Network modelling for high-dimensional data - II

Instructors:

Carel F.W. Peeters & Wessel N. van Wieringen

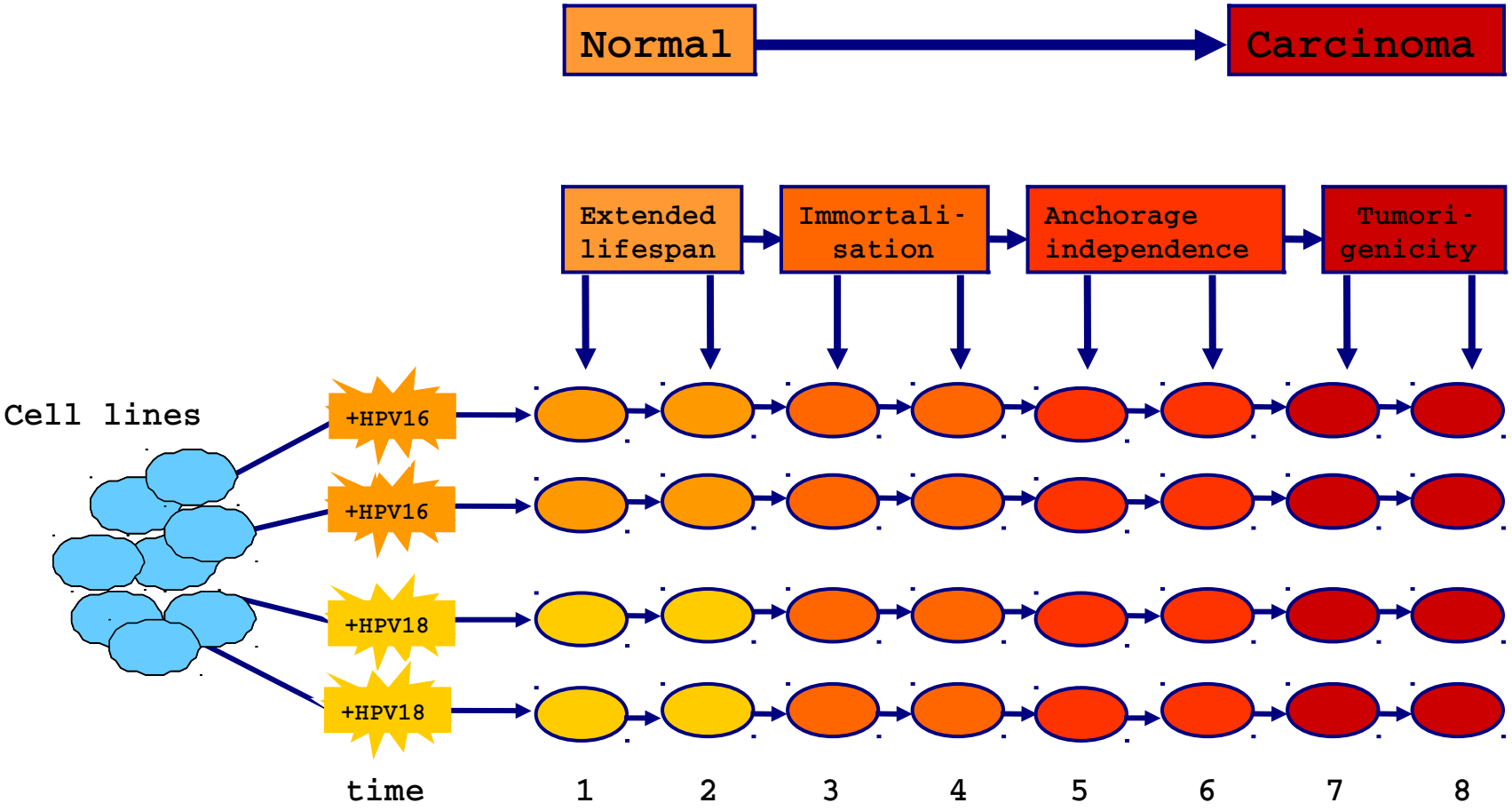
cf.peeters@vumc.nl & w.vanwieringen@vumc.nl

Dept. Epidemiology and Biostatistics, VUmc
& Dept. Mathematics, Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

vrije Universiteit

VU medisch centrum

Experiment


Normal
Carcinoma
Extended lifespan
Immortali-sation
Anchorage independence
Tumori-genicity
Cell lines
+HPV16
+HPV16
+HPV18
+HPV18
time
1
2
3
4
5
6
7
8

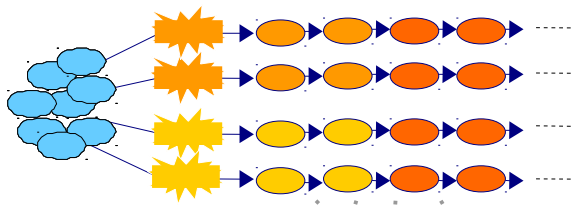
1.1 + 1.2: Libraries + experiment

```
# load libraries
library(Biobase)
library(lattice)
library(longitudinal)
library(rags2ridges)
library(ragt2ridges)
library(SparseTSCGM)

# load data
data(hpvP53)

# reformat data
Y <- longitudinal2array(t(exprs(hpvP53rna)))

# zero center data, variate- and cell line-wise
Y <- centerVAR1data(Y)
```

1.3: Data exploration

```
# plot time-courses of all variates
plotVAR1data(Y)

# plot time-courses of a single variate
plotVAR1data(Y[5, , , drop=FALSE])

# K-means clustering of the variates
cellLine <- 1
kClust   <- kmeans(Y[, , cellLine],
                   centers=7, nstart=100)$cluster

# heatmap of reshuffled data
edgeHeat(Y[unlist(lapply(1:max(kClust),
         function(id, clusters){ which(clusters==id) },
         kClust)), , cellLine])

# plot time-courses of smallest cluster
plotVAR1data(Y[kClust==which.min(table(kClust)), , ,
               drop=FALSE])
```

Aim + model



VAR(1) model:

$$Y_{1,t+1} = + e_{1,t}$$

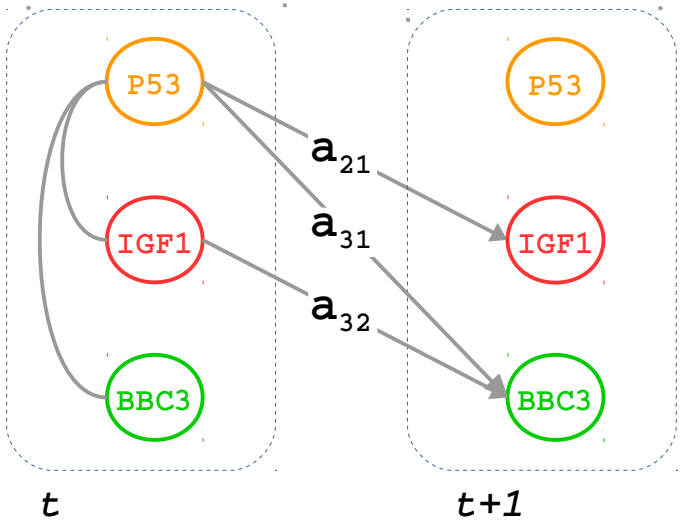
$$Y_{2,t+1} = a_{21}\, Y_{1,t} + e_{2,t}$$

$$Y_{3,t+1} = a_{31}\, Y_{1,t} + a_{32}\, Y_{2,t} + e_{3,t}$$

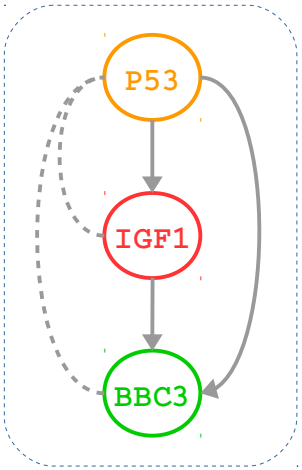
where:

$Y_1 \leftrightarrow$ P53; $Y_2 \leftrightarrow$ IGF1; $Y_3 \leftrightarrow$ BBC3

Unrolled:



Curled up:



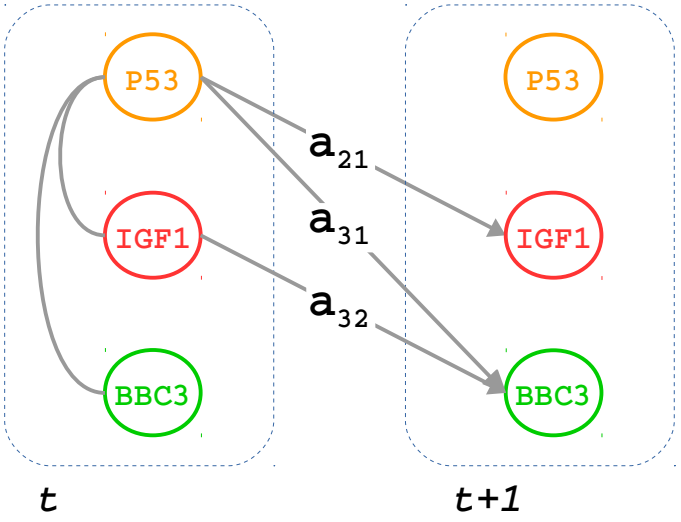
Model

VAR(1) model:

$$Y_{1,t+1} = + e_{1,t}$$
$$Y_{2,t+1} = a_{21}\, Y_{1,t} + e_{2,t}$$
$$Y_{3,t+1} = a_{31}\, Y_{1,t} + a_{32}\, Y_{2,t} + e_{3,t}$$

where:

$Y_1 \leftrightarrow$ P53; $Y_2 \leftrightarrow$ IGF1; $Y_3 \leftrightarrow$ BBC3



In matrix notation:

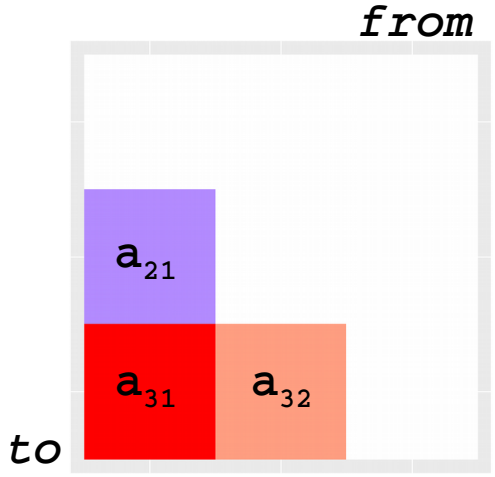
$$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_{t-1} + \boldsymbol{\varepsilon}_t$$

with $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}_{p\times 1}, \boldsymbol{\Omega}_{\varepsilon}^{-1})$

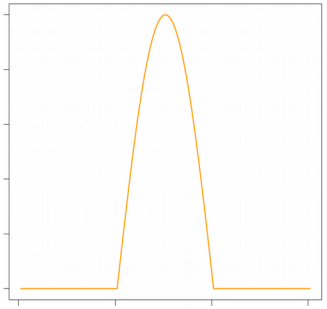
A and ?: propagation of endo- and exogenous signal, resp.

? : innovation of the system.

Parameters as heatmap:



Model



exogenous pulse

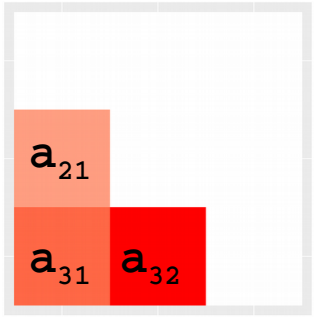
VAR(1) model:

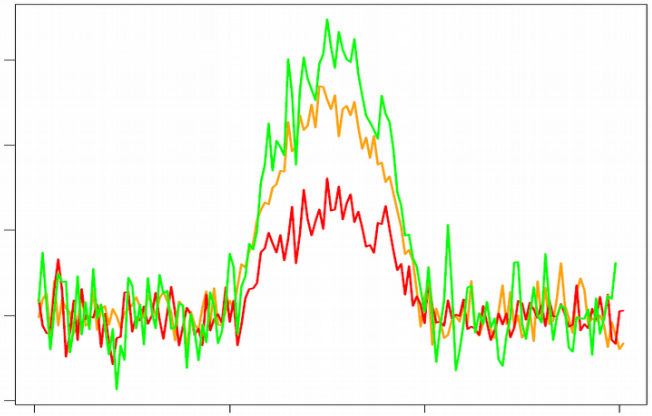
$$Y_{1,t+1} = + e_{1,t}$$

$$Y_{2,t+1} = a_{21} Y_{1,t} + e_{2,t}$$

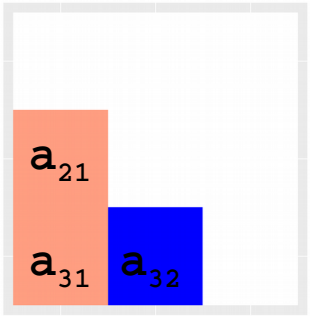
$$Y_{3,t+1} = a_{31} Y_{1,t} + a_{32} Y_{2,t} + e_{3,t}$$

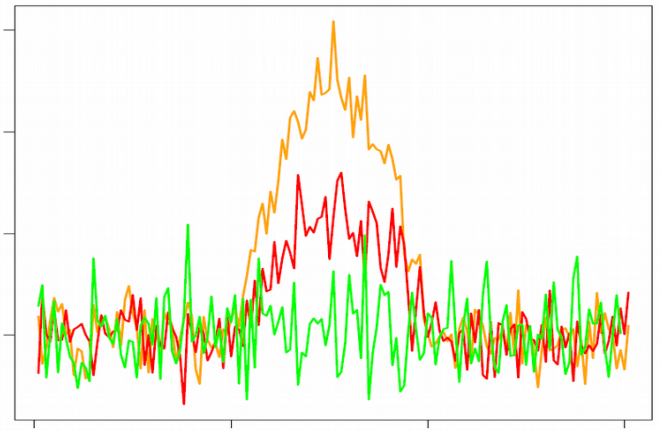
Coherent feed forward loop:



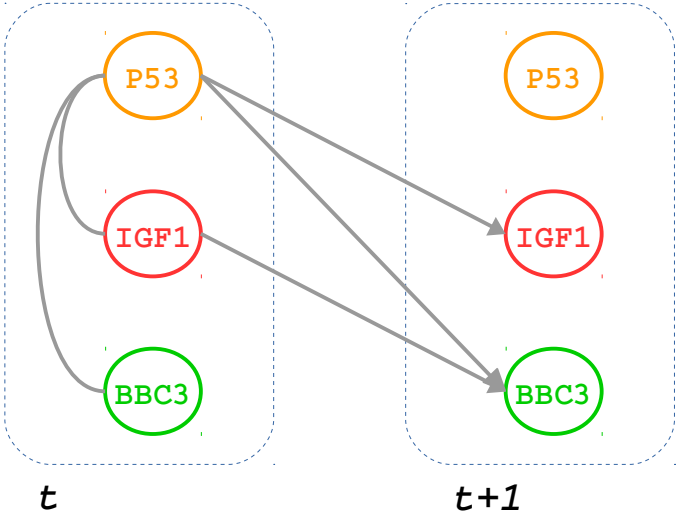


Incoherent feed forward loop:





TSCG



Time Series Chain Graph (TSCG)
→ conditional (in)dependencies

Temporal CI:

$$Y_{j_1,t} \perp\!\!\!\perp Y_{j_2,t+1} \mid \text{other } Y_{j,t}\text{'s}$$

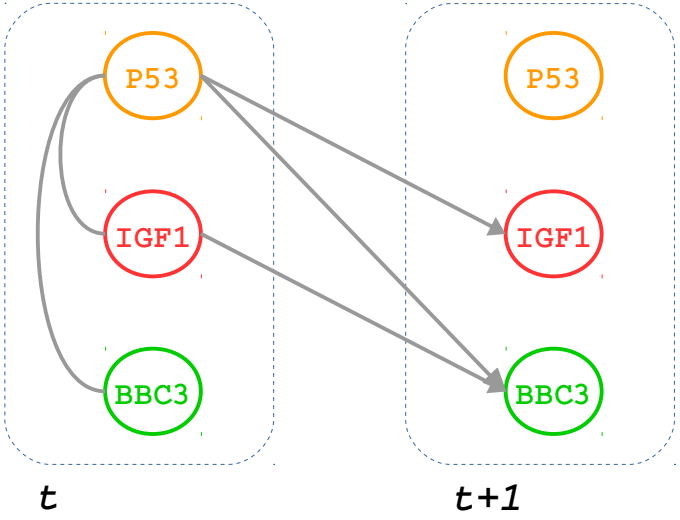
Example:
→ CI: BBC3 & P53 (*no* edge)
→ not CI: P53 & IGF1 (*directed* edge)

Contemporaneous CI:

$$Y_{j_1,t} \perp\!\!\!\perp Y_{j_2,t} \mid Y_{j,t-1}\text{'s, other } Y_{j,t}\text{'s}$$

Example:
→ CI: BBC3 & IGF1 (*no* edge)
→ not CI: P53 & IGF1 (*undirected* edge)

TSCG

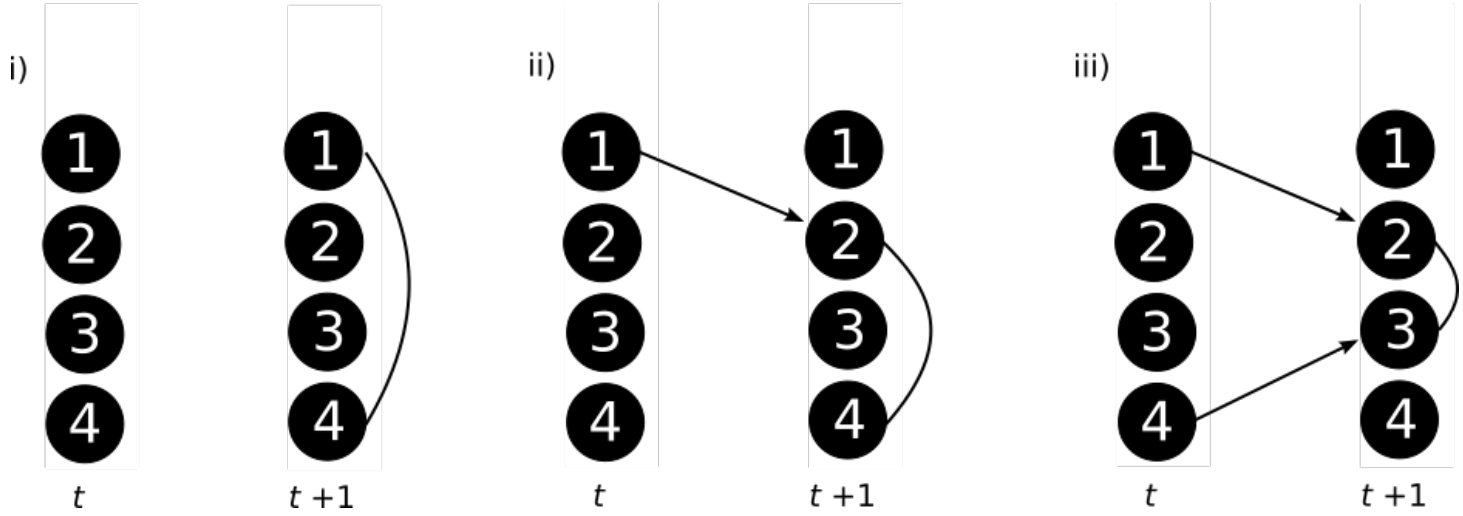


TSCG harbors global conditional (in)dependencies:

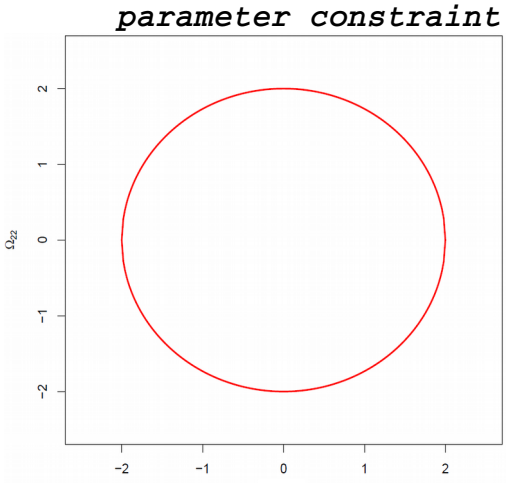
$$Y_{j_1,*} \perp\!\!\!\perp Y_{j_2,*} \mid \text{other } Y_{j,*}\text{'s}$$

Wermuth condition for CI: unconnected nodes may not exert influence on the same node.

CI of nodes 1 and 4 forbids motifs:



Estimation



Augment log-likelihood
with ridge penalty:

$$\lambda_a \|\mathbf{A}\|_2^2 + \lambda_\omega \|\mathbf{\Omega}_\varepsilon\|_2^2$$

Include targets.

Analytic estimators:

$$\mathrm{vec}[\hat{\mathbf{A}}(\lambda_a)] = [\lambda_a \mathbf{I}_{p^2 \times p^2} + \hat{\mathbf{\Gamma}}(0) \otimes \mathbf{\Omega}_\varepsilon]^{-1} \mathrm{vec}[\mathbf{\Omega}_\varepsilon \hat{\mathbf{\Gamma}}(-1)],$$

$$\hat{\mathbf{\Omega}}_\varepsilon(\lambda_\omega) = \left\{ \left[\lambda_\omega \mathbf{I}_{p \times p} + \tfrac{1}{4}\mathbf{S}_\varepsilon^2\right]^{1/2} + \tfrac{1}{2}\mathbf{S}_\varepsilon \right\}^{-1}.$$

Initiate and iterate.

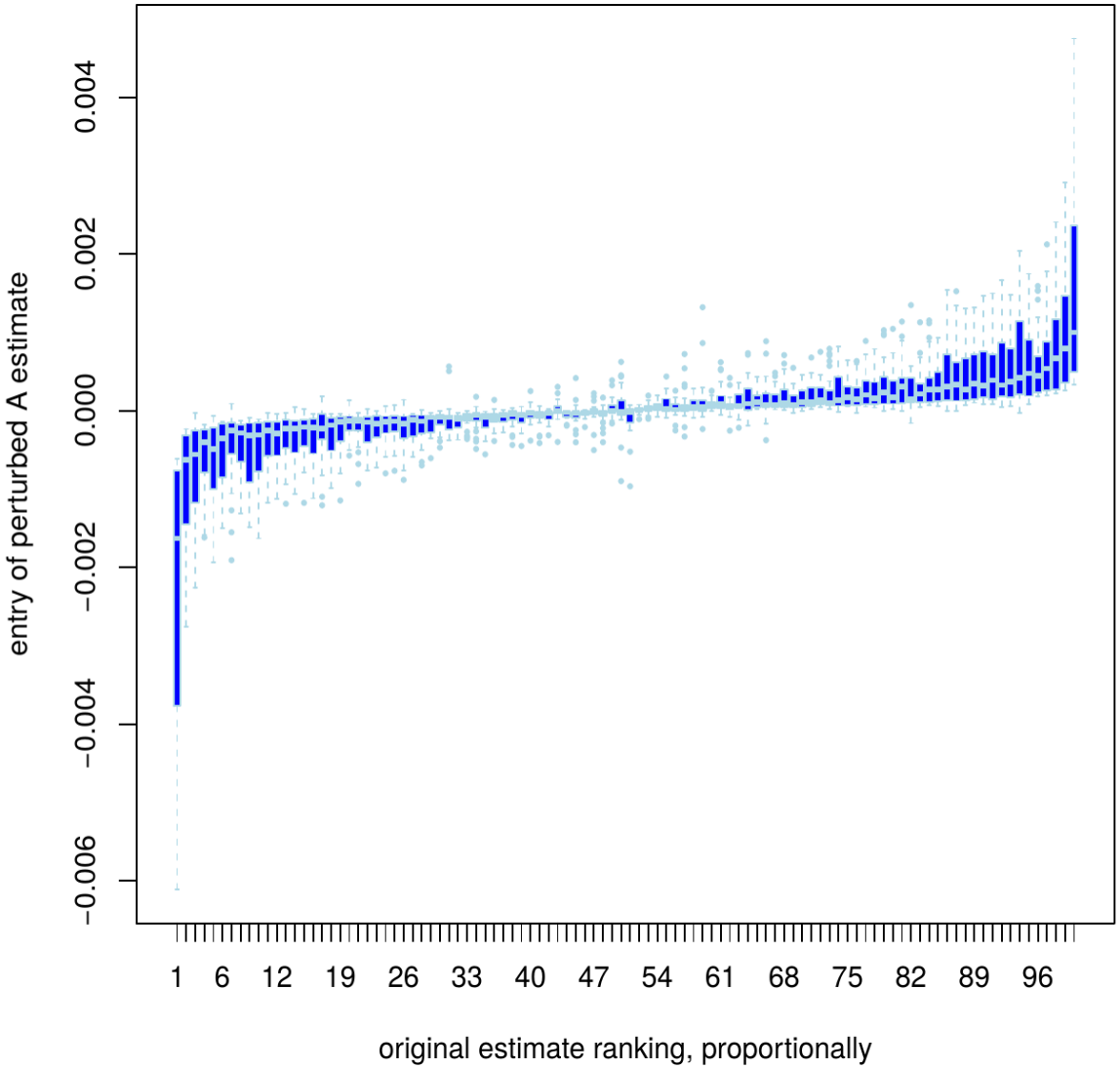
Rewrite vec(A) for computational efficiency

1.4 + 1.5: Penalty + estimation

```
# find optimal penalty parameters
optLambdas <- optPenaltyVAR1(Y, lambdaMin =c(0.01, 0.00001),
                                lambdaMax =c(1000, 1))

# fit the model
VAR1hat <- ridgeVAR1(Y=Y, lambdaA=optLambdas[1],
                          lambdaP=optLambdas[2])

# extract parameter estimates
Ahat <- VAR1hat$A; Phat <- VAR1hat$P

# add row and column names
rownames(Ahat) <- colnames(Ahat) <- rownames(Phat) <-
                  colnames(Phat) <- rownames(hpvP53rna)

# heatmap of estimate of A
edgeHeat(Ahat, main="ridge estimate of A")

# heatmap of partial correlation matrix estimate
edgeHeat(pcor(Phat), main="part. Cor. estimate", diag=FALSE)
```

1.6: Stability

```
# Homework: code takes too long for practical
```



Support

Support from: KEGG

Known

Equality constraints for known absent temporal edges:

$$\max_{\{\mathbf{A}_c : \mathbf{C}\,\mathrm{vec}(\mathbf{A}_c)=\mathbf{d}\}} \mathcal{L}^{\mathrm{pen}}(\mathbf{Y}; \mathbf{A}_c, \mathbf{\Omega}_\varepsilon; \lambda_a, \lambda_\omega)$$

Analytic solution, efficiently evaluable.

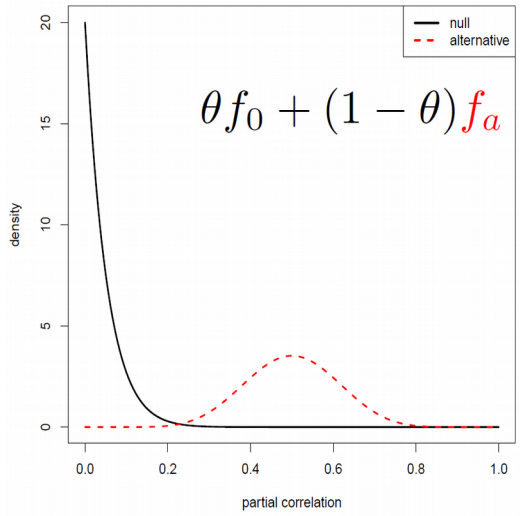
Similarly for known absent contemporaneous edges.

Infer

Sparsification by:

→ empirical Bayes,

→ largest elements.



1.7: Support determination + ...

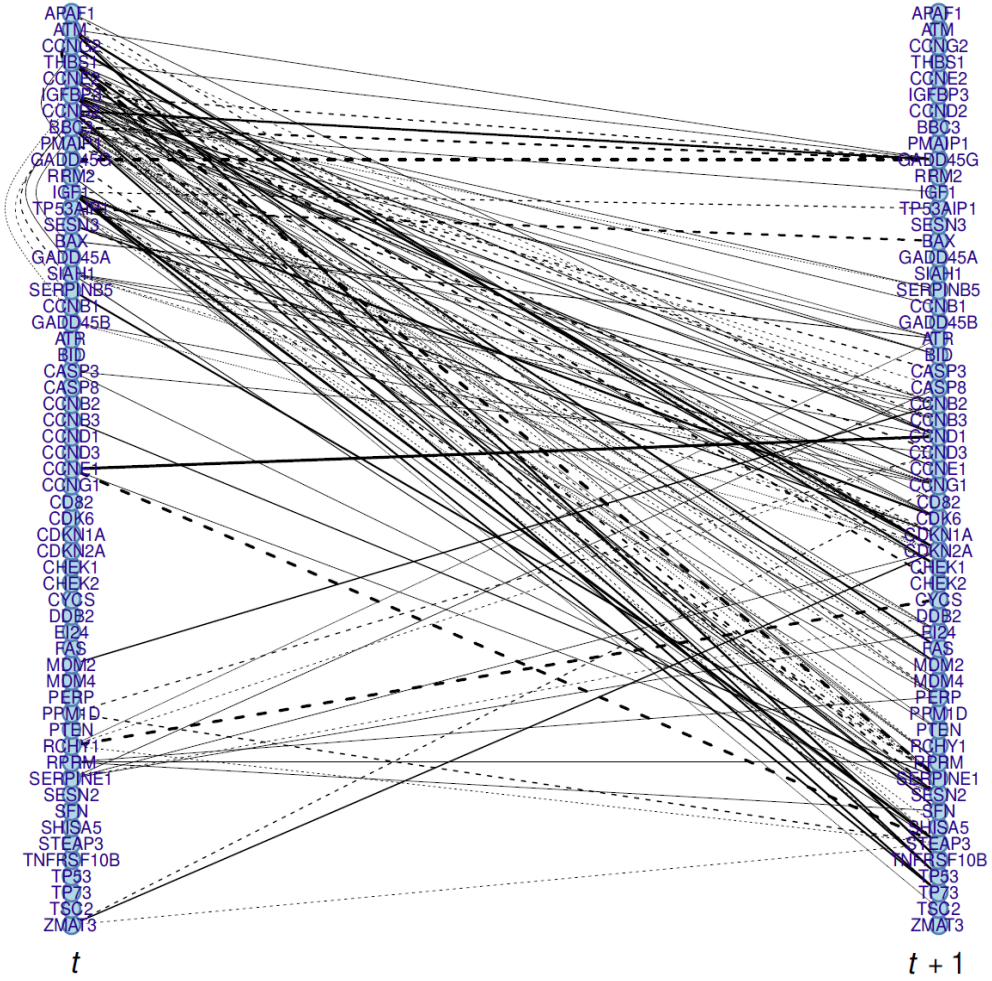
```
# support determination of A
zerosA <- sparsifyVAR1(A=Ahat, SigmaE=symm(solve(Phat)),
                       threshold="localFDR", FDRcut=0.95,
                       statistics=FALSE, verbose=FALSE)$zeros

# support determination of precision matrix
zerosP <- sparsify(Phat, threshold="localFDR", FDRcut=0.95,
                   output="light", verbose=FALSE)$zeros

# format precision support
supportP <- support4ridgeP(zeros=zerosP, nNodes=nrow(Y))

# optimal penalty parameter determination
optLambdas <- optPenaltyVAR1(Y, lambdaMin=rep(10^(-5), 2),
                             lambdaMax=c(10, 0.1),
                             lambdaInit=c(5, 0.01),
                             zerosA=zerosA, zerosP=zerosP,
                             cliquesP=supportP$cliques,
                             separatorsP=supportP$separators,
                             zerosAfit="sparse")
```

1.8: Re-estimation

```
# re-fit the model
VAR1hat         <- ridgeVAR1(Y=Y, lambdaA=optLambdas[1],
                             lambdaP=optLambdas[2],
                             zerosA=zerosA,
                             cliquesP=supportP$cliques,
                             separatorsP=supportP$separators,
                             zerosP=zerosP, zerosAfit="sparse")

# extract parameter estimates
Ahat <- VAR1hat$A; Phat <- VAR1hat$P

# add row and column names
rownames(Ahat) <- colnames(Ahat) <- rownames(Phat) <-
                  colnames(Phat) <- rownames(hpvP53rna)

# time-series chain graph
graphVAR1(Ahat, Phat, nNames=rownames(Ahat), type="TSCG",
          vertex.label.cex=0.5, vertex.label.font=1,
          vertex.size=4, edge.width=1.5)
```

Application

HPV-induced oncogenesis
→ P53 pathway,
→ 64 genes,
→ fit + LOOCV,
→ sparsify A and ?,
→ re-fit + LOOCV.



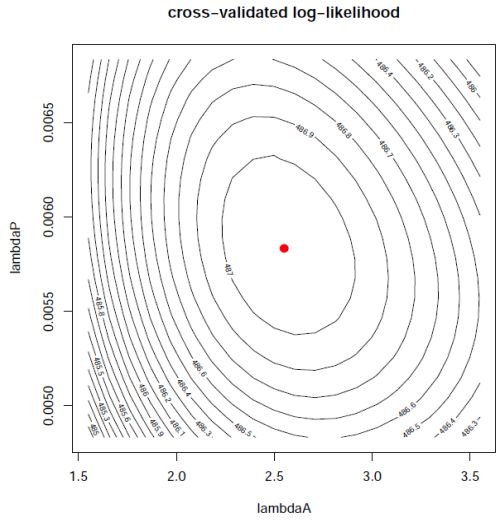
Biology
→ P53 knocked-out by HPV: zero connections
→ HPV-related IGF1/IGFBP3: many connections.

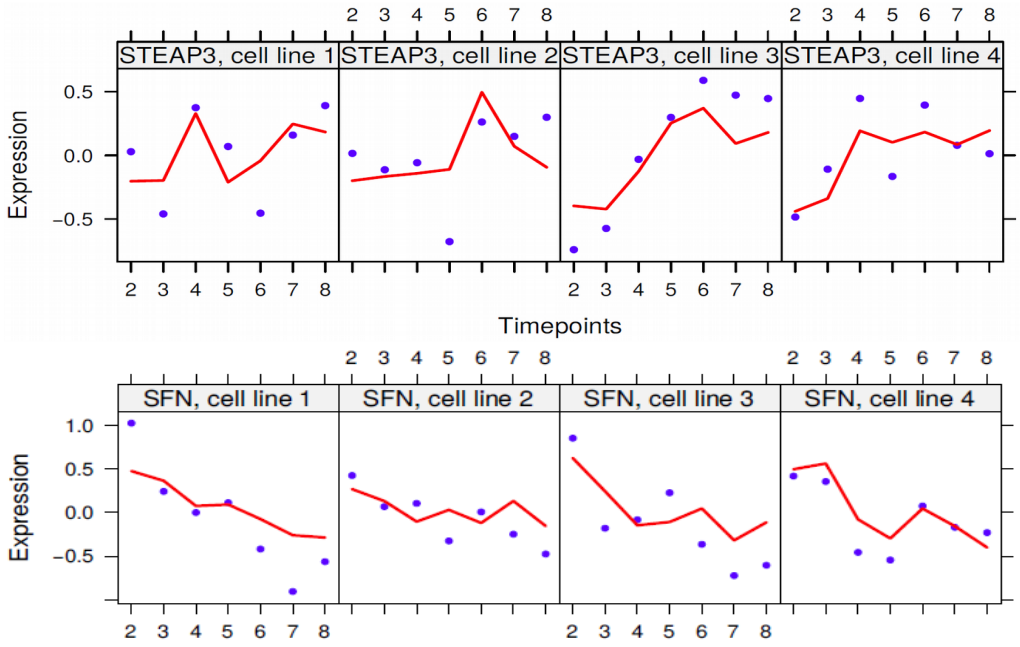
Application

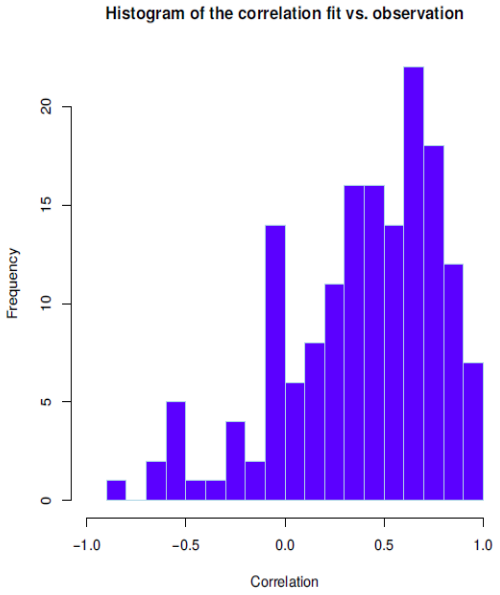
Diagnostics

→ LOOCV contourplot,

→ fit vs. data.







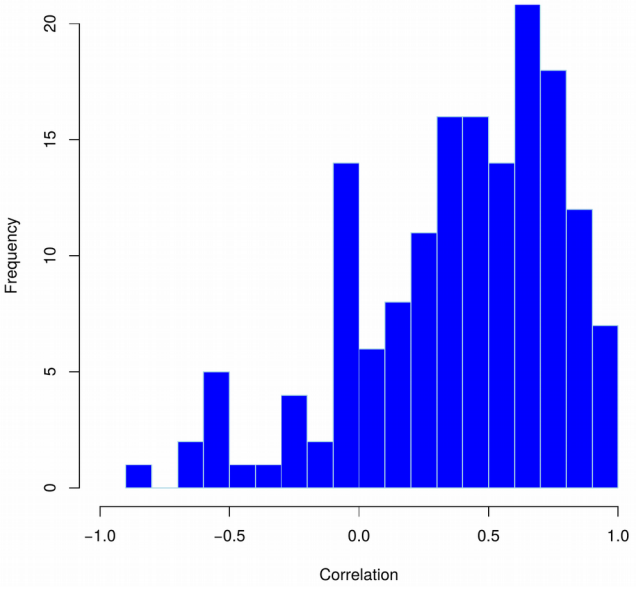
Application

Fit better than:

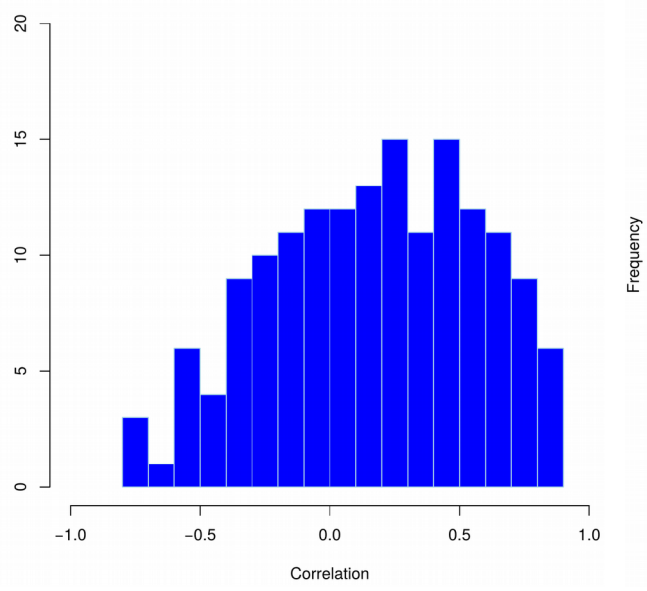
→ Previous observation,

→ Average other cell lines.

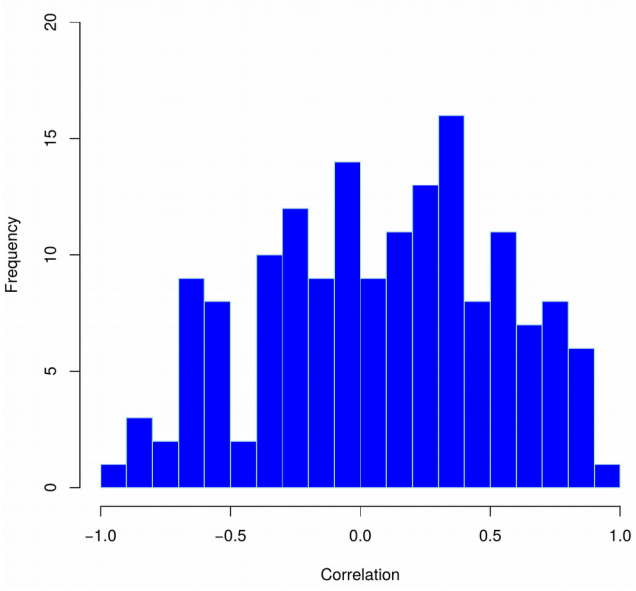
$r(Y_{obs}, Y_{VAR1})$



$r(Y_{obs}, Y_{prev})$



$r(Y_{obs}, Y_{aver.})$



1.9: Evaluate fit

```
# calculate fits and fit-vs-observation correlations
Yhat <- array(dim=dim(Y))
for (i in 1:dim(Y)[3]){
   Yhat[, -1, i] <- Ahat %*% Y[, -dim(Y)[2], i]
}

# calculate fits and fit-vs-observation correlations
for (i in 1:dim(Y)[3]){
   Yhat[, -1, i] <- Ahat %*% Y[, -dim(Y)[2], i]
}
corFit <- numeric()
for (j in 1:dim(Y)[1]){
    slHelper <- numeric()
    for (i in 1:dim(Y)[3]){
      slHelper <- c(slHelper, cor(Yhat[j, -1, i],
                                  Y[j, -1, i], m="s"))
    }
    corFit <- rbind(corFit, slHelper)
}
hist(corFit, xlab="Correlation", ylab="Frequency", n=20)
```

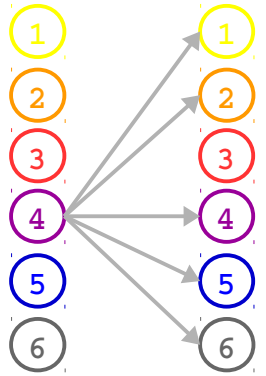
Application

Node analysis

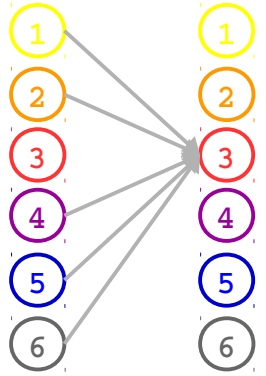
	$\deg^-(\mathbf{A})$	$\deg^+(\mathbf{A})$	between.	close.	eigen centr.
BBC3	0	17	17	0.00192	1.00000
CCND2	0	12	18	0.00187	0.68783
IGF1	1	14	0	0.00191	0.97635
IGFBP3	0	16	7	0.00190	0.87513
THBS1	0	11	0	0.00188	0.87717
CCNG1	6	0	0	0.00177	0.25154
CDKN2A	12	0	0	0.00181	0.49508
SERPINE1	8	4	0	0.00185	0.70869
SESN2	8	0	0	0.00180	0.26759
STEAP3	9	0	0	0.00179	0.36285

```
upper five ? 'regulators'
lower five ? 'regulatees'
```

'regulator'



'regulatee'



Downstream

Mutual information.

A generalized 'correlation' measure, between a variate at some time and all variates at a future time point (given the past):

$$\mathcal{I}(\mathbf{Y}_{*,t+\tau}, Y_{j,t} \,|\, \mathbf{Y}_{*,t-1}) = \mathcal{H}(\mathbf{Y}_{*,t+\tau} \,|\, \mathbf{Y}_{*,t-1}) - \mathcal{H}(\mathbf{Y}_{*,t+\tau} \,|\, Y_{j,t}, \mathbf{Y}_{*,t-1})$$

where:

$$\mathcal{H}(\mathbf{Y}_{*,t+\tau} \,|\, \mathbf{Y}_{*,t-1}) = \log\left(\left|\mathbb{V}(\mathbf{Y}_{*,t+\tau} \,|\, \mathbf{Y}_{*,t-1})\right|\right)$$

E.g.:

'regulator' ($deg^+=17$):

→ $I(all_{t+1}, BBC3_t | ...) = 0.05605$

'regulatee' ($deg^-=12$):

→ $I(all_{t+1}, CDKN2A_t | ...) = 0.00000$

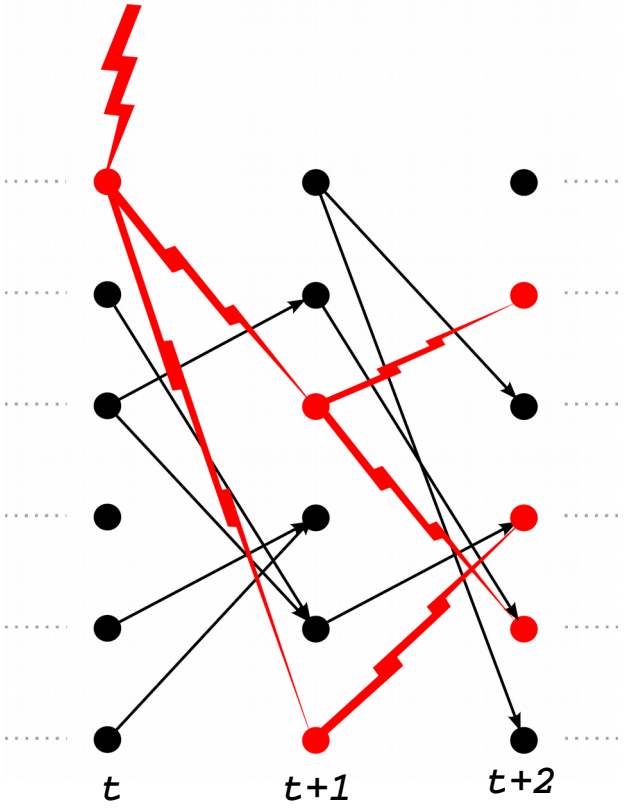
Downstream

Impulse response.

Assess the change in variates at future time points due to a change in a variate at the current time:

$$\frac{\partial \mathbf{Y}_{*,t+\tau}}{\partial \boldsymbol{\varepsilon}_{*,t}} = \mathbf{A}^{\tau}$$

Facilitates prediction of knock-out effect.



E.g.:

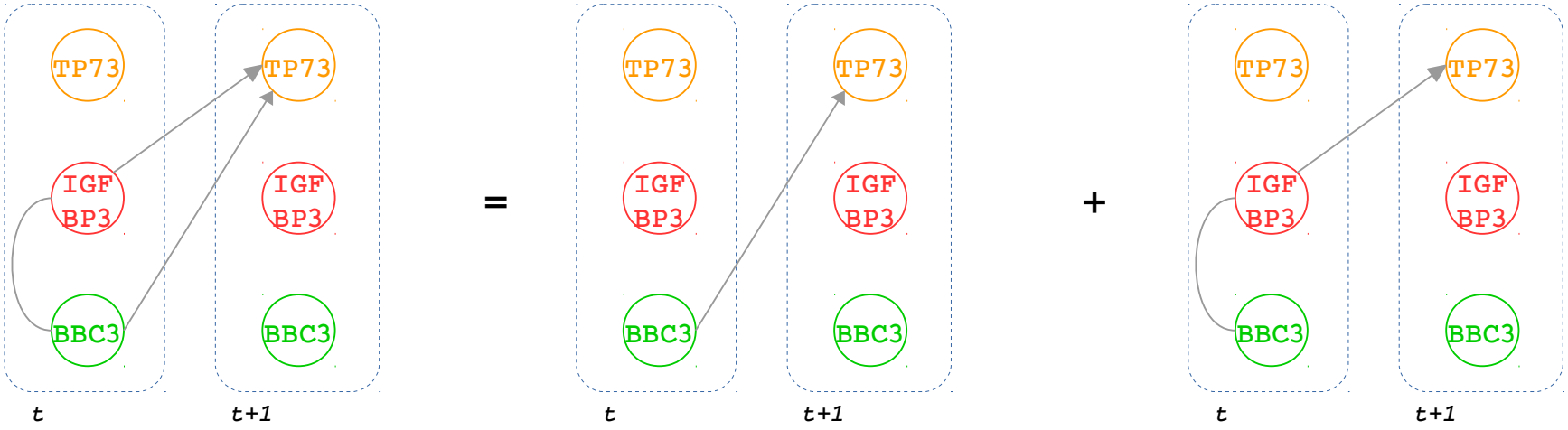
→ 'regulator' (deg^{+}=17): effect of $BBC3_{t}$ on all_{t+1} = 0.01497

→ 'regulatee' (deg^{-}=12): effect of $CDKN2A_{t}$ on all_{t+1} = 0.00000

Downstream

Covariance decomposition.

Given TSCG, decompose conditional covariance in terms of paths:



E.g.:
$Cov(BBC3_t, TP73_{t+1} \mid ...)$:

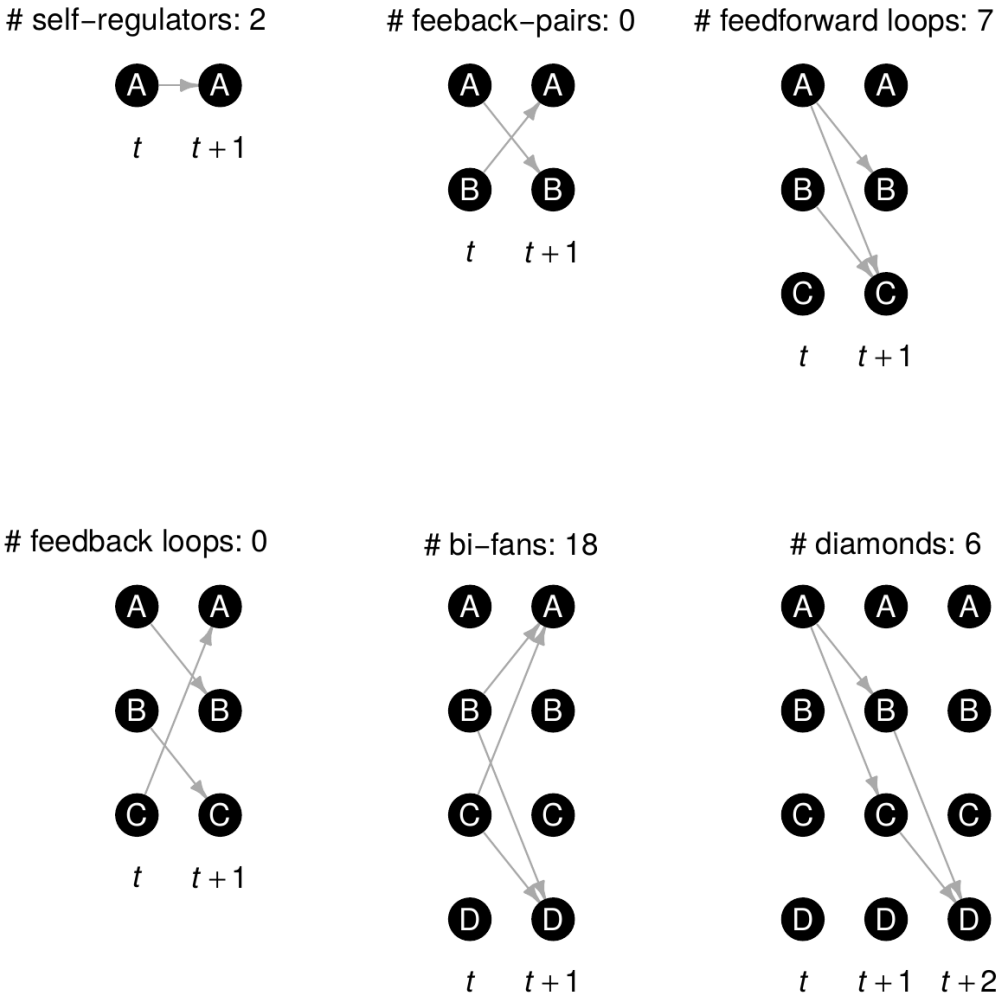
$-0.003168 = -0.002483 + -0.0006845$

1.10: Down-stream analysis

```
# calculate node-wise network stats
nodeStats <- nodeStatsVAR1(Ahat, Phat, as.table=TRUE)

# show node degree table
print(nodeStats[, 1:7])

# specify time lag
lag <- 1

# evaluate mutual informations with specified lag
MIs <- mutualInfoVAR1(Ahat, solve(Phat), lag)

# evaluate impulse response with specified lag
IRs <- impulseResponseVAR1(Ahat, lag)
```

1.10: Down-stream analysis

```
# motif detection
motifStatsVAR1(Ahat)
```



References + contributors

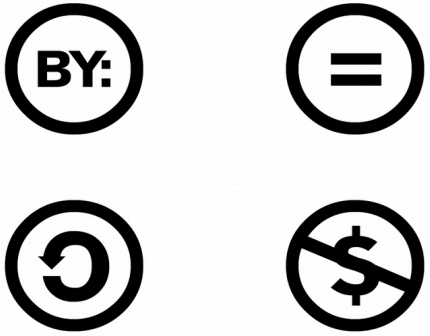
- → Viktorian Miok
- → Renske Steenbergen
- → Saskia Wilting

Based on:

Miok, V., Wilting, S.M., Van Wieringen, W.N. (2017), "Ridge estimation of the VAR(1) model and its time-series chain graph model from multivariate time-course omics data", *Biometrical Journal*, 59(1), 172-191.

vrije Universiteit

VU medisch centrum



This material is provided under the Creative Commons Attribution/Share-Alike/Non-Commercial License.

See http://www.creativecommons.org for details.

Slides by: Wessel N. van Wieringen