Network analysis of the Lymphome proteomics data set

Julien Chiquet

2/4/2019

Contents

Requirements

```
library(tidyverse)
library(corrplot)
library(GGally)
library(igraph)
library(ggfortify)
library(blockmodels)
library(QUIC)
library(stabs)
library(simone)
library(RColorBrewer)
pal <- brewer.pal(10, "Set3") # a fancy palette for graph
```

Basic proteomics data analysis

We first load both proteomics and covariates

```
proteomics <- readRDS("proteins.rds")
covariates <- readRDS("covariates.rds")
```

Let us have a quick look on both data frames:

Covariates

First a quick look,

```
covariates %>% head() %>% knitr::kable()
```

	disease	type	sex	age	bmi
NS001BC	BC	CO	F	60.36413	25.46938
NS001LY	LY	CO	M	40.19165	25.85463
NS002BC	BC	CA	F	60.87885	30.85938
NS002LY	LY	CA	M	39.95072	23.51556
NS003BC	BC	CO	F	50.08350	24.34175
NS003LY	LY	CO	M	40.36413	26.27840

Then a pairs plot:

```
GGally::ggpairs(covariates, columns = 3:5, aes(colour=disease))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Proteomics

Again first a quick look:

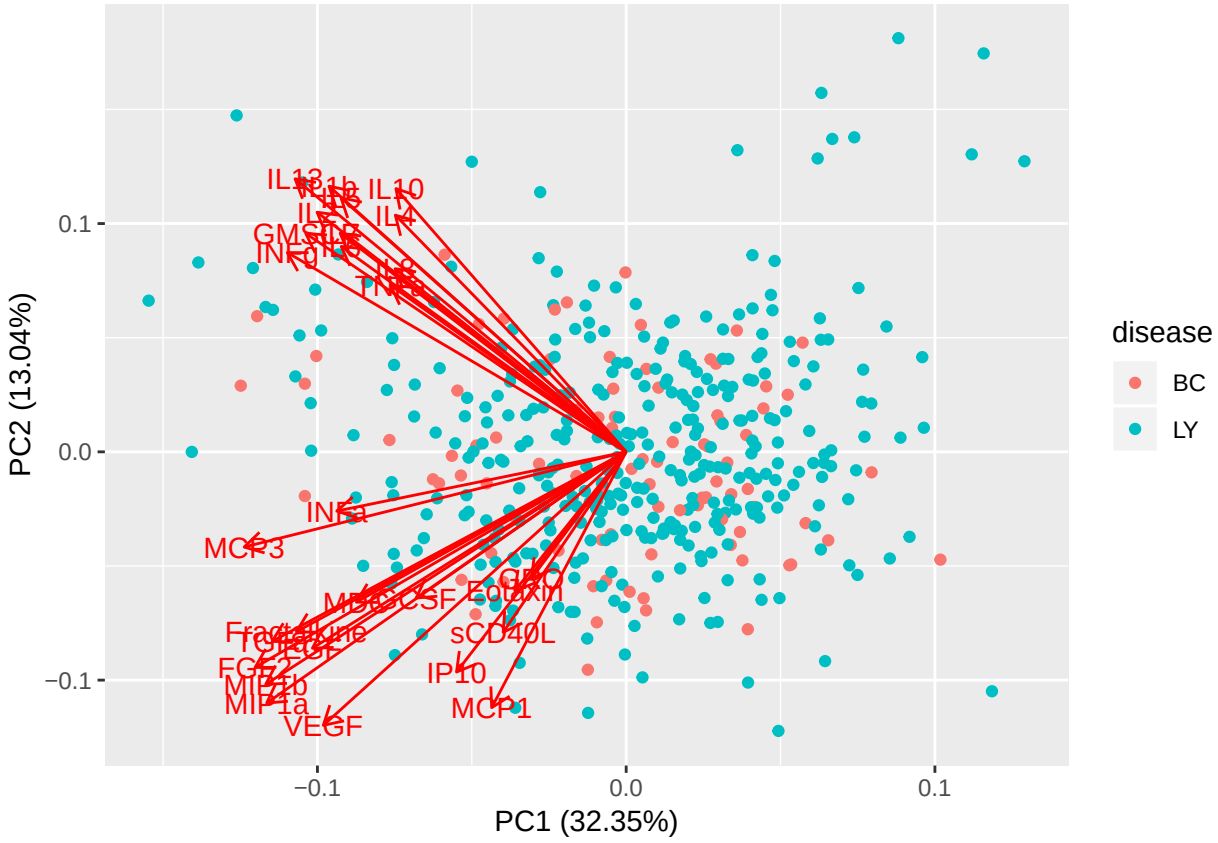
```
proteomics %>% head() %>% knitr::kable()
```

	EGF	Eotaxin	FGF2	Fractalkine	GCSF	GRO	INFa	IP10	M
NS001BC	0.0055664	-0.4618951	-1.4097618	-1.7219935	-1.0901432	0.3206727	-4.062710	-0.3409295	-0.603
NS001LY	-0.0629069	-0.3928171	0.7316576	0.4892188	0.2884048	0.5713894	3.874830	0.3249101	0.347
NS002BC	1.3193453	0.3758365	-1.8407796	2.0508717	-1.6820340	0.4068460	-4.607592	0.0947143	-0.029
NS002LY	1.1888093	0.6687950	-0.7144504	-0.7407218	-0.6017474	-0.3593785	-5.800678	-0.3458286	-0.081
NS003BC	1.3047518	0.5055310	1.2003649	-0.4681025	-0.8453924	0.0313707	-4.073082	-0.1651572	0.272
NS003LY	1.7069147	0.2284819	1.5671801	0.3511166	3.5593701	-0.3491252	3.345324	-1.0905690	-0.402

Simple PCA does not seem to show a strong structuration of the data due to the disease status:

```
pca <- prcomp(proteomics, scale. = TRUE)
for_plot <- as_tibble(proteomics) %>%
  add_column(disease = covariates$disease,
             type    = covariates$type,
```

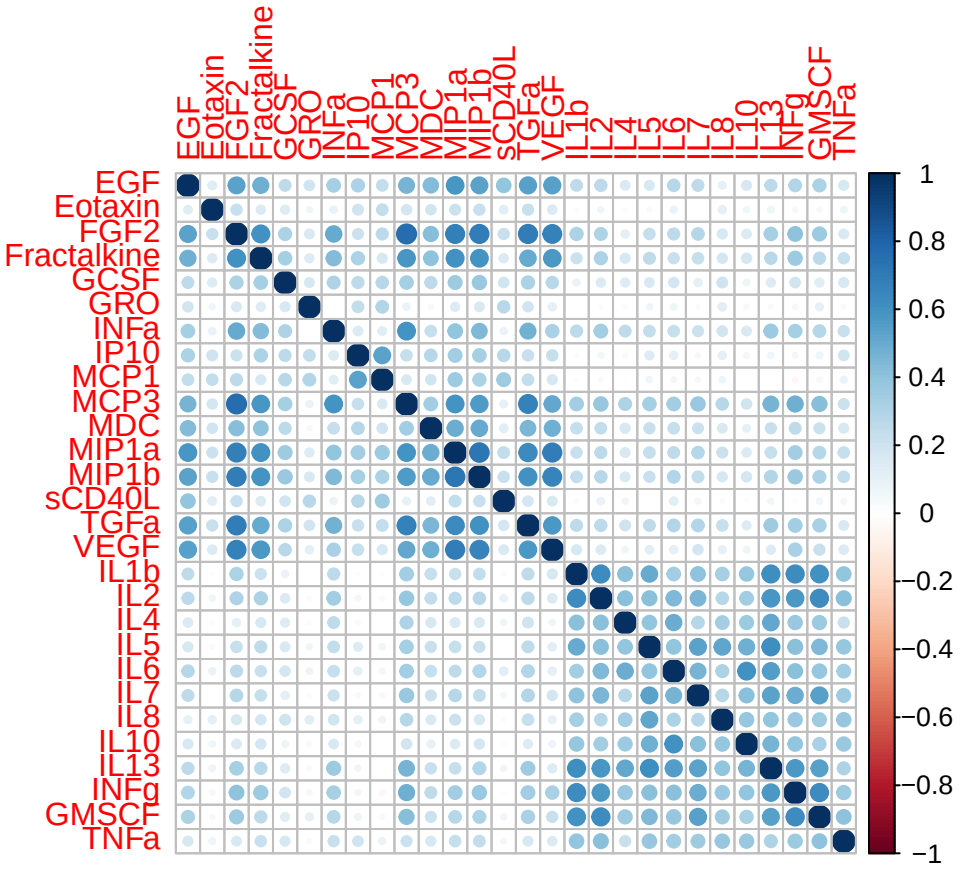
```
            sex     = covariates$sex)
autoplot(pca, loadings = TRUE, loadings.label = TRUE,
         data = for_plot, colour = 'disease')
```




```
# autoplot(pca, loadings = TRUE, loadings.label = TRUE,
#          data = for_plot, colour = 'sex')
# autoplot(pca, loadings = TRUE, loadings.label = TRUE,
#          data = for_plot, colour = 'type')
```

Still, there is some structure between the proteines:

```
corrplot(cor(proteomics))
```



Tutorial: proteomics network reconstruction on pooled data

Neighborhood selection with *simone*

We first retrieve the whole path of inferred networks with *simone*.

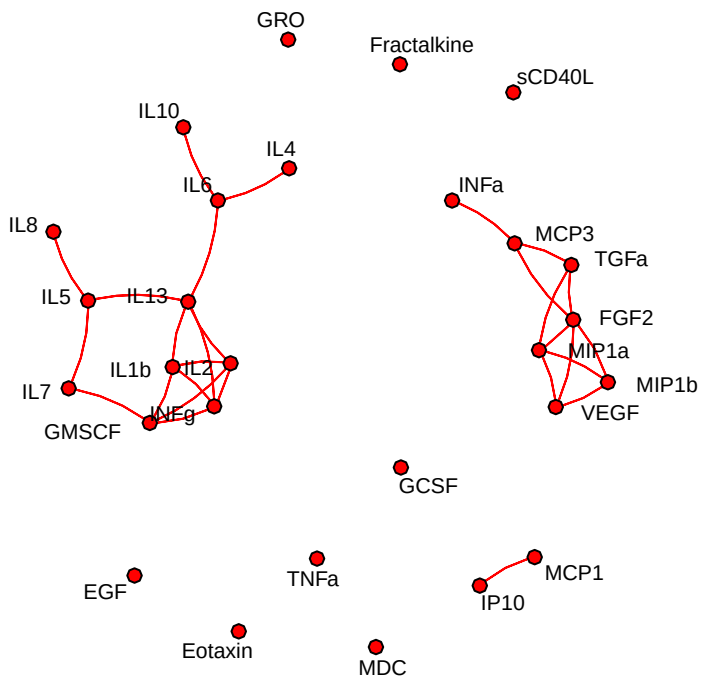
```
pooled_path <- simone(proteomics)
```

And extract one of them:

```
myNetwork <- getNetwork(pooled_path)
```

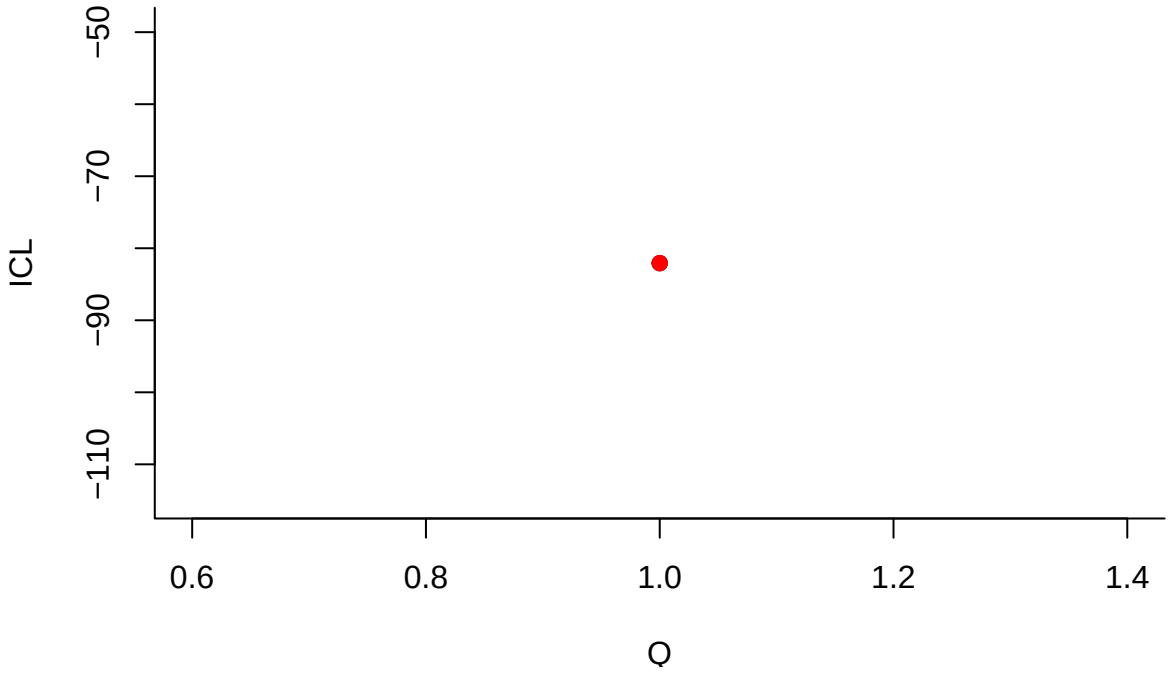
```
##
## Found a network with 28 edges.
```

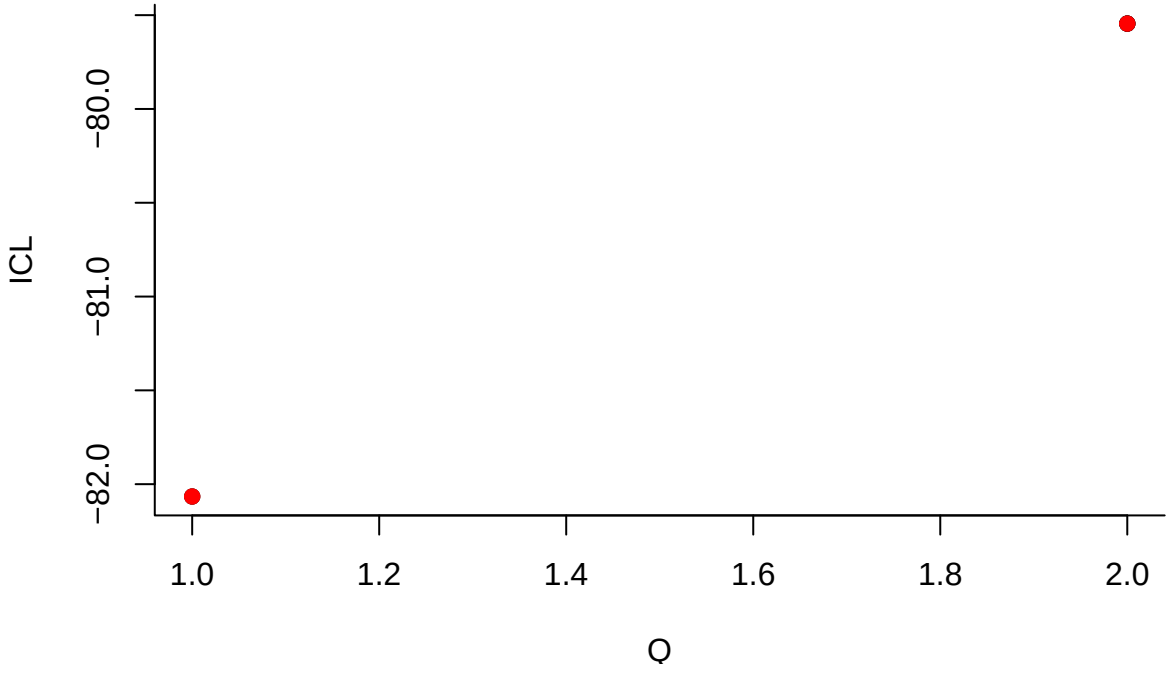
```
plot(myNetwork)
```

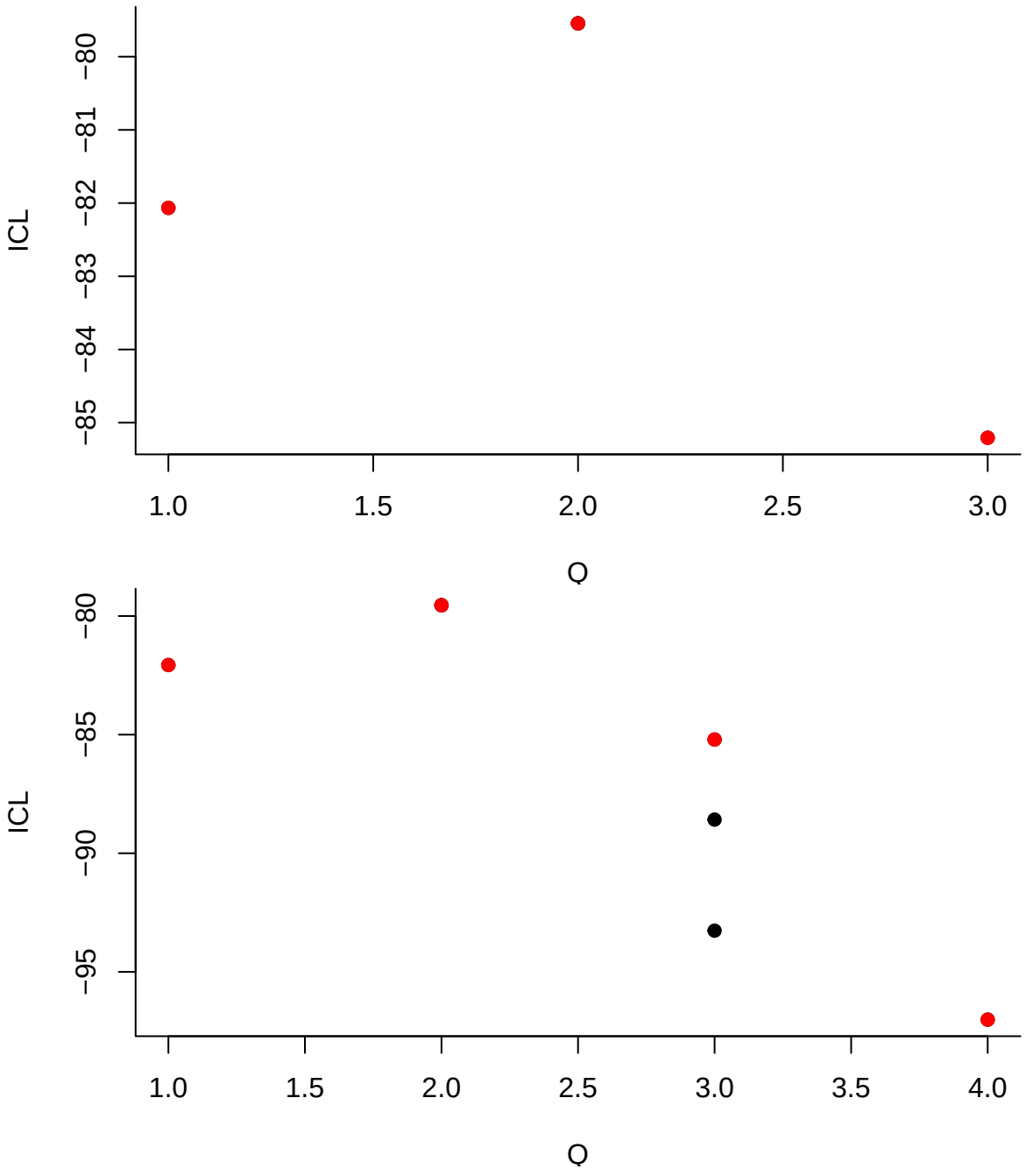


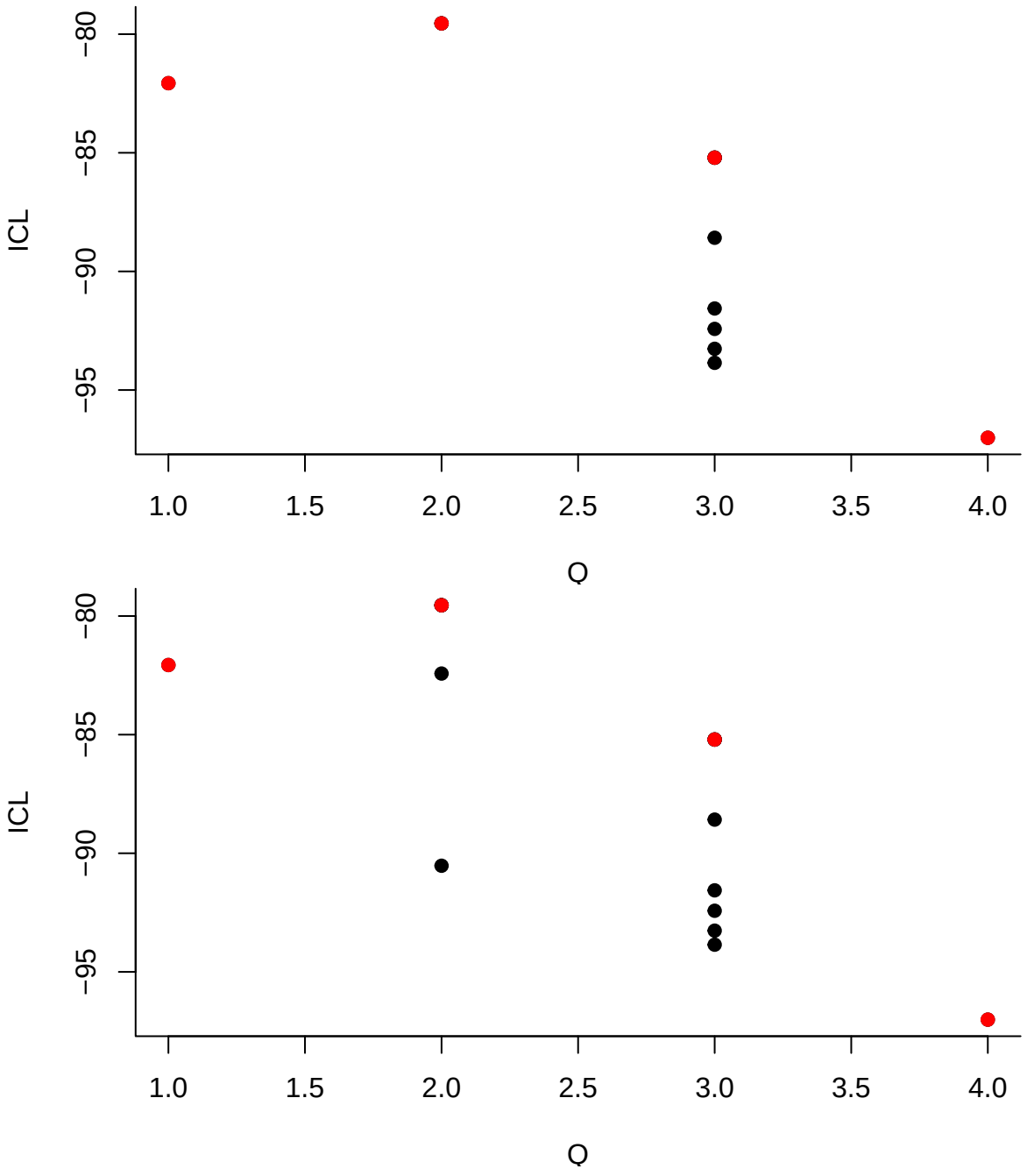
There seems to be two clusters: let us manually adjus a stochastic block model on this network

```
subNetwork <- myNetwork$A
to_remove <- rowSums(subNetwork) == 0
subNetwork <- subNetwork[!to_remove, !to_remove]
mySBM_collection <-
  BM_bernoulli(
    "SBM_sym",
    subNetwork,
    verbosity = 0
  )
mySBM_collection$estimate()
```


ICL
−50
−70
−90
−110
0.6
0.8
1.0
1.2
1.4
Q


ICL
−80.0
−81.0
−82.0
1.0
1.2
1.4
1.6
1.8
2.0
Q


ICL
−80
−81
−82
−83
−84
−85
1.0
1.5
2.0
2.5
3.0
Q
ICL
−80
−85
−90
−95
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Q



The ICL criterion choses 3 clusters.

```
clusters <- apply(mySBM_collection$memberships[[2]]$Z, 1, which.max)
```

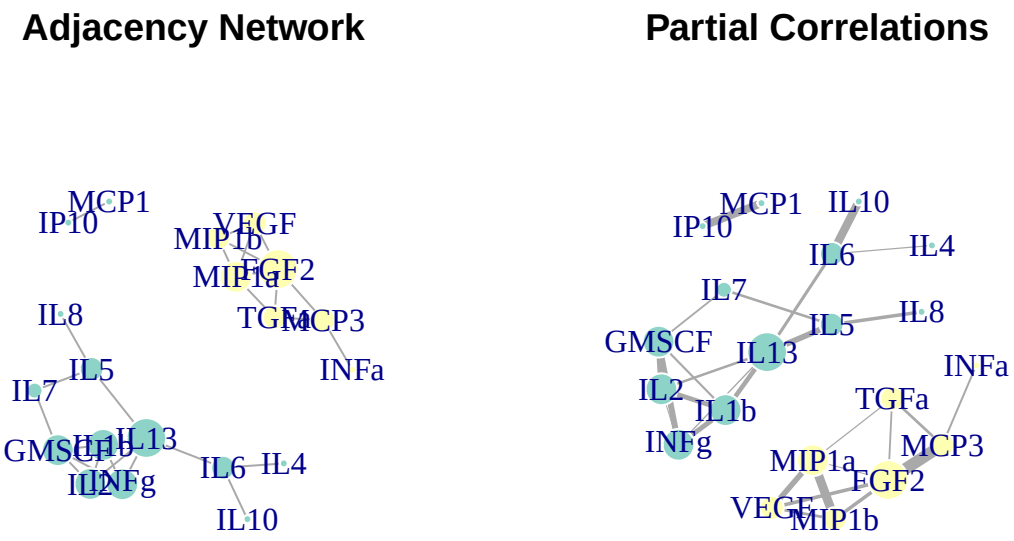
The following piece of code plots a fancy version of this graph thanks to the *igraph* package: first from the adjacency matrix, second from the partial correlation

```
g1 <- graph_from_adjacency_matrix(subNetwork, mode = "undirected", diag = FALSE)
V(g1)$class <- clusters
V(g1)$size <- degree(g1) * 5
V(g1)$frame.color <- "white"
V(g1)$color <- pal[V(g1)$class]
E(g1)$arrow.mode <- 0
```

```
partialCor <- -myNetwork$Theta
partialCor <- partialCor[!to_remove, !to_remove]

g2 <- graph_from_adjacency_matrix(partialCor, weighted = TRUE, mode = "undirected", diag = FALSE)
V(g2)$class <- clusters
V(g2)$size <- degree(g2) * 5
V(g2)$frame.color <- "white"
V(g2)$color <- pal[V(g2)$class]
E(g2)$arrow.mode <- 0

par(mfrow = c(1,2))
plot(g1, main = "Adjacency Network")
plot(g2, edge.width=20*E(g2)$weight, main = "Partial Correlations")
```

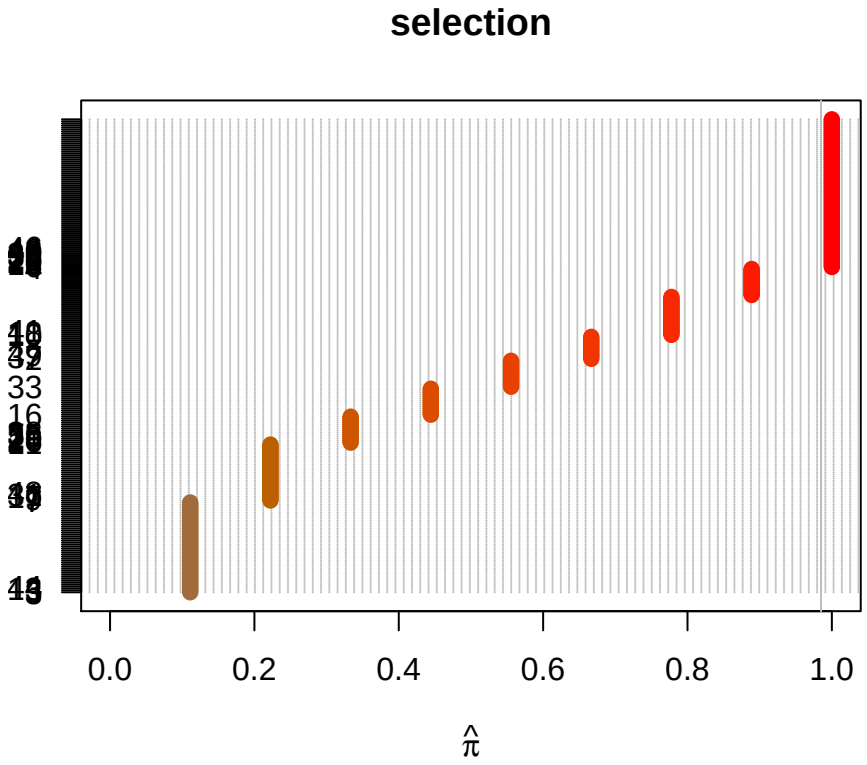


Graphical-Lasso with Stability Selection

The graphical-Lasso is efficiently implemented by the QUIC algorithm. An alternative to BIC to choose the most stables edges in the network is the stability selection approach. It subsamples the original data set, run the algorithm on each subsample and estimate the probability of selection of each edge along the path.

```
Glasso_stability <- stabsel(x = proteomics, fitfun = "quic.graphical_model", cutoff = 0.985, PFER = 5)
```

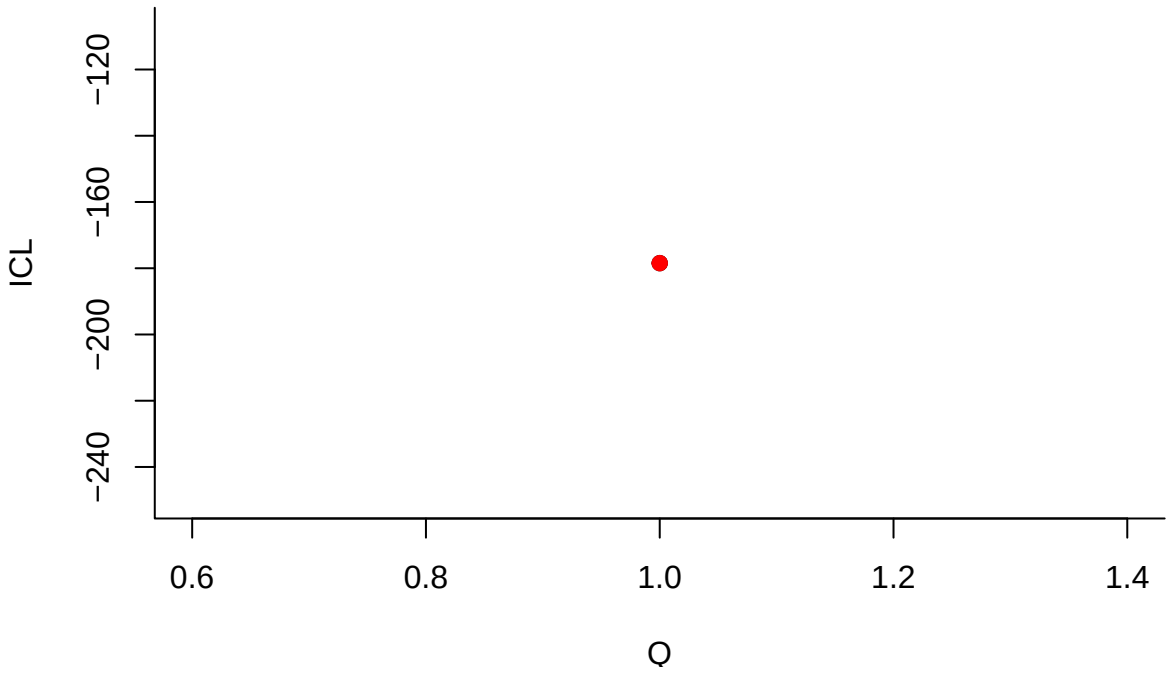
```
plot(Glasso_stability, type = "maxsel", labels = 1:50, main = "selection")
```

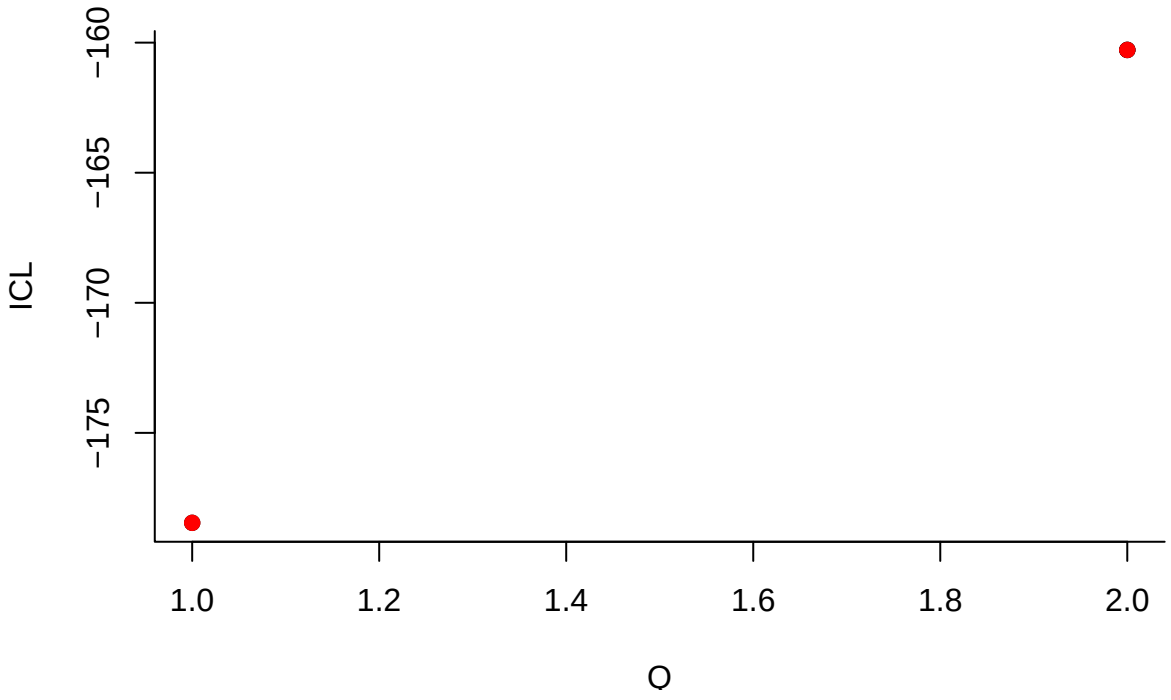


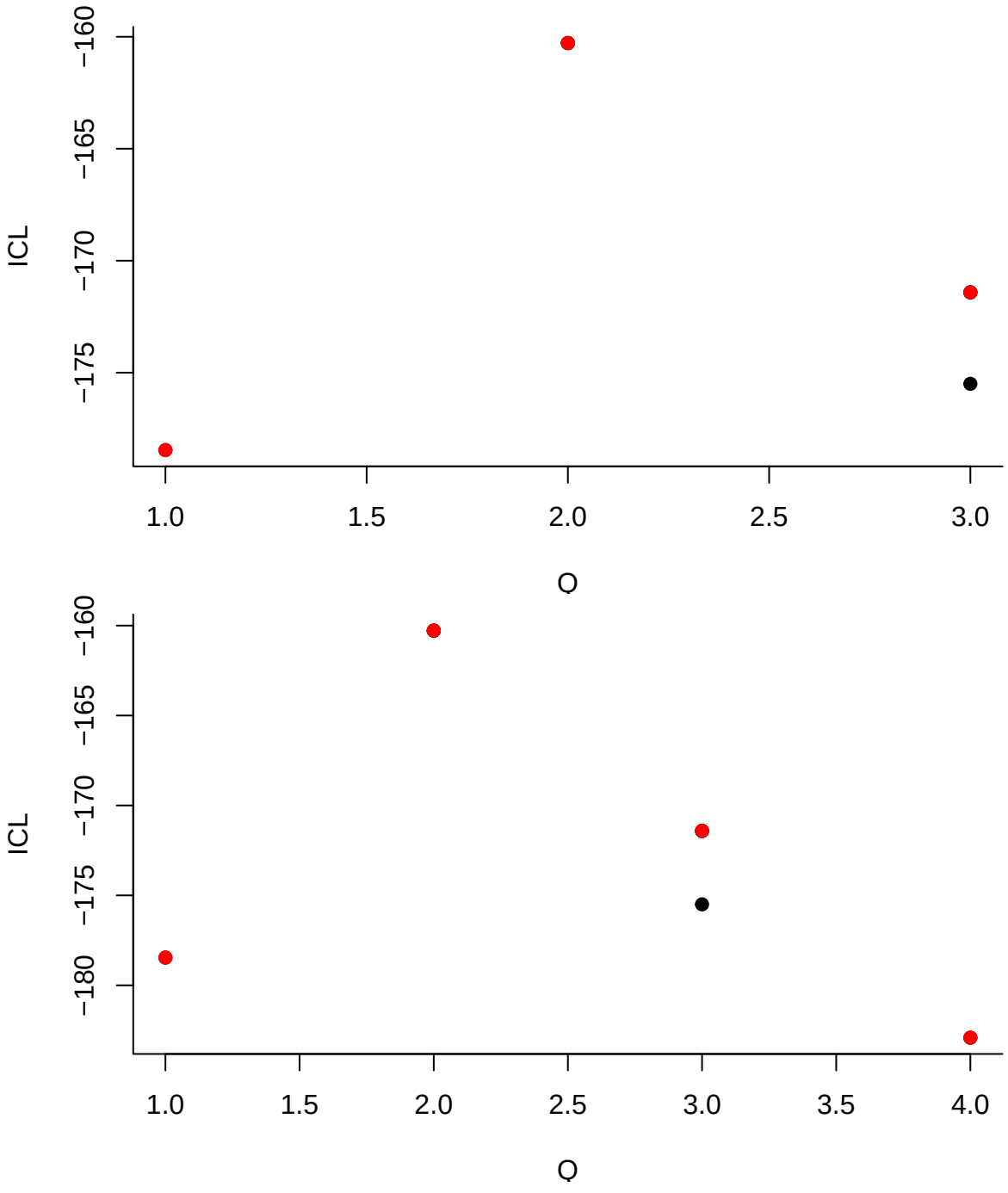

```
g_stabs <-
  graph_from_edgelist(
    do.call(rbind, strsplit(names(Glasso_stability$selected), " : ")),
    directed = FALSE
  )
```

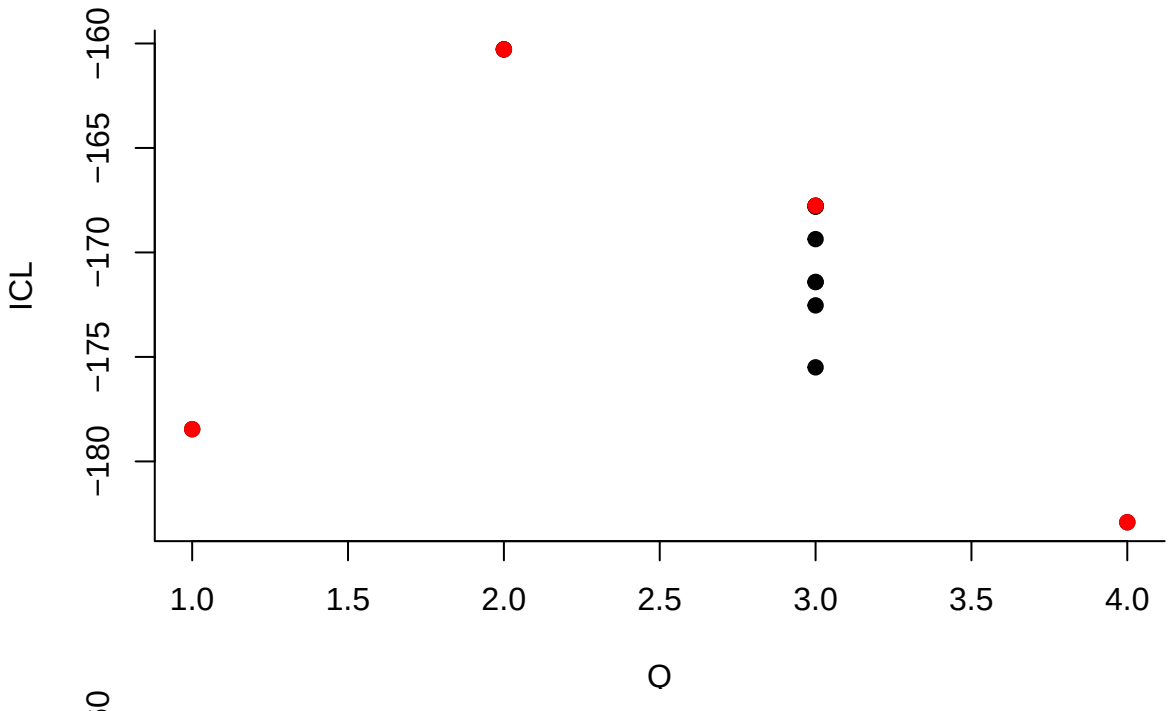
Now try to find some structure thanks to blockmodels:

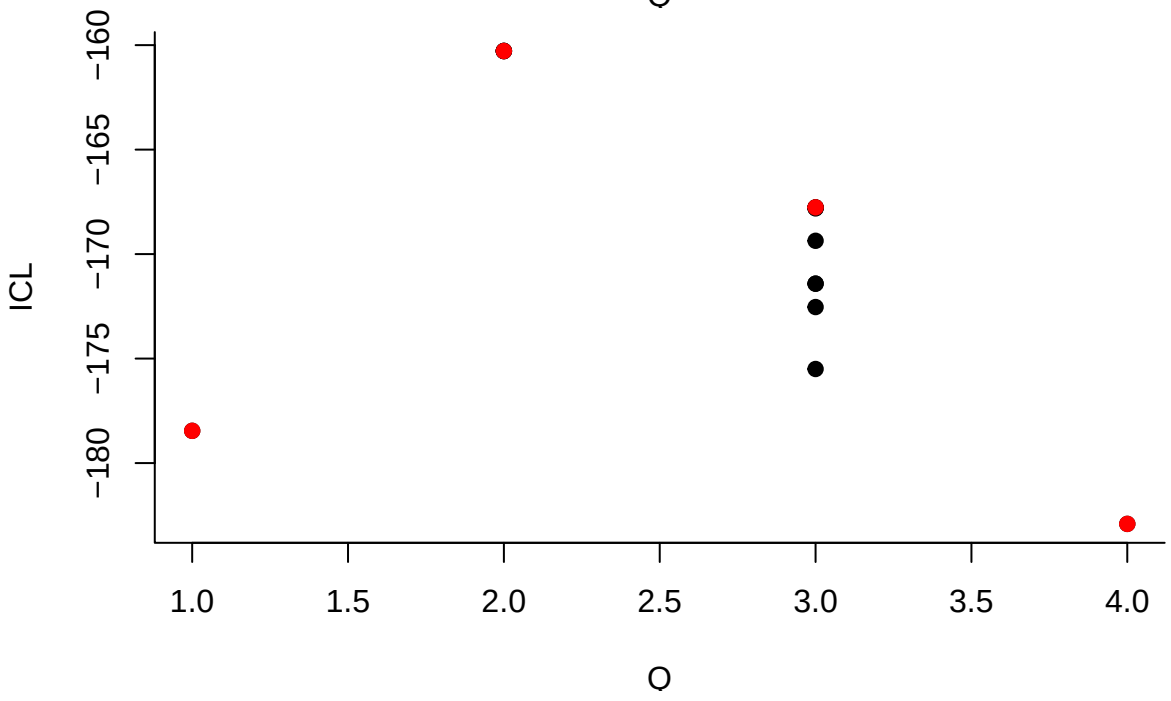
```
mySBM_collection <-
  BM_bernoulli(
    "SBM_sym",
    as_adjacency_matrix(g_stabs, sparse = FALSE),
    verbosity = 0
  )
mySBM_collection$estimate()
```

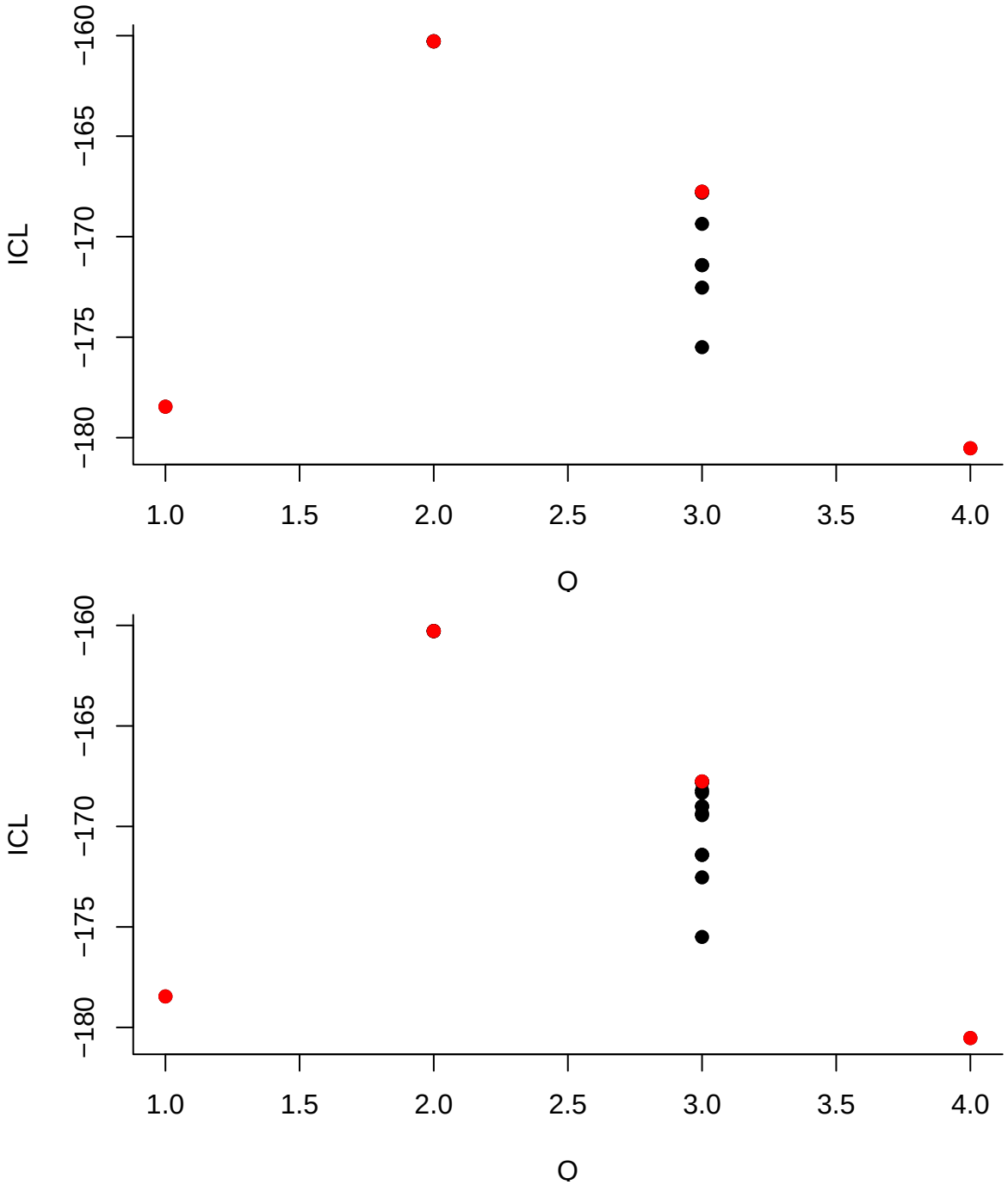

ICL
−120
−160
−200
−240
0.6
0.8
1.0
1.2
1.4
Q


ICL
−160
−165
−170
−175
1.0
1.2
1.4
1.6
1.8
2.0
Q


ICL
−160
−165
−170
−175
1.0
1.5
2.0
2.5
3.0
Q
ICL
−160
−165
−170
−175
−180
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Q


ICL
−160
−165
−170
−175
−180
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Q

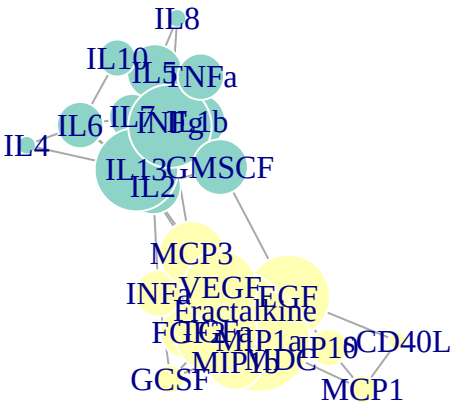

ICL
−160
−165
−170
−175
−180
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Q




```
clusters <- apply(mySBM_collection$memberships[[2]]$Z, 1, which.max)
```

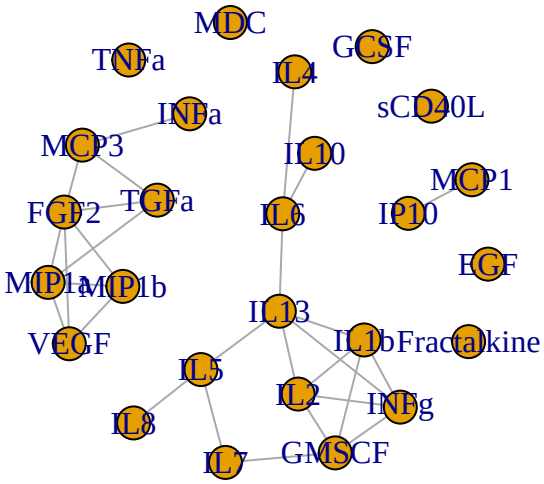
And plot it:

```
V(g_stabs)$class <- clusters
V(g_stabs)$size <- degree(g_stabs) * 5
V(g_stabs)$frame.color <- "white"
V(g_stabs)$color <- pal[V(g_stabs)$class]
E(g_stabs)$arrow.mode <- 0
plot(g_stabs)
```



What is the intersection with plot inferred with neighborhood selection and *simone*?

```
plot(igraph::intersection(g_stabs, g2))
```



Practical: multiple network analysis (accounting for patient status)

Now it is your turn: how about

1. Consensus and differential network from independent network inference

- Infer one network per patient status, with a method of your choice (e.g. *stabsel* with 'quic' fit function, or neighborhood selection with *huge* and STaRS for model selection)
- Compare these two networks according to
  - some descriptive statistics learnt during today course
  - the clusterings obtained with blockmodels
- Build a "differential" and a "consensus" network from the two network associated with each status

2. Robustness of the networks

Let us try a "home-made" stability selection for evaluating the robustness of the differential or consensus network: For each patient status, - create K subsamples with, says 90% of the original data - redo the analysis described in 1. - evaluate the stability of the differential and consensus network

2. Consensus netork with coupled network inference

- Use the *simone* multitask framework to infer the network jointly
- Try to find some edges specific to each patient status
- Compare your results with the first approach (that consider independent network inference)

- If you have time (and multiple cores on your computer. . .), evaluate the robustness of the consensus network with resampling

A.1 Optional 1: weighted penalties

You may put some prior in the network inference model by weighting the penalty associated to each edge:

To this end, build a matrix of weights which depends on the connection probability inferred by the SBM ($\propto (1 - \hat{\pi}_{ij})$) to refine the inferred networks (parameter rho in function 'quic.graphical_model').

A.2 Optional 2: multiattribute network from proteomics and transcriptomics

You may find a consensus network by coupling transcriptomics and proteomics data: first determine the transcripts associated with the set of 20 proteins. Then use the package found here:

```
devtools::install_github("jchiquet/multivarNetwork")
```