36-315: Statistical Graphics and Visualization

Handout 22
Date: April 14, 2003

Finding the variables that explain outliers

Useful functions for examining outliers: `locate.bbox`, `identify`, `points`

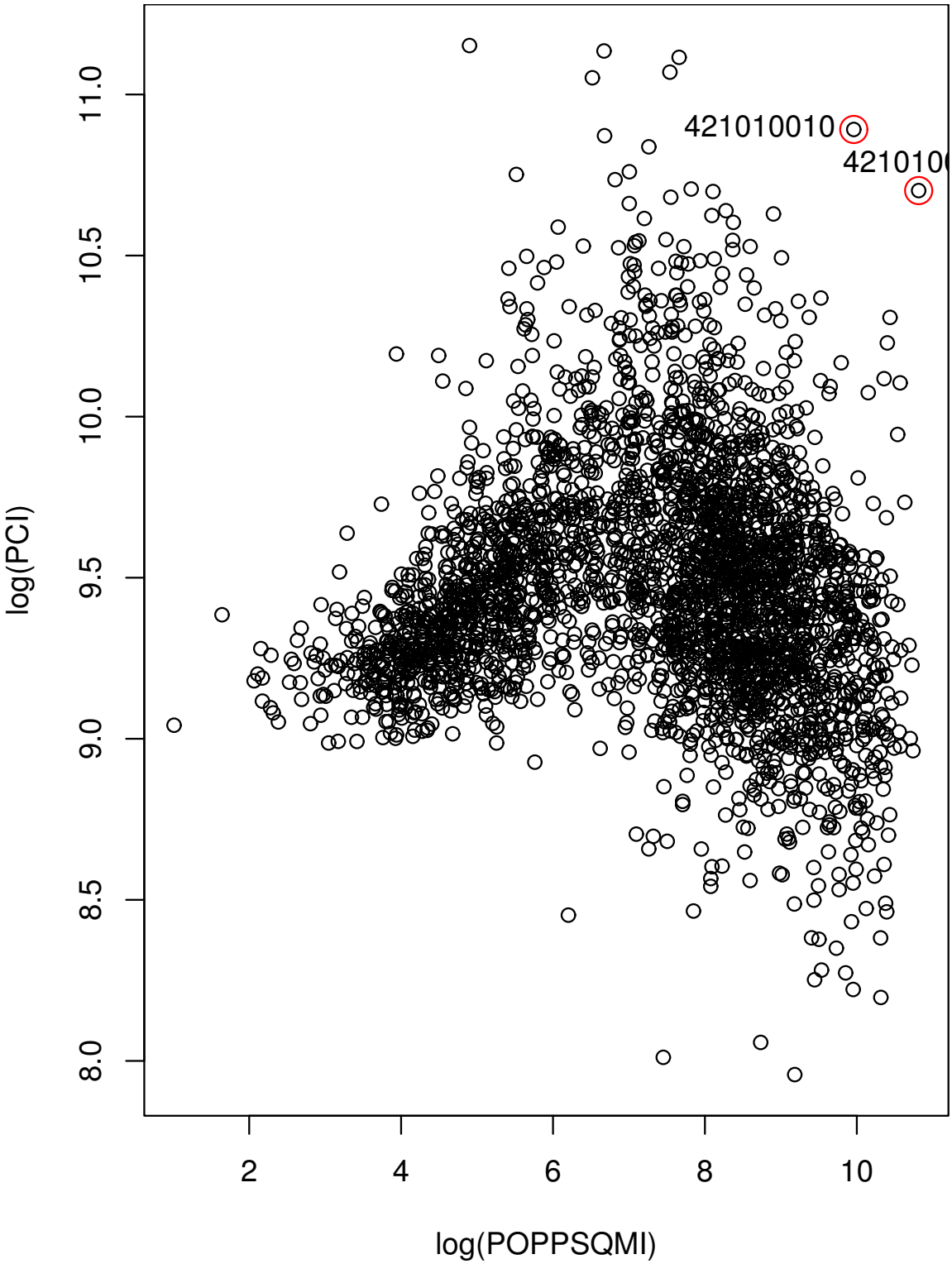
Example of explaining outliers with maps:
`http://www.mimas.ac.uk/argus/Tutorials/CartoViz/PopViz/Figs/C2Outliers.html`

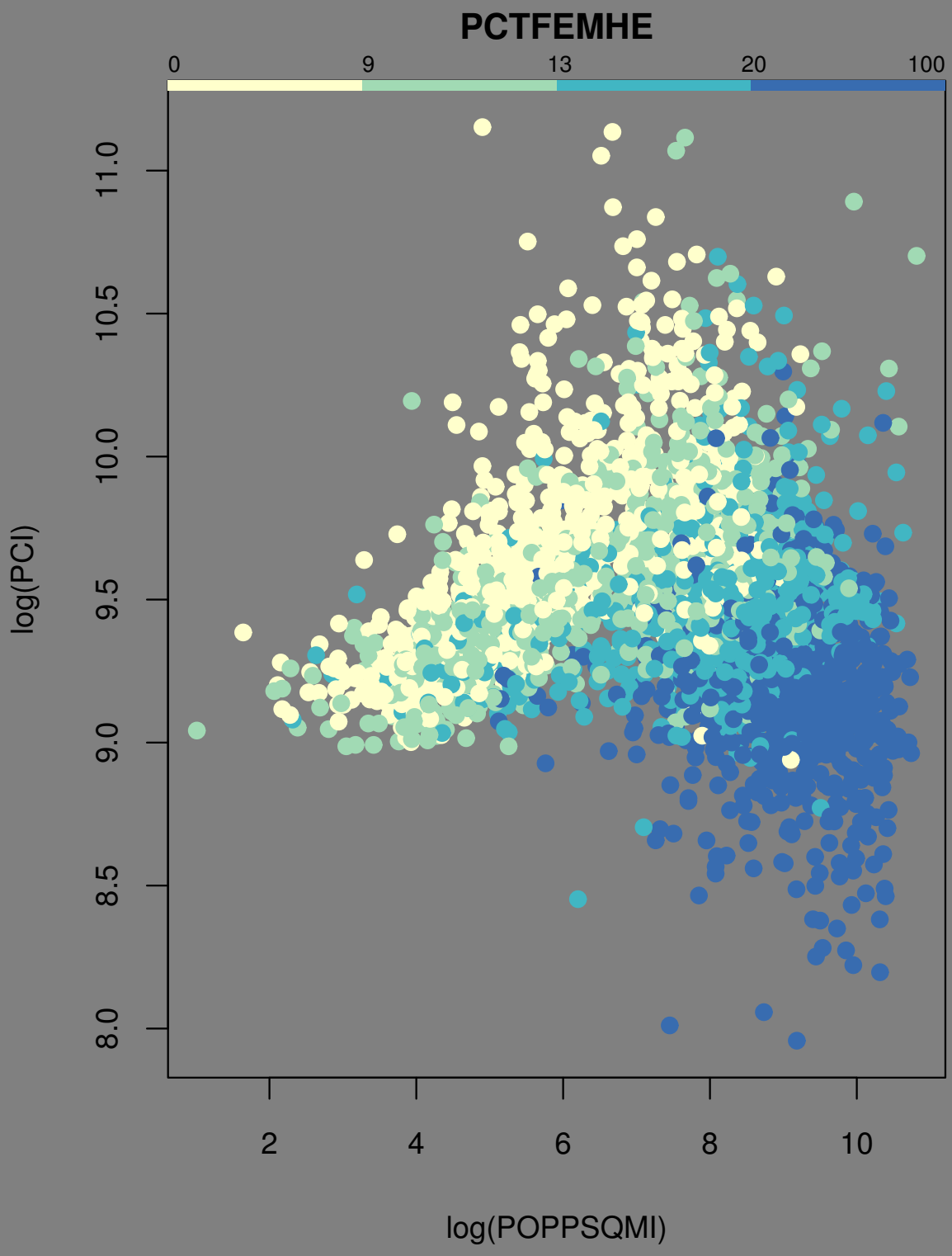
Hypervariate—Having four or more dimensions.

Projection—A method for reducing a dataset of many dimensions to fewer (usually two) dimensions, by taking a weighted sum. Geometrically, it is like taking a picture of a high-dimensional point cloud onto two-dimensional film, where the dataset is rotated and then all dimensions beyond the first two are dropped.

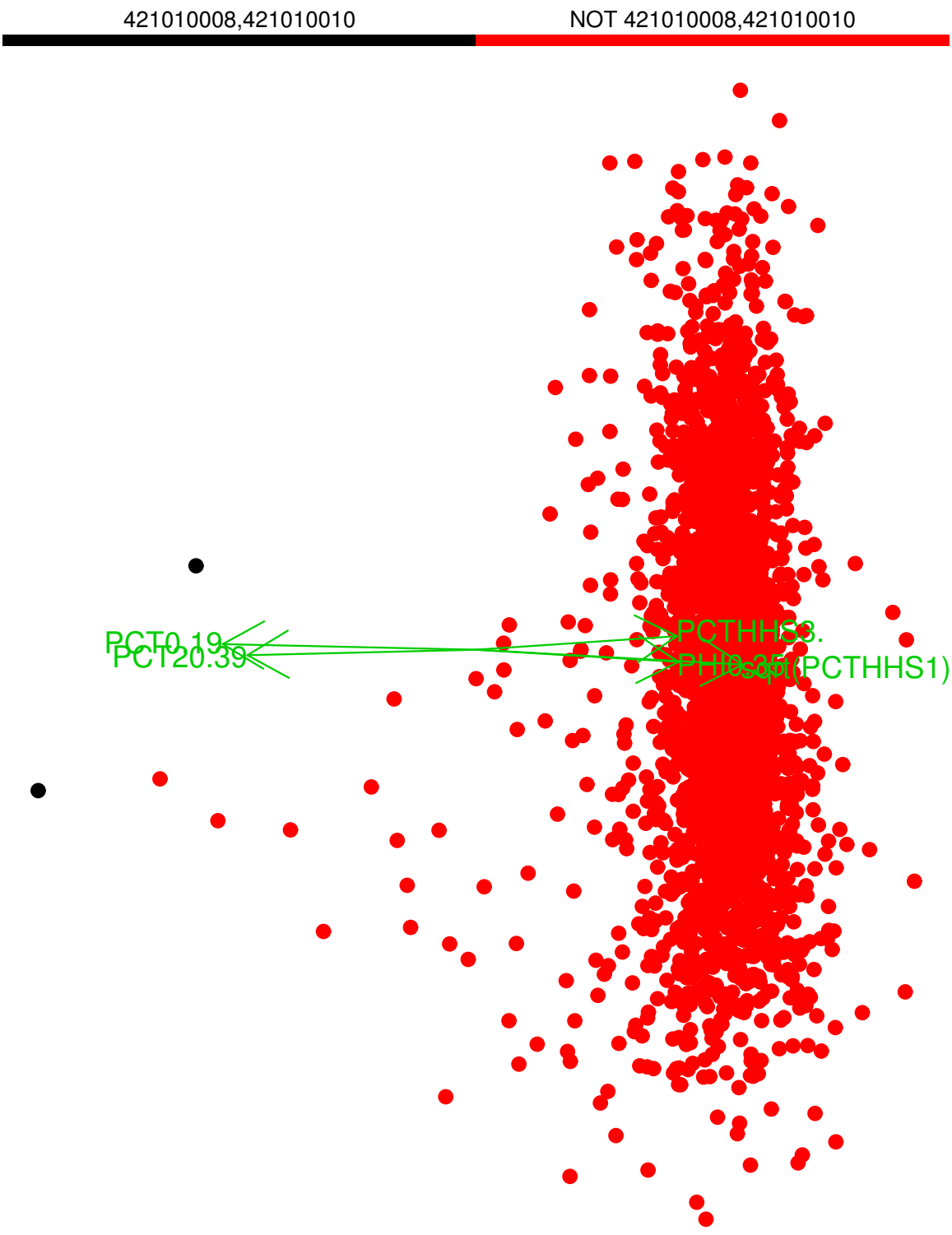
Discriminative projection—A projection which tries to separate data groups are much as possible. Typically we assume the groups have a normal distribution, so "separation" amounts to the distance between the group means, divided by the group standard deviations. This uniquely defines the projection.

To find the variables that make a data point unusual, make a discriminative projection which separates the point from the rest of the data. For explaining an outlier, first exclude the variables that were originally used to find the outlier.


421010010
log(PCI)
log(POPPSQMI)
11.0
10.5
10.0
9.5
9.0
8.5
8.0
2
4
6
8
10



Can we explain the outliers better?

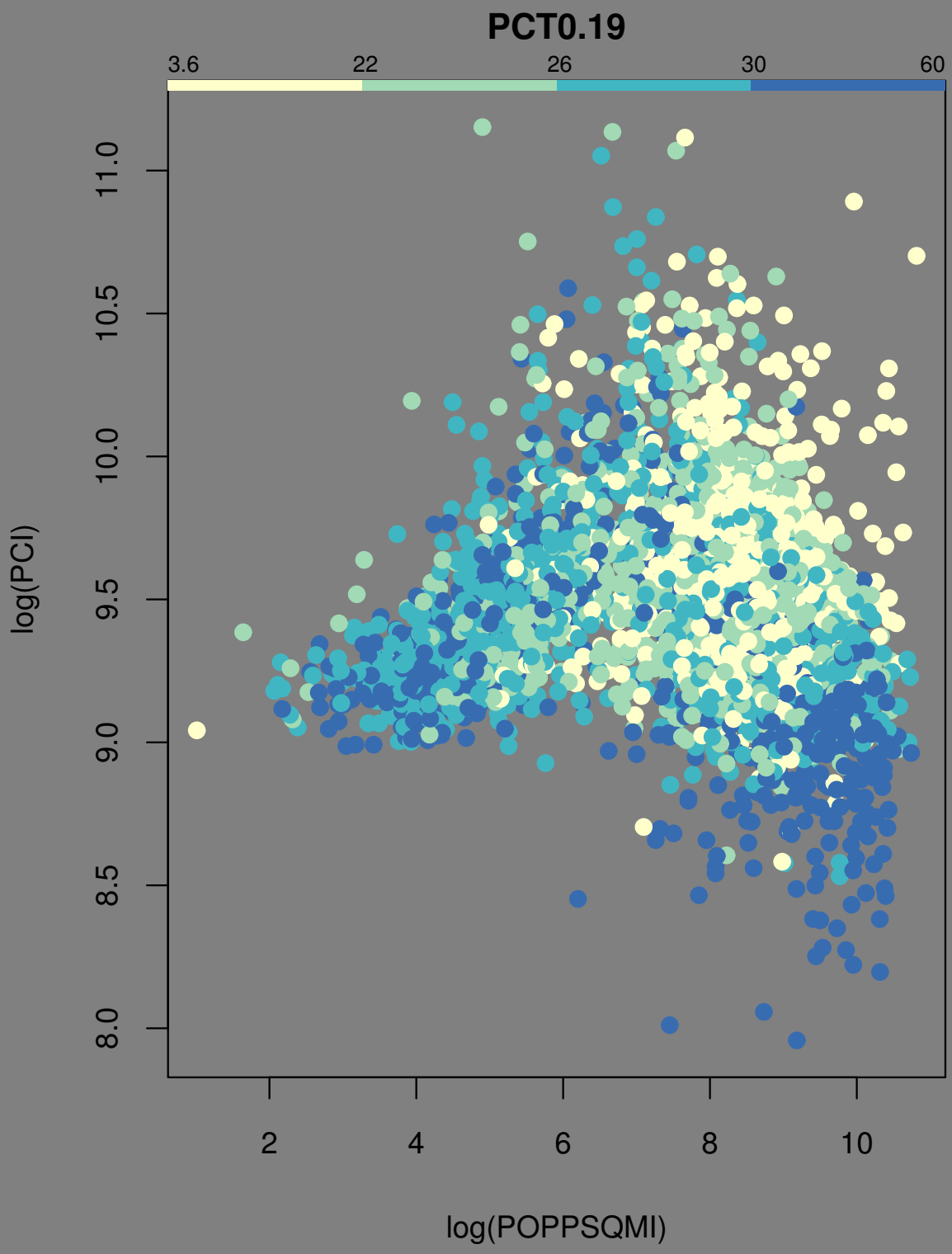


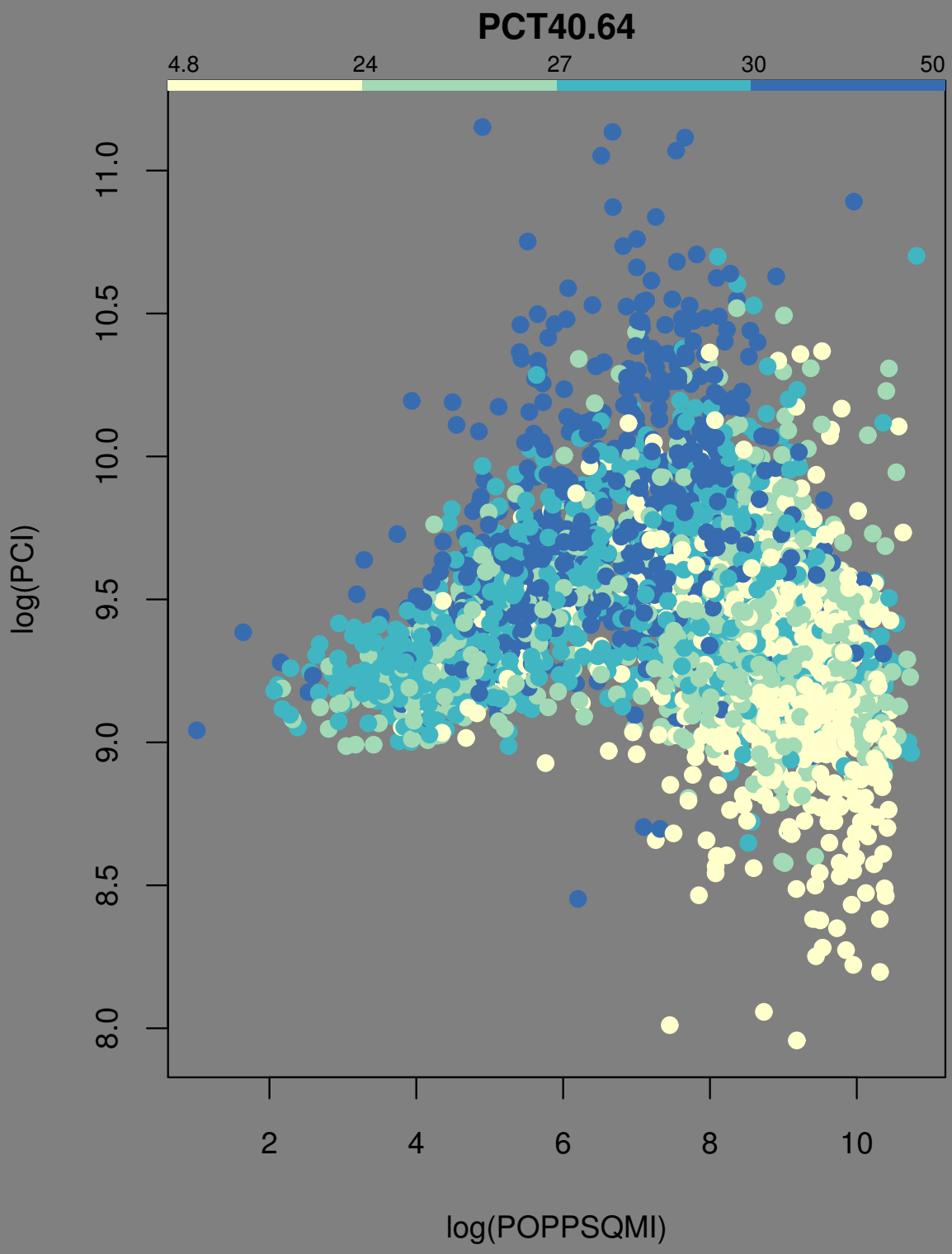
Discriminative projection

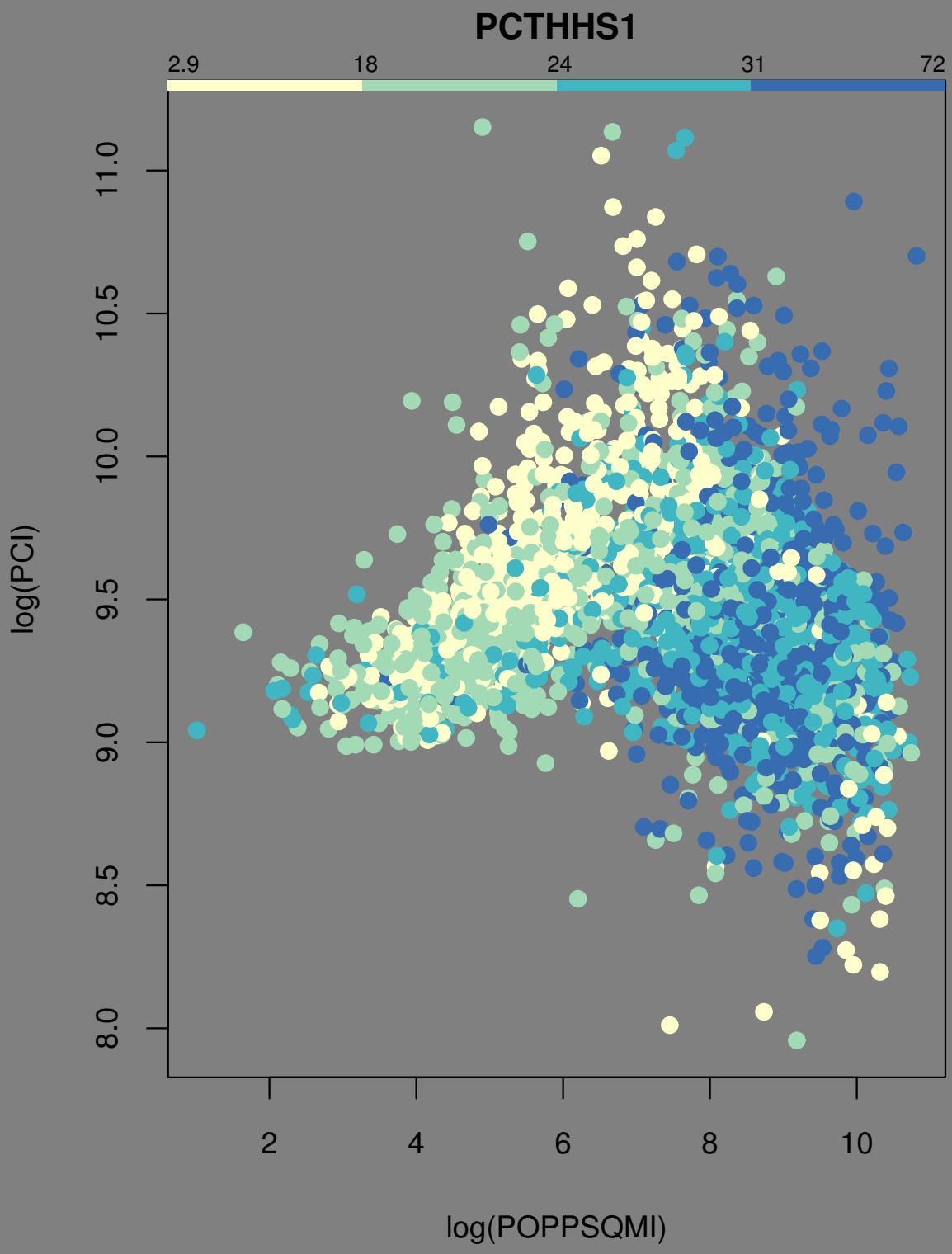
Variable weights used in the projection:

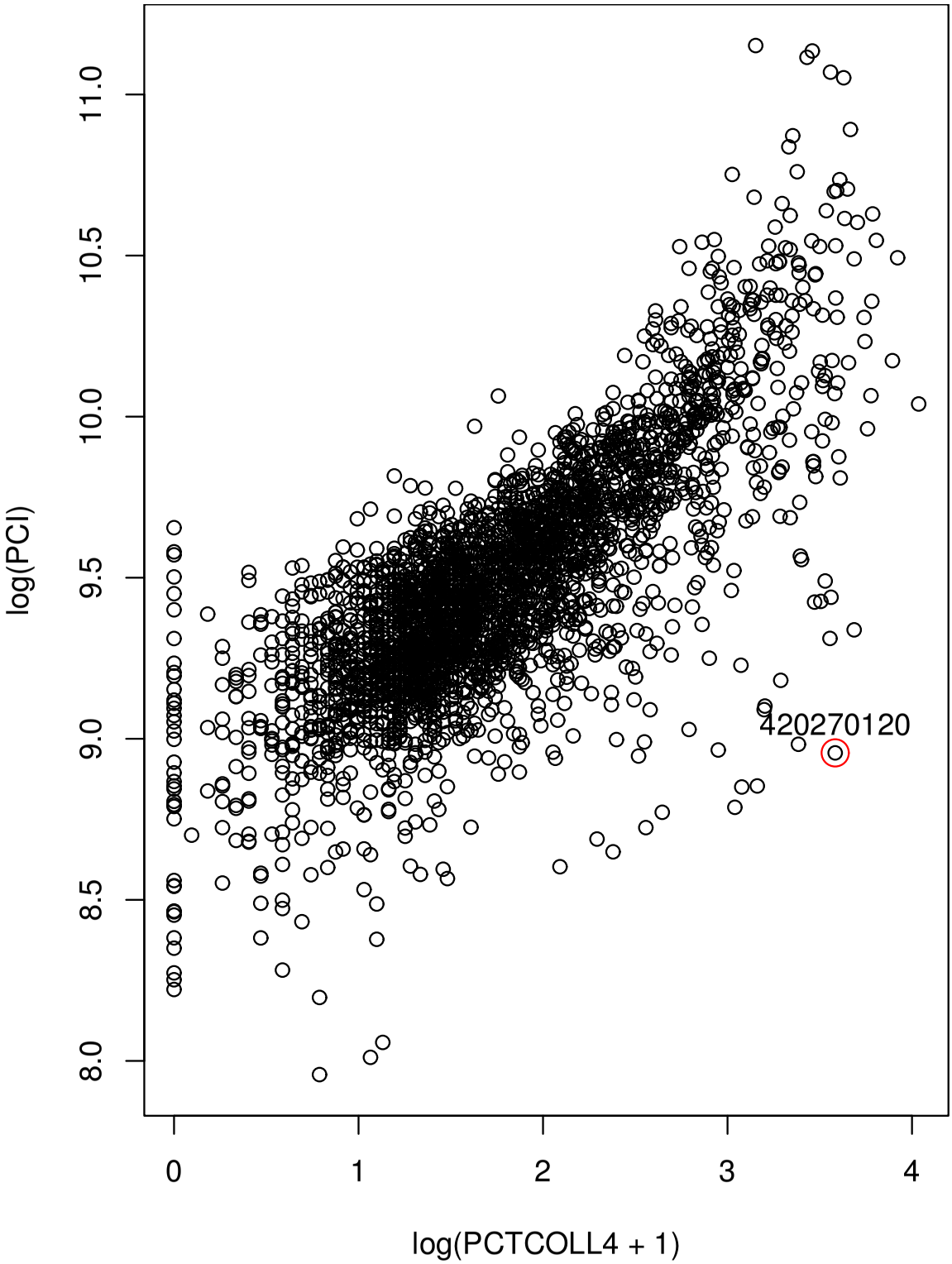
```
PCT0.19               -0.354221126
PCT20.39              -0.319680787
PCT40.64              -0.230885360
PHI0.75               -0.165780582
PCTFAMHH              -0.149873844
sqrt(PCTNFHHS)        -0.130067171
...
sqrt(PCTHHS5.)         0.156165703
sqrt(PHI0.15)          0.162485771
PCT18.24               0.228470905
PCTHHS2                0.240949045
PCTHHS3.               0.280324301
PHI0.35                0.282467035
sqrt(PCTHHS1)          0.371919142
```

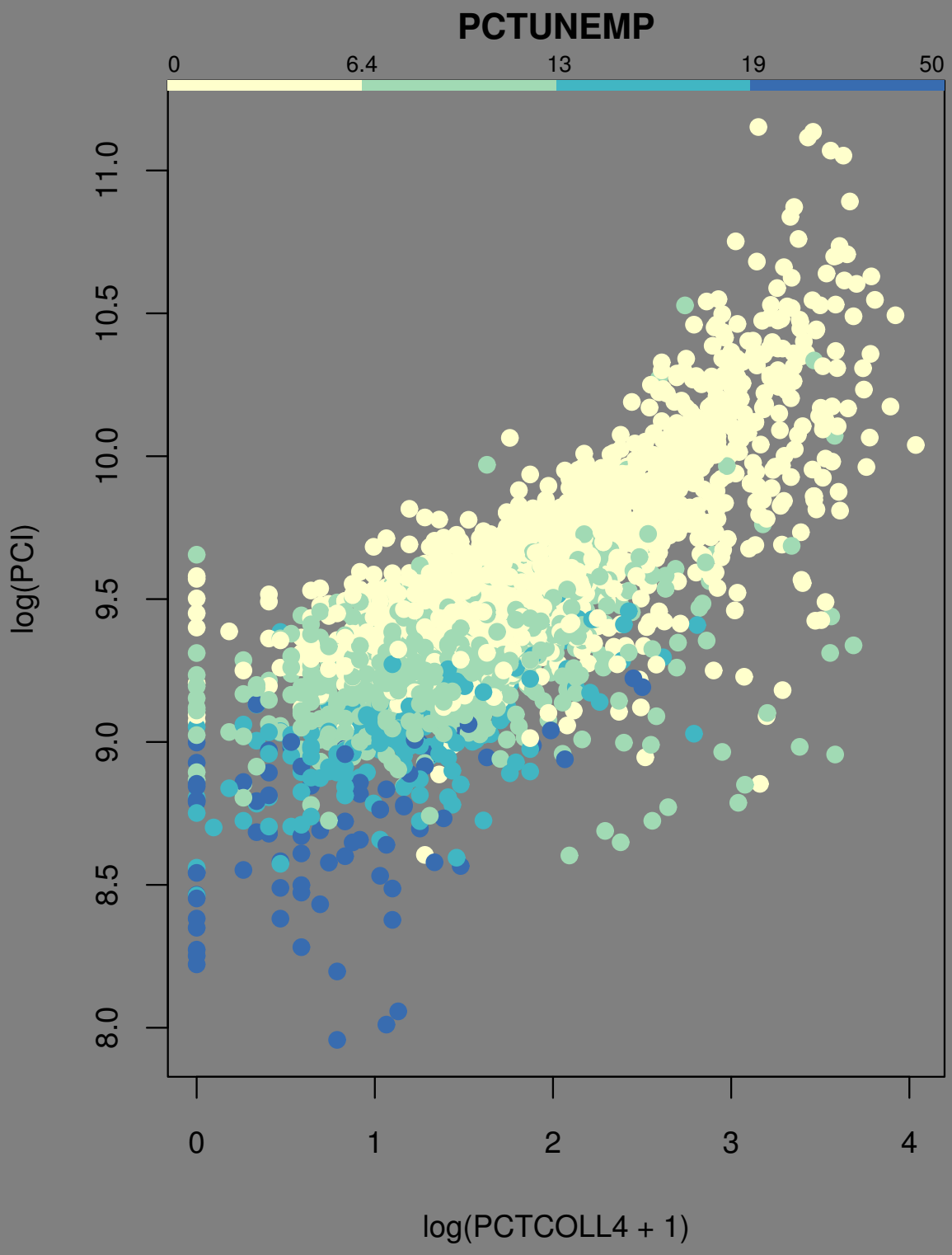
Largest weights (in magnitude) best explain the outliers


PCT0.19
3.6
22
26
30
60
log(PCI)
11.0
10.5
10.0
9.5
9.0
8.5
8.0
2
4
6
8
10
log(POPPSQMI)

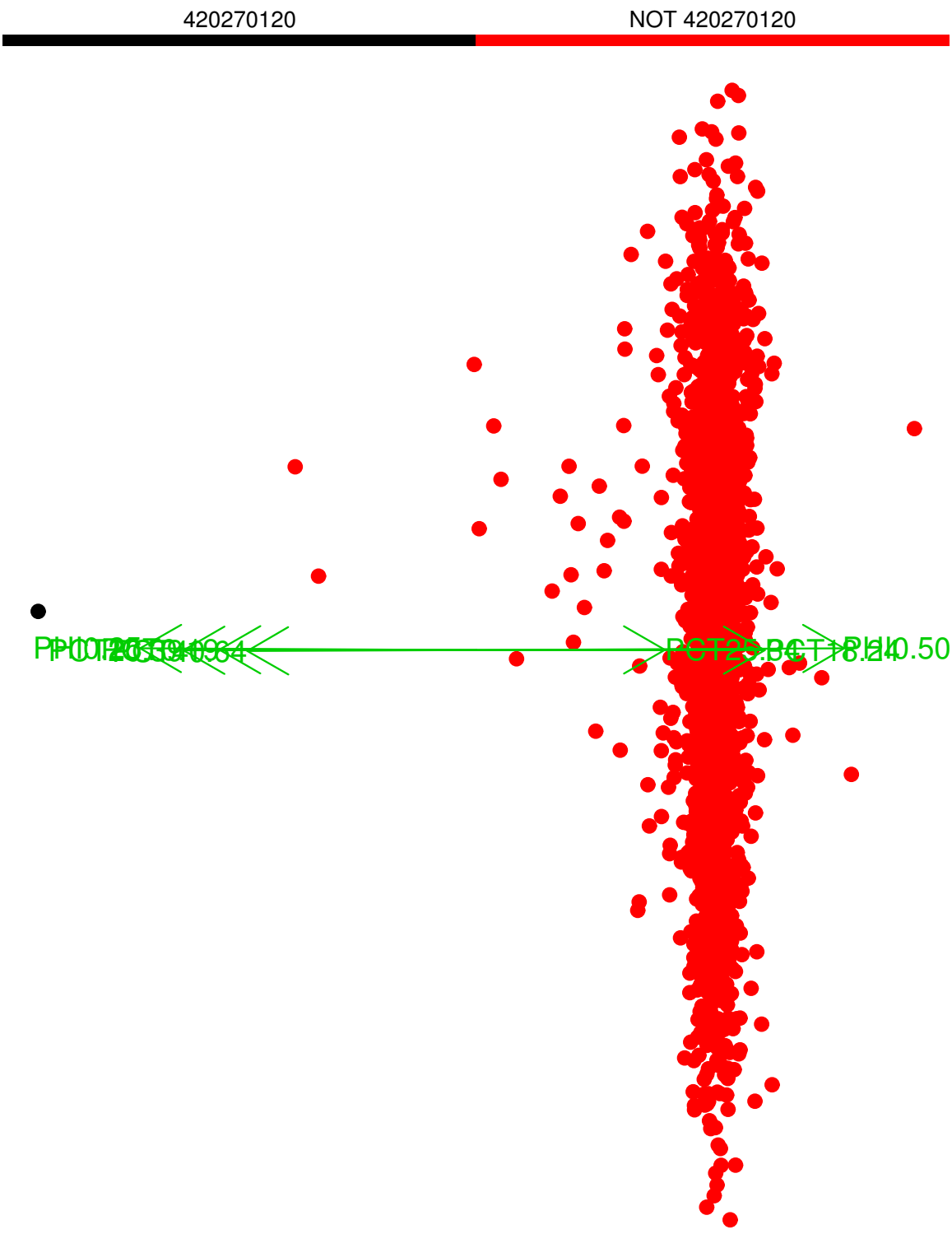

PCT40.64
4.8
24
27
30
50
log(PCI)
11.0
10.5
10.0
9.5
9.0
8.5
8.0
2
4
6
8
10
log(POPPSQMI)


PCTHHS1
2.9
18
24
31
72
log(PCI)
8.0
8.5
9.0
9.5
10.0
10.5
11.0
2
4
6
8
10
log(POPPSQMI)


log(PCI)
11.0
10.5
10.0
9.5
9.0
8.5
8.0
0
1
2
3
4
log(PCTCOLL4 + 1)
420270120



Can we explain the outliers better?

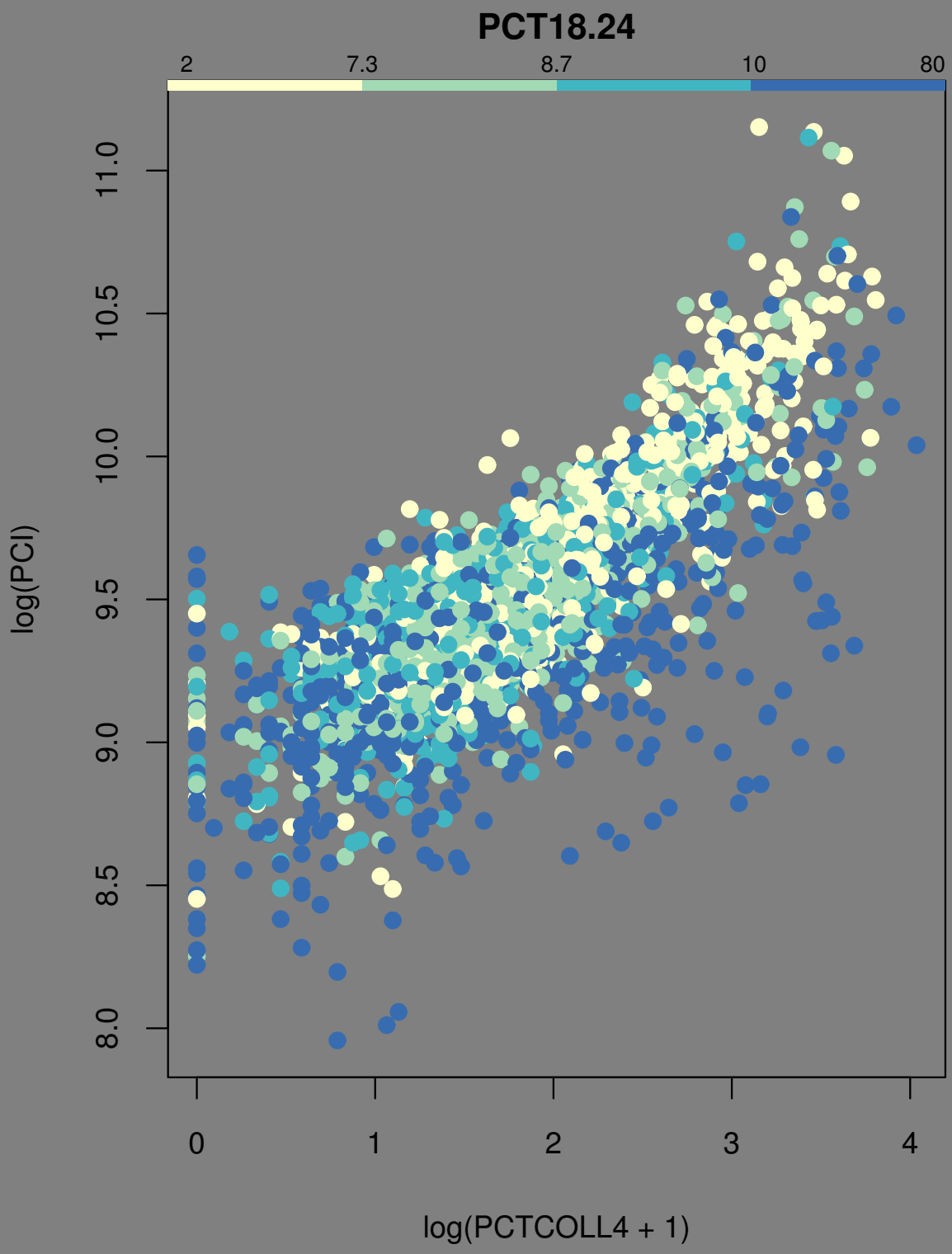


Discriminative projection

Variable weights used in the projection:

```
PHI0.25               -4.089384e-01
PCT20.39              -3.562167e-01
PCT0.19               -3.125527e-01
PCT40.64              -2.785998e-01
PHI35.50              -1.634679e-01
PHI0.35               -1.021199e-01
PHI25.35              -1.018871e-01
PCTHHS3.              -1.213641e-02
...
PCT60.64               1.006973e-01
PCT5.9                 1.044438e-01
PCT0.4                 1.058806e-01
PCT45.54               1.480137e-01
PCT35.44               1.749425e-01
PCT25.34               2.294755e-01
PCT18.24               3.509350e-01
PHI0.50                4.474431e-01
```

Largest weights (in magnitude) best explain the outliers


PCT18.24
2
7.3
8.7
10
80
log(PCI)
8.0
8.5
9.0
9.5
10.0
10.5
11.0
0
1
2
3
4
log(PCTCOLL4 + 1)