Model Searches

KMG & STL

May 22, 2019

Set up Environment

```r
library(mvtnorm)
library(gimme)
```

```
## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2019a.
## 1.0/zoneinfo/America/New_York'
```

```r
library(perturbR)
```

Data generating function

```r
genData <- function(A = NULL,
                    Phi = NULL,
                    obs = 200,
                    n  = 10 ){
  data2ls <- list()
  for (p in 1:n)
  {
    negA1 <- diag(5)-A

    time  <- matrix(0, nrow = 5, ncol = obs+400)

    time1 <- matrix(0, nrow = 5, ncol = obs+400)

    noise <- solve(negA1, matrix(rnorm(5*(obs+400),0,1), nrow = 5, ncol = obs+400))

    time[,1] <- noise[,1]

    time1[,1] <- solve(negA1, Phi) %*% time[,1] + noise[,1]

    time[,2]  <- time1[,1]

    for (i in 2:(obs+400)){
      time1[,i]  <- solve(negA1, Phi) %*% time[,(i)] + noise[,i]
      if (i<(obs+400))
        time[,(i+1)] <- time1[,i]
    }
    data2ls[[p]] <- t(time[,400:600])
    names(data2ls)[p]<- paste0('ind', p)
  }
  return(gen_data = data2ls)
}
```

Data Generation

Now we'll generate data where half of the individuals have one pattern of relations and the other half have a couple of differences in their patterns.

Specifically, pattern A1 has variable 1 predict variable 4 contemporaneously and A2 variable 3 predict variable 5 contemporoeusly. All other relations are the same between the two, with the exception that variable 5 predicted by 4 is positive for on subset of the data (A1) and negative for the other (see A2).

```r
# Generate first pattern #
A1 <- matrix(
    c(0, 0, 0, 0,  0,
      .7,  0, 0, 0,  0,
      0, .7, 0, 0,  0,
      .7,  0, 0, 0, 0,
      0,  0, 0, .7,  0), nrow = 5, ncol = 5, byrow = TRUE)

  Phi1 <- matrix(
    c(.5,   0, 0,  0,   0,
      0,  .5, 0,  0,   0,
      0,   0, .5,  0,   0,
      0,   0, 0, .5,   0,
      0,   0, 0,  0,  .5), nrow = 5, ncol = 5, byrow = TRUE)

Data1 <- genData(A = A1, Phi = Phi1, obs = 200)

# Generate second pattern #
A2 <- matrix(
  c(0, 0, 0, 0,  0,
    .7,  0, 0, 0,  0,
    0, .7, 0, 0,  0,
    0,  0, 0, 0, 0,
    0,  0, .7, -.7,  0), nrow = 5, ncol = 5, byrow = TRUE)

Phi2 <- matrix(
  c(.5,   0, 0,  0,   0,
    0,  .5, 0,  0,   0,
    0,   0, .5,  0,   0,
    0,   0, 0, .5,   0,
    0,   0, 0,  0,  .5), nrow = 5, ncol = 5, byrow = TRUE)

Data2 <- genData(A = A2, Phi = Phi2, obs = 200)

# combine data #
Data_all <- append(Data1, Data2)
# make sure each individual has a unique name
names(Data_all) <- paste0('ind', seq(1,length(Data_all)))
```

Investigate modification indices (MIs)

Let's select one individual and see what happens when we start with a null model wherein only the autoregressive effects are estimated.

Do the MIs that are significant make sense?

```
# Take one participant and time-embed the data
data_test <- Data_all[[1]]
data_test2 <- embed(data_test,2)
# The first variables are time at zero lag, the second set are the
# lag-1 variables. Rearrange.
data_test3 <- matrix(,200,10)
data_test3[,1:5] <- data_test2[,6:10]
data_test3[,6:10] <- data_test2[,1:5]

# This model only has AR effects
modelAR <- '
V1 ~ 0*V6
V2 ~ 0*V7
V3 ~ 0*V8
V4 ~ 0*V9
V5 ~ 0*V10
V6 ~ V1
V7 ~ V2
V8 ~ V3
V9 ~ V4
V10 ~ V5

'
# lavaan will only give MIs for variables that are dependent variables
# in an equation. So we put nonsense paths with "0*" (Stephanie Lane convention).

fit <- lavaan::sem(modelAR, data_test3)

# Get MIs
mi <- lavaan::modindices(fit)
mi[mi$op == "~",][41:80,] # select those where time is predictede (remove time-1)
```

```
##     lhs op rhs     mi    epc sepc.lv sepc.all sepc.nox
## 106  V6  ~  V2  0.050  0.010   0.010    0.013    0.013
## 107  V6  ~  V3  0.022  0.005   0.005    0.009    0.009
## 108  V6  ~  V4  0.186  0.017   0.017    0.026    0.026
## 109  V6  ~  V5  0.237  0.014   0.014    0.029    0.029
## 110  V6  ~  V7 25.339  0.216   0.216    0.301    0.301
## 111  V6  ~  V8  6.735  0.080   0.080    0.155    0.155
## 112  V6  ~  V9 45.887  0.266   0.266    0.405    0.405
## 113  V6  ~ V10 15.081  0.114   0.114    0.232    0.232
## 114  V7  ~  V1  9.352  0.215   0.215    0.155    0.155
## 115  V7  ~  V3  0.069 -0.010  -0.010   -0.013   -0.013
## 116  V7  ~  V4  1.578  0.058   0.058    0.064    0.064
## 117  V7  ~  V5  0.683  0.029   0.029    0.042    0.042
## 118  V7  ~  V6 54.482  0.521   0.521    0.375    0.375
## 119  V7  ~  V8 29.957  0.201   0.201    0.278    0.278
## 120  V7  ~  V9 24.733  0.231   0.231    0.252    0.252
## 121  V7  ~ V10 10.148  0.111   0.111    0.162    0.162
## 122  V8  ~  V1 24.597  0.424   0.424    0.221    0.221
## 123  V8  ~  V2  9.548  0.191   0.191    0.138    0.138
## 124  V8  ~  V4 10.076  0.179   0.179    0.142    0.142
## 125  V8  ~  V5  5.618  0.100   0.100    0.106    0.106
## 126  V8  ~  V6 35.617  0.514   0.514    0.266    0.266
## 127  V8  ~  V7 74.029  0.532   0.532    0.384    0.384
## 128  V8  ~  V9 19.922  0.252   0.252    0.199    0.199
## 129  V8  ~ V10 13.249  0.154   0.154    0.163    0.163
## 130  V9  ~  V1  4.287  0.163   0.163    0.108    0.108
## 131  V9  ~  V2  1.931  0.079   0.079    0.072    0.072
## 132  V9  ~  V3  0.903  0.039   0.039    0.049    0.049
## 133  V9  ~  V5  0.009 -0.004  -0.004   -0.005   -0.005
## 134  V9  ~  V6 73.173  0.678   0.678    0.445    0.445
## 135  V9  ~  V7 24.882  0.284   0.284    0.260    0.260
## 136  V9  ~  V8  7.411  0.112   0.112    0.142    0.142
## 137  V9  ~ V10 34.321  0.228   0.228    0.305    0.305
## 138 V10  ~  V1 17.582  0.368   0.368    0.182    0.182
## 139 V10  ~  V2  5.300  0.146   0.146    0.100    0.100
## 140 V10  ~  V3  1.992  0.065   0.065    0.061    0.061
## 141 V10  ~  V4  7.321  0.157   0.157    0.118    0.118
## 142 V10  ~  V6 49.723  0.623   0.623    0.307    0.307
## 143 V10  ~  V7 20.800  0.289   0.289    0.198    0.198
## 144 V10  ~  V8  8.485  0.133   0.133    0.127    0.127
## 145 V10  ~  V9 79.755  0.518   0.518    0.388    0.388
```

The MIs point us to paths that, if added, would significantly improve the model fit. We see here that the MIs true (i.e., data-generating) paths tend to be high. For instance, V10~V9 is variable 5 at time regressed on variable 4 at time, which is in both groups' models.

Sequentially adding paths

One way to build a model is to start by adding the path with the highest MI for that individual, re-estimating this model to get new MIs, selecting the next path to add from these MIs, and so on until the model is considered a good fit.

This is what the function "indSEM" does in the *gimmme* R package. It does individual-level model searches in this manner and then provides summative results (as well as individual-level estimates).

Let's explore.

Results from offering no information

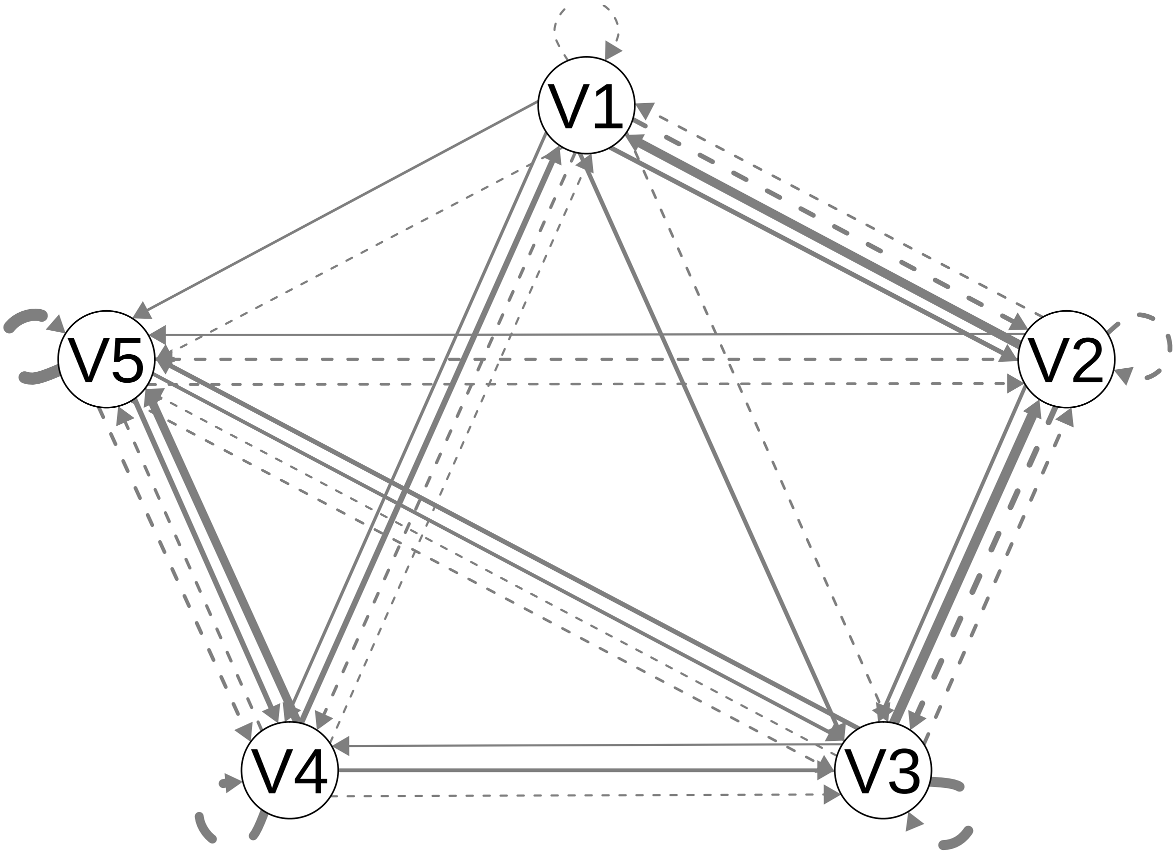
In what follows we provide no starting information and let the model search start from a null model where no paths are estimated - not even the AR effects.

```
ind_out <-indSEM(data = Data_all,
                 ar = FALSE)
```

The plot below depicts the results. The width of the lines indicates the proportion of individuals for whom the path was found to exist in their model.

```
plot(ind_out)
```

```
## Please specify a file id for individual plots. Otherwise, summary plot is presented.
```



We see a lot of variability in the paths, despite there being only 2 patterns for all individuals. It seems we have a lot of spurious relations here.

Results from starting with AR effects estimated

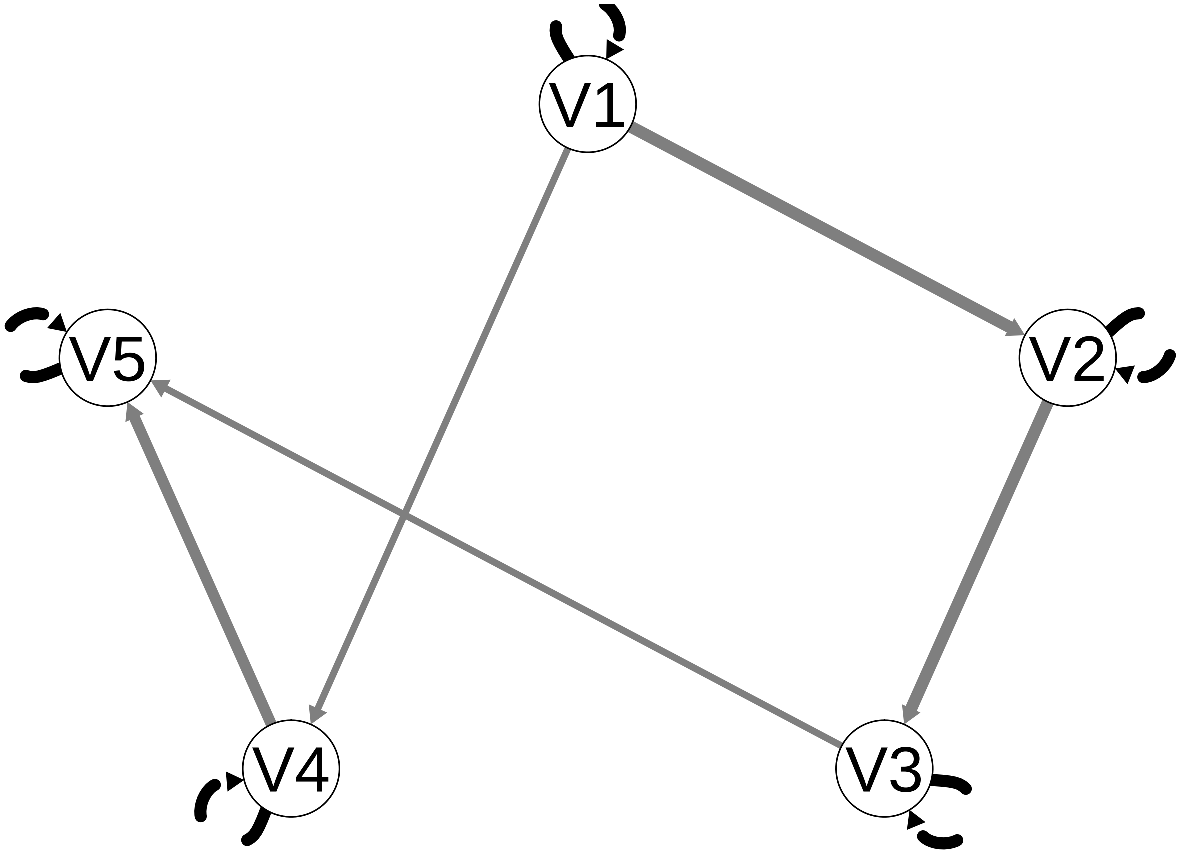
Now let's give the model some information. If we start with the AR effects estimated, the search procedure will probably work better. This is because by starting with the AR effects estimated then the directionality of the relation can be obtained.

```
ind_out_AR <-indSEM(data = Data_all,
                    ar = TRUE)
```

Now this looks much better. We see that only the paths that were in the data-generating structures are present here. Note that the AR effects (the dashed recursive lines on each variable) are black. This is because they are estimated for each individual.

```
plot(ind_out_AR)
```

```
## Please specify a file id for individual plots. Otherwise, summary plot is presented.
```

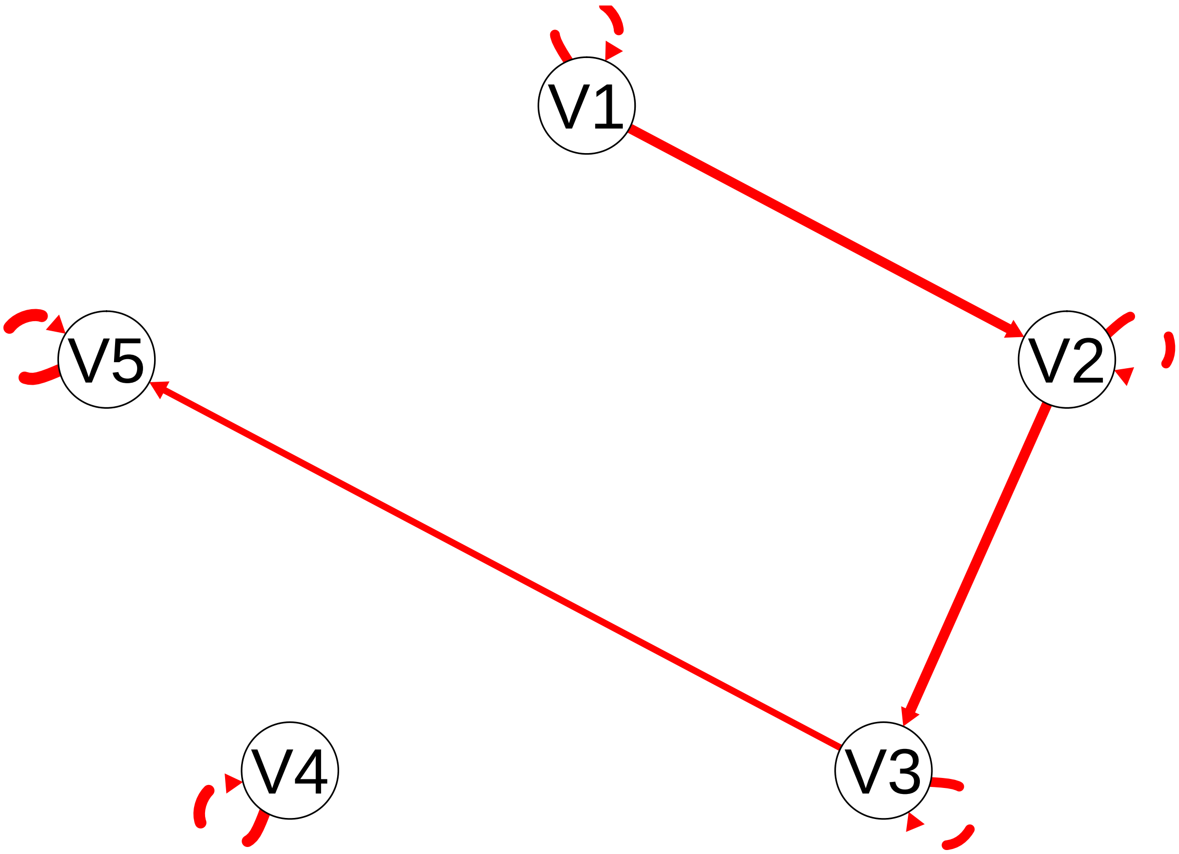

V1
V2
V3
V4
V5

What would happen if we aggregated the data? We can concatenate the data and arrive at results. This is similar to averaging lag-zero correlation matrices, a common practice in functional MRI research.

In the function aggSEM, paths are iteratively added according to the MIs.

Note that aggregating in this way is not endorsed by the gimme creators.

```r
agg_out <- aggSEM(data = Data_all)
plot(agg_out)
```



We see that the path from V4 to V5 is missing. This path was positive for half the individuals and negative for the other half.

The approach did appropriately find the paths fro V1 to V2 and V2 to V3. It included one of the subgroup-level paths.

GIMME

The above investigations have suggested a few things. 1. When the AR effects are not modeled spurious relations may emerge. 2. When directed contemporaneous relations exist, they might surface as lagged.

Oftentime individual-level searches lead to spurious relations as well. This is because the starting model that contains only AR effects might be far from the final model, and one wrong path selection early on will have deletrious downstream effects.

GIMME was developed to help individual-level model searches by first looking for relations that are consistent across people. It then starts the individual-model searches with these paths included.

Running basic gimme.

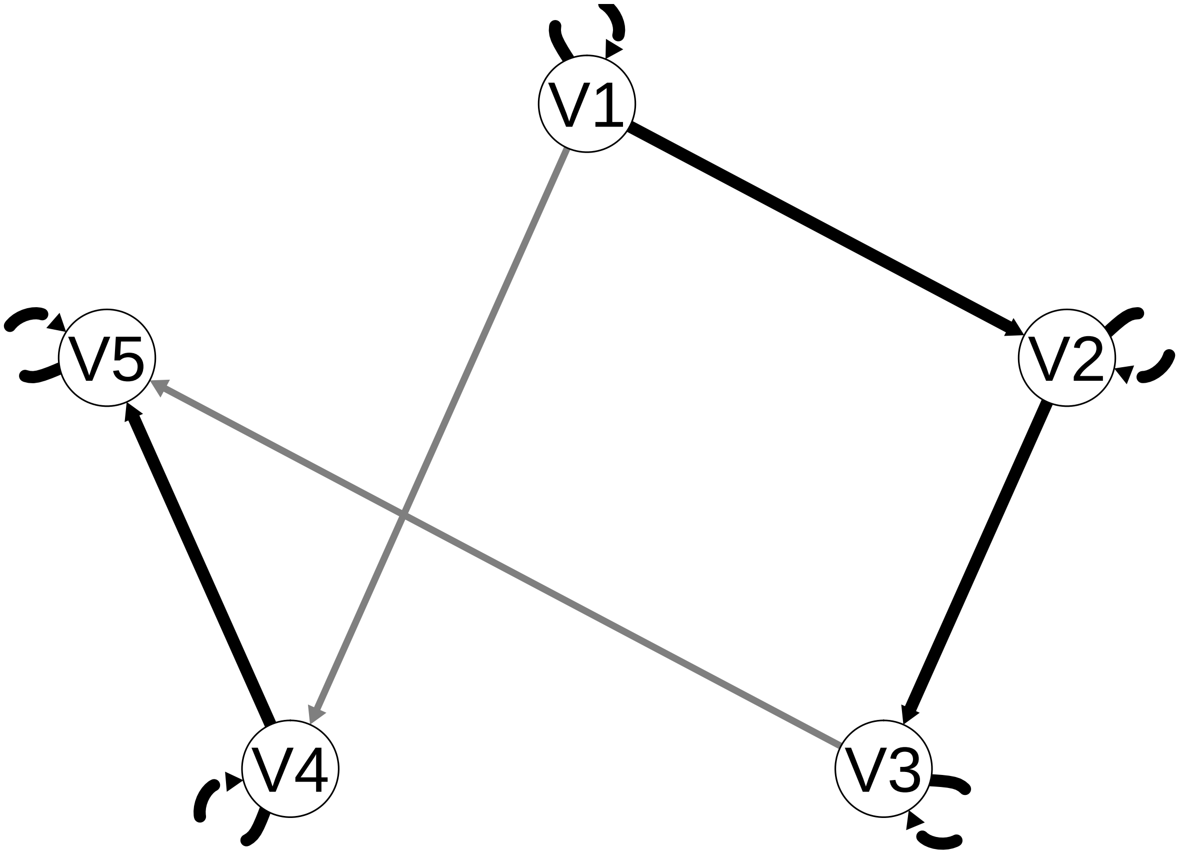
The default in gimme is to have AR effects estimatd as a starting point for the model. It then identifies which path is the best path to add when looking across the majority of individuals.

There are lots of options for gimme, but if your data are in list form then you only need to provide the name of the data. Here we also save results to an optional output folder as well as to an R object.

```
gimme_out <- gimme(data = Data_all,
                   out = "~/Desktop/output")
```

```
plot(gimme_out)
```

```
## Please specify a file id for individual plots. Otherwise, summary plot is presented.
```

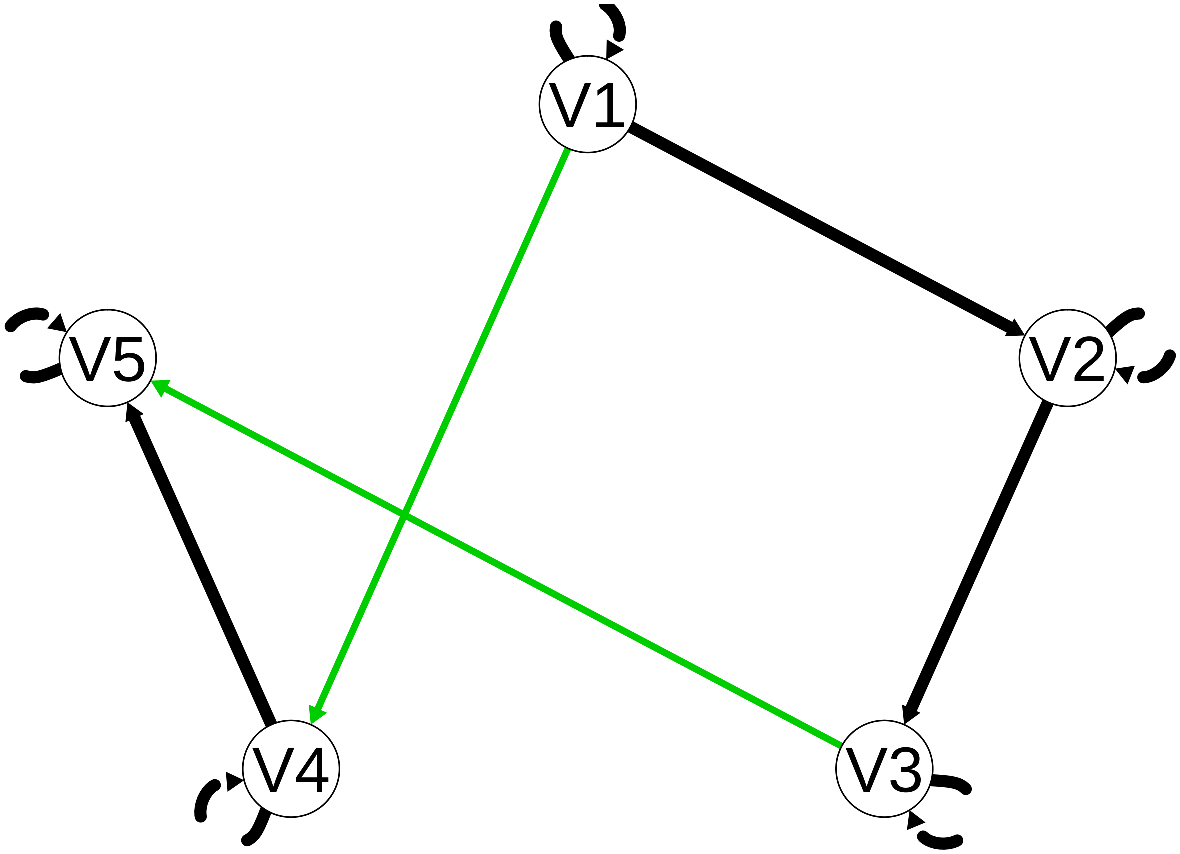


# Subgrouping GIMME We know there are two subgroups in our data; can subgrouping GIMME find them?

```
subgroup_out <- gimme(data = Data_all,
                subgroup = TRUE)
```

```
plot(subgroup_out)
```

```
## Please specify a file id for individual plots. Otherwise, summary plot is presented.
```



Phew, we got 2 subgroups, just as expected. We also see that the subgroup-level paths (green) are the ones that do in fact differ according to subgroups in our data-generation code.

Let's take a look at the composition of these subgroups to make sure people are placed in the right one.

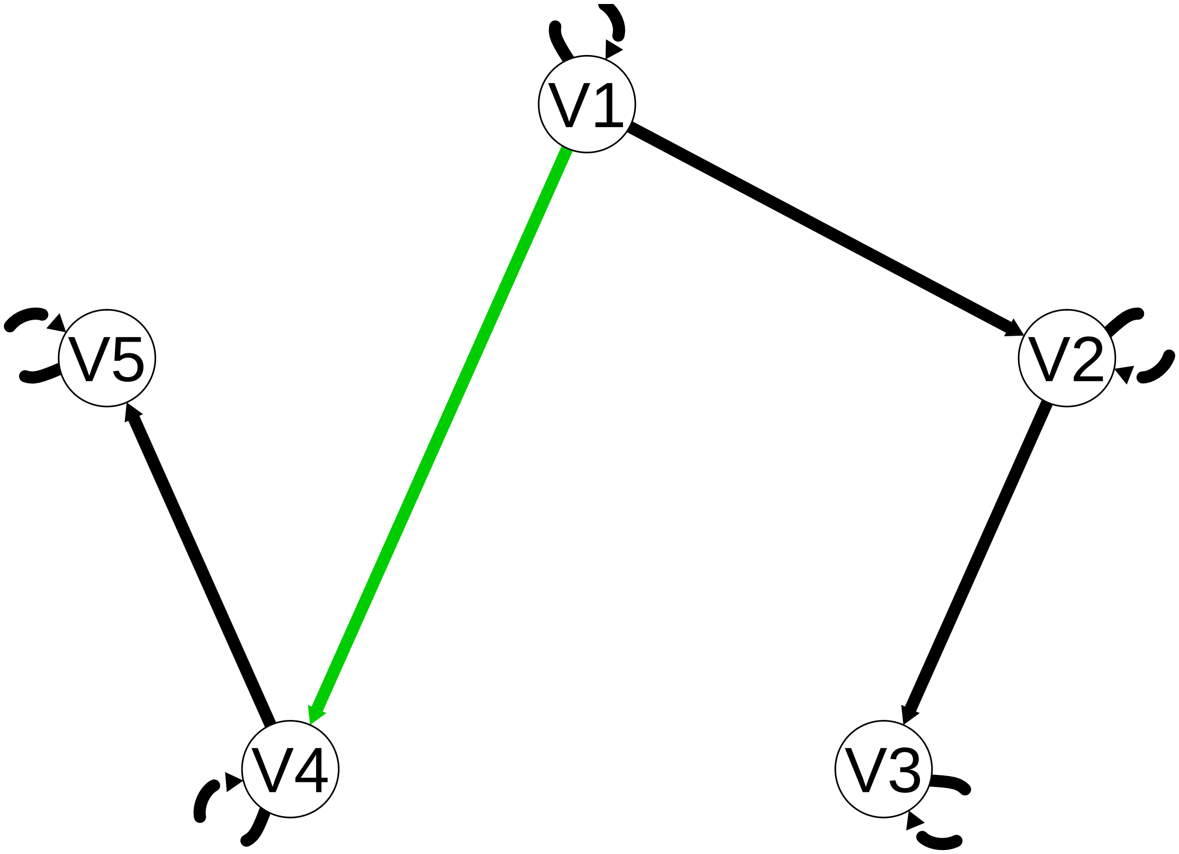
```
new <- cbind(subgroup_out$fit$file,subgroup_out$fit$sub_membership)
new <- new[order(new[,2]),]
new
```

```
##       [,1]    [,2]
##  [1,] "ind1"  "1"
##  [2,] "ind10" "1"
##  [3,] "ind2"  "1"
##  [4,] "ind3"  "1"
##  [5,] "ind4"  "1"
##  [6,] "ind5"  "1"
##  [7,] "ind6"  "1"
##  [8,] "ind7"  "1"
##  [9,] "ind8"  "1"
## [10,] "ind9"  "1"
## [11,] "ind11" "2"
## [12,] "ind12" "2"
## [13,] "ind13" "2"
## [14,] "ind14" "2"
## [15,] "ind15" "2"
## [16,] "ind16" "2"
## [17,] "ind17" "2"
## [18,] "ind18" "2"
## [19,] "ind19" "2"
## [20,] "ind20" "2"
```

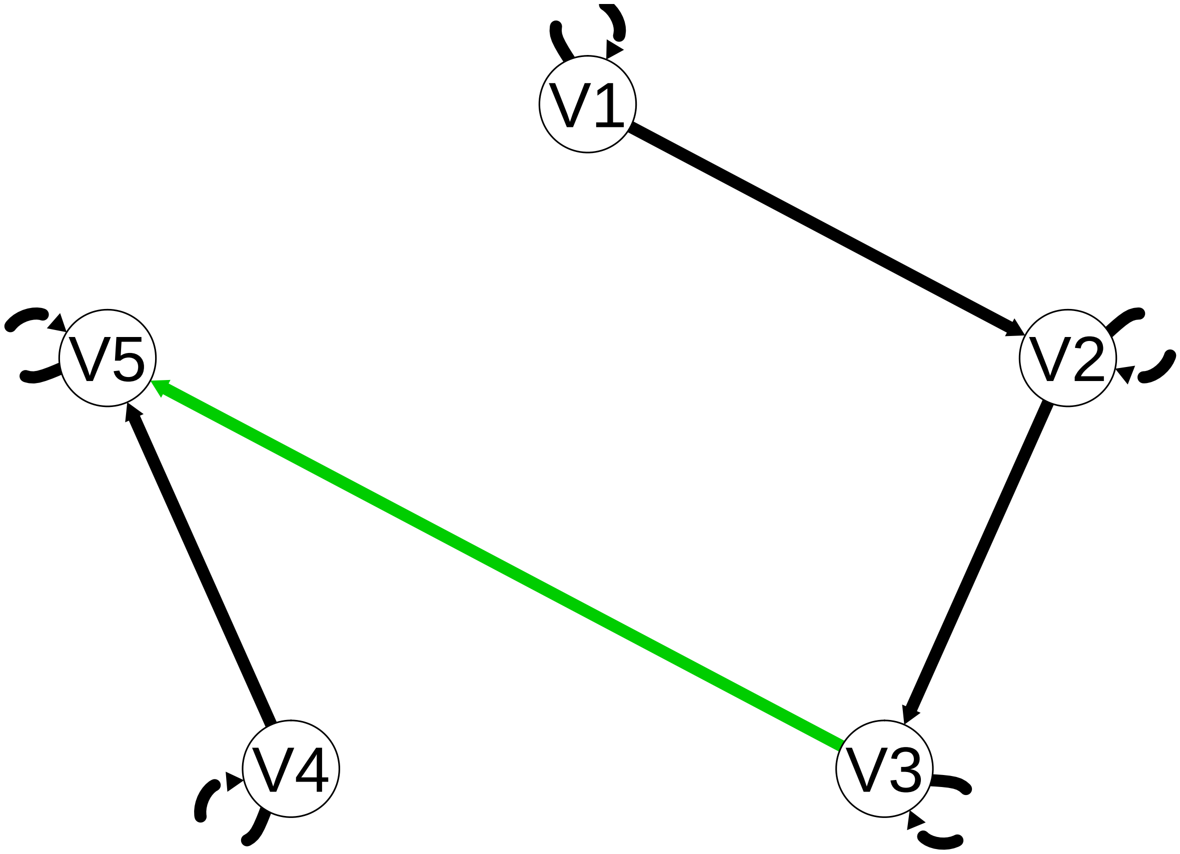
Perfect!

Let's plot the subgroups to see the paths a bit clearer.

```
plot(subgroup_out$sub_plots[[1]])
```

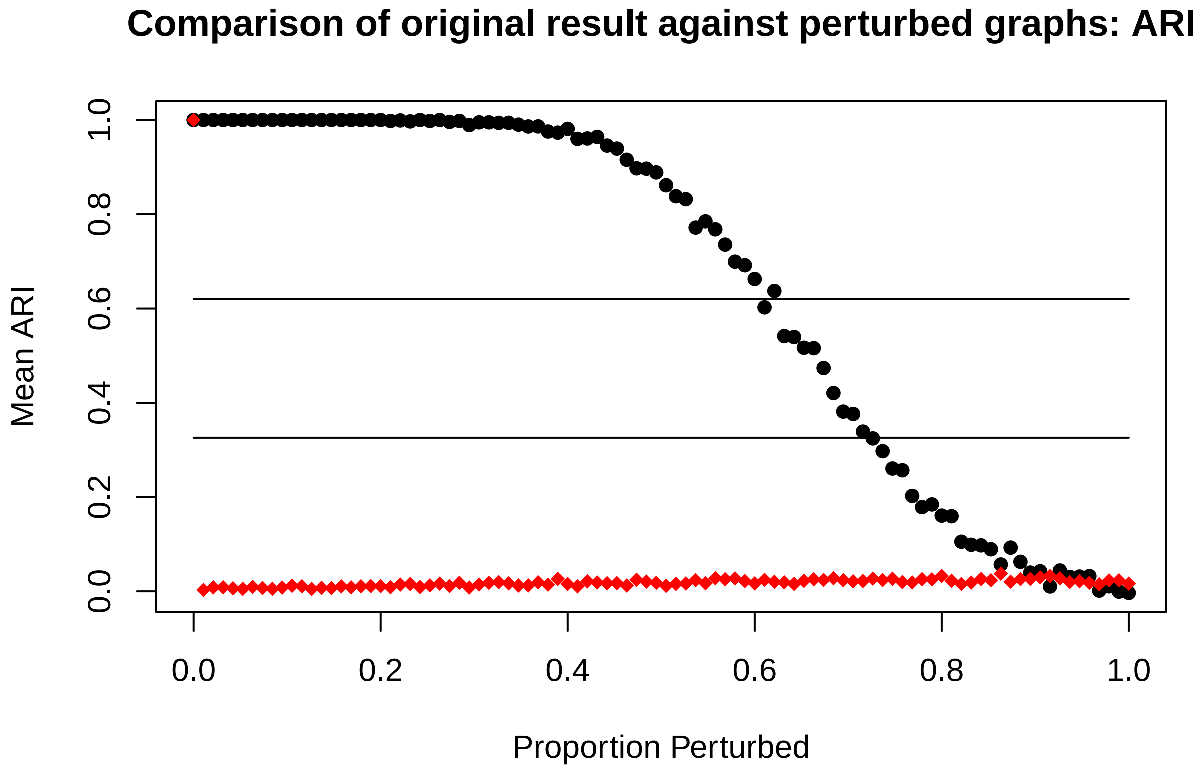


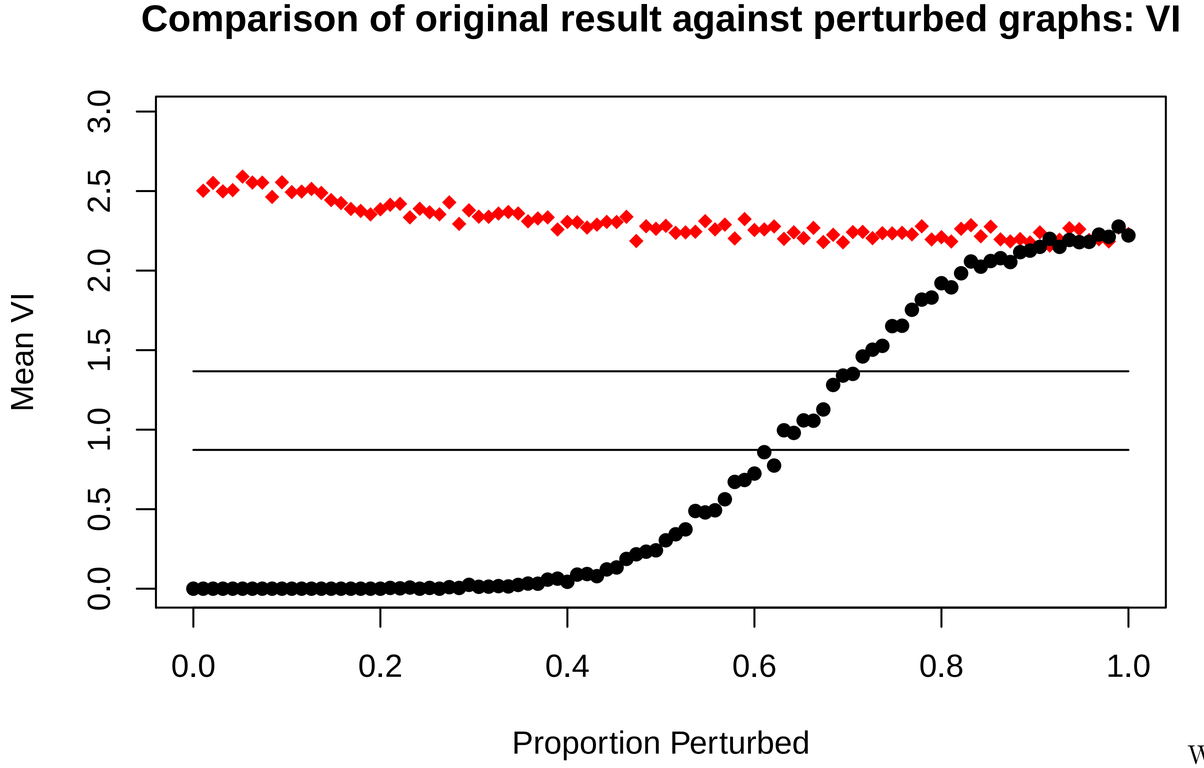
```
plot(subgroup_out$sub_plots[[2]])
```


V1
V2
V3
V4
V5

Robustness of subgroup classification Let’s check to see if our subgroups are robust.

```
test <- perturbR(subgroup_out$sim_matrix)
```





Wow, this solution is robust. The black dots above indicate the similarity between the original subgroup solution (i.e., which indivdiuals were in which subgroup) with the solution obtained by increasingly perturbing the matrix used to subgroup.

Robust results are ones where the black dots take a long time - at least past the alpha = 0.20 point - to intersect with the horizontal lines.

It must be noted that some solutions may not evidence robustness but still may be interesting. It is important to keep in mind your goal through subgrouping - do you want to (1) arrive at distinc subsets of people who differ in their patterns of relations, or (2) identify patterns of relations that tend to exist in some people? For the latter goal, it is not very necessary that the solutions be robust - the goal of detecting similar patterns across some subset of people can still aid in interpretation of the indivdiual-level nuances.