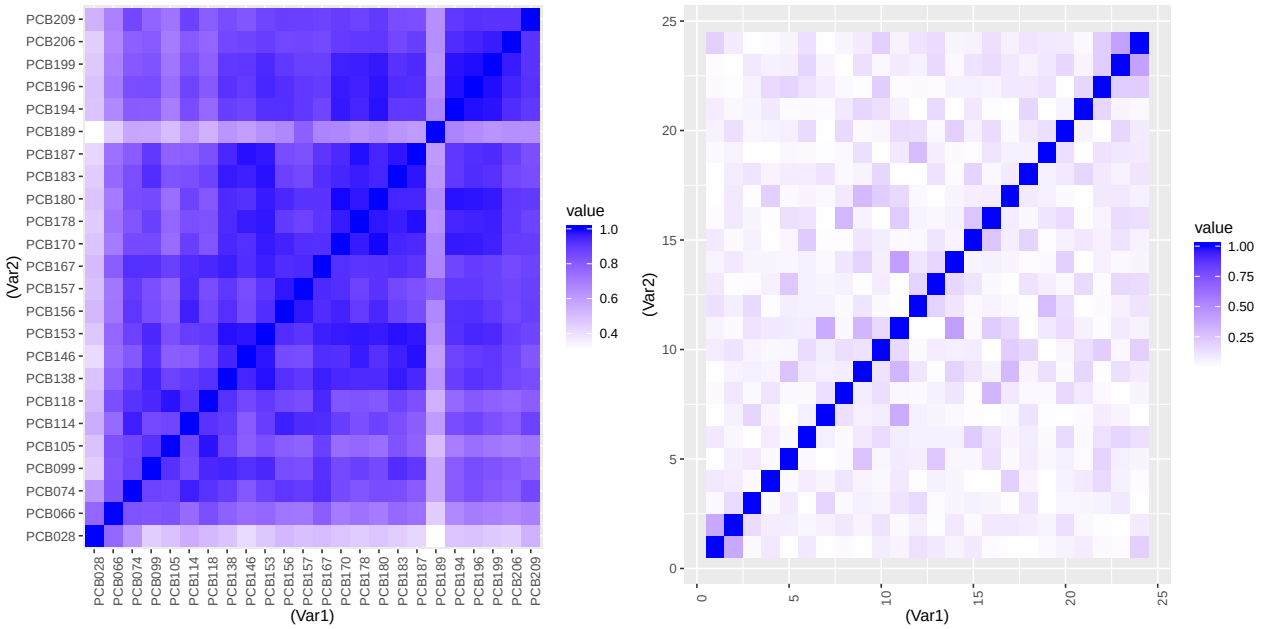
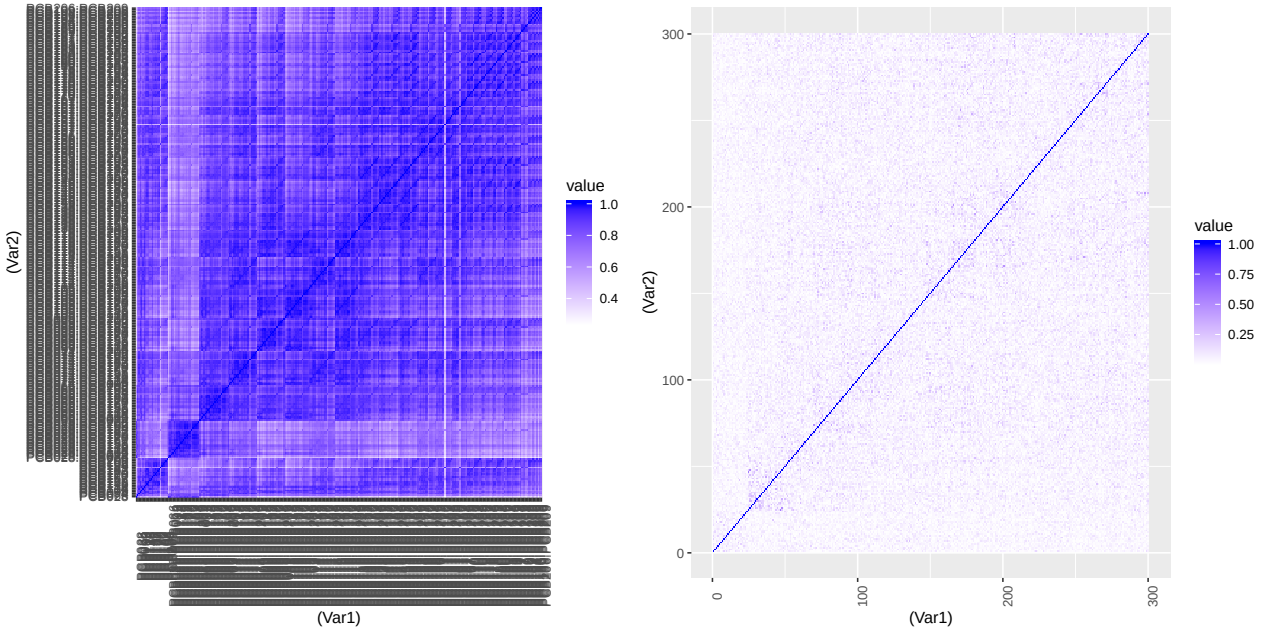


Figure 19: 2013-2014



gender Followings are the heat-maps of correlation-matrix for different gender

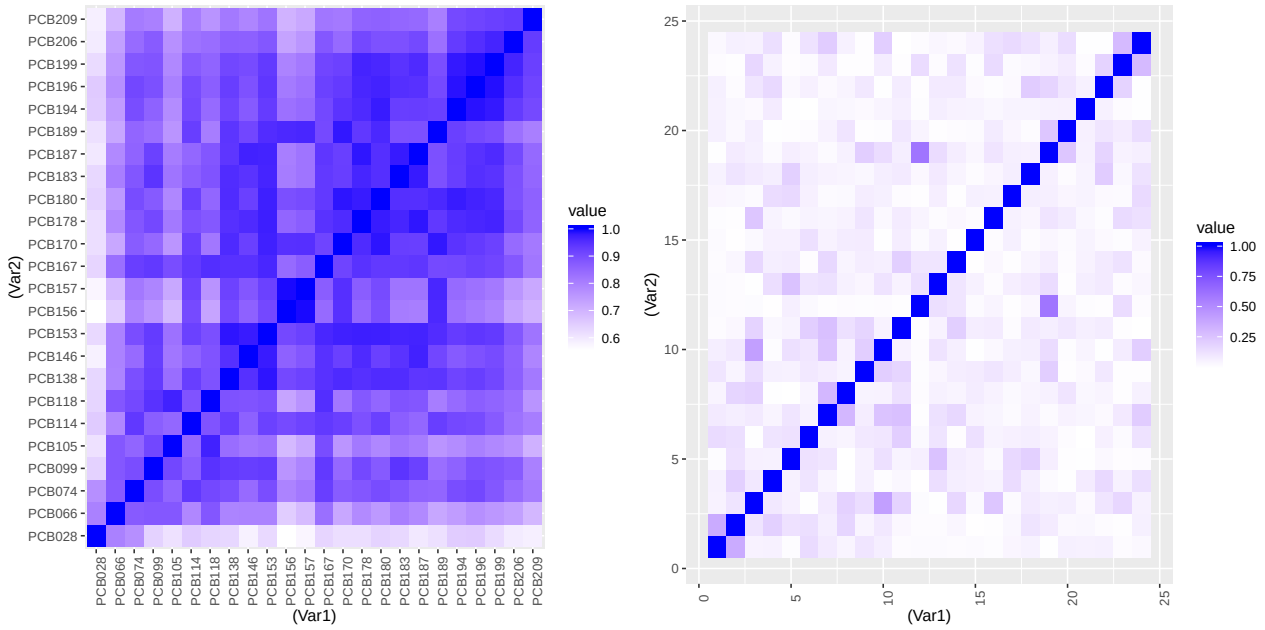


Figure 20: Male

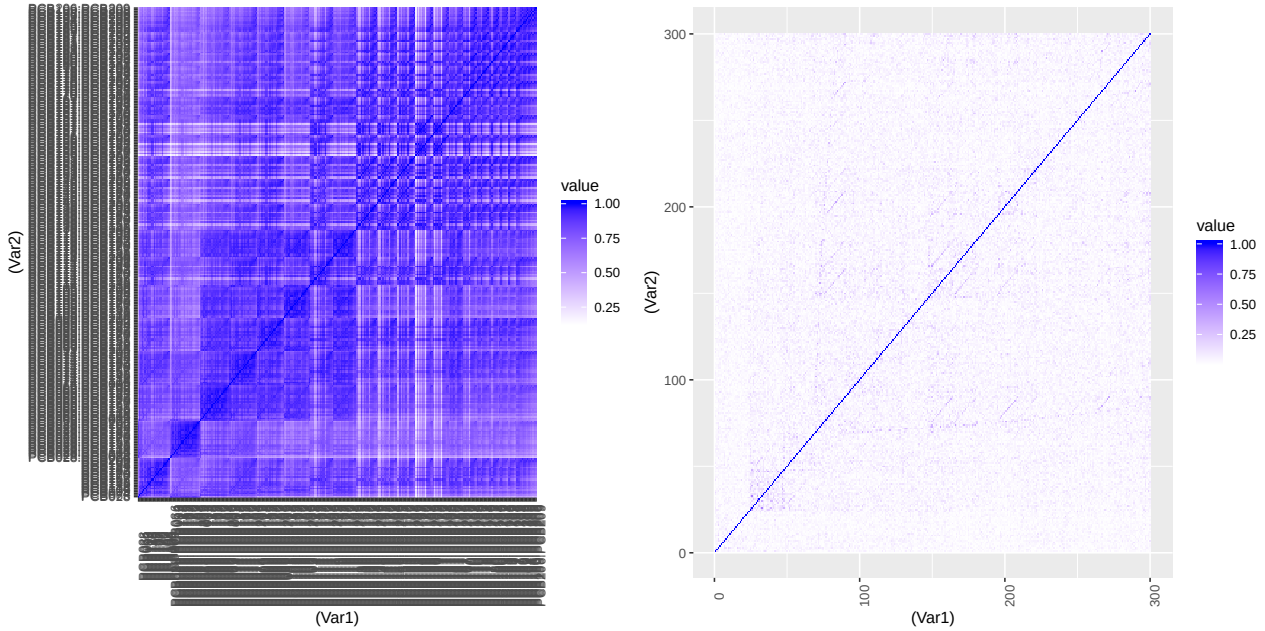


Figure 21: Combined main and interaction Male

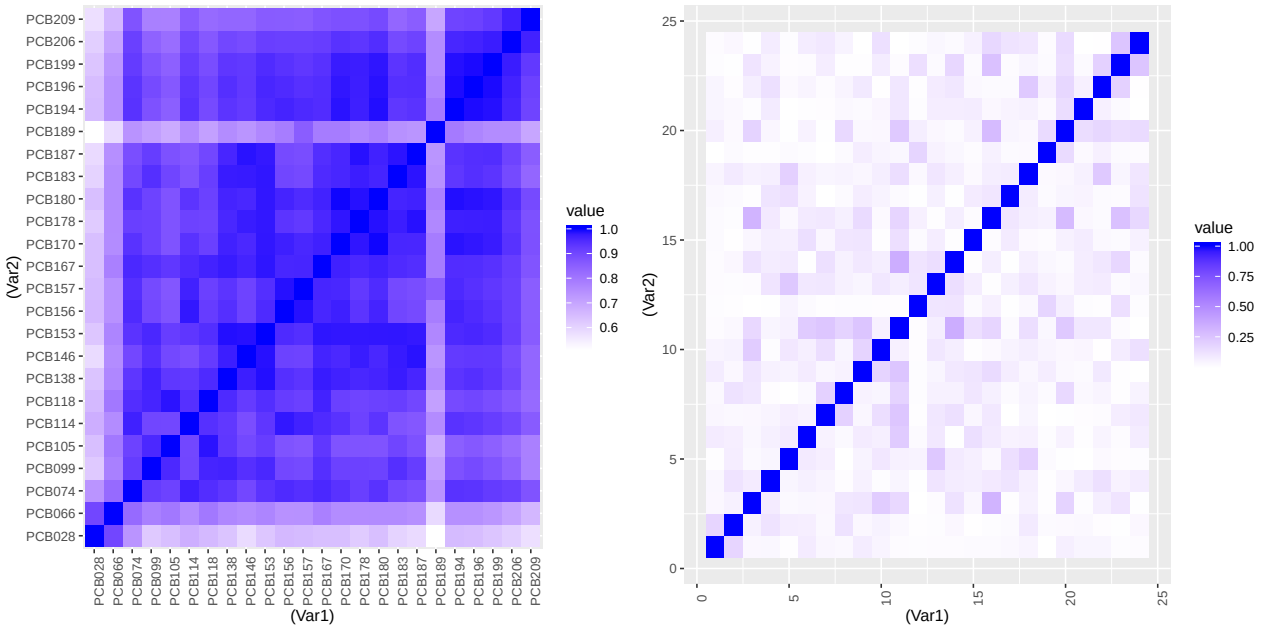


Figure 22: Female