Using R6causal

Juha Karvanen

2024-03-14

Overview

The R package `R6causal` implements an R6 class called `SCM`. The class aims to simplify working with structural causal models. The missing data mechanism can be defined as a part of the structural model.

The class contains methods for

- defining a structural causal model via functions, text or conditional probability tables
- printing basic information on the model
- plotting the graph for the model using packages `igraph` or `qgraph`
- simulating data from the model
- applying an intervention
- checking the identifiability of a query using the R packages `causaleffect` and `dosearch`
- defining the missing data mechanism
- simulating incomplete data from the model according to the specified missing data mechanism
- checking the identifiability in a missing data problem using the R package `dosearch`
- checking the identifiability of a counterfactual query using the R package `cfid`

In addition, there are functions for

- running experiments
- counterfactual inference using simulation
- evaluating fairness of a prediction model

The class `ParallelWorld` inherits `SCM` and defines a structural causal model that describes parallel worlds for counterfactual inference.

The class `LinearGaussianSCM` inherits `SCM` and defines a structural causal model where all functions are linear and all background variables follow Gaussian distribution.

Setup

```
library(R6causal)
library(data.table)
library(stats)
data.table::setDTthreads(2)
```

Defining the model

Structural causal model (SCM) for a backdoor situation can be defined as follows

```
backdoor <- SCM$new("backdoor",
  uflist = list(
```

```
    uz = function(n) {return(runif(n))},
    ux = function(n) {return(runif(n))},
    uy = function(n) {return(runif(n))}
  ),
  vflist = list(
    z = function(uz) {
      return(as.numeric(uz < 0.4))},
    x = function(ux, z) {
      return(as.numeric(ux < 0.2 + 0.5*z))},
    y = function(uy, z, x) {
      return(as.numeric(uy < 0.1 + 0.4*z + 0.4*x))}
  )
)
```

A shortcut notation for this is

```
backdoor_text <- SCM$new("backdoor",
  uflist = list(
    uz = "n : runif(n)",
    ux = "n : runif(n)",
    uy = "n : runif(n)"
  ),
  vflist = list(
    z = "uz : as.numeric(uz < 0.4)",
    x = "ux, z : as.numeric(ux < 0.2 + 0.5*z)",
    y = "uy, z, x : as.numeric(uy < 0.1 + 0.4*z + 0.4*x)"
  )
)
```

Alternatively the functions of SCM can be specified via conditional probability tables

```
backdoor_condprob <- SCM$new("backdoor",
  uflist = list(
    uz = function(n) {return(runif(n))},
    ux = function(n) {return(runif(n))},
    uy = function(n) {return(runif(n))}
  ),
  vflist = list(
    z = function(uz) {
      return( generate_condprob( ycondx = data.table(z = c(0,1),
                                                     prob = c(0.6,0.4)),
                                 x = data.table(uz = uz),
                                 Umerge_expr = "uz"))},
    x = function(ux, z) {
      return( generate_condprob( ycondx = data.table(x = c(0,1,0,1),
                                                     z = c(0,0,1,1),
                                                     prob = c(0.8,0.2,0.3,0.7)),
                                            x = data.table(z = z, ux = ux),
                                            Umerge_expr = "ux"))},
    y = function(uy, z, x) {
      return( generate_condprob( ycondx = data.table(y= rep(c(0,1), 4),
                                                     z = c(0,0,1,1,0,0,1,1),
                                                     x = c(0,0,0,0,1,1,1,1),
                                                     prob = c(0.9,0.1,0.5,0.5,
                                                              0.5,0.5,0.1,0.9)),
```

```
                                                x = data.table(z = z, x = x, uy = uy),
                                                Umerge_expr = "uy"))}
  )
)
```

It is possible to mix the styles and define some elements of a function list as functions, some as text and some as conditional probability tables.

Defining a linear Gaussian SCM

A linear Gaussian SCM can be defined giving the coefficients for the structural equations:

```
lgbackdoor <- LinearGaussianSCM$new("Linear Gaussian Backdoor",
                                    linear_gaussian = list(
                                      uflist = list(ux = function(n) {rnorm(n)},
                                                    uy = function(n) {rnorm(n)},
                                                    uz = function(n) {rnorm(n)}),
                                      vnames = c("x","y","z"),
                                      vcoefmatrix = matrix(c(0,0.4,0,0,0,0,0.6,0.8,0),3,3),
                                      ucoefvector = c(1,1,1),
                                      ccoefvector = c(0,0,0)))
print(lgbackdoor)
#> Name of the model:  Linear Gaussian Backdoor
#>
#> Graph:
#>  z -> x
#>  x -> y
#>  z -> y
#>
#> Functions of background (exogenous) variables:
#>
#> $ux
#> function(n) {rnorm(n)}
#>
#> $uy
#> function(n) {rnorm(n)}
#>
#> $uz
#> function(n) {rnorm(n)}
#>
#> Functions of endogenous variables:
#>
#> $x
#> function (z, ux)
#> {
#>     return(0 + 0.6 * z + 1 * ux)
#> }
#> <environment: 0x000001d0766a0928>
#>
#> $y
#> function (x, z, uy)
#> {
#>     return(0 + 0.4 * x + 0.8 * z + 1 * uy)
#> }
```

```
#> <environment: 0x000001d076693898>
#>
#> $z
#> function (uz)
#> {
#>     return(0 + 1 * uz)
#> }
#> <environment: 0x000001d0766acb80>
#>
#> Topological order of endogenous variables:
#> [1] "z" "x" "y"
#>
#> No missing data mechanism
```

It is also possible to generate the underlying DAG and the coefficients randomly:

```
randomlg <- LinearGaussianSCM$new("Random Linear Gaussian",
                                  random_linear_gaussian = list(
                                  nv = 6,
                                  edgeprob=0.5,
                                  vcoefdistr = function(n) {rnorm(n)},
                                  ccoefdistr = function(n) {rnorm(n)},
                                  ucoefdistr = function(n) {rnorm(n)}))
print(randomlg)
#> Name of the model:  Random Linear Gaussian
#>
#> Graph:
#>  v3 -> v1
#>  v4 -> v2
#>  v3 -> v4
#>  v2 -> v5
#>  v1 -> v6
#>  v4 -> v6
#>
#> Functions of background (exogenous) variables:
#>
#> $u1
#> function (n)
#> {
#>     return(rnorm(n))
#> }
#> <environment: 0x000001d076d789a0>
#>
#> $u2
#> function (n)
#> {
#>     return(rnorm(n))
#> }
#> <environment: 0x000001d076d83f60>
#>
#> $u3
#> function (n)
#> {
#>     return(rnorm(n))
```

```
#> }
#> <environment: 0x000001d076d815b0>
#>
#> $u4
#> function (n)
#> {
#>     return(rnorm(n))
#> }
#> <environment: 0x000001d076d7cbf0>
#>
#> $u5
#> function (n)
#> {
#>     return(rnorm(n))
#> }
#> <environment: 0x000001d076d7e260>
#>
#> $u6
#> function (n)
#> {
#>     return(rnorm(n))
#> }
#> <environment: 0x000001d076d8d640>
#>
#> Functions of endogenous variables:
#>
#> $v1
#> function (v3, u1)
#> {
#>     return(-0.194452020142582 + 0.340822645556077 * v3 + 1.19572099912747 *
#>         u1)
#> }
#> <environment: 0x000001d076d92330>
#>
#> $v2
#> function (v4, u2)
#> {
#>     return(1.14436900375859 + 0.348478944040261 * v4 + -0.132040473085734 *
#>         u2)
#> }
#> <environment: 0x000001d076d8b550>
#>
#> $v3
#> function (u3)
#> {
#>     return(0.549220880109555 + -1.03735874103047 * u3)
#> }
#> <environment: 0x000001d076d948f0>
#>
#> $v4
#> function (v3, u4)
#> {
#>     return(-1.11379680798657 + -2.53563294243635 * v3 + -0.61458842511206 *
```

```
#>         u4)
#> }
#> <environment: 0x000001d076d89ce8>
#>
#> $v5
#> function (v2, u5)
#> {
#>     return(-0.930116389868676 + 0.626790759039911 * v2 + 0.0258670170501744 *
#>         u5)
#> }
#> <environment: 0x000001d076d99048>
#>
#> $v6
#> function (v1, v4, u6)
#> {
#>     return(0.34941419530480 4 + -0.382204636867487 * v1 + 1.1015778200197 *
#>         v4 + 1.44327064430964 * u6)
#> }
#> <environment: 0x000001d076da8418>
#>
#> Topological order of endogenous variables:
#> [1] "v3" "v1" "v4" "v2" "v6" "v5"
#>
#> No missing data mechanism
```

Printing the model

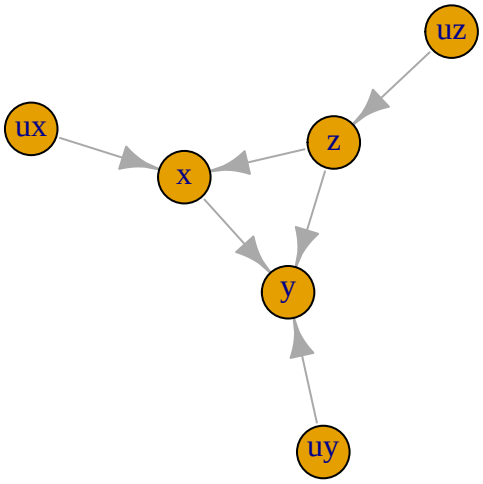
The print method presents the basic information on the model

```
backdoor
#> Name of the model:  backdoor
#>
#> Graph:
#>  z -> x
#>  z -> y
#>  x -> y
#>
#> Functions of background (exogenous) variables:
#>
#> $uz
#> function(n) {return(runif(n))}
#>
#> $ux
#> function(n) {return(runif(n))}
#>
#> $uy
#> function(n) {return(runif(n))}
#>
#> Functions of endogenous variables:
#>
#> $z
#> function(uz) {
#>       return(as.numeric(uz < 0.4))}
#>
#> $x
#> function(ux, z) {
#>       return(as.numeric(ux < 0.2 + 0.5*z))}
#>
#> $y
#> function(uy, z, x) {
#>       return(as.numeric(uy < 0.1 + 0.4*z + 0.4*x))}
#>
#> Topological order of endogenous variables:
#> [1] "z" "x" "y"
#>
#> No missing data mechanism
```

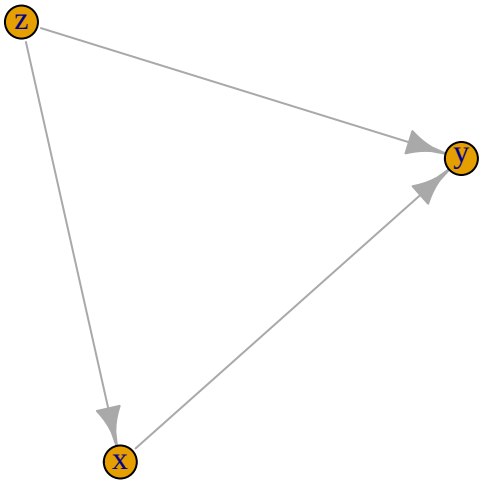
Plotting the graph

The plotting method of the package `igraph` is used by default. If `qgraph` is available, its plotting method can be used as well. The argument `subset` controls which variables are plotted. Plotting parameters are passed to the plotting method.

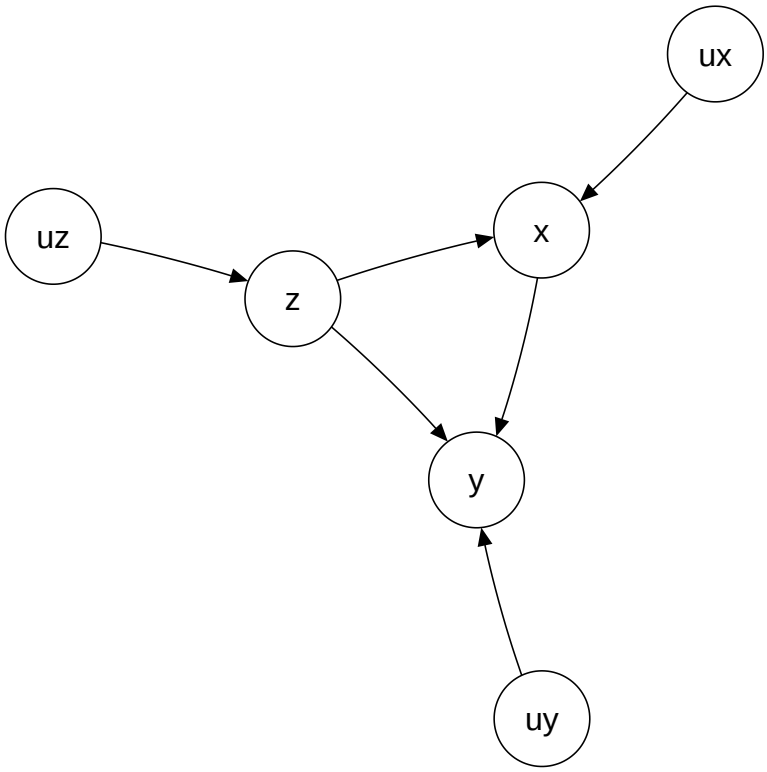
```
backdoor$plot(vertex.size = 25) # with package 'igraph'
```




```
backdoor$plot(subset = "v") # only observed variables
```




```
if (requireNamespace("qgraph", quietly = TRUE)) backdoor$plot(method = "qgraph")
```




```
# alternative look with package 'qgraph'
```

Simulating data

Calling method `simulate()` creates or updates data table `simdata`.

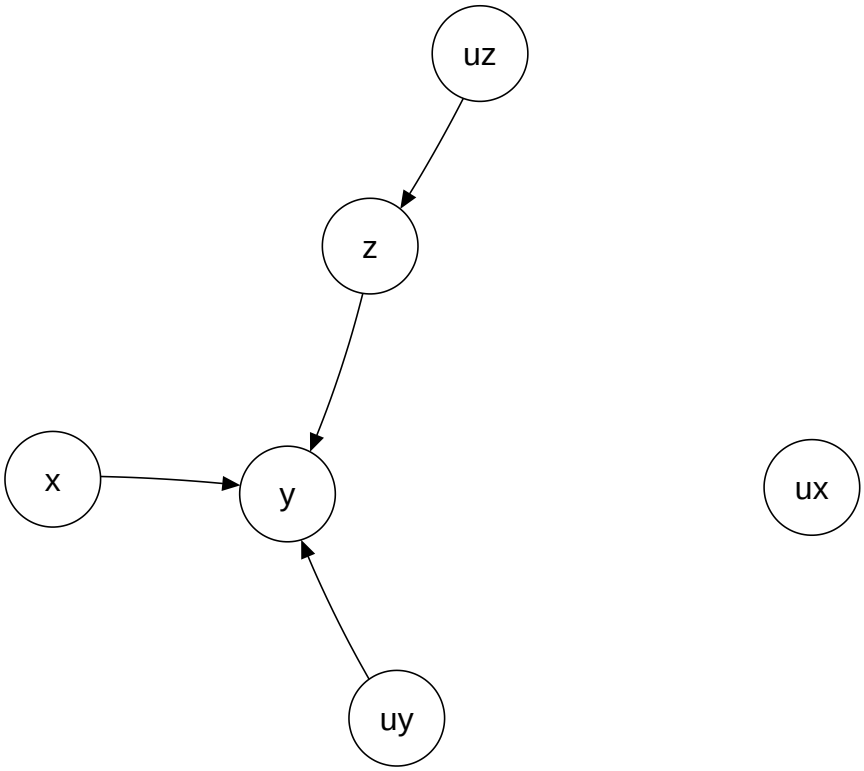
```
backdoor$simulate(10)
backdoor$simdata
#>              uz          ux          uy     z     x     y
#>          <num>       <num>       <num> <num> <num> <num>
#>  1: 0.93706919 0.65222882 0.04931049     0     0     1
#>  2: 0.62029457 0.16501685 0.85497557     0     1     0
#>  3: 0.99339531 0.34660394 0.83191445     0     0     0
#>  4: 0.15499748 0.09027895 0.80703810     1     1     1
#>  5: 0.01558495 0.99805482 0.58878679     1     0     0
#>  6: 0.23510216 0.61778566 0.17032293     1     1     1
#>  7: 0.51754335 0.79296122 0.16970637     0     0     0
#>  8: 0.01453718 0.04206458 0.53460183     1     1     1
#>  9: 0.32152004 0.54656103 0.43992911     1     1     1
#> 10: 0.53686887 0.24877851 0.74216326     0     0     0
backdoor$simulate(8)
backdoor$simdata
#>            uz        ux        uy     z     x     y
#>         <num>     <num>     <num> <num> <num> <num>
#> 1: 0.2562507 0.9516747 0.6994977     1     0     0
#> 2: 0.1340398 0.3607472 0.9519103     1     1     0
#> 3: 0.3436197 0.3637142 0.5748896     1     1     1
#> 4: 0.3953156 0.4762140 0.2458309     1     1     1
#> 5: 0.7310678 0.7387591 0.7152488     0     0     0
```

```
#> 6: 0.8274389 0.1857632 0.4729871     0     1     1
#> 7: 0.9219097 0.6306178 0.9444048     0     0     0
#> 8: 0.3736709 0.9560987 0.5207959     1     0     0
backdoor_text$simulate(20)
backdoor_condprob$simulate(30)
```

Applying an intervention

In an intervention, the structural equation of the target variable is changed.

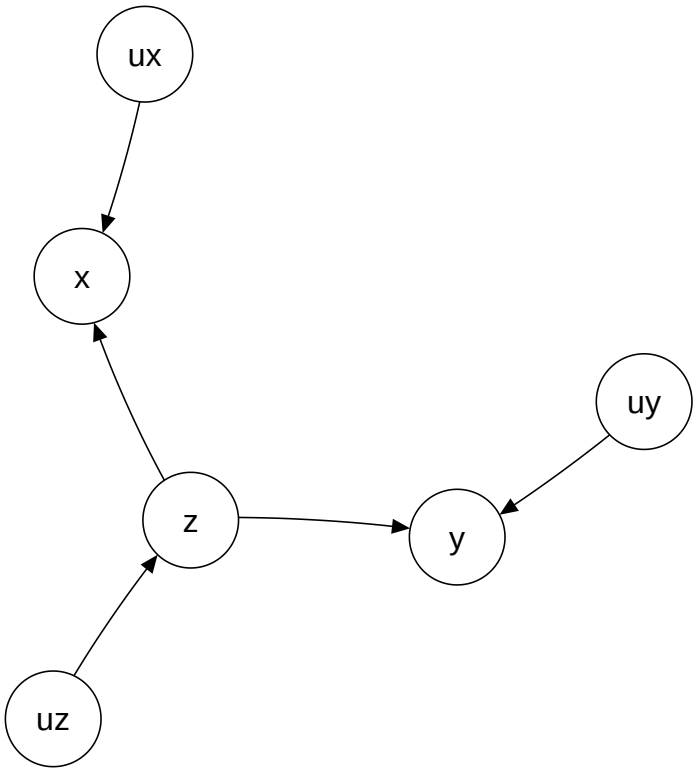
```
backdoor_x1 <- backdoor$clone()  # making a copy
backdoor_x1$intervene("x",1) # applying the intervention
backdoor_x1$plot(method = "qgraph") # to see that arrows incoming to x are cut
```




```
backdoor_x1$simulate(10) # simulating from the intervened model
backdoor_x1$simdata
#>             uz        ux        uy     z     x     y
#>          <num>     <num>     <num> <num> <num> <num>
#>  1: 0.39082100 0.9539338 0.7929557     1     1     1
#>  2: 0.07793962 0.9535986 0.7090539     1     1     1
#>  3: 0.90840393 0.7279363 0.8306501     0     1     0
#>  4: 0.48153531 0.3638536 0.4111566     0     1     1
#>  5: 0.40750634 0.8713156 0.3888855     0     1     1
#>  6: 0.65715288 0.8966490 0.6824894     0     1     0
#>  7: 0.28080946 0.6511335 0.5866035     1     1     1
#>  8: 0.26734228 0.7359989 0.4273321     1     1     1
#>  9: 0.05194049 0.7779869 0.5385563     1     1     1
#> 10: 0.06848496 0.8221791 0.8765795     1     1     1
```

An intervention can redefine a structural equation

```
backdoor_yz <- backdoor$clone()  # making a copy
backdoor_yz$intervene("y",
  function(uy, z) {return(as.numeric(uy < 0.1 + 0.8*z ))}) # making y a function of z only
backdoor_yz$plot(method = "qgraph") # to see that arrow x -> y is cut
```



Running an experiment (set of interventions)

The function `run_experiment` applies a set of interventions, simulates data and collects the results.

```
backdoor_experiment <- run_experiment(backdoor,
                                      intervene = list(x = c(0,1)),
                                      response = "y",
                                      n = 10000)
str(backdoor_experiment)
#> List of 2
#>  $ interventions:Classes 'data.table' and 'data.frame':  2 obs. of  1 variable:
#>   ..$ x: num [1:2] 0 1
#>   ..- attr(*, ".internal.selfref")=<externalptr>
#>   ..- attr(*, "sorted")= chr "x"
#>  $ response_list:List of 1
#>   ..$ y:Classes 'data.table' and 'data.frame':   10000 obs. of  2 variables:
#>   .. ..$ V1: num [1:10000] 0 0 0 0 0 0 0 0 1 1 ...
#>   .. ..$ V2: num [1:10000] 1 0 0 1 0 1 1 1 1 1 ...
#>   .. ..- attr(*, ".internal.selfref")=<externalptr>
colMeans(backdoor_experiment$response_list$y)
#>     V1     V2
#> 0.2614 0.6652
```

Applying the ID algorithm, Do-search and cfid

There are direct plugins to R packages `causaleffect`, `dosearch` and `cfid` that can be used to solve identifiability problems.

```
backdoor$causal.effect(y = "y", x = "x")
#> [1] "\\sum_{z}P(y|z,x)P(z)"
backdoor$dosearch(data = "p(x,y,z)", query = "p(y|do(x))")
#> \sum_{z}\left(p(z)p(y|x,z)\right)
backdoor$cfid(gamma = cfid::conj(cfid::cf("Y",0), cfid::cf("X",0, c(Z=1))) )
#> The query P(y /\ x_{z'}) is not identifiable from P_*.
```

Counterfactual inference (a simple case)

Let us assume that intervention do(X=0) was applied and the response Y = 0 was recorded. What is the probability that in this situation the intervention do(X=1) would have led to the response Y = 1? We estimate this probability by means of simulation.

```
cfdata <- counterfactual(backdoor, situation = list(do = list(target = "x", ifunction = 0),
                                                    condition = data.table( x = 0, y = 0)),
                         target = "x", ifunction = 1, n = 100000,
                         method = "rejection")
mean(cfdata$y)
#> [1] 0.53843
```

The result differs from P(Y = 1 | do(X = 1))

```
backdoor_x1$simulate(100000)
mean(backdoor_x1$simdata$y)
#> [1] 0.66093
```

Counterfactual inference (parallel worlds)

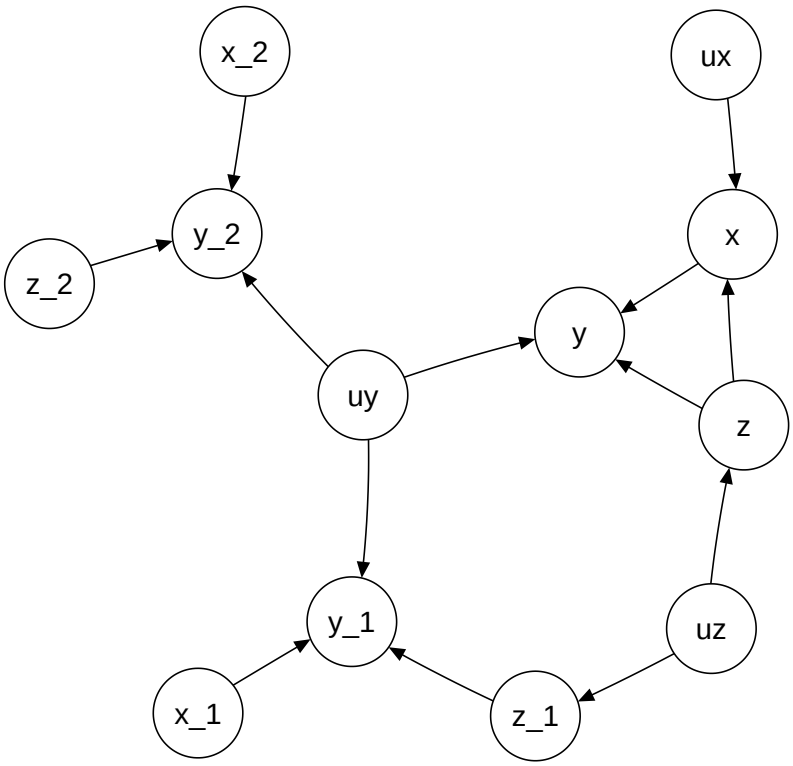
Parallel world graphs (a generalization of a twin graph) are used for counterfactual inference with several counterfactual interventions . The package implements class `ParallelWorld` which heritates class `SCM`. A `ParallelWorld` object is created from an `SCM` object by specifying the interventions for each world. By default the variables of the parallel worlds are named with suffixes "_1", "_2", ...

In the example below, we have the original world (variables `x`, `z`, `y`) and its two variants. In the variant 1 (variables `x_1`, `z_1`, `y_1`), the value of `x` (variable `x_1` in the object) is set to be 0. In the variant 2 (variables `x_2`, `z_2`, `y_2`), the value of `x` (variable `x_2` in the object) is set to be 0 and the value of `z` (variable `z_2` in the object) is set to be 1.

```
backdoor_parallel <- ParallelWorld$new(
                          backdoor,
                         dolist=list(
                              list(target = "x",
                                   ifunction = 0),
                              list(target = list("z","x"),
                                   ifunction = list(1,0))
                          )
 )
backdoor_parallel
#> Name of the model:  backdoor
#>
```

```
#> Graph:
#>  uz -> z
#>  z -> x
#>  uy -> y
#>  z -> y
#>  x -> y
#>  uz -> z_1
#>  uy -> y_1
#>  z_1 -> y_1
#>  x_1 -> y_1
#>  uy -> y_2
#>  z_2 -> y_2
#>  x_2 -> y_2
#>
#> Functions of background (exogenous) variables:
#>
#> $uz
#> function(n) {return(runif(n))}
#> <bytecode: 0x000001d00769cba8>
#>
#> $ux
#> function(n) {return(runif(n))}
#> <bytecode: 0x000001d00772d340>
#>
#> $uy
#> function(n) {return(runif(n))}
#> <bytecode: 0x000001d0077bdad8>
#>
#> Functions of endogenous variables:
#>
#> $z
#> function(uz) {
#>       return(as.numeric(uz < 0.4))}
#> <bytecode: 0x000001d007879a60>
#>
#> $x
#> function(ux, z) {
#>       return(as.numeric(ux < 0.2 + 0.5*z))}
#> <bytecode: 0x000001d00799948a0>
#>
#> $y
#> function(uy, z, x) {
#>       return(as.numeric(uy < 0.1 + 0.4*z + 0.4*x))}
#> <bytecode: 0x000001d007b10ee8>
#>
#> $z_1
#> function (uz)
#> {
#>     return(as.numeric(uz < 0.4))
#> }
#>
#> $x_1
#> function (...)
```

```
#> {
#>     return(constant)
#> }
#> <environment: 0x000001d007f06470>
#>
#> $y_1
#> function (uy, z_1, x_1)
#> {
#>     return(as.numeric(uy < 0.1 + 0.4 * z_1 + 0.4 * x_1))
#> }
#>
#> $z_2
#> function (...)
#> {
#>     return(constant)
#> }
#> <environment: 0x000001d004725850>
#>
#> $x_2
#> function (...)
#> {
#>     return(constant)
#> }
#> <environment: 0x000001d0047227d8>
#>
#> $y_2
#> function (uy, z_2, x_2)
#> {
#>     return(as.numeric(uy < 0.1 + 0.4 * z_2 + 0.4 * x_2))
#> }
#>
#> Topological order of endogenous variables:
#> [1] "x_1" "z_2" "x_2" "z"   "z_1" "y_2" "x"   "y_1" "y"
#>
#> No missing data mechanism
if (requireNamespace("qgraph", quietly = TRUE)) backdoor_parallel$plot(method = "qgraph")
```



Counterfactual data can be simulated with function `counterfactual`. In the example below, we know that variable `y` obtained value 0 in the original world as well as variants 1 and 2. We are interested in the counterfactual distribution of `y` if `x` had been set to 1.

```
cfdata <- counterfactual(backdoor_parallel,
                         situation = list(
                            do = NULL,
                            condition = data.table::data.table( y = 0, y_1 = 0, y_2 = 0)),
                         target = "x",
                         ifunction = 1,
                         n = 100000,
                         method = "rejection")
mean(cfdata$y)
#> [1] 0.12464
```

The printed value is a simulation based estimate for the counterfactual probability $P(Y = 1)$.

An alternative way for answering the same question defines the case of interest as one of the parallel worlds (here variant 3).

```
backdoor_parallel2 <- ParallelWorld$new(
                          backdoor,
                         dolist=list(
                              list(target = "x",
                                  ifunction = 0),
                              list(target = list("z","x"),
                                   ifunction = list(1,0)),
                               list(target = "x",
                                  ifunction = 1)
                          )
 )
```

```
cfdata <- counterfactual(backdoor_parallel2,
                         situation = list(
                            do = NULL,
                            condition = data.table::data.table( y = 0, y_1 = 0, y_2 = 0)),
                         n = 100000,
                         method = "rejection")
mean(cfdata$y_3)
#> [1] 0.12431
```

The printed value is a simulation based estimate for the counterfactual probability $P(Y = 1)$.

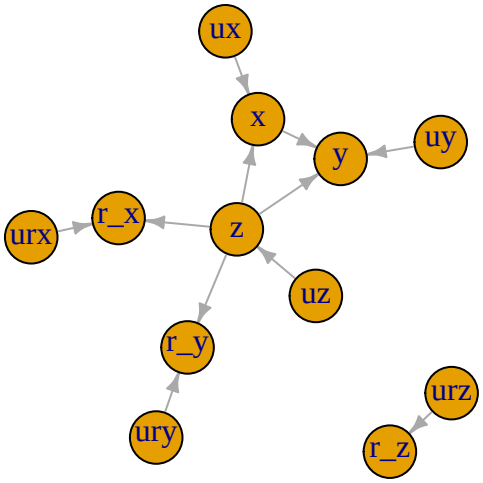
A model with a missing data mechanism

The missing data mechanism is defined in similar manner as the other variables.

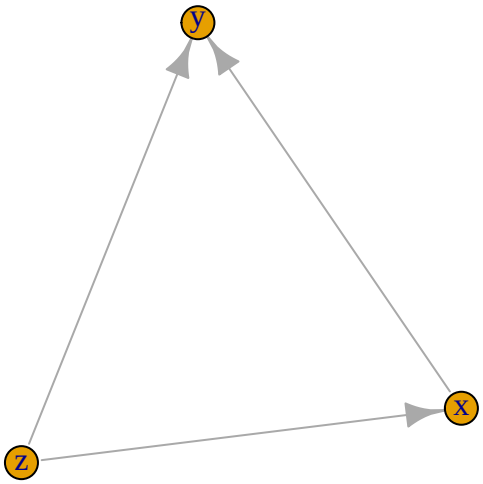
```
backdoor_md <- SCM$new("backdoor_md",
                       uflist = list(
                         uz = "n : runif(n)",
                         ux = "n : runif(n)",
                         uy = "n : runif(n)",
                         urz = "n : runif(n)",
                         urx = "n : runif(n)",
                         ury = "n : runif(n)"
                       ),
                       vflist = list(
                         z = "uz : as.numeric(uz < 0.4)",
                         x = "ux, z : as.numeric(ux < 0.2 + 0.5*z)",
                         y = "uy, z, x : as.numeric(uy < 0.1 + 0.4*z + 0.4*x)"
                       ),
                       rflist = list(
                         z = "urz : as.numeric( urz < 0.9)",
                         x = "urx, z : as.numeric( (urx + z)/2 < 0.9)",
                         y = "ury, z : as.numeric( (ury + z)/2 < 0.9)"
                       ),
                       rprefix = "r_"
)
```

Plotting the graph for a model with missing data mechanism

```
backdoor_md$plot(vertex.size = 25, edge.arrow.size=0.5) # with package 'igraph'
```




```
backdoor_md$plot(subset = "v") # only observed variables a
```




```
if (!requireNamespace("qgraph", quietly = TRUE)) backdoor_md$plot(method = "qgraph")
# alternative look with package 'qgraph'
```

Simulating incomplete data

By default both complete data and incomplete data are simulated. The incomplete dataset is named as `$simdata_obs`.

```
backdoor_md$simulate(100)
summary(backdoor_md$simdata)
#>        uz                  ux                 uy                urz
#>  Min.   :0.007739   Min.   :0.01717   Min.   :0.04045   Min.   :0.004756
#>  1st Qu.:0.229987   1st Qu.:0.34379   1st Qu.:0.19353   1st Qu.:0.222194
#>  Median :0.470774   Median :0.54481   Median :0.47960   Median :0.603977
#>  Mean   :0.498856   Mean   :0.54419   Mean   :0.49486   Mean   :0.538054
#>  3rd Qu.:0.788599   3rd Qu.:0.78421   3rd Qu.:0.74903   3rd Qu.:0.808703
#>  Max.   :0.997104   Max.   :0.99674   Max.   :0.99756   Max.   :0.994521
#>
#>       urx                ury                z              x
#>  Min.   :0.004847   Min.   :0.003778   Min.   :0.00   Min.   :0.00
#>  1st Qu.:0.236167   1st Qu.:0.298850   1st Qu.:0.00   1st Qu.:0.00
```

```
#>  Median :0.452273   Median :0.505153   Median :0.00   Median :0.00
#>  Mean   :0.502744   Mean   :0.509611   Mean   :0.43   Mean   :0.38
#>  3rd Qu.:0.771682   3rd Qu.:0.742523   3rd Qu.:1.00   3rd Qu.:1.00
#>  Max.   :0.995291   Max.   :0.995081   Max.   :1.00   Max.   :1.00
#>
#>        y             z_md             x_md             y_md
#>  Min.   :0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
#>  1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
#>  Median :0.00   Median :0.0000   Median :0.0000   Median :0.0000
#>  Mean   :0.44   Mean   :0.4157   Mean   :0.3511   Mean   :0.4066
#>  3rd Qu.:1.00   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
#>  Max.   :1.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
#>                 NA's   :11       NA's   :6        NA's   :9
#>       r_z             r_x             r_y
#>  Min.   :0.00   Min.   :0.00   Min.   :0.00
#>  1st Qu.:1.00   1st Qu.:1.00   1st Qu.:1.00
#>  Median :1.00   Median :1.00   Median :1.00
#>  Mean   :0.89   Mean   :0.94   Mean   :0.91
#>  3rd Qu.:1.00   3rd Qu.:1.00   3rd Qu.:1.00
#>  Max.   :1.00   Max.   :1.00   Max.   :1.00
#>
summary(backdoor_md$simdata_obs)
#>       z_md             x_md             y_md             r_z
#>  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00
#>  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.00
#>  Median :0.0000   Median :0.0000   Median :0.0000   Median :1.00
#>  Mean   :0.4157   Mean   :0.3511   Mean   :0.4066   Mean   :0.89
#>  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.00
#>  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00
#>  NA's   :11       NA's   :6        NA's   :9
#>       r_x             r_y
#>  Min.   :0.00   Min.   :0.00
#>  1st Qu.:1.00   1st Qu.:1.00
#>  Median :1.00   Median :1.00
#>  Mean   :0.94   Mean   :0.91
#>  3rd Qu.:1.00   3rd Qu.:1.00
#>  Max.   :1.00   Max.   :1.00
#>
```

By using the argument `fixedvars` one can keep the complete data unchanged and re-simulate the missing data mechanism.

```
backdoor_md$simulate(100, fixedvars = c("x","y","z","ux","uy","uz"))
summary(backdoor_md$simdata)
#>        uz                ux                uy               urz
#>  Min.   :0.007739   Min.   :0.01717   Min.   :0.04045   Min.   :0.004579
#>  1st Qu.:0.229987   1st Qu.:0.34379   1st Qu.:0.19353   1st Qu.:0.273413
#>  Median :0.470774   Median :0.54481   Median :0.47960   Median :0.542744
#>  Mean   :0.498856   Mean   :0.54419   Mean   :0.49486   Mean   :0.508264
#>  3rd Qu.:0.788599   3rd Qu.:0.78421   3rd Qu.:0.74903   3rd Qu.:0.727928
#>  Max.   :0.997104   Max.   :0.99674   Max.   :0.99756   Max.   :0.997830
#>
#>       urx               ury               z              x
#>  Min.   :0.003913   Min.   :0.0238   Min.   :0.00   Min.   :0.00
```

```
#>  1st Qu.:0.158391   1st Qu.:0.3319   1st Qu.:0.00   1st Qu.:0.00
#>  Median :0.363550   Median :0.5091   Median :0.00   Median :0.00
#>  Mean   :0.400148   Mean   :0.5268   Mean   :0.43   Mean   :0.38
#>  3rd Qu.:0.616533   3rd Qu.:0.7519   3rd Qu.:1.00   3rd Qu.:1.00
#>  Max.   :0.931812   Max.   :0.9916   Max.   :1.00   Max.   :1.00
#>
#>        y              z_md            x_md             y_md            r_z
#>  Min.   :0.00   Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.00
#>  1st Qu.:0.00   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.00
#>  Median :0.00   Median :0.000   Median :0.0000   Median :0.000   Median :1.00
#>  Mean   :0.44   Mean   :0.427   Mean   :0.3542   Mean   :0.382   Mean   :0.89
#>  3rd Qu.:1.00   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.00
#>  Max.   :1.00   Max.   :1.000   Max.   :1.0000   Max.   :1.000   Max.   :1.00
#>                 NA's   :11      NA's   :4        NA's   :11
#>       r_x             r_y
#>  Min.   :0.00   Min.   :0.00
#>  1st Qu.:1.00   1st Qu.:1.00
#>  Median :1.00   Median :1.00
#>  Mean   :0.96   Mean   :0.89
#>  3rd Qu.:1.00   3rd Qu.:1.00
#>  Max.   :1.00   Max.   :1.00
#>
summary(backdoor_md$simdata_obs)
#>       z_md            x_md             y_md            r_z             r_x
#>  Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.00   Min.   :0.00
#>  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.00   1st Qu.:1.00
#>  Median :0.000   Median :0.0000   Median :0.000   Median :1.00   Median :1.00
#>  Mean   :0.427   Mean   :0.3542   Mean   :0.382   Mean   :0.89   Mean   :0.96
#>  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.00   3rd Qu.:1.00
#>  Max.   :1.000   Max.   :1.0000   Max.   :1.000   Max.   :1.00   Max.   :1.00
#>  NA's   :11      NA's   :4        NA's   :11
#>       r_y
#>  Min.   :0.00
#>  1st Qu.:1.00
#>  Median :1.00
#>  Mean   :0.89
#>  3rd Qu.:1.00
#>  Max.   :1.00
#>
```

Applying Do-search to a missing data problem

```
backdoor_md$dosearch(data = "p(x*,y*,z*,r_x,r_y,r_z)", query = "p(y|do(x))")
#> \sum_{z}\left(\frac{p(z,r_z = 1)}{p(r_z = 1)}p(y|z,r_z = 1,x,r_x = 1,r_y = 1)\right)
```

It is automatically recognized that the problem is a missing data problem when `rflist != NULL`.