Causal Inference - Assignment Part1

Emilia Löscher Kyuri Park

24 February, 2022

1. Draw a DAG representing a simple and fairly plausible causal system from your preferred topic of choice. Describe briefly the substantive motivation behind your DAG.

Many people with Obsessive-compulsive disorder (OCD) become depressed. According to Millet et al. (2004), the lifetime rates of major depression in OCD patients is about 81.2%. Several studies attempted to identify the causal mechanism in the comorbidity of OCD and depression (McNally et al.2017; Zandberg et al. 2015) and speculated that OCD symptoms often precede and correspondingly activate the depression symptoms.

Given this background, here we proceed to look into the causal relationship between OCD and depression symptoms. We hypothesize that *distress associated with obsession* (OCD symptom) causes *feeling of guilt* (depression symptom) via several paths. Having been inspired by the DAG model from McNally et al. (2017), we construct our DAG with total 7 variables (nodes), which consist of OCD symptoms as well as depression symptoms. The specifics of the variables in our DAG is as follows:

< OCD symptoms >

- **ocdis**: distress caused by obsessions/compulsions
- **ocint**: interference due to obsessions/compulsions
- **occon**: difficulty controlling obsessions/compulsions

< Depression symptoms >

- **sad**: sadness
- **insom**: insomnia/sleeping problems
- **concen**: concentration/decision-making impairment
- **guilt**: guilt and self-blame

```
# varnames <- c("ocint", "ocdis", "occon", "sad", "insom", "concen", "guilt")
# Adj <- matrix(c(0,1,0,0,0,1,0,
#                 0,0,1,1,0,0,0,
#                 0,0,0,1,0,0,0,
#                 0,0,0,0,1,0,1,
#                 0,0,0,0,0,1,0,
#                 0,0,0,0,0,0,1,
#                  0,0,0,0,0,0,0), 7,7, byrow = TRUE, dimnames = list(varnames, varnames))
#
# qgraph(Adj, labels = varnames, layout= "groups", theme="classic")

ocddep <- dagify(
  ocdis ~ ocint,
  occon ~ ocdis,
  sad ~ ocdis + occon,
  insom ~ sad,
  concen ~ ocint + insom,
  guilt ~ concen + sad,
  exposure = "ocdis", # cause variable we are interested in
  outcome = "guilt", # effect variable we are interested in
  # set the coordinates
  coords = list(x = c(ocdis = -1, ocint=-0.5, occon=-0.5, sad = 0, insom=0, concen=0.5, guilt =
                y = c(ocdis = 0, ocint=1, occon=-1, sad = 0, insom=0.5, concen=1, guilt = 0))
)

ggdag_status(ocddep) + theme_dag()
```

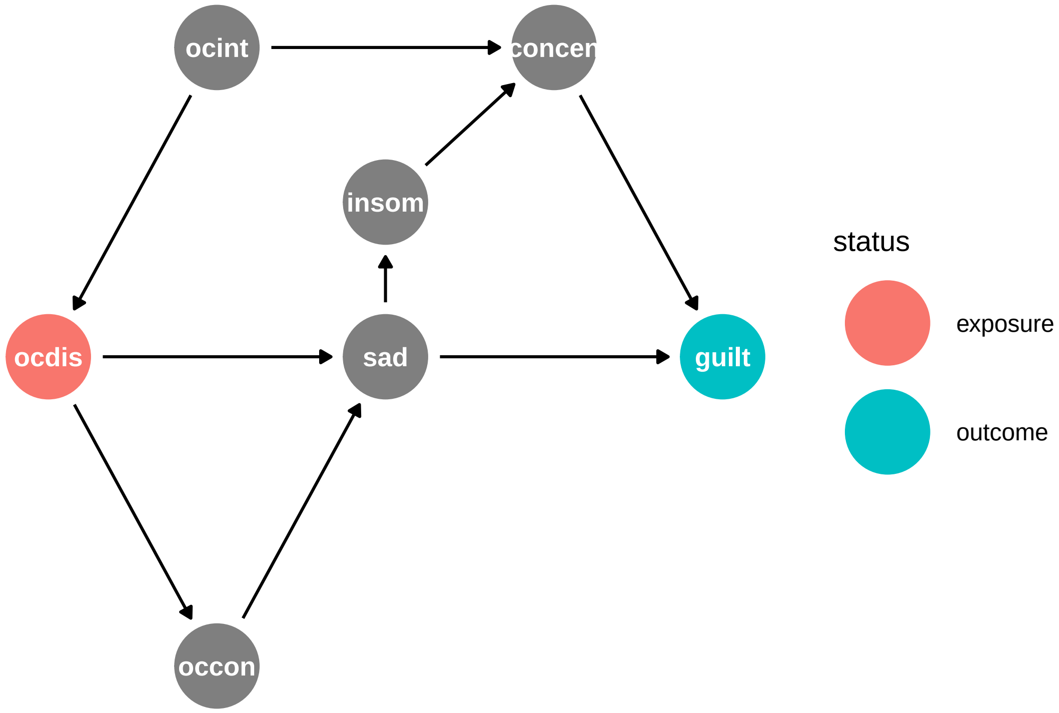

Figure 1: OCD-Depression DAG

2. Specify a structural causal model for your DAG. Assume all relationships between the variables are linear, and all the variables have nrmally distributed residuals.

In the following, the structural causal model for our DAG is specified:

$$ocint := 2.69 + \epsilon_{ocint}$$

$$ocdis := 2.81 + 3.52 \cdot ocint + \epsilon_{ocdis}$$

$$occon := 2.67 + 3.38 \cdot ocdis + \epsilon_{occon}$$

$$sad := 1.55 + 4.33 \cdot ocdis + 2 \cdot occon + \epsilon_{sad}$$

$$insom := 0.81 + 2 \cdot sad + \epsilon_{insom}$$

$$concen := 1.48 + 2 \cdot ocint + 3 \cdot insom + \epsilon_{concen}$$

$$guilt := 1.56 + 1.5 \cdot concen + 3.3 \cdot sad + \epsilon_{guilt}$$

,where $\epsilon_{ocint}, \ldots, \epsilon_{guilt} \overset{iid}{\sim} N(0, SD_i)$ with $i \in ocint, \ldots, guilt$.

We base our estimates of the intercepts and residual standard deviations (SD) on the data from McNally et al. (2017), where they interviewed 408 patients for the OCD and depression symptoms. Intercepts and residual SDs are set to the mean and SD values of the corresponding item (symptom) that they found. The regression coefficients are decided such that it could represent the relative strength (effect) of each item we have thought of, considering the size of SD for each item.

3. Generate data from your SCM with a sample size of 500 units.

```r
# set the seed
set.seed(1000)

# sample size (n) = 500
n <- 500

# generate the data
ocint <- 2.69 + rnorm(n, 0, 0.82)
ocdis <- 2.81 + 3.52 * ocint + rnorm(n, 0, 0.76)
occon <- 2.67 + 3.38 * ocdis + rnorm(n, 0, 0.76)
sad <- 1.55 + 4.33 * ocdis + 2.98 * occon + rnorm(n, 0, 0.94)
insom <- 0.81 + 2.17 * sad + rnorm(n, 0, 1.07)
concen <-  1.48 + 2.54 * ocint + 3.46 * insom + rnorm(n, 0, 0.87)
guilt <- 1.56 + 1.57 * concen + 3.31 * sad + rnorm(n, 0, 1.17)

# putting them in a dataframe called "OCDDEP"
OCDDEP <- data.frame(ocint, ocdis, occon, sad, insom, concen, guilt)
```

4. Use the PC-algorithm on the generated data to ‘discover’ the structure of your causal system

a. Is your true DAG covered in the Markov equivalence class provided by the algorithm?

No, our true DAG is not covered in the Markov equivalence class. A couple of edges are missing (e.g., *ocint* –> *concen*, *ocdis* –> *occon*).

```
# sufficient statistics: correlation matrix + sample size
suffStat <- list(C= cor(OCDDEP), n = nrow(OCDDEP))

# run the pc algorithm
varnames <- colnames(OCDDEP)
pc_fit <- pc(suffStat = suffStat, indepTest = gaussCItest, alpha = 0.01,
              labels = varnames)

# # plot the Markov equivalence class
# if (require(Rgraphviz)) {
#   p1 <- plot(pc_fit, main = "Inferred CPDAG using PC algorithm")
# }
laymat <- rbind(c(-0.5, 1),
                c(-1, 0),
                c(-0.5, -1),
                c(0, 0),
                c(0, 0.5),
                c(0.5, 1),
                c(1, 0))

par(mar=c(5, 4, 3.5, 2) + 0.1)

cpdag_mat <- as(pc_fit, "matrix")
qgraph(t(cpdag_mat), bidirectional=TRUE, color=c("white","salmon", "white",
                                                 "white", "white", "white", "lightblue"),
       layout = laymat, labels = varnames)
title("Inferred CPDAG using PC algorithm", line=3, cex.main = 1)
```

b. Provide the CP-DAG. To what extent did the procedure correctly recover which relationship were absent/present/directed?

```
# extract the DAG adjacency matrix in a vector form (by rows)
res1 <- pdag2allDags(cpdag_mat)
```

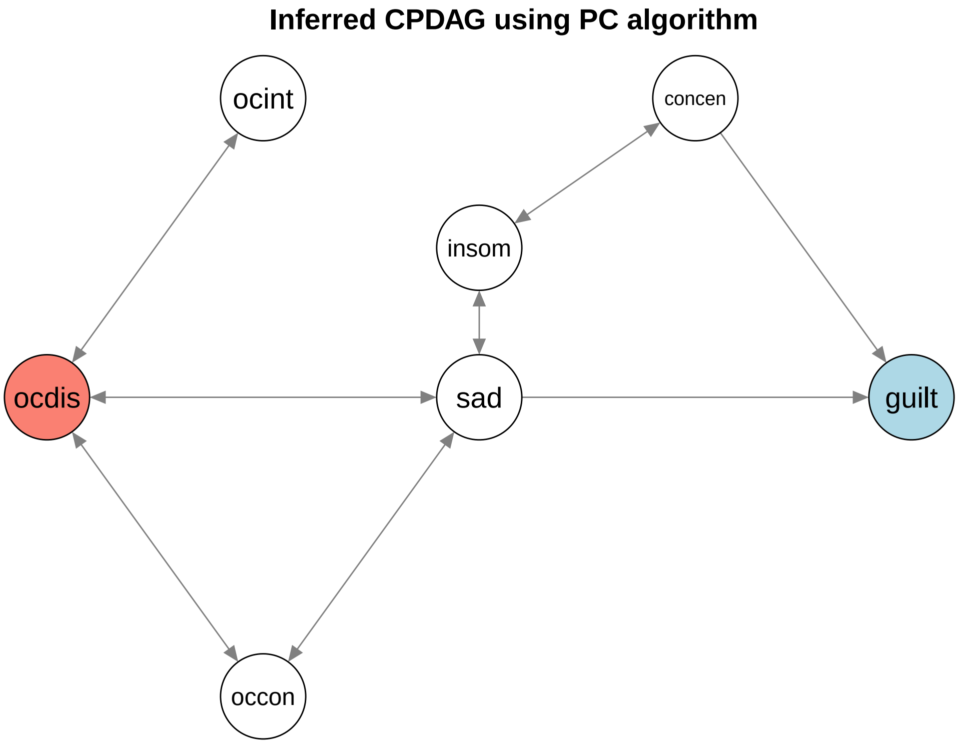

Figure 2: Inferred CPDAG using pcalg

```r
# get the adjacency matrix of an individual DAG
res1_dags <- list()
for(i in 1 :nrow(res1$dags)){
  res1_dags[[i]] <- t(matrix(res1$dags[i,], 7, 7, byrow = T))
}

# plot the DAG for each adj.matrix
par(mfrow=c(1,2))
for (i in 1:length(res1_dags)){
  qgraph(res1_dags[[i]], bidirectional=TRUE, color=c("white","salmon", "white",
                                                 "white", "white", "white", "lightblue"),
       layout = laymat, labels = varnames)
}
```

5. data. Choose two variables for which they should estimate the causal relationship

a. Cause variable: *ocdis*

b. Outcome variable: *guilt*

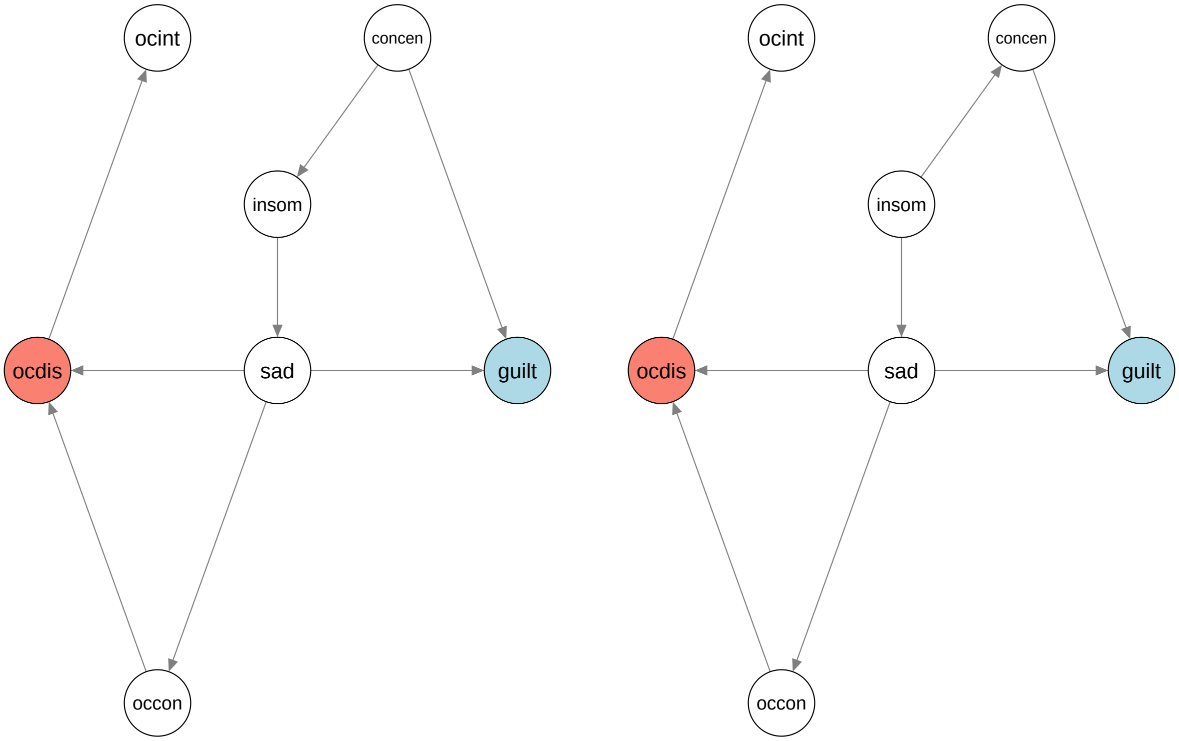

Figure 3: Markov-Equivalence Set

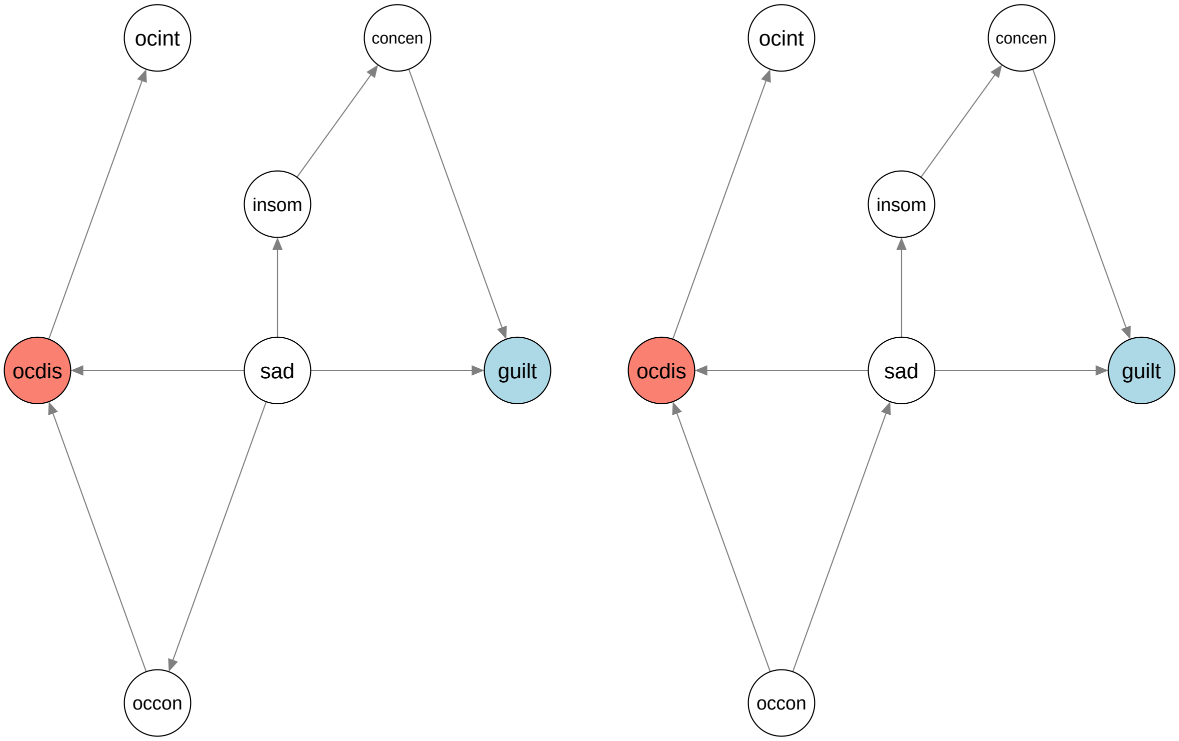

Figure 4: Markov-Equivalence Set

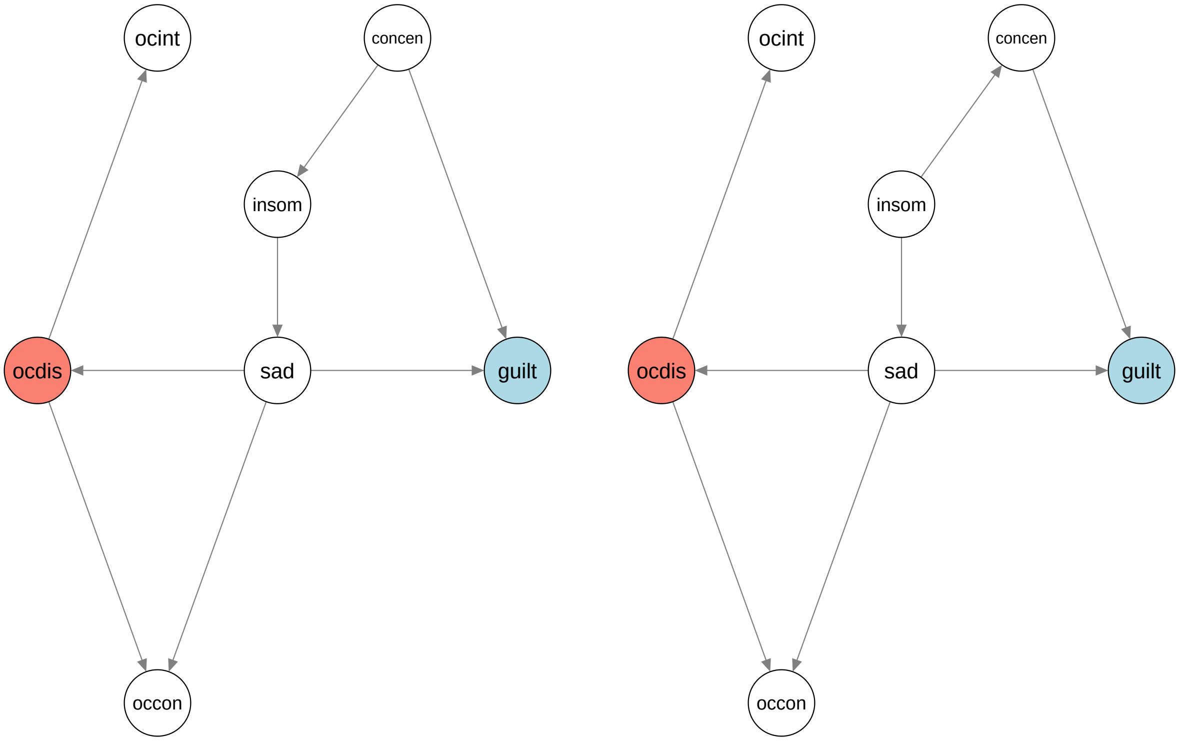

Figure 5: Markov-Equivalence Set

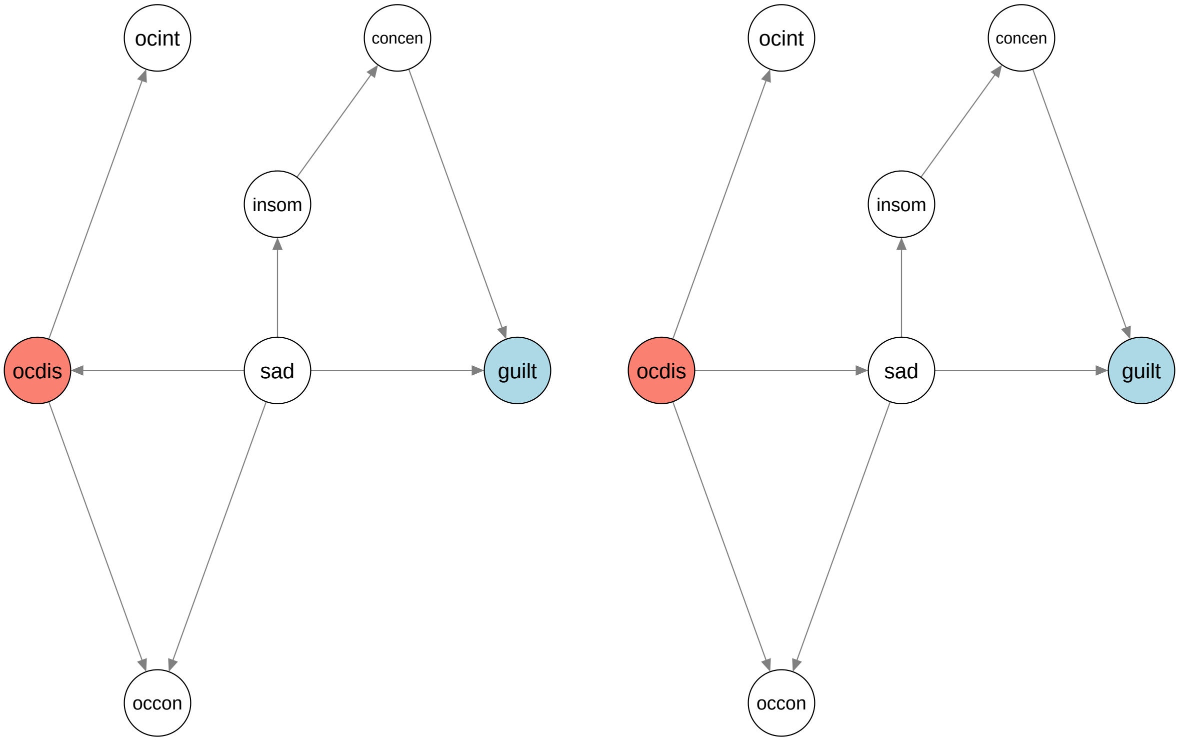

Figure 6: Markov-Equivalence Set

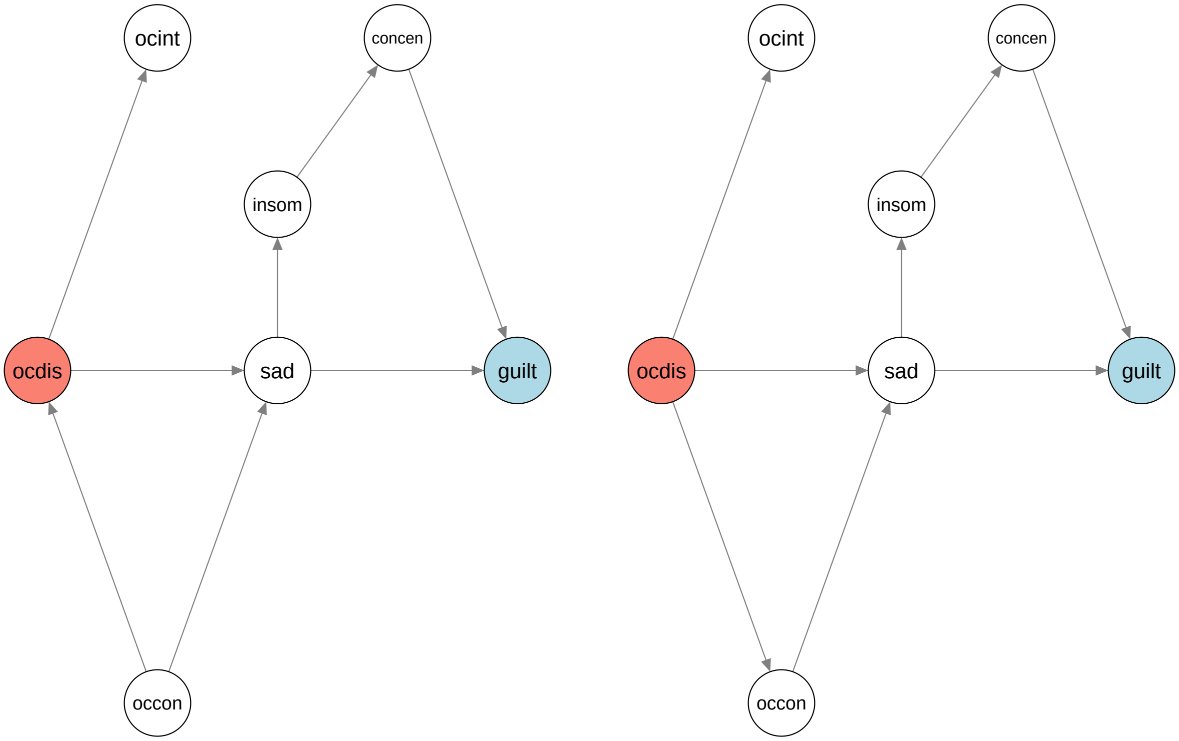

Figure 7: Markov-Equivalence Set

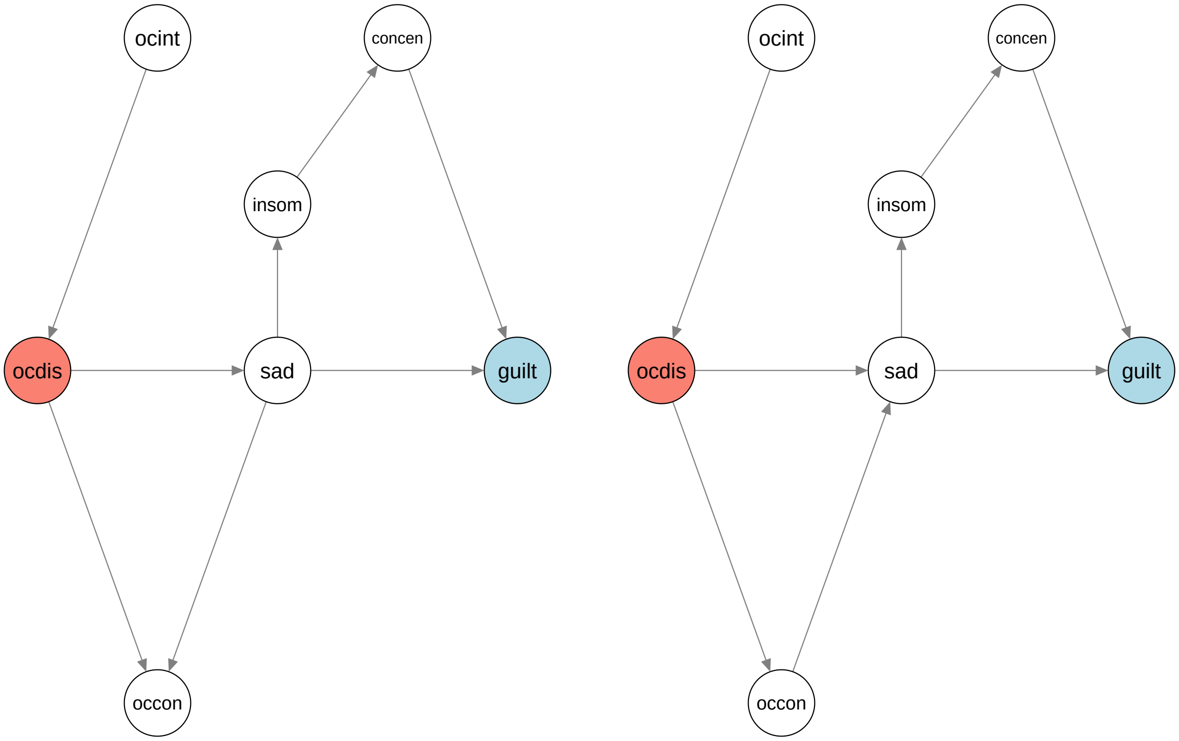

Figure 8: Markov-Equivalence Set

6. For the causal effect specified in 5

a. What is the true causal effect of the cause on the outcome variable based on your SCM?

- path1) ocdis –> sad –> guilt
- path2) ocdis –> occon –> sad –> guilt
- pa.6th3) ocdis –> sad –> insom –> concen –> guilt

```
# path 1
path1 <- 4.33 * 3.31
# path 2
path2 <- 3.38 * 2.98 * 3.31
# path 3
path3 <- 4.32 * 2.17 * 3.46 * 1.57

# total causal effect
path1 + path2 + path3
```

```
## [1] 98.59556
```

b. Based on the true DAG, what linear regression model should be used to estimate the causal effect correctly?

We block the backdoor path via controlling for *ocint*.

$$guilt = intercept + ocdis + ocint + \epsilon$$

c. Estimate the causal effect with this regression model based on your generated data. To what extent is the true effect recovered?

for now it is very different… something to do with the intercept?

```
mod <- lm(guilt ~ ocdis + ocint, data = OCDDEP)
summary(mod)
```

```
##
## Call:
## lm(formula = guilt ~ ocdis + ocint, data = OCDDEP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.588  -25.108    0.275   23.568  104.698
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  152.192      8.457  17.996   <2e-16 ***
## ocdis        216.171      2.157 100.232   <2e-16 ***
## ocint          9.873      7.953   1.242    0.215    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 37.45 on 497 degrees of freedom
## Multiple R-squared:  0.9964, Adjusted R-squared:  0.9964
## F-statistic: 6.863e+04 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
# cpdag_mat_OCD <- as(pc_OCD,"matrix")
#
# # Each row is a DAG adjacency matrix in vector form (by rows)
# res_OCD <- pdag2allDags(cpdag_mat_OCD)
#
# # We can get the adjacency matrix of an individual DAG using
# resOCD_dags <- list()
# for(i in 1:nrow(res_OCD$dags)){
#   resOCD_dags[[i]] <- t(matrix(res_OCD$dags[i,],7,7,byrow = TRUE))
# }
#
#
# ida(2,7, cov(OCDdata), pc_OCD@graph, verbose = TRUE)
```

The third CP-DAG does a good job?!

```
# res_Adj <- resOCD_dags[[3]]
#
# qgraph(res_Adj)
```

7. For assignment Part II, we will need a dichotomous cause variable.

a. Make a dichotomized version of the cause variable from question 5&6 in your dataset, for example by assigning scores lower than the mean a 0 and the rest a 1.

```
#Dichtomizing the effect variable OCDdist
mean_ocdis <- mean(ocdis)
ocdis_d <- rep(1, length(ocdis))
ocdis_d[which(ocdis < mean_ocdis)] <- 0

OCDDEP <- cbind(OCDDEP, ocdis_d)
```

b. Use the correct model for the causal effect (the one from 6b) to estimate the causal effect again, but now with your dichotomized cause variable. Discuss the results.

```
mod_OCD_d <- lm(guilt ~ ocdis_d + ocint, data = OCDDEP)
summary(mod_OCD_d)
```

```
##
## Call:
## lm(formula = guilt ~ ocdis_d + ocint, data = OCDDEP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -440.65 -102.44   11.84  110.62  422.17
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   885.66      32.63  27.142  < 2e-16 ***
## ocdis_d       174.40      22.49   7.756 5.02e-14 ***
## ocint         690.21      14.61  47.230  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 162.9 on 497 degrees of freedom
## Multiple R-squared:  0.9317, Adjusted R-squared:  0.9315
## F-statistic:  3391 on 2 and 497 DF,  p-value: < 2.2e-16
```

8. Prepare the data for the second causal inference assignment

a. Create a data frame with all your variables

```
# save(OCDdata, file = "LoscherPark_Assign1_Data.RData")
```

b. Make a brief .txt file See LoscherPark_Assign1.txt.

Reference

McNally, R.J., Mair, P., Mugno, B.L., & Riemann, B.C. (2017). Co-Morbid Obsessive–Compulsive Disorder and Depression: A Bayesian Network Approach. *Psychological Medicine*. *47*(7): 1204–14. https://doi.org/10.1017/S0033291716003287.

Millet, B., Kochman, F., Gallarda, T., Krebs, M.O., Demonfaucon, F., Barrot, I., Bourdel, M.C., Olie, J.P., Loo, H., & Hantouche, E.G. (2004). Phenomenological and Comorbid Features Associated in Obsessive–Compulsive Disorder: Influence of Age of Onset. *Journal of Affective Disorders. 79*(1-3): 241–46.

Zandberg, L.J., Zang, Y., McLean, C.P., Yeh, R., Simpson, H.B., & Foa, E.B. (2015). Change in Obsessive-Compulsive Symptoms Mediates Subsequent Change in Depressive Symptoms During Exposure and Response Prevention. *Behaviour Research and Therapy. 68*:76–81.