

Graphical models and Bayesian networks with R

Tutorial at University of Oslo, November 2012

1. Task: Start out from the saturated model; do stepwise backward model selection among decomposable models to obtain "reduced" model.

```
library(gRim)
data(cadl, package = "gRbase")
# specify saturated model
m.sat <- dmod(~.^., data = cadl)
# Stepwise backward model selection among decomposable models using AIC as
# selection criterion
m.back1 <- stepwise(m.sat, details = 1)
```

```
## STEPWISE:
## criterion: aic ( k = 2 )
## direction: backward
## type      : decomposable
## search    : all
## steps     : 1000
## . BACKWARD: type=decomposable search=all, criterion=aic(2.00), alpha=0.00
## . Initial model: is graphical=TRUE is decomposable=TRUE
## change.AIC   -7.2946 Edge deleted: Hyperchol SuffHeartF
## change.AIC   -9.9724 Edge deleted: SuffHeartF STchange
## change.AIC   -8.9869 Edge deleted: Hyperchol STchange
## change.AIC   -6.9429 Edge deleted: Hyperchol AMI
## change.AIC   -5.0911 Edge deleted: Inherit SuffHeartF
## change.AIC   -8.3854 Edge deleted: SuffHeartF AngPec
## change.AIC   -5.8491 Edge deleted: Inherit AMI
## change.AIC   -3.3619 Edge deleted: SuffHeartF AMI
## change.AIC   -4.7562 Edge deleted: SuffHeartF Sex
## change.AIC   -4.8269 Edge deleted: SuffHeartF Smoker
## change.AIC   -3.8263 Edge deleted: AMI Hypertrophi
## change.AIC   -3.1899 Edge deleted: STchange Inherit
## change.AIC   -2.6369 Edge deleted: CAD Hyperchol
## change.AIC   -7.6756 Edge deleted: Inherit CAD
## change.AIC   -2.1665 Edge deleted: SuffHeartF QWavecode
## change.AIC   -2.9359 Edge deleted: SuffHeartF QWave
## change.AIC   -1.7686 Edge deleted: QWavecode AMI
## change.AIC   -0.6979 Edge deleted: Inherit Hyperchol
## change.AIC   -2.5723 Edge deleted: Inherit QWavecode
## change.AIC   -1.5910 Edge deleted: Hyperchol Hypertrophi
## change.AIC   -3.7745 Edge deleted: Hypertrophi QWavecode
## change.AIC   -1.1525 Edge deleted: Inherit QWave
## change.AIC   -3.9161 Edge deleted: QWave Hypertrophi
```

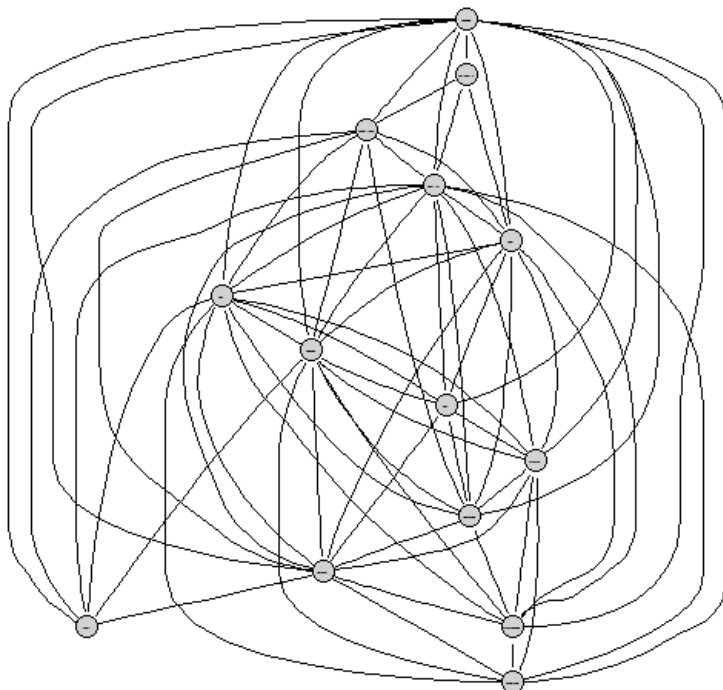
```
m.back1
```

```
## Model: A dModel with 14 variables
## graphical : TRUE decomposable : TRUE
## -2logL      :      2669.64 mdim : 1647 aic :      5963.64
## ideviance  :      1346.81 idf  : 1632 bic :     11668.57
## deviance   :       276.63 df   : 22928
## Notice: Table is sparse
## Asymptotic chi2 distribution may be questionable.
## Degrees of freedom can not be trusted.
## Model dimension adjusted for sparsity : 167
```

```
summary(m.back1)
```

```
## is graphical=TRUE; is decomposable=TRUE
## generators (glist):
## : "STcode" "SuffHeartF" "Hypertrophi" "Heartfail" ...
## : "Sex" "AngPec" "AMI" "QWave" ...
## : "Sex" "AngPec" "QWave" "QWavecode" ...
## : "Sex" "AngPec" "QWave" "QWavecode" ...
## : "Sex" "AngPec" "STcode" "Hypertrophi" ...
## : "Sex" "AngPec" "STcode" "STchange" ...
```

```
plot(m.back1)
```



1. Task: Start out from the independence model; do stepwise forward model selection among decomposable models to obtain “expanded” model.

```
# Specify independence model
m.ind <- dmod(~.1, data = cadl)
# Stepwise forward selection among decomposable models using AIC as
# selection criterion.
m.forw1 <- stepwise(m.ind, direction = "forward", details = 1)
```

```
## STEPWISE:
## criterion: aic ( k = 2 )
## direction: forward
## type : decomposable
## search : all
## steps : 1000
## . FORWARD: type=decomposable search=all, criterion=aic(2.00), alpha=0.00
## . Initial model: is graphical=TRUE is decomposable=TRUE
## change.AIC -119.2448 Edge added: STchange STcode
## change.AIC -76.5649 Edge added: CAD AngPec
## change.AIC -64.5275 Edge added: Heartfail Hypertrophi
## change.AIC -57.6549 Edge added: STcode SuffHeartF
## change.AIC -43.9098 Edge added: CAD AMI
## change.AIC -43.8447 Edge added: CAD QWave
## change.AIC -32.2871 Edge added: CAD Hyperchol
## change.AIC -24.2890 Edge added: CAD STchange
## change.AIC -21.2167 Edge added: SuffHeartF Heartfail
## change.AIC -17.1157 Edge added: CAD Inherit
## change.AIC -15.1030 Edge added: QWave AMI
## change.AIC -13.7300 Edge added: CAD Smoker
## change.AIC -13.4485 Edge added: STcode QWavecode
## change.AIC -5.6746 Edge added: CAD Sex
## change.AIC -4.3187 Edge added: AngPec STchange
## change.AIC -12.6785 Edge added: AngPec STcode
## change.AIC -4.0464 Edge added: STchange AMI
## change.AIC -2.3670 Edge added: SuffHeartF Hypertrophi
## change.AIC -1.9598 Edge added: Heartfail STcode
## change.AIC -6.3689 Edge added: Heartfail STchange
## change.AIC -1.7043 Edge added: AngPec QWavecode
## change.AIC -1.5107 Edge added: STcode CAD
## change.AIC -20.0465 Edge added: Heartfail CAD
## change.AIC -4.4059 Edge added: Heartfail AMI
## change.AIC -1.3905 Edge added: AngPec Inherit
## change.AIC -0.9806 Edge added: AMI Hyperchol
## change.AIC -0.4710 Edge added: AngPec Heartfail
## change.AIC -4.6389 Edge added: AngPec AMI
## change.AIC -14.9313 Edge added: AngPec QWave
## change.AIC -7.4308 Edge added: AMI STcode
## change.AIC -4.4333 Edge added: QWavecode Heartfail
## change.AIC -3.1649 Edge added: STchange QWave
## change.AIC -0.2581 Edge added: Heartfail Smoker
```

```
m.forw1
```

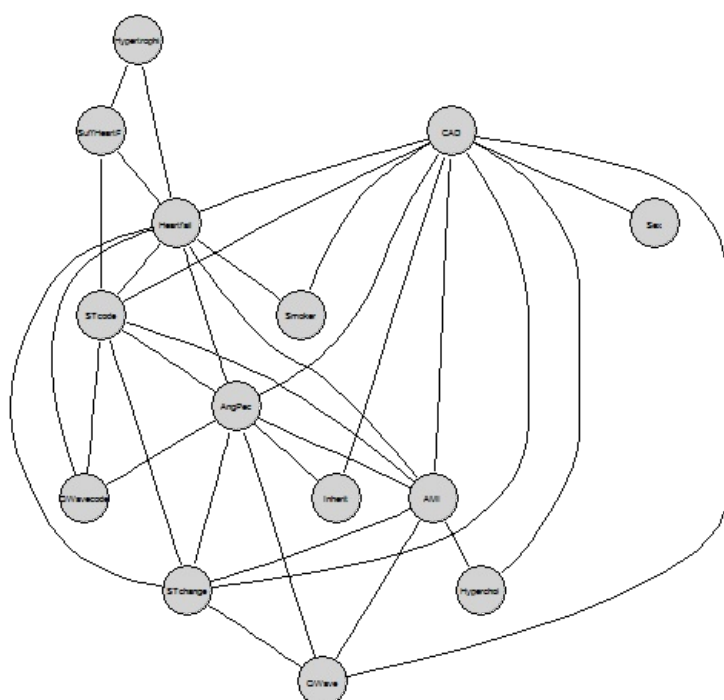
```
## Model: A dModel with 14 variables
```

```
## graphical : TRUE decomposable : TRUE
## -2logL : 3176.73 mdim : 155 aic : 3486.73
## ideviance : 839.72 idf : 140 bic : 4023.62
## deviance : 783.72 df : 24420
## Notice: Table is sparse
## Asymptotic chi2 distribution may be questionable.
## Degrees of freedom can not be trusted.
## Model dimension adjusted for sparsity : 77
```

```
summary(m.forw1)
```

```
## is graphical=TRUE; is decomposable=TRUE
## generators (glist):
## : "CAD" "Sex"
## : "Hypertrophi" "SuffHeartF" "Heartfail"
## : "Heartfail" "STcode" "SuffHeartF"
## : "AngPec" "Inherit" "CAD"
## : "AMI" "Hyperchol" "CAD"
## : "AMI" "STcode" "AngPec" "Heartfail" ...
## : "Heartfail" "QWavecode" "STcode" "AngPec"
## : "QWave" "STchange" "AngPec" "AMI" ...
## : "Heartfail" "Smoker" "CAD"
```

```
plot(m.forw1)
```



1. Task: Can you identify direct and indirect “risk factors” from these models?

We can at least identify variables that are directly and indirectly related to CAD in the two models.

```
x <- terms(m.back1)[!is.na(sapply(terms(m.back1), function(g) match("CAD", g)))]
setdiff(unique(unlist(x)), "CAD")
```

```
## [1] "STcode" "SuffHeartF" "Hypertrophi" "Heartfail" "Sex"
## [6] "AngPec" "AMI" "QWave" "STchange" "Smoker"
## [11] "QWavecode"
```

```
x <- terms(m.forw1)[!is.na(sapply(terms(m.forw1), function(g) match("CAD", g)))]
setdiff(unique(unlist(x)), "CAD")
```

```
## [1] "Sex" "AngPec" "Inherit" "AMI" "Hyperchol"
## [6] "STcode" "Heartfail" "STchange" "QWave" "Smoker"
```

1. “Convert” these two models to a Bayesian network.

```
bn.back1 <- compile(grain(ugList(terms(m.back1)), data = cad1, smooth = 0.001))
bn.forw1 <- compile(grain(ugList(terms(m.forw1)), data = cad1, smooth = 0.001))
```

1. Predict the disease variable CAD in the validation dataset cad2 (which has incomplete observations).

Validation data looks as follows (notice: missing values):

```
data(cad2, package = "gRbase")
table(cad2$CAD)
```

```
##
## No Yes
## 41 26
```

```
classMat <- function(trueLabel, classLabel) {
  tt <- table(trueLabel, classLabel)
  sweep(tt, 1, apply(tt, 1, sum), FUN = "/")
}
class.back1 <- predict(bn.back1, resp = "CAD", newdata = cad2, type = "class")
class.forw1 <- predict(bn.forw1, resp = "CAD", newdata = cad2, type = "class")
prob.back1 <- predict(bn.back1, resp = "CAD", newdata = cad2, type = "dist")
prob.forw1 <- predict(bn.forw1, resp = "CAD", newdata = cad2, type = "dist")
classMat(cad2$CAD, class.back1$pred$CAD)
```

```
##          classLabel
## trueLabel      No      Yes
##      No 0.97561 0.02439
##      Yes 0.34615 0.65385
```

```
classMat(cad2$CAD, class.forw1$pred$CAD)
```

```
##          classLabel
## trueLabel      No      Yes
##      No 0.95122 0.04878
##      Yes 0.38462 0.61538
```

1. How good prediction results can you obtain? Can you improve your result by changing the model selection criterion, for example going from AIC to BIC?

```
m.back2 <- stepwise(m.sat, k = log(nrow(cad1)), details = 1)
```

```
## STEPWISE:
## criterion: aic ( k = 5.46 )
## direction: backward
## type      : decomposable
## search    : all
## steps     : 1000
## . BACKWARD: type=decomposable search=all, criterion=aic(5.46), alpha=0.00
## . Initial model: is graphical=TRUE is decomposable=TRUE
## change.AIC -68.3685 Edge deleted: Inherit AngPec
## change.AIC -75.2553 Edge deleted: Inherit Hyperchol
## change.AIC -67.1031 Edge deleted: Hyperchol AngPec
## change.AIC -71.8786 Edge deleted: AngPec Smoker
## change.AIC -66.8927 Edge deleted: AngPec Sex
## change.AIC -62.9244 Edge deleted: AngPec QWave
## change.AIC -61.4196 Edge deleted: Hyperchol Smoker
## change.AIC -69.4553 Edge deleted: Hyperchol Sex
## change.AIC -59.8683 Edge deleted: Inherit Sex
## change.AIC -66.5041 Edge deleted: Inherit QWave
## change.AIC -73.8916 Edge deleted: Inherit Smoker
## change.AIC -56.9040 Edge deleted: Inherit AMI
## change.AIC -56.6416 Edge deleted: Hyperchol QWave
## change.AIC -53.5747 Edge deleted: QWave Sex
## change.AIC -51.2656 Edge deleted: Smoker Sex
## change.AIC -51.0938 Edge deleted: Hyperchol Heartfail
## change.AIC -47.7515 Edge deleted: Inherit Hypertroph
## change.AIC -47.0986 Edge deleted: Hyperchol AMI
## change.AIC -45.5477 Edge deleted: Smoker QWave
## change.AIC -53.8504 Edge deleted: Smoker CAD
## change.AIC -46.4669 Edge deleted: QWave Heartfail
## change.AIC -45.4139 Edge deleted: Sex CAD
## change.AIC -45.9671 Edge deleted: Sex AMI
## change.AIC -45.3351 Edge deleted: SuffHeartF AngPec
## change.AIC -53.9770 Edge deleted: AngPec Heartfail
## change.AIC -57.3531 Edge deleted: AngPec Hypertroph
## change.AIC -45.2526 Edge deleted: Smoker Heartfail
## change.AIC -40.1283 Edge deleted: Sex Hypertroph
## change.AIC -39.9618 Edge deleted: Hyperchol Hypertroph
## change.AIC -37.9974 Edge deleted: AngPec AMI
## change.AIC -36.8479 Edge deleted: Heartfail Inherit
## change.AIC -42.4289 Edge deleted: Heartfail CAD
## change.AIC -36.7664 Edge deleted: Smoker AMI
## change.AIC -32.2058 Edge deleted: CAD AMI
## change.AIC -26.5987 Edge deleted: AMI Heartfail
## change.AIC -34.8624 Edge deleted: AMI SuffHeartF
## change.AIC -32.7167 Edge deleted: QWave SuffHeartF
## change.AIC -25.7426 Edge deleted: AngPec QWavecode
## change.AIC -24.5883 Edge deleted: Smoker SuffHeartF
## change.AIC -21.4370 Edge deleted: SuffHeartF Sex
## change.AIC -20.4183 Edge deleted: Heartfail SuffHeartF
## change.AIC -19.9289 Edge deleted: Hyperchol SuffHeartF
## change.AIC -19.6613 Edge deleted: Inherit SuffHeartF
```

```
## change.AIC -20.4700 Edge deleted: SuffHeartF STchange
## change.AIC -17.8226 Edge deleted: Hyperchol QWavecode
## change.AIC -16.2269 Edge deleted: Hypertrophi AMI
## change.AIC -29.3572 Edge deleted: Hypertrophi QWave
## change.AIC -16.1274 Edge deleted: STchange AngPec
## change.AIC -15.2012 Edge deleted: Sex QWavecode
## change.AIC -14.7831 Edge deleted: Smoker QWavecode
## change.AIC -14.2829 Edge deleted: Inherit QWavecode
## change.AIC -10.8741 Edge deleted: Inherit STchange
## change.AIC -10.6194 Edge deleted: AMI QWavecode
## change.AIC -10.3788 Edge deleted: Inherit STcode
## change.AIC -10.2034 Edge deleted: Heartfail QWavecode
## change.AIC -9.5411 Edge deleted: Sex STchange
## change.AIC -9.7628 Edge deleted: Heartfail STchange
## change.AIC -8.7195 Edge deleted: Smoker STcode
## change.AIC -8.4443 Edge deleted: SuffHeartF QWavecode
## change.AIC -9.3392 Edge deleted: Hypertrophi QWavecode
## change.AIC -11.6157 Edge deleted: QWavecode CAD
## change.AIC -8.6113 Edge deleted: QWavecode STchange
## change.AIC -6.8791 Edge deleted: Hyperchol STchange
## change.AIC -7.2023 Edge deleted: Hyperchol STcode
## change.AIC -5.0708 Edge deleted: AMI STcode
## change.AIC -5.5476 Edge deleted: AMI STchange
## change.AIC -9.1727 Edge deleted: QWave STchange
## change.AIC -5.0630 Edge deleted: Sex STcode
## change.AIC -2.6745 Edge deleted: SuffHeartF CAD
## change.AIC -0.2916 Edge deleted: Sex Heartfail
```

```
m.forw2 <- stepwise(m.ind, k = log(nrow(cad1)), direction = "forward", details = 1)
```

```
## STEPWISE:
## criterion: aic ( k = 5.46 )
## direction: forward
## type : decomposable
## search : all
## steps : 1000
## . FORWARD: type=decomposable search=all, criterion=aic(5.46), alpha=0.00
## . Initial model: is graphical=TRUE is decomposable=TRUE
## change.AIC -115.7810 Edge added: STchange STcode
## change.AIC -69.6372 Edge added: CAD AngPec
## change.AIC -61.0636 Edge added: Heartfail Hypertrophi
## change.AIC -54.1911 Edge added: STcode SuffHeartF
## change.AIC -40.4460 Edge added: CAD AMI
## change.AIC -40.3808 Edge added: CAD QWave
## change.AIC -28.8232 Edge added: CAD Hyperchol
## change.AIC -20.8251 Edge added: CAD STchange
## change.AIC -17.7528 Edge added: SuffHeartF Heartfail
## change.AIC -13.6518 Edge added: CAD Inherit
## change.AIC -10.2661 Edge added: CAD Smoker
## change.AIC -9.9847 Edge added: STcode QWavecode
## change.AIC -8.1753 Edge added: QWave AMI
## change.AIC -2.2108 Edge added: CAD Sex
```

```
bn.back2 <- compile(grain(ugList(terms(m.back2)), data = cad1, smooth = 0.001))
bn.forw2 <- compile(grain(ugList(terms(m.forw2)), data = cad1, smooth = 0.001))

class.back2 <- predict(bn.back2, resp = "CAD", newdata = cad2, type = "class")
class.forw2 <- predict(bn.forw2, resp = "CAD", newdata = cad2, type = "class")

classMat(cad2$CAD, class.back2$pred$CAD)
```

```
##          classLabel
## trueLabel    No    Yes
##      No  0.90244 0.09756
##      Yes 0.50000 0.50000
```

```
classMat(cad2$CAD, class.forw2$pred$CAD)
```

```
##          classLabel
## trueLabel    No    Yes
##      No  0.8293 0.1707
##      Yes 0.4231 0.5769
```

1. Task: Create a "naive Bayes model": All predictor variables are independent given CAD. How does this model perform in predicting the validation cases?

```
g.naive <- lapply(setdiff(names(cad1), "CAD"), function(x) c(x, "CAD"))
str(g.naive)
```

```
## List of 13
## $ : chr [1:2] "Sex" "CAD"
## $ : chr [1:2] "AngPec" "CAD"
## $ : chr [1:2] "AMI" "CAD"
## $ : chr [1:2] "QWave" "CAD"
## $ : chr [1:2] "QWavecode" "CAD"
```

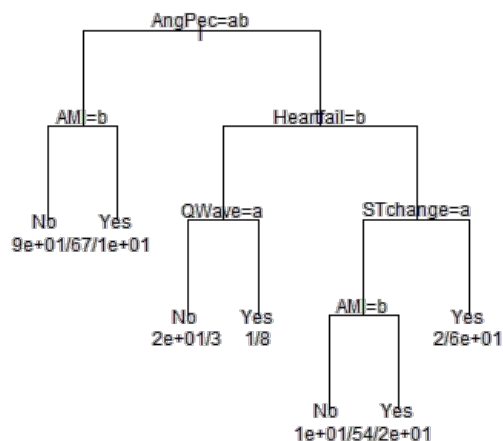
```
## $ : chr [1:2] "STcode" "CAD"
## $ : chr [1:2] "STchange" "CAD"
## $ : chr [1:2] "SuffHeartF" "CAD"
## $ : chr [1:2] "Hypertrophi" "CAD"
## $ : chr [1:2] "Hyperchol" "CAD"
## $ : chr [1:2] "Smoker" "CAD"
## $ : chr [1:2] "Inherit" "CAD"
## $ : chr [1:2] "Heartfail" "CAD"
```

```
bn.naive <- compile(grain(ugList(g.naive), data = cad1))
class.naive <- predict(bn.naive, resp = "CAD", newdata = cad2, type = "class")
classMat(cad2$CAD, class.naive$pred$CAD)
```

```
##          classLabel
## trueLabel      No      Yes
##      No  0.90244  0.09756
##      Yes 0.50000  0.50000
```

1. For the curious: Create a recursive partitioning tree using rpart (see supplementary notes).

```
library(rpart)
treel <- rpart(CAD ~ ., data = cad1)
plot(treel, uniform = T, margin = 0.2)
text(treel, use.n = TRUE)
```



```
class.tree <- predict(treel, newdata = cad2, type = "class")
classMat(cad2$CAD, class.tree)
```

```
##          classLabel
## trueLabel      No      Yes
##      No  0.90244  0.09756
##      Yes 0.42308  0.57692
```

1. Does such trees perform better than the graphical models?

The approaches perform similarly.

1. Discuss pros and cons of the approaches

One notable difference is that with the graphical model / BN approach, predictions can always be made, also if there are missing values among the predictor variables.

```
cad3 <- subset(cad2, select = c("CAD", "AngPec", "Smoker"))
class.tree <- predict(treel, newdata = cad3, type = "class")
```

```
## Error: object 'Sex' not found
```

```
class.bn <- predict(bn.back2, resp = "CAD", newdata = cad3, type = "class")
```