Basic R tutorial

February 22, 2024

Contents

Problem 3: Learning Bayesian networks from protein data

Preparation to run the code

Setting seed for reproducibility and loading packages.

```
set.seed(42)
library("GGally")
library("BiDAG")
```
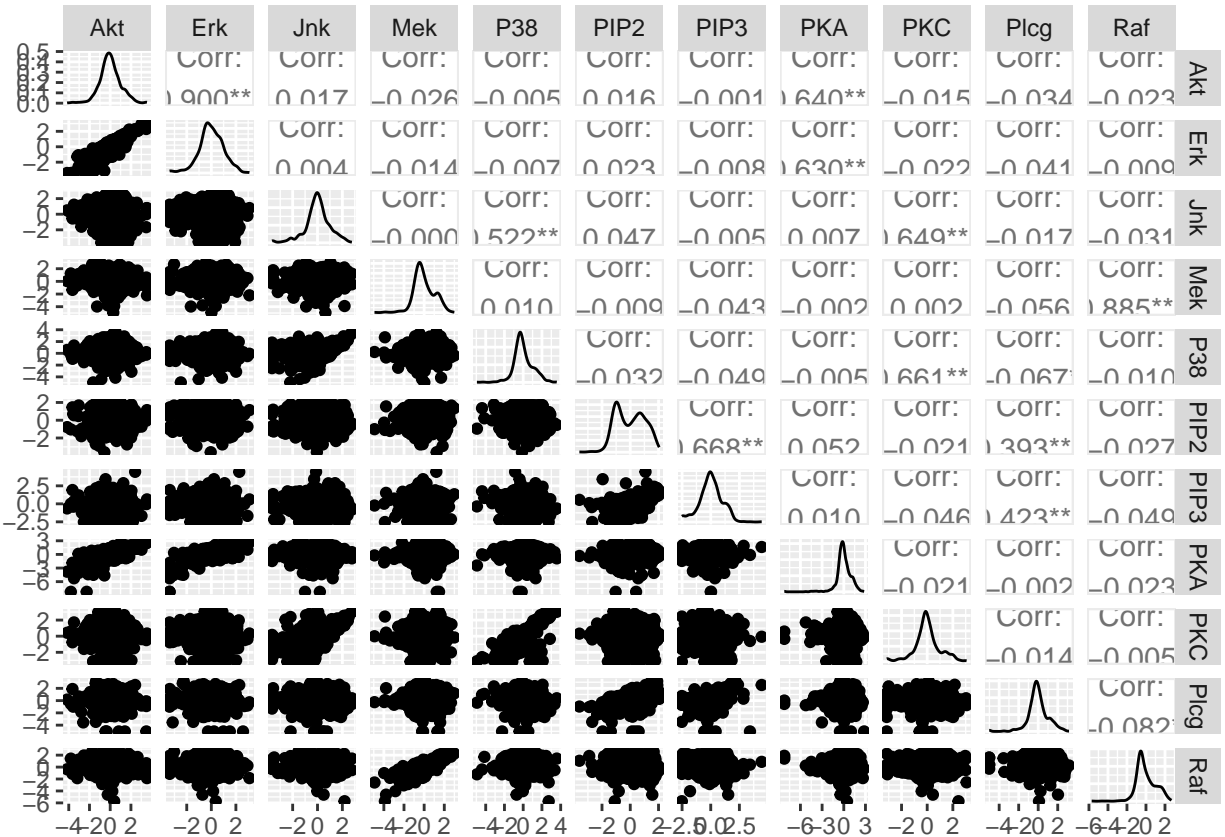
Loading data

```
data <- read.table("2005_sachs_2_cd3cd28icam2_log_std.csv", sep=",", header=TRUE)
```

Variables and observations

```
num_variables <- ncol(data)
num_observations <- nrow(data)
```

Visualisation of transformed data

```
pair_plot <- ggpairs(data, progress=FALSE)
pair_plot
```

Corr:
	Akt	Erk	Jnk	Mek	P38	PIP2	PIP3	PKA	PKC	Plcg	Raf
Akt		0.900**	0.017	−0.026	−0.005	0.016	−0.001	0.640**	−0.015	−0.034	−0.023
Erk			0.004	−0.014	−0.007	0.023	−0.008	0.630**	−0.022	−0.041	−0.009
Jnk				−0.000	0.522**	0.047	−0.005	0.007	0.649**	−0.017	−0.031
Mek					0.010	−0.009	−0.043	−0.002	0.002	−0.056	0.885**
P38						−0.032	−0.049	−0.005	0.661**	−0.067	−0.010
PIP2							0.668**	0.052	−0.021	0.393**	−0.027
PIP3								0.010	−0.046	0.423**	−0.049
PKA									−0.021	−0.002	−0.023
PKC										−0.014	−0.005
Plcg											−0.082
Raf											

Splitting data

TODO: shuffling might be needed as well

```r
train_size <- ceiling(num_observations*0.8)
train_data <- data[row.names(data) %in% 1:train_size, ]
test_data <- data[row.names(data) %in% (train_size+1):nrow(data), ]
```

```r
init_score <- scoreparameters("bge", train_data)
```

```r
learnt_BN <- iterativeMCMC(init_score)
```

```
## maximum parent set size is 2
## core space defined, score table are being computed
## score tables completed, iterative MCMC is running
```

Render this .rmd into a pdf

```r
library(rmarkdown)
render("1/1.Rmd", pdf_document(TRUE), "1.pdf") # TRUE adds table of content
```