

SIS NMF Final: Diagnosis to DSD

November 30, 2014

1 Preparation

```
options(java.parameters = "-Xmx4G")

library(survival)

## Loading required package: splines

library(energy)
library(NMF)

## Loading required package: methods
## Loading required package: pkgmaker
## Loading required package: registry
## Loading required package: rngtools
## Loading required package: cluster
## NMF - BioConductor layer [OK] | Shared memory capabilities [OK] | Cores 7/8

library(nnlsl)

library(glmulti)

## Loading required package: rJava
##
## Attaching package: 'glmulti'
##
## The following object is masked from 'package:NMF':
##
##     consensus

library(glmnet)

## Loading required package: Matrix
## Loaded glmnet 1.9-8

library(RColorBrewer)
library(gplots)

##
## Attaching package: 'gplots'
##
## The following object is masked from 'package:stats':
##
##     lowess
```

```

library(xtable)
library(stargazer)

##
## Please cite as:
##
## Hlavac, Marek (2014). stargazer: LaTeX code and ASCII text for well-formatted regression
and summary statistics tables.
## R package version 5.1. http://CRAN.R-project.org/package=stargazer

load("image.rda")

```

2 Cohort characteristics

```

cpvs.diag_dsd$Path.TumourLocation[cpvs.diag_dsd$Path.TumourLocation == ""] = NA
cpvs.diag_dsd$Path.Nodes.Regional.Involved.Fraction = cpvs.diag_dsd$Path.Nodes.Regional.Involved/cpvs.d
cpvs.diag_dsd$Treat.Surgery.ExcisionStatus.Coarse = ordered(ifelse(cpvs.diag_dsd$Treat.Surgery.ExcisionS
"R0", "Clear", "Involved"), levels = c("Clear", "Involved"))
cpvs.diag_dsd$Path.Grade.Coarse = ordered(ifelse(cpvs.diag_dsd$Path.Grade %in%
c("1", "2"), "1or2", "3or4"), levels = c("1or2", "3or4"))
cpvs.diag_dsd$Path.TumourLocation.Coarse = factor(ifelse(cpvs.diag_dsd$Path.TumourLocation %in%
c("Head", "Head (Uncinate)", "Head", "Other")))

summary(cpvs.diag_dsd)

## Patient.ID Patient.Gender Patient.Ethnicity
## Length:110 Female:50 Asian : 5
## Class :character Male :60 Asian, White/Caucasian : 0
## Mode :character Black/African : 0
## Black/African, White/Caucasian: 0
## White/Caucasian :104
## NA's : 1
##
## Patient.Country History.LastFollowup.Date
## Australia :110 Min. :2007-06-29
## Italy : 0 1st Qu.:2011-08-19
## New Zealand : 0 Median :2013-03-12
## Puerto Rico : 0 Mean :2012-10-16
## United Kingdom : 0 3rd Qu.:2014-04-24
## United States of America: 0 Max. :2014-09-23
## NA's :1
## History.Smoking.PackYears History.Diagnosis.Date
## Min. : 0.75 Min. :2007-06-04
## 1st Qu.: 9.00 1st Qu.:2010-01-28
## Median :22.50 Median :2011-01-04
## Mean :26.89 Mean :2011-01-14
## 3rd Qu.:43.75 3rd Qu.:2012-02-15
## Max. :70.00 Max. :2012-10-17
## NA's :68
## History.Diagnosis.AgeAtYears History.Surgery.Date
## Min. :36.0 Min. :2007-05-29
## 1st Qu.:61.0 1st Qu.:2010-01-22

```

```

## Median :67.0                      Median :2011-01-01
## Mean   :66.4                      Mean   :2011-01-13
## 3rd Qu.:73.0                      3rd Qu.:2012-02-13
## Max.   :87.0                      Max.   :2012-10-17
##
##
##                               Treat.Surgery.Procedure
## Classic Whipple                               :79
## Splenectomy, Subtotal Panc/L sided Panc or distal Panc: 6
## Cholecystectomy, Classic Whipple              : 5
## Subtotal Panc/L sided Panc or distal Panc      : 4
## Classic Whipple, Exploratory laparotomy        : 3
## PPPD                                           : 3
## (Other)                                       :10
## Treat.Surgery.ExcisionStatus Treat.Surgery.Margin.Pancreatic
## R0:69                      <2 mm   : 4
## R1:35                      Clear    :88
## R2: 6                      Involved: 9
##                               NA's    : 9
##
##
##
## Treat.Surgery.MarginSizeMm.Pancreatic Treat.Surgery.Margin.Periunc
## Min.   : 0.0                      <2 mm   :20
## 1st Qu.: 5.0                      Clear    :52
## Median :10.0                     Involved:15
## Mean   :10.6                     NA's    :23
## 3rd Qu.:10.2
## Max.   :40.0
## NA's   :30
## Treat.Surgery.MarginSizeMm.Periunc Treat.Surgery.Margin.PVGroove
## Min.   : 0.00                     <2 mm   :23
## 1st Qu.: 1.00                     Clear    :55
## Median : 3.00                     Involved:12
## Mean   : 6.21                     NA's    :20
## 3rd Qu.:10.00
## Max.   :40.00
## NA's   :43
## Treat.Surgery.MarginSizeMm.PVGroove Treat.Surgery.Margin.Retrop
## Min.   : 0.00                     <2 mm   :21
## 1st Qu.: 1.00                     Clear    :68
## Median : 3.00                     Involved: 9
## Mean   : 4.08                     NA's    :12
## 3rd Qu.: 5.00
## Max.   :30.00
## NA's   :45
## Treat.Surgery.MarginSizeMm.Retrop Treat.Surgery.Margin.CBD
## Min.   : 0.10                     <2 mm   : 1
## 1st Qu.: 1.75                     Clear    :83
## Median : 3.00                     Involved: 0
## Mean   : 5.62                     NA's    :26
## 3rd Qu.:10.00
## Max.   :25.00
## NA's   :31
## Treat.Surgery.MarginSizeMm.CBD Treat.Surgery.Margin.Duodenal

```

```

## Min.      : 1.0                      Clear      :60
## 1st Qu.:11.8                      Involved: 1
## Median :20.0                      NA's       :49
## Mean    :23.6
## 3rd Qu.:32.5
## Max.    :55.0
## NA's    :47
## Treat.Surgery.MarginSizeMm.Duodenal Treat.Surgery.Margin.Gastric
## Min.      : 10.0                      Clear:59
## 1st Qu.: 40.0                      NA's :51
## Median : 80.0
## Mean    : 86.2
## 3rd Qu.:132.5
## Max.    :190.0
## NA's    :102
## Treat.Surgery.MarginSizeMm.Gastric Treat.Surgery.Margin.Comments
## Min.      : 10.0                      Length:110
## 1st Qu.: 50.0                      Class :character
## Median : 70.0                      Mode  :character
## Mean    : 67.9
## 3rd Qu.: 97.5
## Max.    :100.0
## NA's    :103
##
##                               Path.HistoType
## Pancreatic Ductal Adenocarcinoma:110
## Acinar Cell Carcinoma             : 0
## Ampullary Adenocarcinoma          : 0
## Carcinoid Tumour                  : 0
## Cholangiocarcinoma                : 0
## Clear Cell Carcinoma              : 0
## (Other)                           : 0
##
##                               Path.HistoType.Subtype Path.Grade
## Gastric                         : 0             1: 8
## Intestinal                      : 0             2:71
## Mixed                           : 0             3:30
## Not otherwise Specified (NOS):31             4: 1
## Pancreatobiliary                :13
## Squamous                        : 0
## NA's                             :66
##
## Path.TumourLocation Path.TumourSizeMm Path.Invasion.PN
## Head                 :83             Min.      :10.0      Absent :13
## Head (Uncinate):10             1st Qu.:28.0      Present:96
## Tail                 : 9             Median :35.0      NA's   : 1
## Body                 : 7             Mean    :37.6
##                      : 0             3rd Qu.:45.0
## (Other)              : 0             Max.    :90.0
## NA's                 : 1             NA's    :1
##
## Path.Invasion.VS Path.Nodes.Regional.Total Path.Nodes.Regional.Involved
## Absent :34             Min.      : 0.0             Min.      : 0.00
## Present:72             1st Qu.:11.0             1st Qu.: 1.00
## NA's   : 4             Median :16.0             Median : 2.00
##                      Mean    :18.1             Mean    : 3.18
##                      3rd Qu.:24.0             3rd Qu.: 4.00
##                      Max.    :46.0             Max.    :18.00

```

```

##
## Path.Nodes.SepRec.Total Path.Nodes.SepRec.Involved
## Min. : 0.0 Min. : 0.00
## 1st Qu.:11.0 1st Qu.: 1.00
## Median :16.0 Median : 2.00
## Mean :18.1 Mean : 3.18
## 3rd Qu.:24.0 3rd Qu.: 4.00
## Max. :46.0 Max. :18.00
##
##
## Staging.Version Staging.pM Staging.pN
## pTNM AJCC 6th Ed 2002 :14 M0 : 2 N0 :25
## pTNM AJCC 7th Ed 2010 :96 M1 : 6 N1 :84
## pTNM AJCC 7th Ed 2010 (Ampulla) : 0 NA's:102 NA's: 1
## pTNM AJCC 7th Ed 2010 (Cholangiocarcinoma): 0
## pTNM AJCC 7th Ed 2010 (Neuroendocrine) : 0
##
##
## Staging.pT Staging.Stage History.Recurrence History.Recurrence.Date
## Tis : 0 IA : 0 Not observed:24 Min. :2007-10-14
## T1 : 0 IB : 3 Suspected : 4 1st Qu.:2010-12-11
## T2 : 6 IIA:20 Confirmed :78 Median :2012-02-22
## T3 :102 IIB:80 NA's : 4 Mean :2012-01-21
## T4 : 1 III: 1 3rd Qu.:2012-12-29
## NA's: 1 IV : 6 Max. :2014-08-27
## NA's :29
##
## History.Recurrence.Site.Stomach History.Recurrence.Site.Peritoneum
## Mode :logical Mode :logical
## FALSE:110 FALSE:94
## NA's :0 TRUE :16
## NA's :0
##
##
##
## History.Recurrence.Site.PancRemnant History.Recurrence.Site.PancBed
## Mode :logical Mode :logical
## FALSE:106 FALSE:91
## TRUE :4 TRUE :19
## NA's :0 NA's :0
##
##
##
## History.Recurrence.Site.Other History.Recurrence.Site.Omentum
## Mode :logical Mode :logical
## FALSE:102 FALSE:109
## TRUE :8 TRUE :1
## NA's :0 NA's :0
##
##
##
## History.Recurrence.Site.Mesentery History.Recurrence.Site.LymphNodes
## Mode :logical Mode :logical
## FALSE:108 FALSE:88
## TRUE :2 TRUE :22
## NA's :0 NA's :0

```

```

##
##
##
## History.Recurrence.Site.Lung History.Recurrence.Site.Liver
## Mode :logical Mode :logical
## FALSE:88 FALSE:72
## TRUE :22 TRUE :38
## NA's :0 NA's :0
##
##
##
## History.Recurrence.Site.Brain History.Recurrence.Site.Bone
## Mode :logical Mode :logical
## FALSE:109 FALSE:104
## TRUE :1 TRUE :6
## NA's :0 NA's :0
##
##
##
## History.Status History.Death.Date
## Alive - With Disease :15 Min. :2007-11-21
## Alive - Without Disease :22 1st Qu.:2011-01-14
## Deceased - Of Disease :70 Median :2012-03-07
## Deceased - Of Other Cause : 3 Mean :2012-02-21
## Deceased - Of Unknown Cause: 0 3rd Qu.:2013-03-17
## Max. :2014-06-17
## NA's :37
##
## History.Death.Cause Surv.Event.Death
## Cancer Death (Pancreatic) :69 Min. :0.000
## Cancer Death (Other) - Lung ca : 1 1st Qu.:0.000
## Died of Treatment Complication : 1 Median :1.000
## Other (please specify) : 1 Mean :0.664
## Other (please specify) - Suicide: 1 3rd Qu.:1.000
## (Other) : 0 Max. :1.000
## NA's :37
## Surv.EventTimeFromDiag.Death Surv.EventTimeFromSurg.Death
## Min. : 36 Min. : 36
## 1st Qu.: 402 1st Qu.: 406
## Median : 632 Median : 634
## Mean : 674 Mean : 676
## 3rd Qu.: 912 3rd Qu.: 917
## Max. :1778 Max. :1779
##
## Surv.EventTimeFromRec.Death Surv.Event.DSDeath
## Min. : 7 Min. :0.000
## 1st Qu.: 68 1st Qu.:0.000
## Median : 183 Median :1.000
## Mean : 250 Mean :0.636
## 3rd Qu.: 338 3rd Qu.:1.000
## Max. :1333 Max. :1.000
## NA's :29
## Surv.EventTimeFromDiag.DSDeath Surv.EventTimeFromSurg.DSDeath
## Min. : 36 Min. : 36
## 1st Qu.: 402 1st Qu.: 406

```

```

## Median : 632                      Median : 634
## Mean   : 673                      Mean    : 675
## 3rd Qu.: 912                      3rd Qu.: 917
## Max.   :1778                      Max.    :1779
##
## Surv.EventTimeFromRec.DSDeath Surv.Event.Recurrence
## Min.    : 7                      Min.     :0.000
## 1st Qu.: 68                      1st Qu.:0.000
## Median : 183                     Median :1.000
## Mean    : 250                     Mean    :0.736
## 3rd Qu.: 338                     3rd Qu.:1.000
## Max.    :1333                     Max.    :1.000
## NA's    :29                      NA's     :4
## Surv.EventTimeFromDiag.Recurrence Surv.EventTimeFromSurg.Recurrence
## Min.    : 34                      Min.     : 34
## 1st Qu.: 240                      1st Qu.: 240
## Median : 392                     Median : 398
## Mean    : 511                     Mean    : 512
## 3rd Qu.: 697                     3rd Qu.: 699
## Max.    :1778                     Max.    :1779
## NA's    :6                      NA's     :6
## Path.Nodes.Regional.Involved.Fraction Treat.Surgery.ExcisionStatus.Coarse
## Min.    :0.0000                      Clear    :69
## 1st Qu.:0.0435                      Involved:41
## Median :0.1667
## Mean    :0.2026
## 3rd Qu.:0.2727
## Max.    :1.0000
## NA's    :1
## Path.Grade.Coarse Path.TumourLocation.Coarse
## 1or2:79           Head :93
## 3or4:31           Other:17
##
##
##
##
##
##
sort(apply(is.na(cpvs.diag_dsd), 2, sum))

## Patient.ID
## 0
## Patient.Gender
## 0
## Patient.Country
## 0
## History.Diagnosis.Date
## 0
## History.Diagnosis.AgeAtYears
## 0
## History.Surgery.Date
## 0
## Treat.Surgery.Procedure
## 0

```

```

##      Treat.Surgery.ExcisionStatus
##                                0
##      Treat.Surgery.Margin.Comments
##                                0
##                                Path.HistoType
##                                0
##                                Path.Grade
##                                0
##      Path.Nodes.Regional.Total
##                                0
##      Path.Nodes.Regional.Involved
##                                0
##      Path.Nodes.SepRec.Total
##                                0
##      Path.Nodes.SepRec.Involved
##                                0
##      Staging.Version
##                                0
##      Staging.Stage
##                                0
##      History.Recurrence.Site.Stomach
##                                0
##      History.Recurrence.Site.Peritoneum
##                                0
##      History.Recurrence.Site.PancRemnant
##                                0
##      History.Recurrence.Site.PancBed
##                                0
##      History.Recurrence.Site.Other
##                                0
##      History.Recurrence.Site.Omentum
##                                0
##      History.Recurrence.Site.Mesentery
##                                0
##      History.Recurrence.Site.LymphNodes
##                                0
##      History.Recurrence.Site.Lung
##                                0
##      History.Recurrence.Site.Liver
##                                0
##      History.Recurrence.Site.Brain
##                                0
##      History.Recurrence.Site.Bone
##                                0
##      History.Status
##                                0
##      Surv.Event.Death
##                                0
##      Surv.EventTimeFromDiag.Death
##                                0
##      Surv.EventTimeFromSurg.Death
##                                0
##      Surv.Event.DSDeath
##                                0

```



```

##      Surv.EventTimeFromDiag.DSDeath
##                                0
##      Surv.EventTimeFromSurg.DSDeath
##                                0
##      Treat.Surgery.ExcisionStatus.Coarse
##                                0
##              Path.Grade.Coarse
##                                0
##      Path.TumourLocation.Coarse
##                                0
##              Patient.Ethnicity
##                                1
##      History.LastFollowup.Date
##                                1
##              Path.TumourLocation
##                                1
##              Path.TumourSizeMm
##                                1
##              Path.Invasion.PN
##                                1
##              Staging.pN
##                                1
##              Staging.pT
##                                1
## Path.Nodes.Regional.Involved.Fraction
##                                1
##              Path.Invasion.VS
##                                4
##              History.Recurrence
##                                4
##              Surv.Event.Recurrence
##                                4
##      Surv.EventTimeFromDiag.Recurrence
##                                6
##      Surv.EventTimeFromSurg.Recurrence
##                                6
##      Treat.Surgery.Margin.Pancreatic
##                                9
##      Treat.Surgery.Margin.Retrop
##                               12
##      Treat.Surgery.Margin.PVGroove
##                               20
##      Treat.Surgery.Margin.Periunc
##                               23
##      Treat.Surgery.Margin.CBD
##                               26
##      History.Recurrence.Date
##                               29
##      Surv.EventTimeFromRec.Death
##                               29
##      Surv.EventTimeFromRec.DSDeath
##                               29
##      Treat.Surgery.MarginSizeMm.Pancreatic
##                               30

```

```
##      Treat.Surgery.MarginSizeMm.Retrop
##                                31
##      History.Death.Date
##                                37
##      History.Death.Cause
##                                37
##      Treat.Surgery.MarginSizeMm.Periunc
##                                43
##      Treat.Surgery.MarginSizeMm.PVGroove
##                                45
##      Treat.Surgery.MarginSizeMm.CBD
##                                47
##      Treat.Surgery.Margin.Duodenal
##                                49
##      Treat.Surgery.Margin.Gastric
##                                51
##      Path.HistoType.Subtype
##                                66
##      History.Smoking.PackYears
##                                68
##      Treat.Surgery.MarginSizeMm.Duodenal
##                                102
##      Staging.pM
##                                102
##      Treat.Surgery.MarginSizeMm.Gastric
##                                103
```

3 Conventional prognostic model

Transform the covariates by median centering where required.

```
print(median(cpvs.diag_dsd$History.Diagnosis.AgeAtYears, na.rm = TRUE))

## [1] 67

cpvs.diag_dsd$History.Diagnosis.AgeAtYears.MedCent = cpvs.diag_dsd$History.Diagnosis.AgeAtYears -
  median(cpvs.diag_dsd$History.Diagnosis.AgeAtYears, na.rm = TRUE)
print(median(cpvs.diag_dsd$Path.TumourSizeMm, na.rm = TRUE))

## [1] 35

cpvs.diag_dsd$Path.TumourSizeMm.MedCent = cpvs.diag_dsd$Path.TumourSizeMm -
  median(cpvs.diag_dsd$Path.TumourSizeMm, na.rm = TRUE)
```

3.1 Covariate forms: deviance residuals

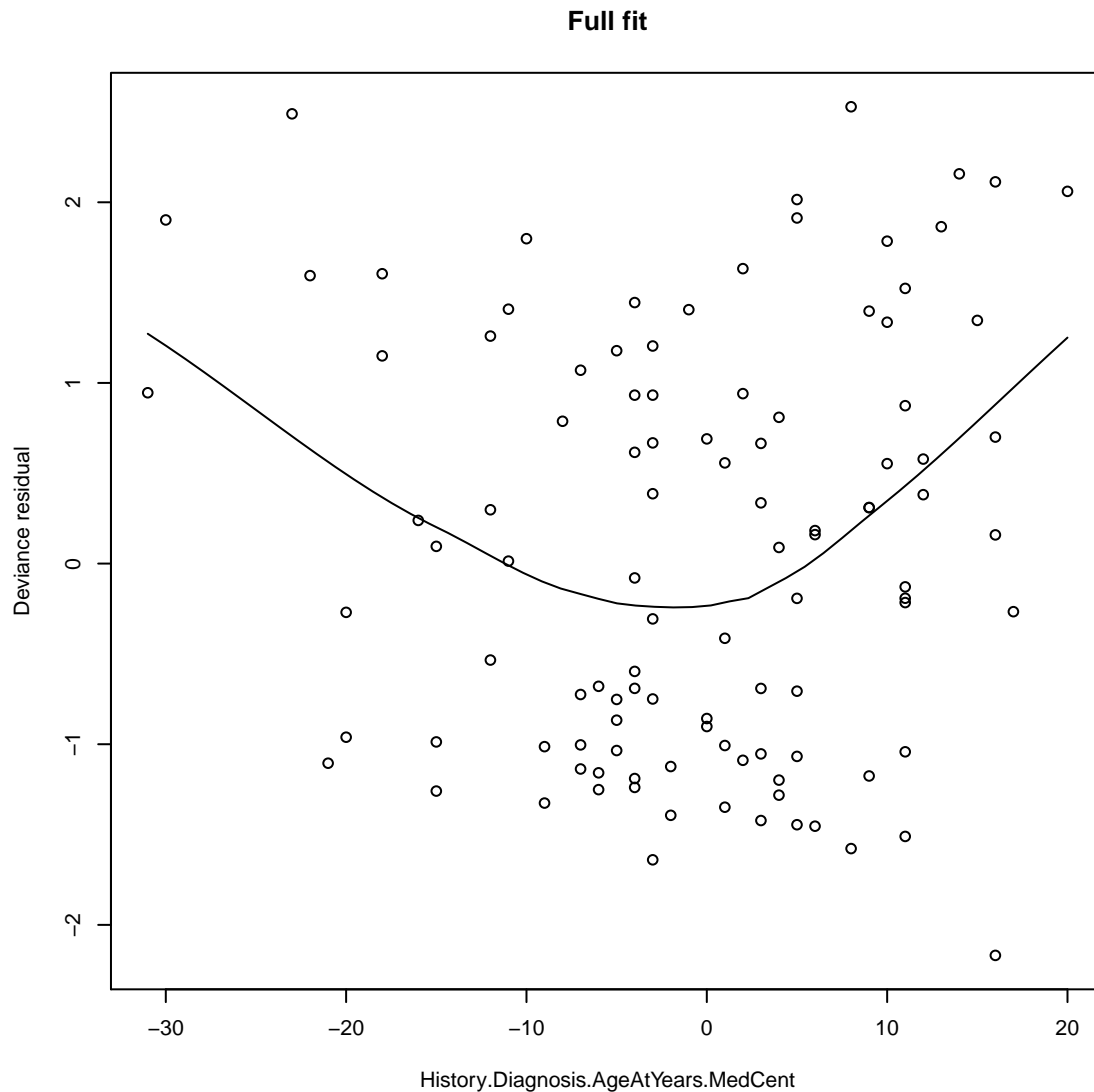
```
# History.Smoking.PackYears + # Remove as it's mostly missing
temp = coxph(Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath) ~ Patient.Gender +
  History.Diagnosis.AgeAtYears.MedCent + Treat.Surgery.ExcisionStatus.Coarse +
  Path.Grade.Coarse + Path.TumourLocation.Coarse + Path.TumourSizeMm.MedCent +
  Path.Invasion.PN + Path.Invasion.VS + Path.Nodes.Regional.Involved.Fraction +
```

```

Staging.pN, data = cpvs.diag_dsd)

# scatter.smooth(cpvs.diag_dsd[names(resid(temp,
# 'deviance'))], l$History.Smoking.PackYears, resid(temp, 'deviance'), xlab =
# 'History.Smoking.PackYears', ylab = 'Deviance residual', main = 'Full
# fit')
scatter.smooth(cpvs.diag_dsd[names(resid(temp, "deviance"))], l$History.Diagnosis.AgeAtYears.MedCent,
  resid(temp, "deviance"), xlab = "History.Diagnosis.AgeAtYears.MedCent",
  ylab = "Deviance residual", main = "Full fit")

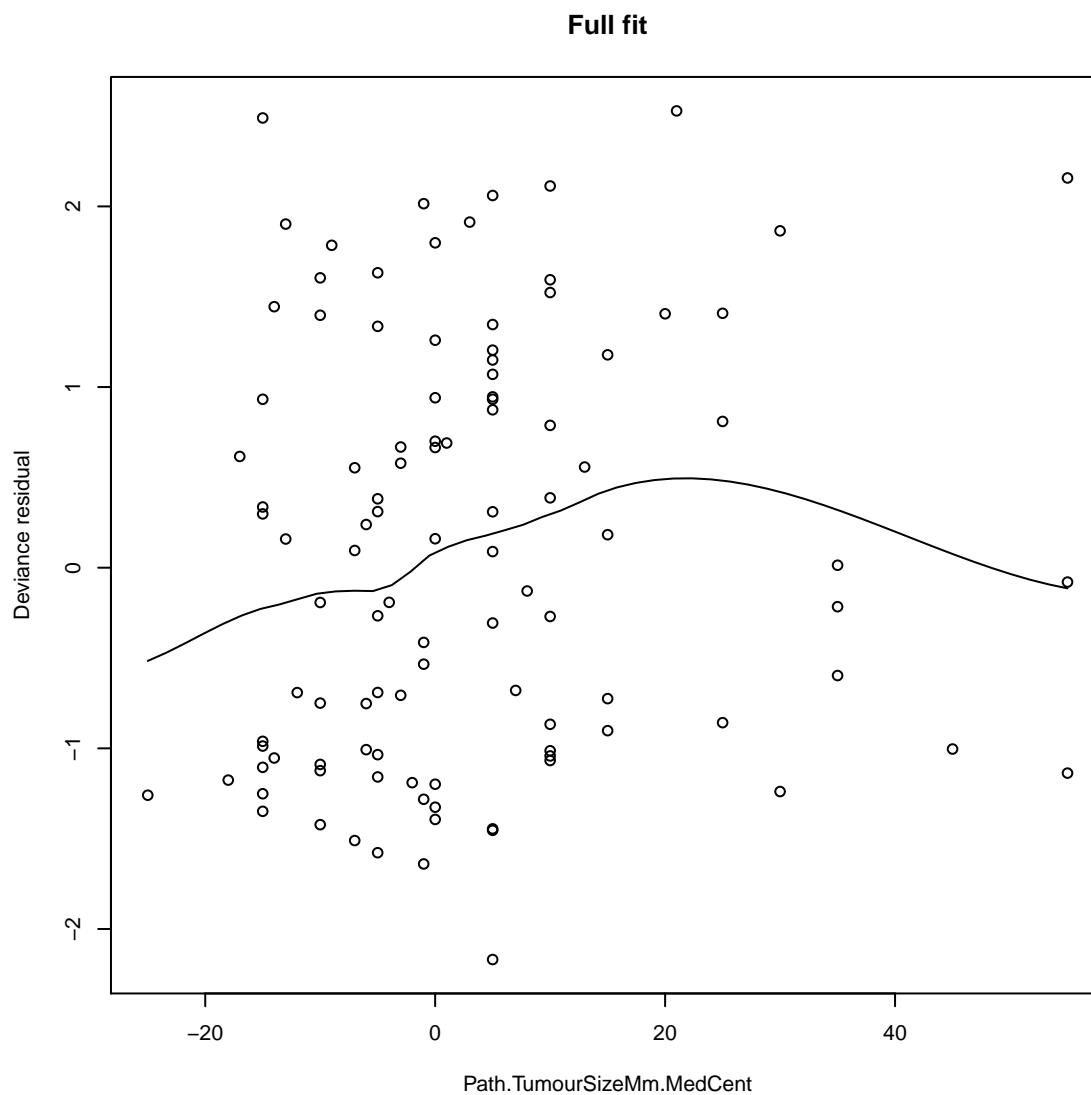
```



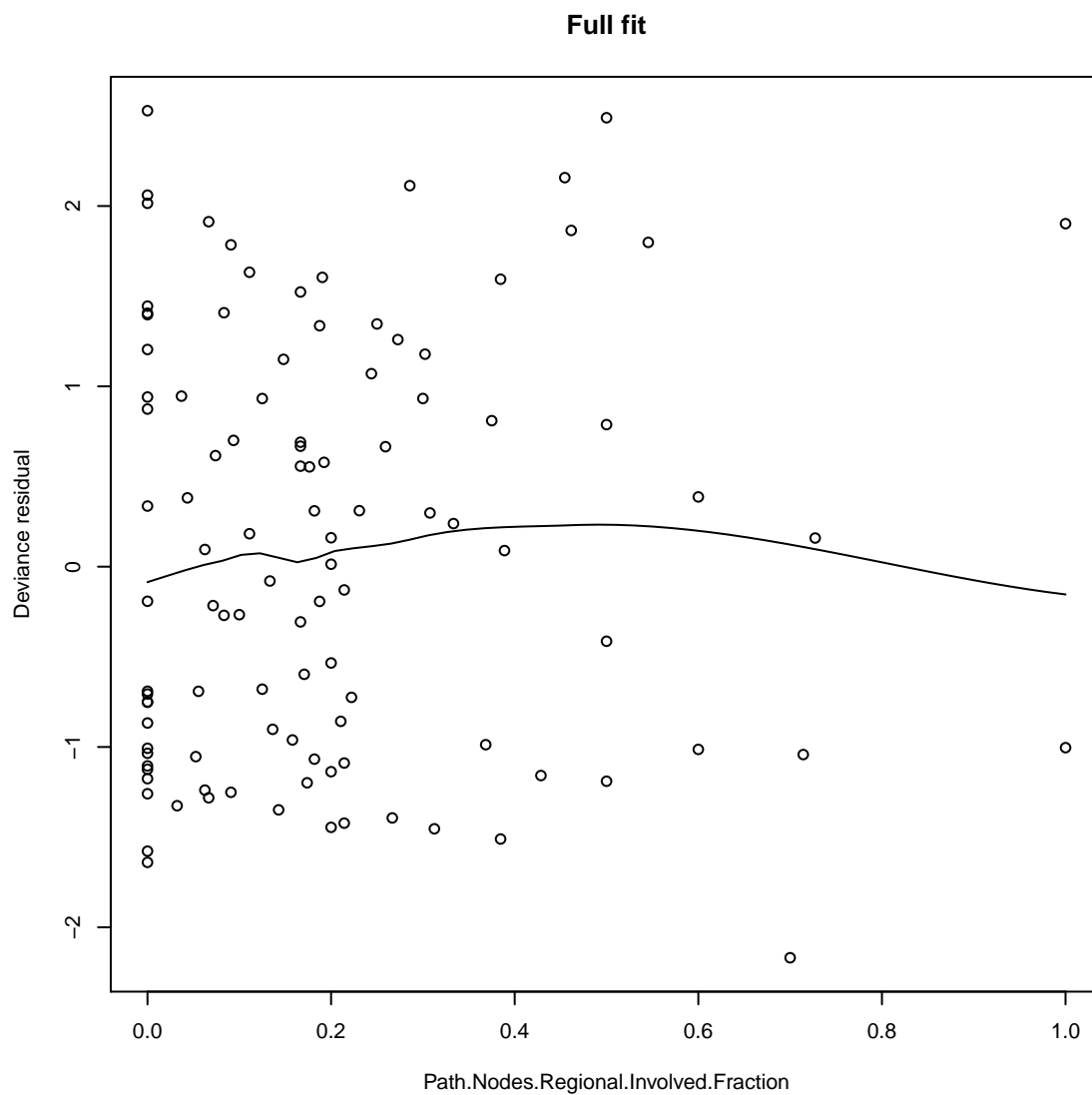
```

scatter.smooth(cpvs.diag_dsd[names(resid(temp, "deviance"))], l$Path.TumourSizeMm.MedCent,
  resid(temp, "deviance"), xlab = "Path.TumourSizeMm.MedCent", ylab = "Deviance residual",
  main = "Full fit")

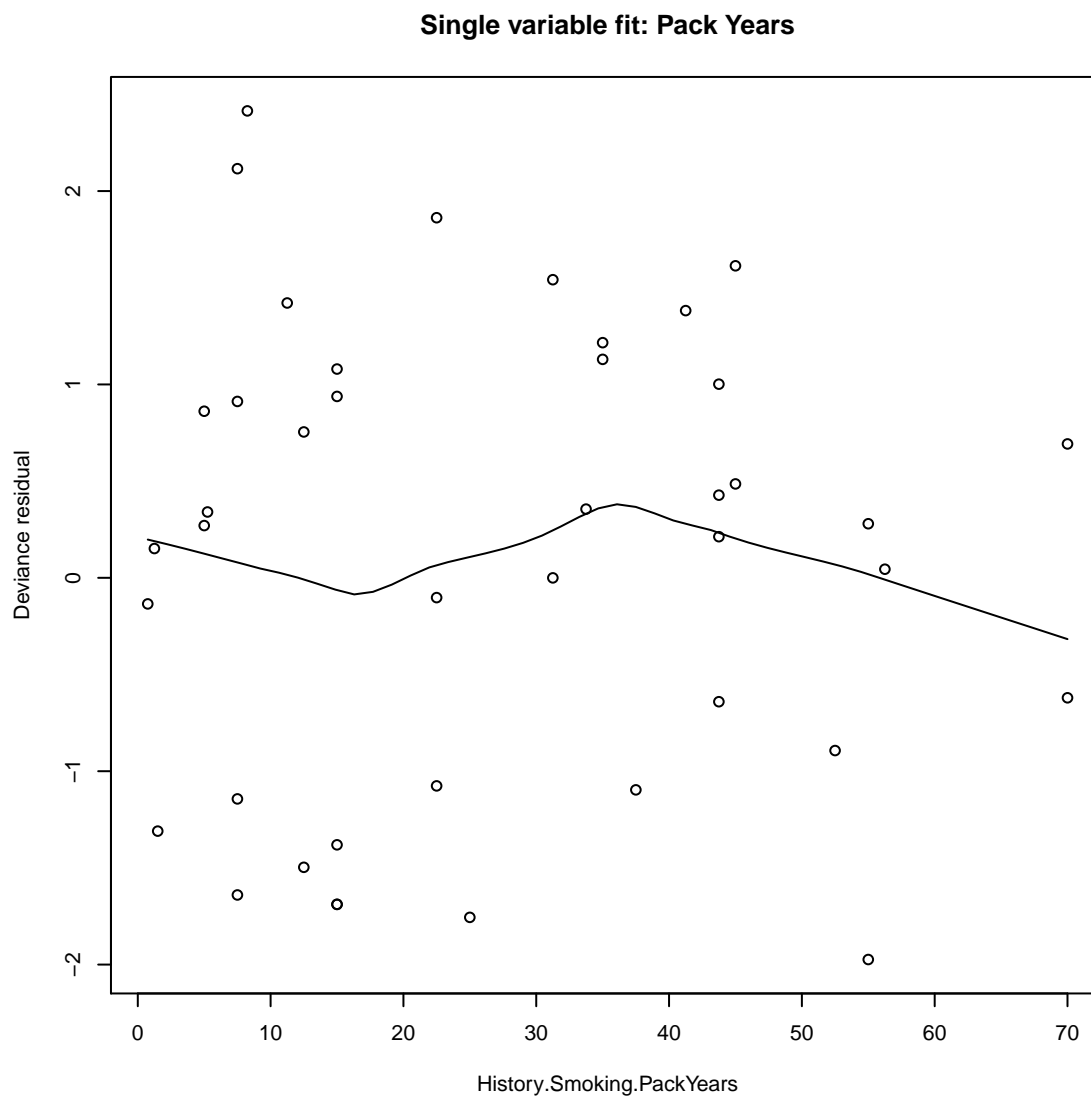
```



```
scatter.smooth(cpv$diag_dsd[names(resid(temp, "deviance")), ]$Path.Nodes.Regional.Involved.Fraction,
  resid(temp, "deviance"), xlab = "Path.Nodes.Regional.Involved.Fraction",
  ylab = "Deviance residual", main = "Full fit")
```

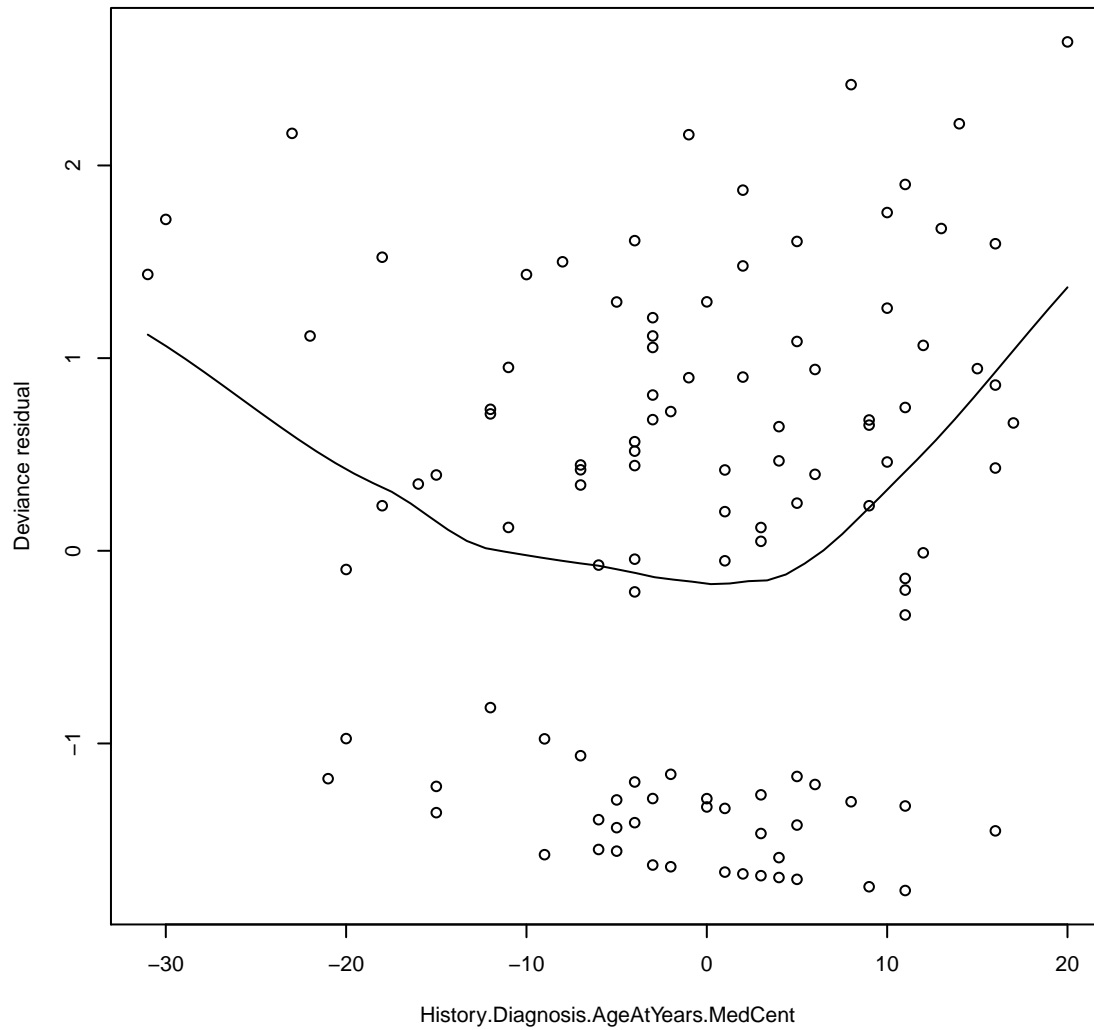


```
temp = coxph(Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath) ~ History.Smoking.PackYears,
  data = cpvs.diag_dsd)
scatter.smooth(cpvs.diag_dsd[names(resid(temp, "deviance")), ]$History.Smoking.PackYears,
  resid(temp, "deviance"), xlab = "History.Smoking.PackYears", ylab = "Deviance residual",
  main = "Single variable fit: Pack Years")
```



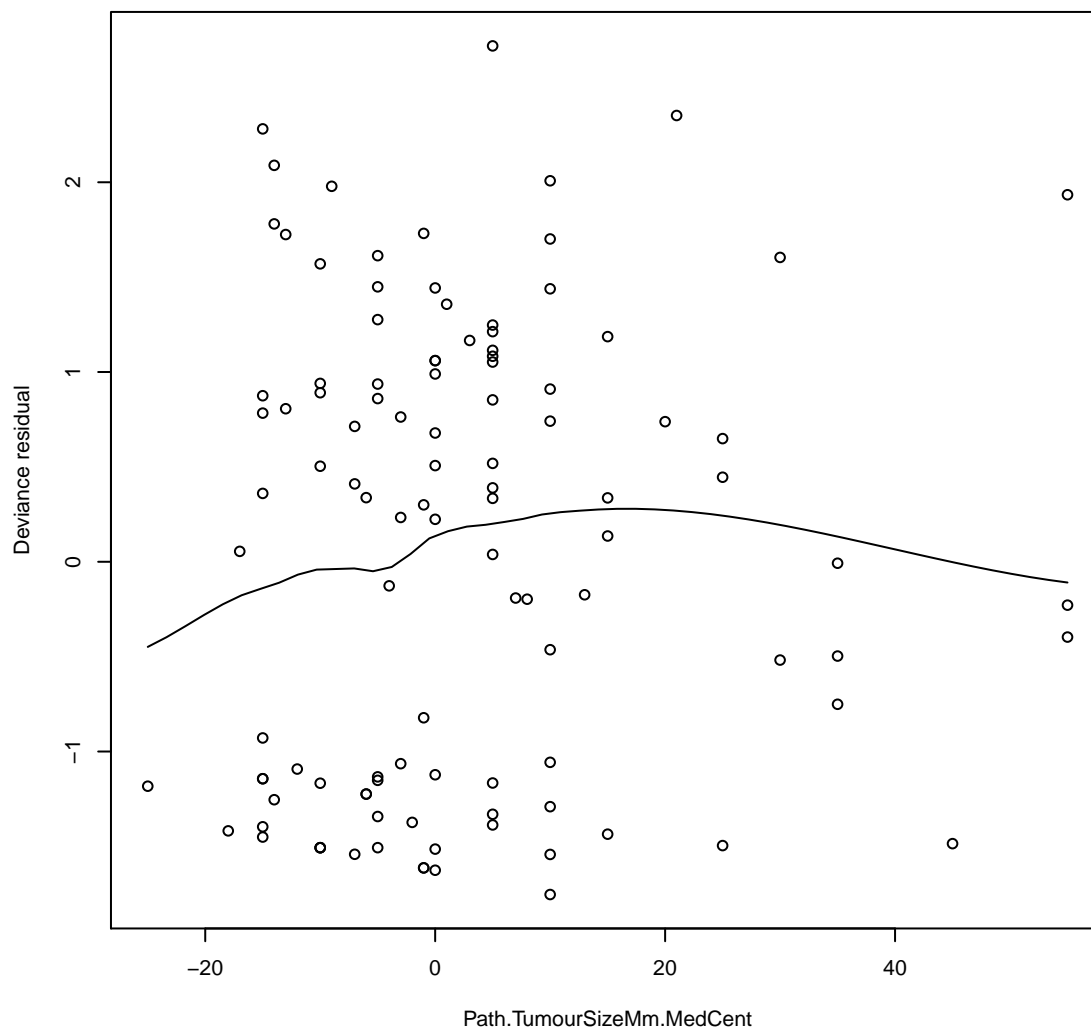
```
temp = coxph(Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath) ~ History.Diagnosis.AgeAtYears.MedCent,
  data = cpvs.diag_dsd)
scatter.smooth(cpvs.diag_dsd[names(resid(temp, "deviance"))], ]$History.Diagnosis.AgeAtYears.MedCent,
  resid(temp, "deviance"), xlab = "History.Diagnosis.AgeAtYears.MedCent",
  ylab = "Deviance residual", main = "Single variable fit: Age at diagnosis")
```

Single variable fit: Age at diagnosis



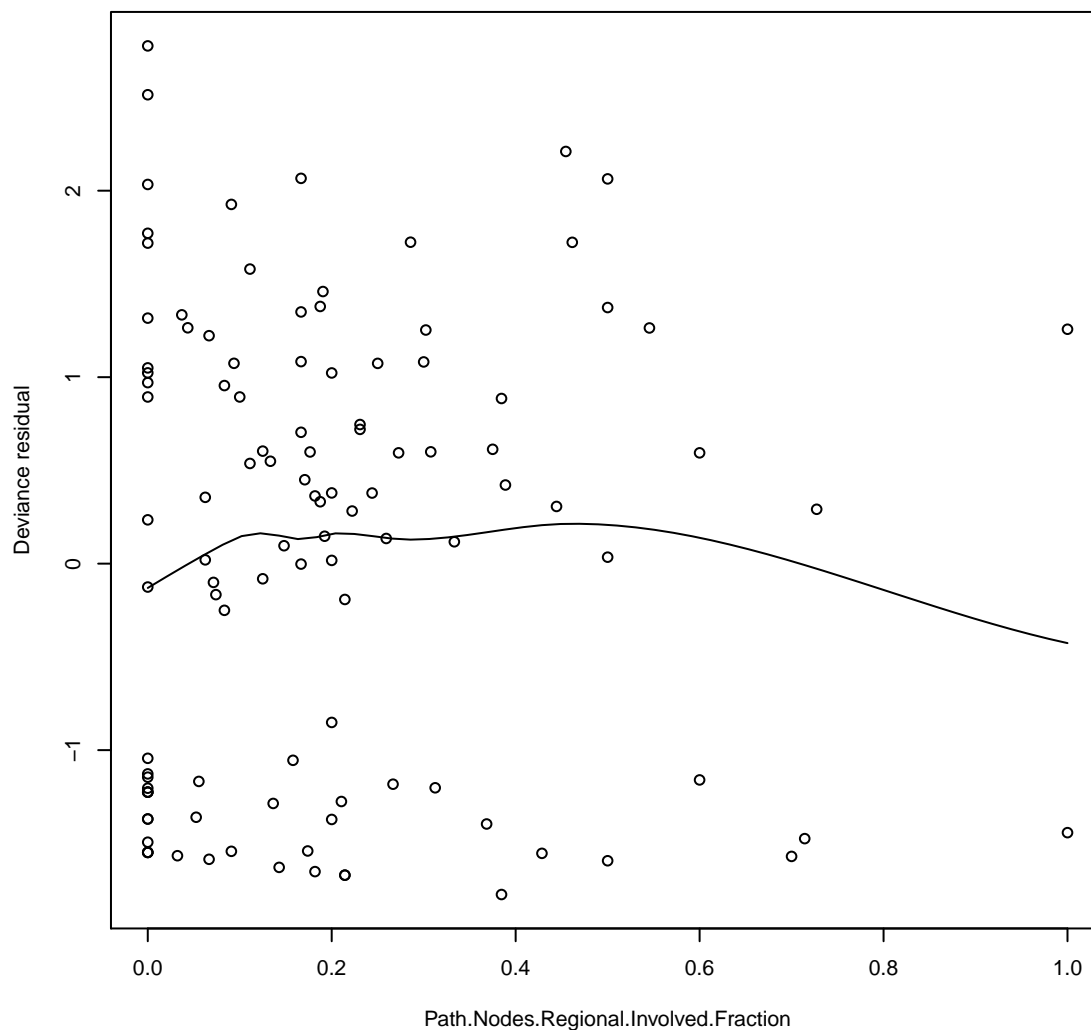
```
temp = coxph(Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath) ~ Path.TumourSizeMm.MedCent,
  data = cpvs.diag_dsd)
scatter.smooth(cpvs.diag_dsd[names(resid(temp, "deviance")), ]$Path.TumourSizeMm.MedCent,
  resid(temp, "deviance"), xlab = "Path.TumourSizeMm.MedCent", ylab = "Deviance residual",
  main = "Single variable fit: Tumour size")
```

Single variable fit: Tumour size



```
temp = coxph(Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath) ~ Path.Nodes.Regional.Involved.Fra
data = cpvs.diag_dsd)
scatter.smooth(cpvs.diag_dsd[names(resid(temp, "deviance")), ]$Path.Nodes.Regional.Involved.Fraction,
resid(temp, "deviance"), xlab = "Path.Nodes.Regional.Involved.Fraction",
ylab = "Deviance residual", main = "Single variable fit: Node involvement fraction")
```


Single variable fit: Node involvement fraction



Evidence of nonlinearity for age of diagnosis – very young diagnosis, or very old, is hazardous. This makes sense. Some indication of nonlinearity for size but it's weak. All other predictor residuals are consistent with a linear effect. Refit the model taking this into account.

```
fit.cpv.withsmoking = coxph(Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath) ~
  Patient.Gender + pspline(History.Diagnosis.AgeAtYears.MedCent, 3) + Path.Grade.Coarse +
  Path.TumourLocation.Coarse + pspline(Path.TumourSizeMm.MedCent, 3) +
  Path.Invasion.PN + Path.Invasion.VS + Path.Nodes.Regional.Involved.Fraction +
  Staging.pN + History.Smoking.PackYears + Treat.Surgery.ExcisionStatus.Coarse,
  data = cpvs.diag_dsd)
summary(fit.cpv.withsmoking)

## Call:
## coxph(formula = Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath) ~
##       Patient.Gender + pspline(History.Diagnosis.AgeAtYears.MedCent,
##       3) + Path.Grade.Coarse + Path.TumourLocation.Coarse +
##       pspline(Path.TumourSizeMm.MedCent, 3) + Path.Invasion.PN +
##       Path.Invasion.VS + Path.Nodes.Regional.Involved.Fraction +
```

```
##      Staging.pN + History.Smoking.PackYears + Treat.Surgery.ExcisionStatus.Coarse,
##      data = cpvs.diag_dsd)
##
##      n= 42, number of events= 30
##      (68 observations deleted due to missingness)
##
##      coef      se(coef) se2      Chisq DF      p
## Patient.GenderMale      -1.0104 0.5963      0.5820 2.87      1.00 0.0900
## pspline(History.Diagnosis -0.0378 0.0211      0.0209 3.21      1.00 0.0730
## pspline(History.Diagnosis      8.52      1.97 0.0140
## Path.Grade.Coarse.L      0.8958 0.4304      0.4056 4.33      1.00 0.0370
## Path.TumourLocation.Coars 1.0384 1.0217      0.9186 1.03      1.00 0.3100
## pspline(Path.TumourSizeMm -0.0307 0.0220      0.0218 1.96      1.00 0.1600
## pspline(Path.TumourSizeMm      2.63      1.96 0.2600
## Path.Invasion.PNPresent -2.8910 1.2657      1.2170 5.22      1.00 0.0220
## Path.Invasion.VSPresent 1.6460 0.7271      0.7174 5.13      1.00 0.0240
## Path.Nodes.Regional.Invol 3.6429 1.4158      1.3707 6.62      1.00 0.0100
## Staging.pN.L      -2.3360 0.9670      0.9294 5.84      1.00 0.0160
## History.Smoking.PackYears 0.0429 0.0143      0.0136 8.95      1.00 0.0028
## Treat.Surgery.ExcisionSta 1.4049 0.4581      0.4328 9.41      1.00 0.0022
##
##      exp(coef) exp(-coef) lower .95 upper .95
## Patient.GenderMale      0.36407      2.7467      1.13e-01      1.172
## ps(History.Diagnosis.AgeA 0.26340      3.7965      1.42e-02      4.900
## ps(History.Diagnosis.AgeA 0.06697      14.9328      7.81e-04      5.744
## ps(History.Diagnosis.AgeA 0.01870      53.4805      1.30e-04      2.684
## ps(History.Diagnosis.AgeA 0.00852      117.3847      6.48e-05      1.120
## ps(History.Diagnosis.AgeA 0.00519      192.8153      5.16e-05      0.521
## ps(History.Diagnosis.AgeA 0.00435      229.7202      4.51e-05      0.421
## ps(History.Diagnosis.AgeA 0.00476      210.0484      4.54e-05      0.499
## ps(History.Diagnosis.AgeA 0.01010      99.0030      9.16e-05      1.114
## ps(History.Diagnosis.AgeA 0.02863      34.9277      1.75e-04      4.674
## ps(History.Diagnosis.AgeA 0.08278      12.0799      1.14e-04      60.241
## Path.Grade.Coarse.L      2.44925      0.4083      1.05e+00      5.693
## Path.TumourLocation.Coars 2.82474      0.3540      3.81e-01      20.924
## ps(Path.TumourSizeMm.MedC 1.47319      0.6788      3.80e-02      57.102
## ps(Path.TumourSizeMm.MedC 2.17028      0.4608      6.20e-03      759.609
## ps(Path.TumourSizeMm.MedC 2.95484      0.3384      4.48e-03      1948.189
## ps(Path.TumourSizeMm.MedC 3.25116      0.3076      5.26e-03      2010.321
## ps(Path.TumourSizeMm.MedC 2.32997      0.4292      3.84e-03      1412.998
## ps(Path.TumourSizeMm.MedC 1.34289      0.7447      1.92e-03      939.456
## ps(Path.TumourSizeMm.MedC 0.56485      1.7704      6.67e-04      478.374
## ps(Path.TumourSizeMm.MedC 0.24119      4.1462      2.42e-04      240.177
## ps(Path.TumourSizeMm.MedC 0.16081      6.2186      1.40e-04      184.700
## ps(Path.TumourSizeMm.MedC 0.12003      8.3315      4.58e-05      314.604
## Path.Invasion.PNPresent 0.05552      18.0113      4.65e-03      0.663
## Path.Invasion.VSPresent 5.18612      0.1928      1.25e+00      21.564
## Path.Nodes.Regional.Invol 38.20095      0.0262      2.38e+00      612.581
## Staging.pN.L      0.09672      10.3394      1.45e-02      0.644
## History.Smoking.PackYears 1.04382      0.9580      1.01e+00      1.074
## Treat.Surgery.ExcisionSta 4.07518      0.2454      1.66e+00      10.002
##
##      Iterations: 3 outer, 12 Newton-Raphson
##      Theta= 0.4087
```

```
##      Theta= 0.3577
## Degrees of freedom for terms= 1.0 3.0 0.9 0.8 3.0 0.9 1.0 0.9 0.9 0.9 0.9
## Concordance= 0.827 (se = 0.059 )
## Likelihood ratio test= 50.7 on 14.1 df, p=5.13e-06

anova(fit.cpv.withsmoking)

## Analysis of Deviance Table
## Cox model: response is Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath)
## Terms added sequentially (first to last)
##
##
##              loglik Chisq Df
## NULL                      -93.6
## Patient.Gender              -93.6  0.09  1
## pspline(History.Diagnosis.AgeAtYears.MedCent, 3) -86.5 14.18 10
## Path.Grade.Coarse           -84.8  3.26  1
## Path.TumourLocation.Coarse  -84.5  0.77  1
## pspline(Path.TumourSizeMm.MedCent, 3)          -82.3  4.34 10
## Path.Invasion.PN            -79.5  5.64  1
## Path.Invasion.VS            -78.6  1.81  1
## Path.Nodes.Regional.Involved.Fraction          -78.4  0.24  1
## Staging.pN                  -76.5  3.97  1
## History.Smoking.PackYears    -72.7  7.60  1
## Treat.Surgery.ExcisionStatus.Coarse            -68.3  8.77  1
##
##              Pr(>|Chi|)
## NULL
## Patient.Gender              0.7607
## pspline(History.Diagnosis.AgeAtYears.MedCent, 3) 0.1651
## Path.Grade.Coarse           0.0710
## Path.TumourLocation.Coarse  0.3803
## pspline(Path.TumourSizeMm.MedCent, 3)          0.9309
## Path.Invasion.PN            0.0176
## Path.Invasion.VS            0.1782
## Path.Nodes.Regional.Involved.Fraction          0.6207
## Staging.pN                  0.0463
## History.Smoking.PackYears    0.0058
## Treat.Surgery.ExcisionStatus.Coarse            0.0031
```

The problem with the above is that smoking status is poorly known, but it nonetheless is a significant predictor. It may be possible to impute smoking from mutation profiles, but that would be complex and uncertain. It's certainly unlikely to be usefully imputed just from the other CPVs. So fit a working model that does not include smoking status, but keep in mind this omission when making conclusions.

```
fit.cpv.withoutsmoking = coxph(Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath) ~
  Patient.Gender + pspline(History.Diagnosis.AgeAtYears.MedCent, 3) + Path.Grade.Coarse +
  Path.TumourLocation.Coarse + pspline(Path.TumourSizeMm.MedCent, 3) +
  Path.Invasion.PN + Path.Invasion.VS + Path.Nodes.Regional.Involved.Fraction +
  Staging.pN + Treat.Surgery.ExcisionStatus.Coarse, data = cpvs.diag_dsd)
summary(fit.cpv.withoutsmoking)

## Call:
## coxph(formula = Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath) ~
##      Patient.Gender + pspline(History.Diagnosis.AgeAtYears.MedCent,
##      3) + Path.Grade.Coarse + Path.TumourLocation.Coarse +
##      pspline(Path.TumourSizeMm.MedCent, 3) + Path.Invasion.PN +
```

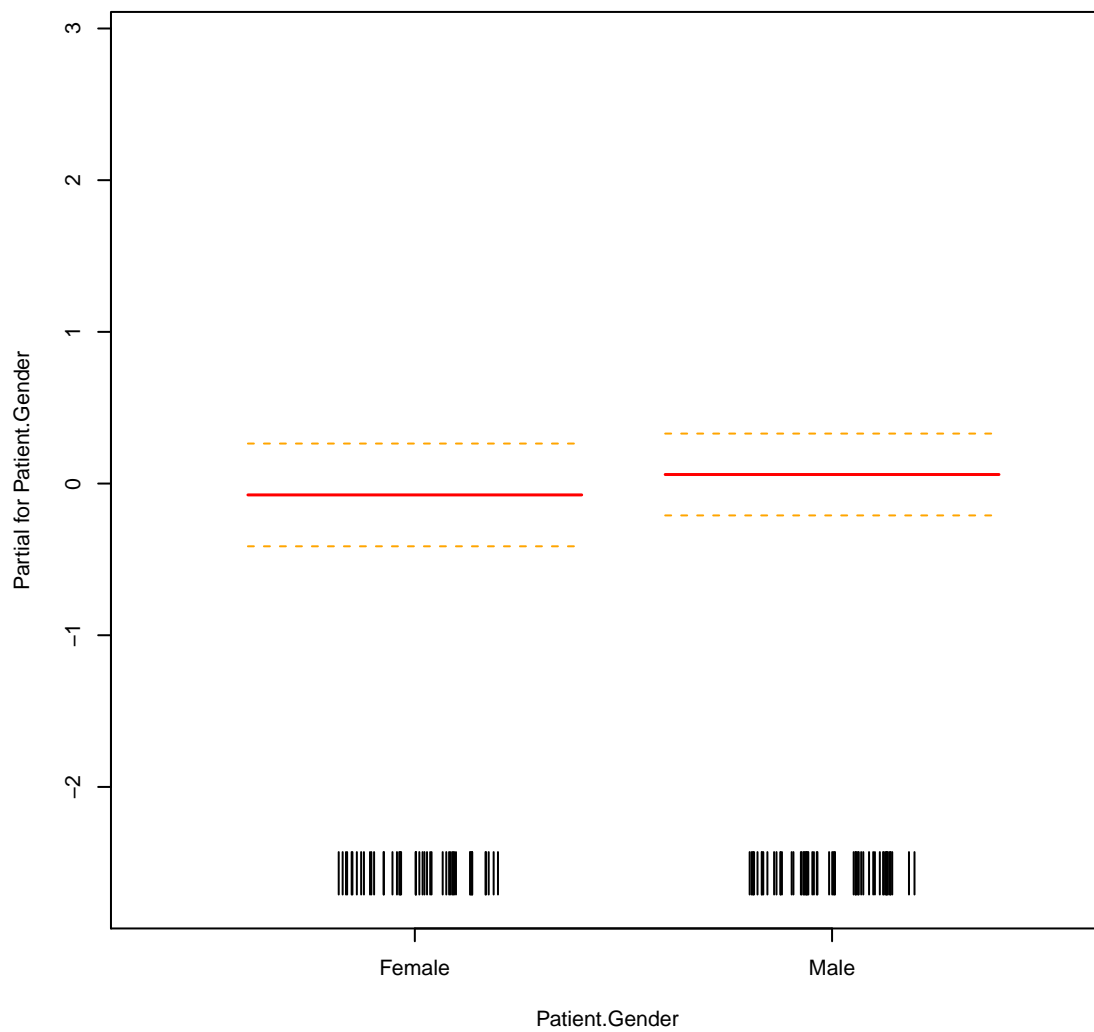
```
##          Path.Invasion.VS + Path.Nodes.Regional.Involved.Fraction +
##          Staging.pN + Treat.Surgery.ExcisionStatus.Coarse, data = cpvs.diag_dsd)
##
## n= 106, number of events= 66
## (4 observations deleted due to missingness)
##
##          coef          se(coef) se2      Chisq DF      p
## Patient.GenderMale      0.13460 0.3042    0.2980 0.20   1.00 0.6600
## pspline(History.Diagnosis 0.00338 0.0126    0.0124 0.07   1.00 0.7900
## pspline(History.Diagnosis          9.71  2.04 0.0082
## Path.Grade.Coarse.L      0.52606 0.2207    0.2157 5.68   1.00 0.0170
## Path.TumourLocation.Coars 0.57318 0.4468    0.4346 1.65   1.00 0.2000
## pspline(Path.TumourSizeMm 0.00458 0.0107    0.0106 0.18   1.00 0.6700
## pspline(Path.TumourSizeMm          2.22  2.05 0.3400
## Path.Invasion.PNPresent  -0.02745 0.4976    0.4906 0.00   1.00 0.9600
## Path.Invasion.VSPresent   0.53760 0.4075    0.4009 1.74   1.00 0.1900
## Path.Nodes.Regional.Invol 0.29582 0.7931    0.7717 0.14   1.00 0.7100
## Staging.pN.L             -0.26863 0.4039    0.3983 0.44   1.00 0.5100
## Treat.Surgery.ExcisionSta 0.66113 0.2228    0.2178 8.80   1.00 0.0030
##
##          exp(coef) exp(-coef) lower .95 upper .95
## Patient.GenderMale      1.144      0.874   0.63022    2.08
## ps(History.Diagnosis.AgeA 0.574      1.744   0.11162    2.95
## ps(History.Diagnosis.AgeA 0.325      3.077   0.02616    4.04
## ps(History.Diagnosis.AgeA 0.192      5.214   0.01096    3.35
## ps(History.Diagnosis.AgeA 0.131      7.639   0.00728    2.35
## ps(History.Diagnosis.AgeA 0.101      9.914   0.00626    1.62
## ps(History.Diagnosis.AgeA 0.113      8.860   0.00757    1.68
## ps(History.Diagnosis.AgeA 0.168      5.954   0.01140    2.47
## ps(History.Diagnosis.AgeA 0.280      3.573   0.01772    4.42
## ps(History.Diagnosis.AgeA 0.466      2.144   0.02201    9.88
## ps(History.Diagnosis.AgeA 0.814      1.229   0.01643   40.30
## Path.Grade.Coarse.L      1.692      0.591   1.09805    2.61
## Path.TumourLocation.Coars 1.774      0.564   0.73888    4.26
## ps(Path.TumourSizeMm.MedC 1.624      0.616   0.28939    9.12
## ps(Path.TumourSizeMm.MedC 2.533      0.395   0.16856   38.08
## ps(Path.TumourSizeMm.MedC 3.275      0.305   0.15860   67.62
## ps(Path.TumourSizeMm.MedC 4.068      0.246   0.19743   83.82
## ps(Path.TumourSizeMm.MedC 4.794      0.209   0.23539   97.62
## ps(Path.TumourSizeMm.MedC 4.467      0.224   0.21112   94.52
## ps(Path.TumourSizeMm.MedC 3.283      0.305   0.14497   74.35
## ps(Path.TumourSizeMm.MedC 2.754      0.363   0.10768   70.44
## ps(Path.TumourSizeMm.MedC 2.651      0.377   0.08821   79.65
## ps(Path.TumourSizeMm.MedC 2.618      0.382   0.05060  135.40
## Path.Invasion.PNPresent   0.973      1.028   0.36686    2.58
## Path.Invasion.VSPresent   1.712      0.584   0.77019    3.81
## Path.Nodes.Regional.Invol 1.344      0.744   0.28406    6.36
## Staging.pN.L             0.764      1.308   0.34637    1.69
## Treat.Surgery.ExcisionSta 1.937      0.516   1.25158    3.00
##
## Iterations: 3 outer, 9 Newton-Raphson
##      Theta= 0.7167
##      Theta= 0.7183
## Degrees of freedom for terms= 1.0 3.0 1.0 0.9 3.0 1.0 1.0 0.9 1.0 1.0
```

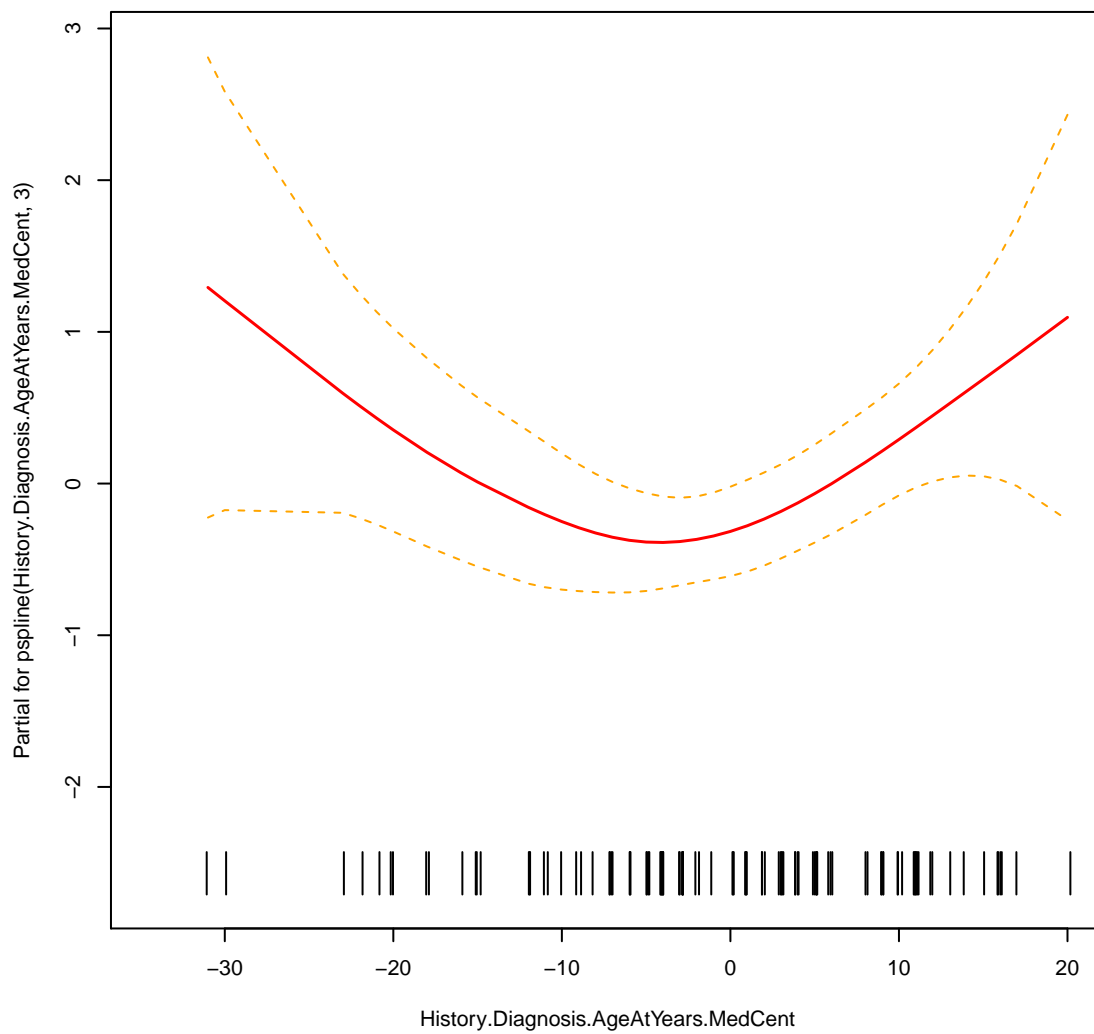
```
## Concordance= 0.734 (se = 0.039 )
## Likelihood ratio test= 45.1 on 13.8 df, p=3.41e-05

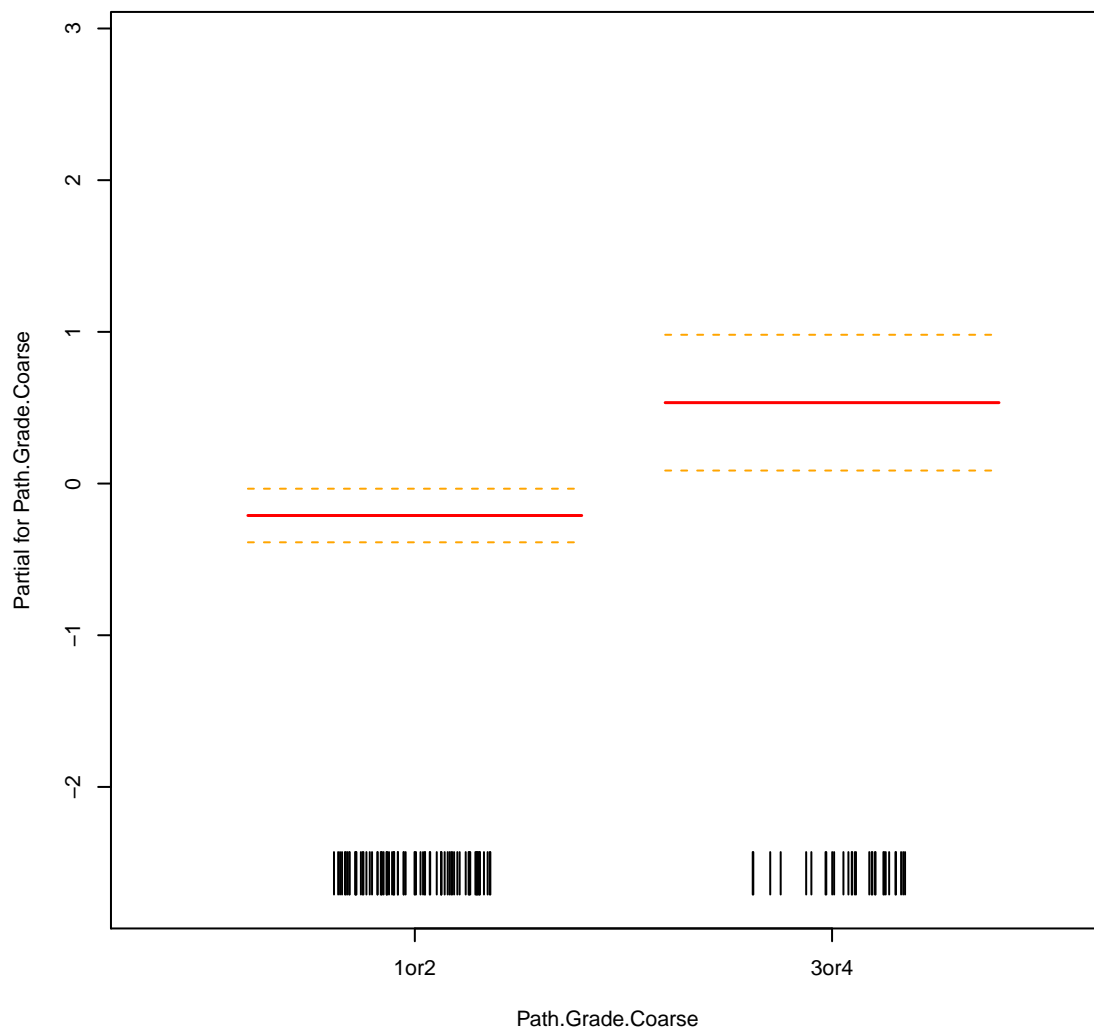
anova(fit.cpv.withoutsmoking)

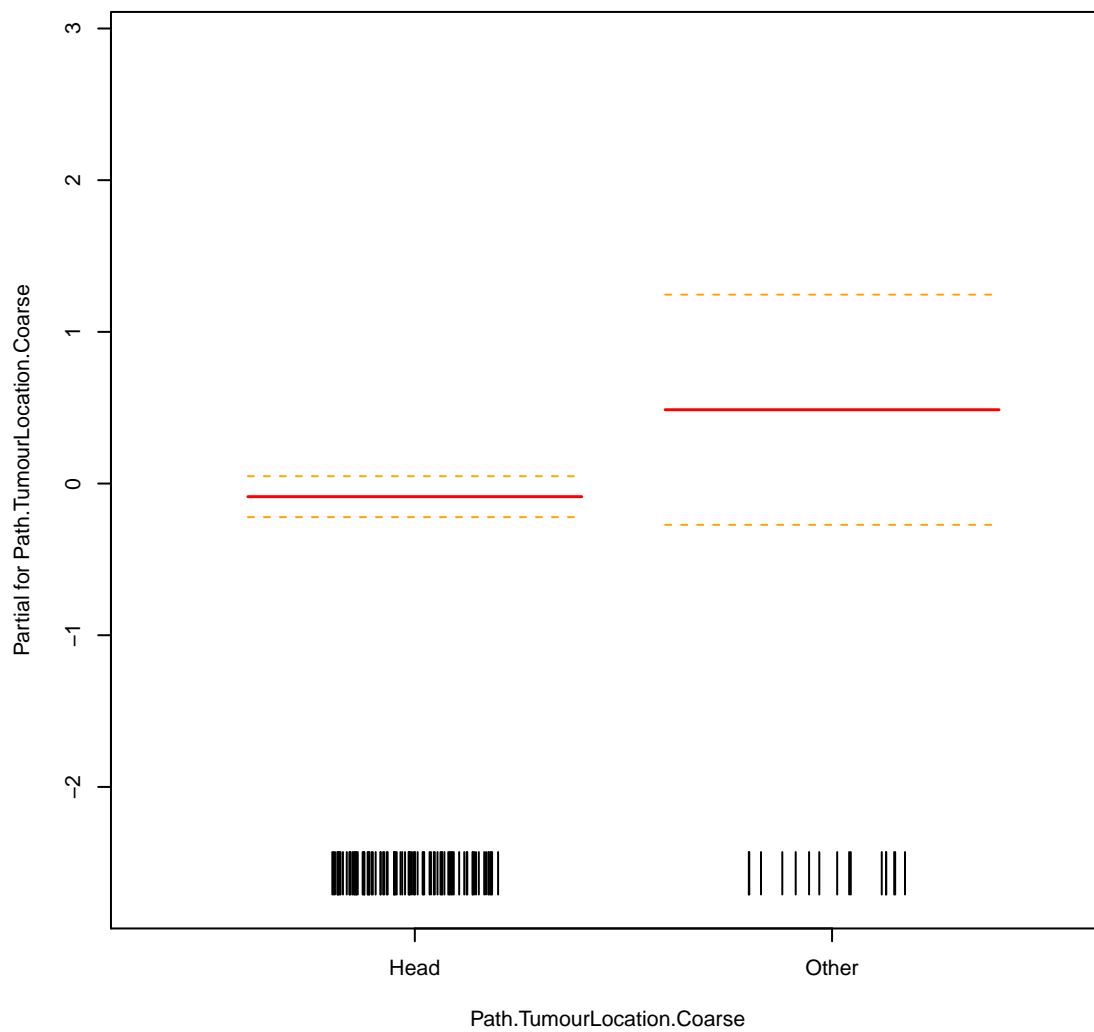
## Analysis of Deviance Table
## Cox model: response is Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath)
## Terms added sequentially (first to last)
##
##
## loglik Chisq Df
## NULL -270
## Patient.Gender -269 1.05 1
## pspline(History.Diagnosis.AgeAtYears.MedCent, 3) -262 15.63 10
## Path.Grade.Coarse -259 4.82 1
## Path.TumourLocation.Coarse -258 2.53 1
## pspline(Path.TumourSizeMm.MedCent, 3) -253 10.06 10
## Path.Invasion.PN -252 0.91 1
## Path.Invasion.VS -252 1.58 1
## Path.Nodes.Regional.Involved.Fraction -252 0.09 1
## Staging.pN -252 0.13 1
## Treat.Surgery.ExcisionStatus.Coarse -247 8.27 1
## Pr(>|Chi|)
## NULL
## Patient.Gender 0.305
## pspline(History.Diagnosis.AgeAtYears.MedCent, 3) 0.111
## Path.Grade.Coarse 0.028
## Path.TumourLocation.Coarse 0.112
## pspline(Path.TumourSizeMm.MedCent, 3) 0.435
## Path.Invasion.PN 0.339
## Path.Invasion.VS 0.208
## Path.Nodes.Regional.Involved.Fraction 0.763
## Staging.pN 0.722
## Treat.Surgery.ExcisionStatus.Coarse 0.004

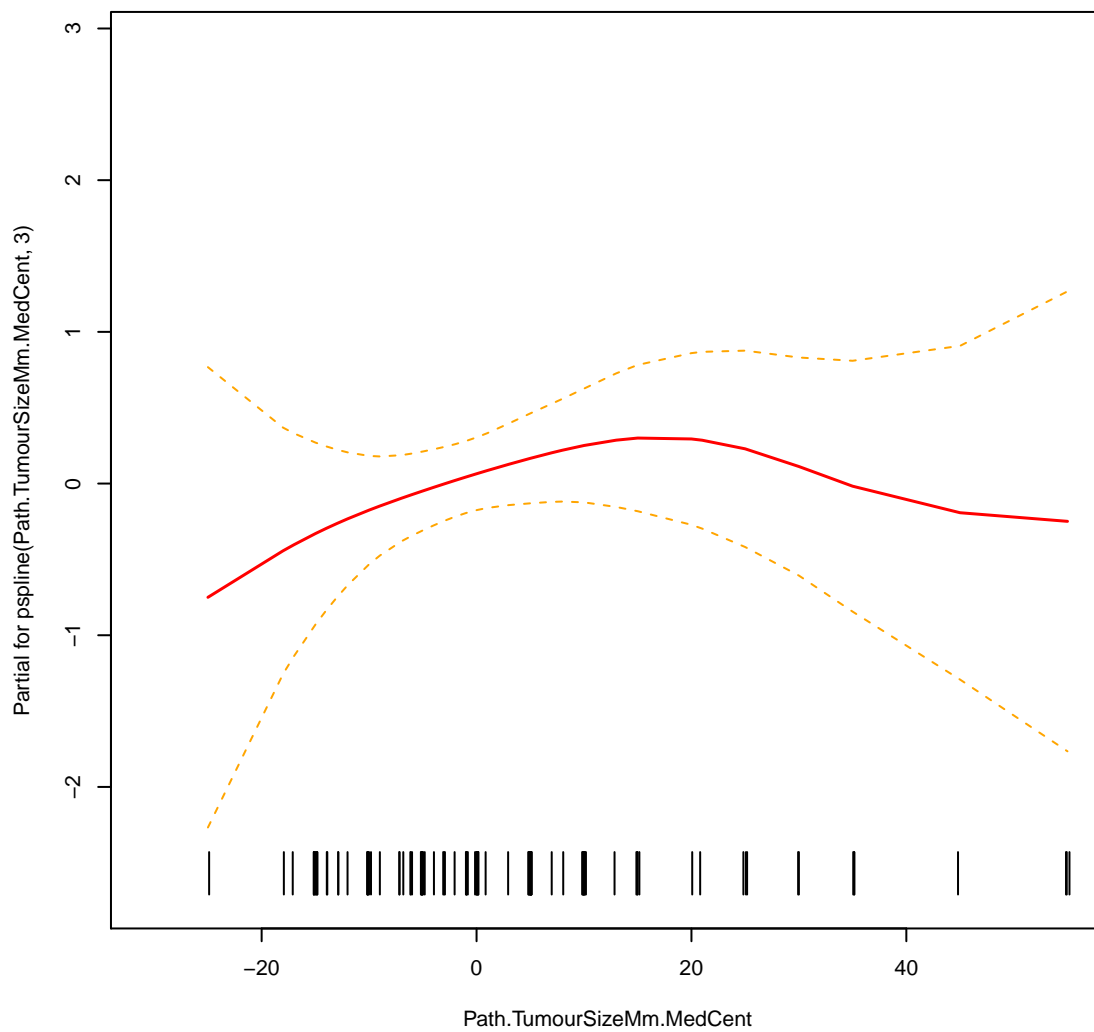
termplot(fit.cpv.withoutsmoking, rug = TRUE, se = TRUE)
```

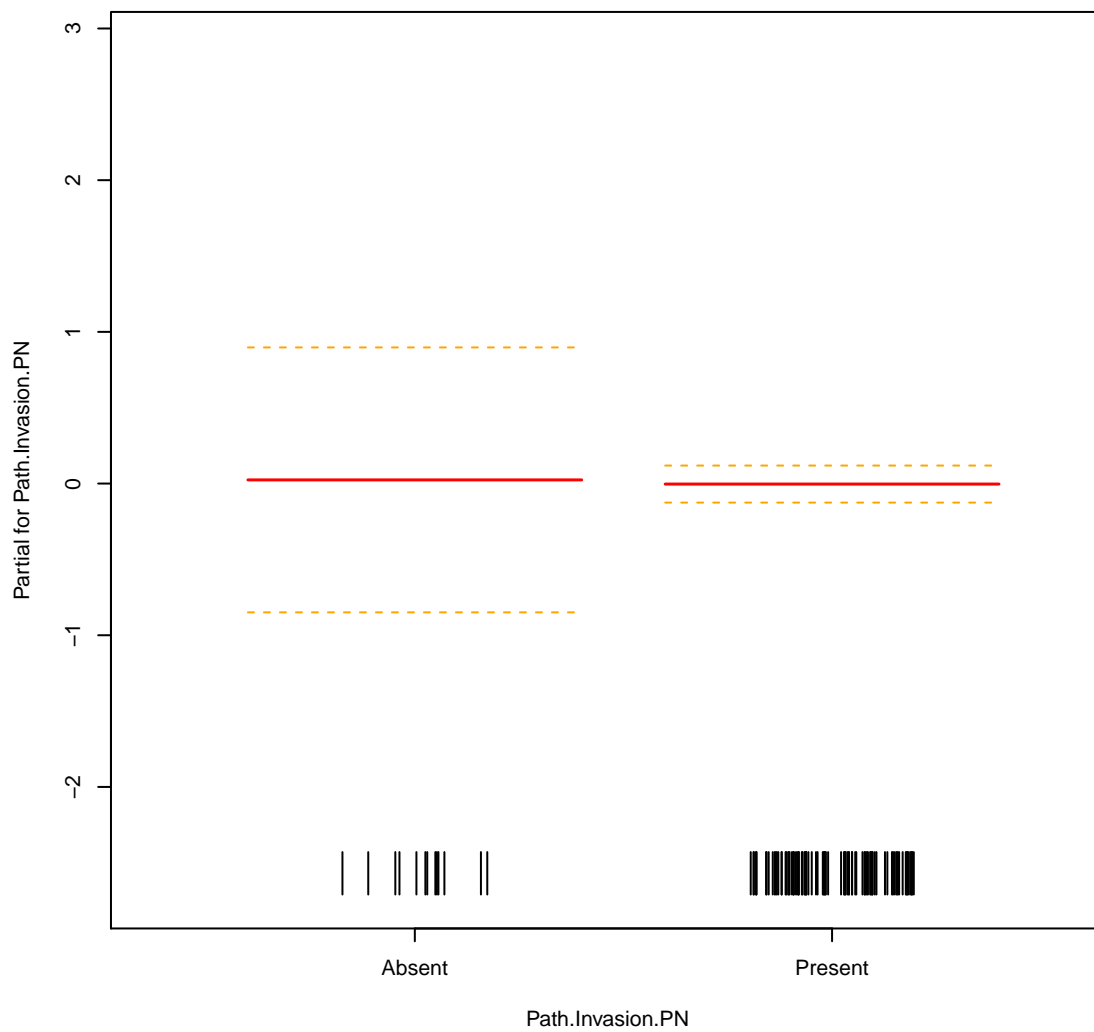


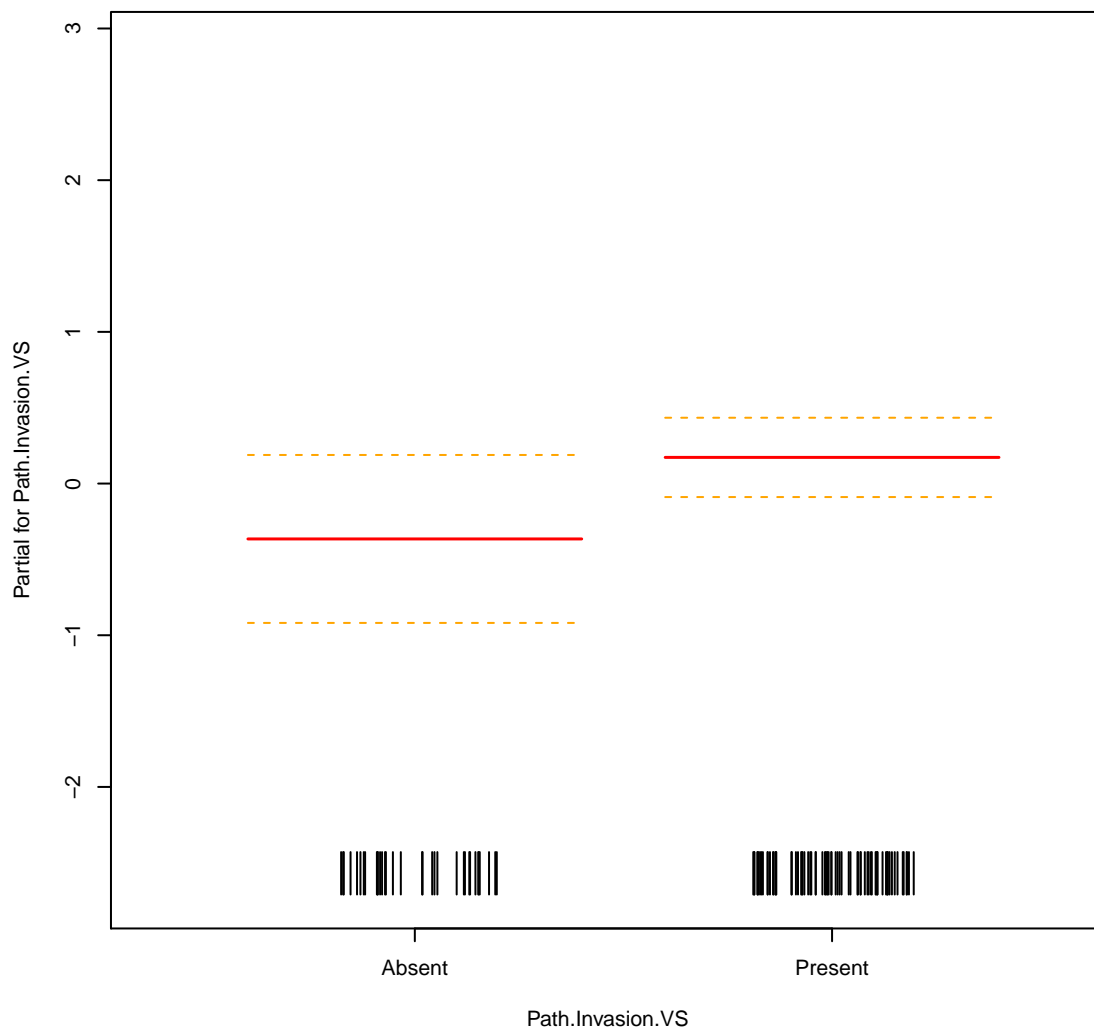


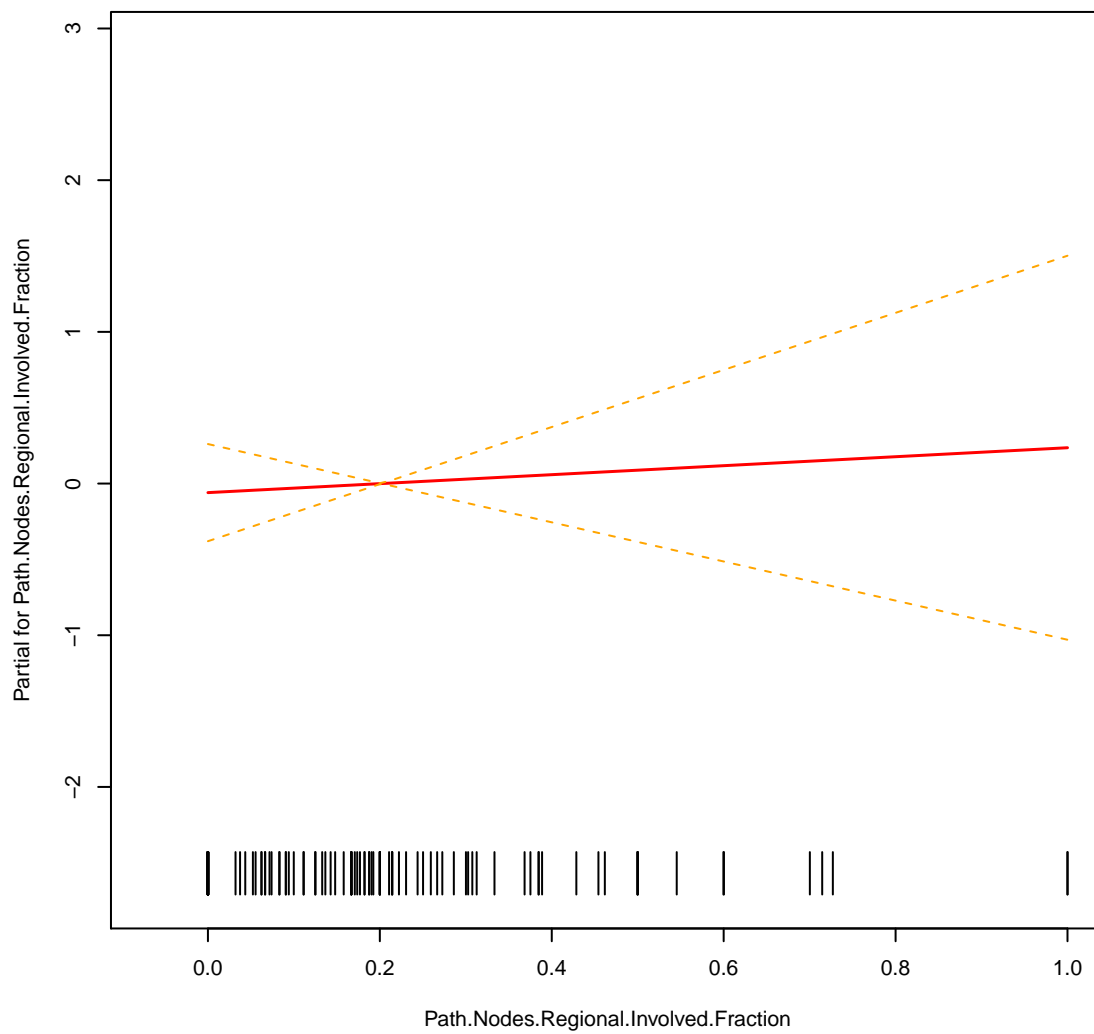


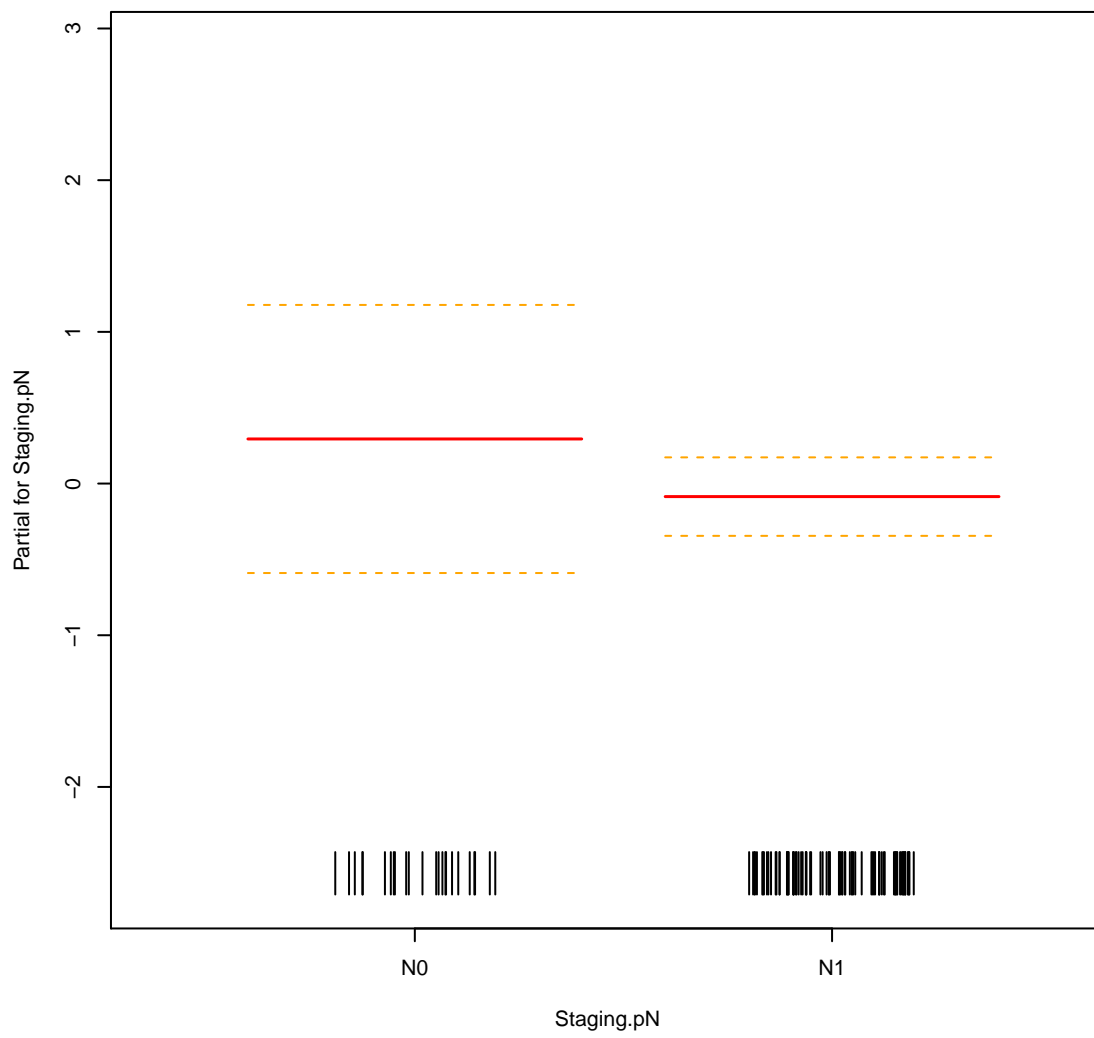


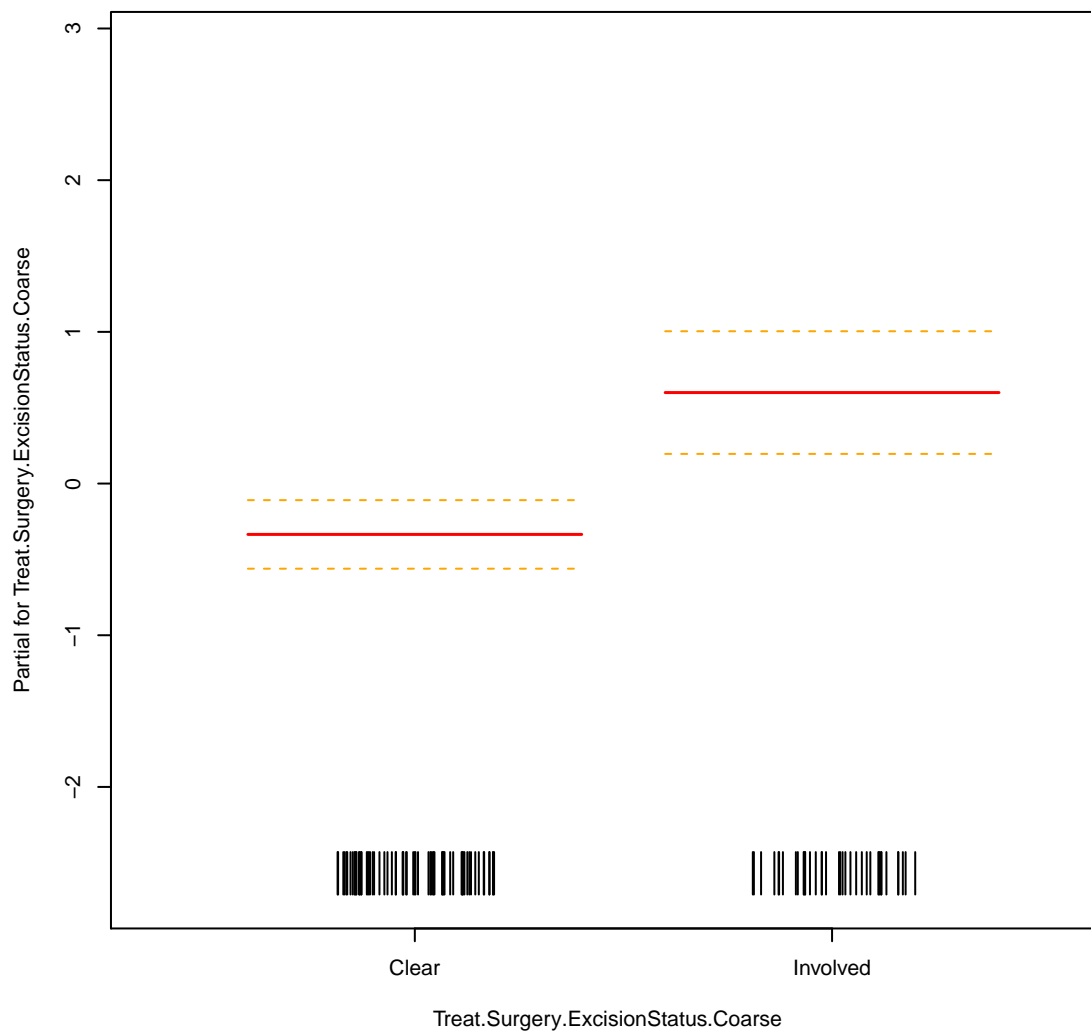












4 Probe selection

```
table(cpss.sis$sel)
```

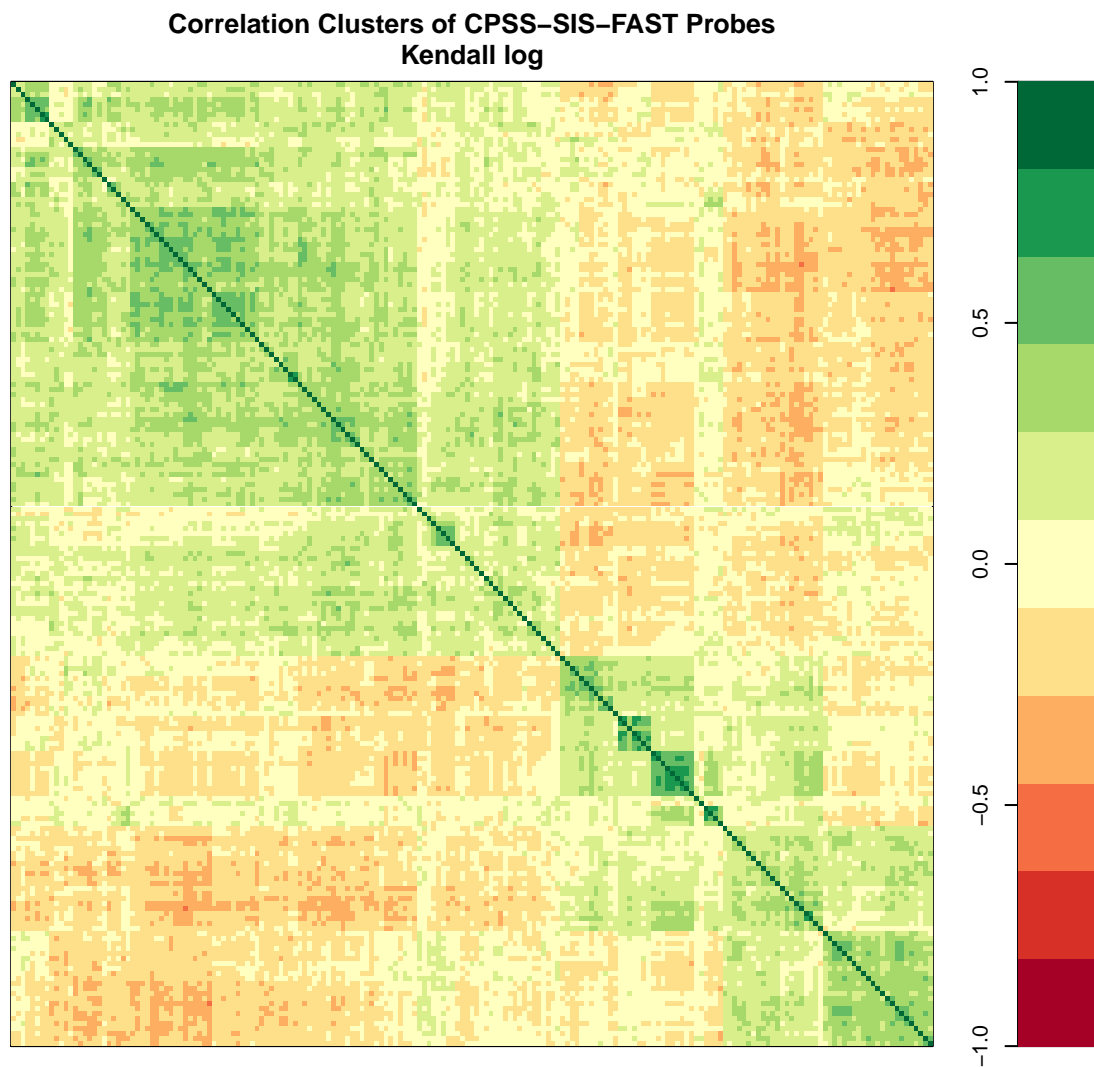
```
##
## FALSE  TRUE
## 12807   193
```

```
mean(cpss.sis$sel)
```

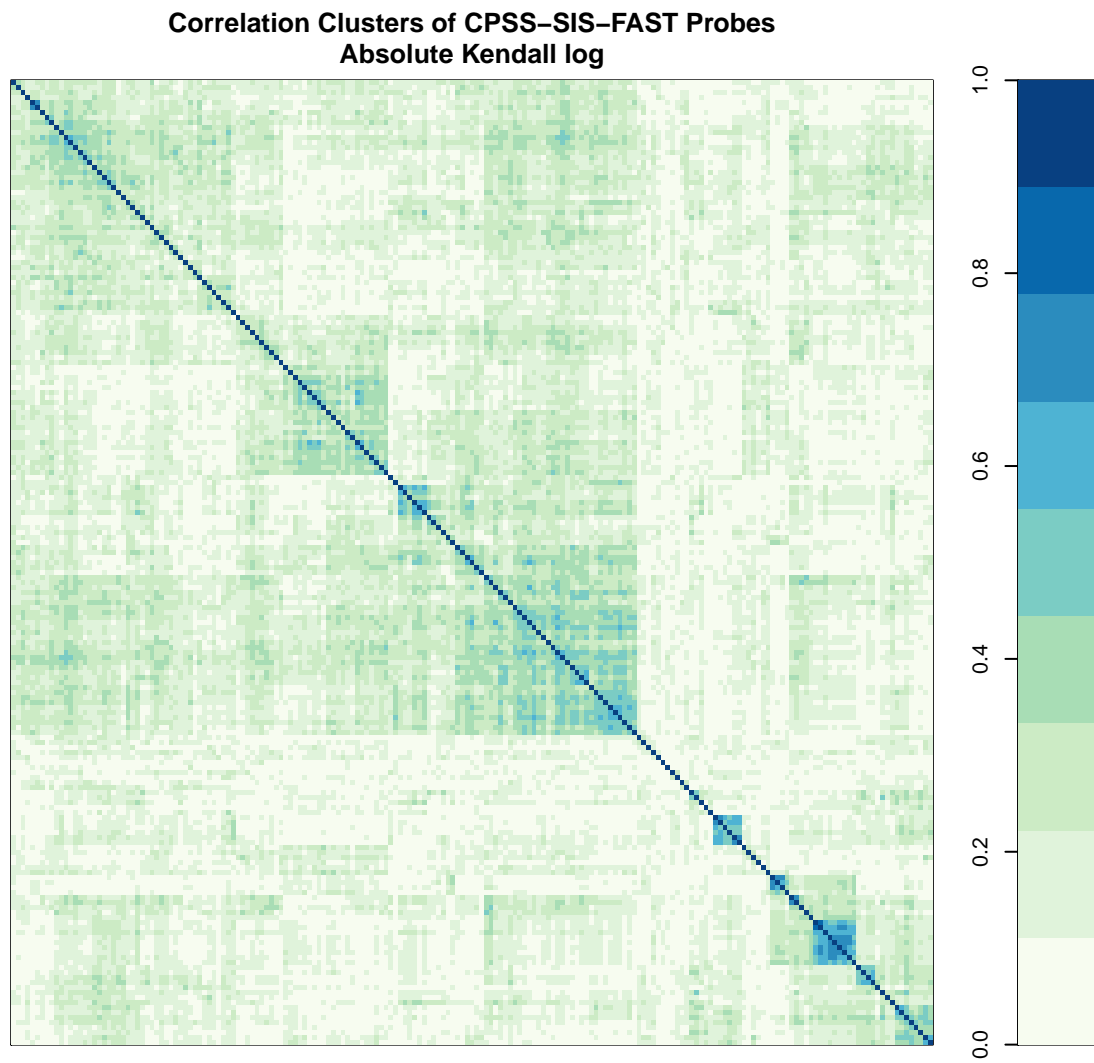
```
## [1] 0.01485
```

5 Expression correlation

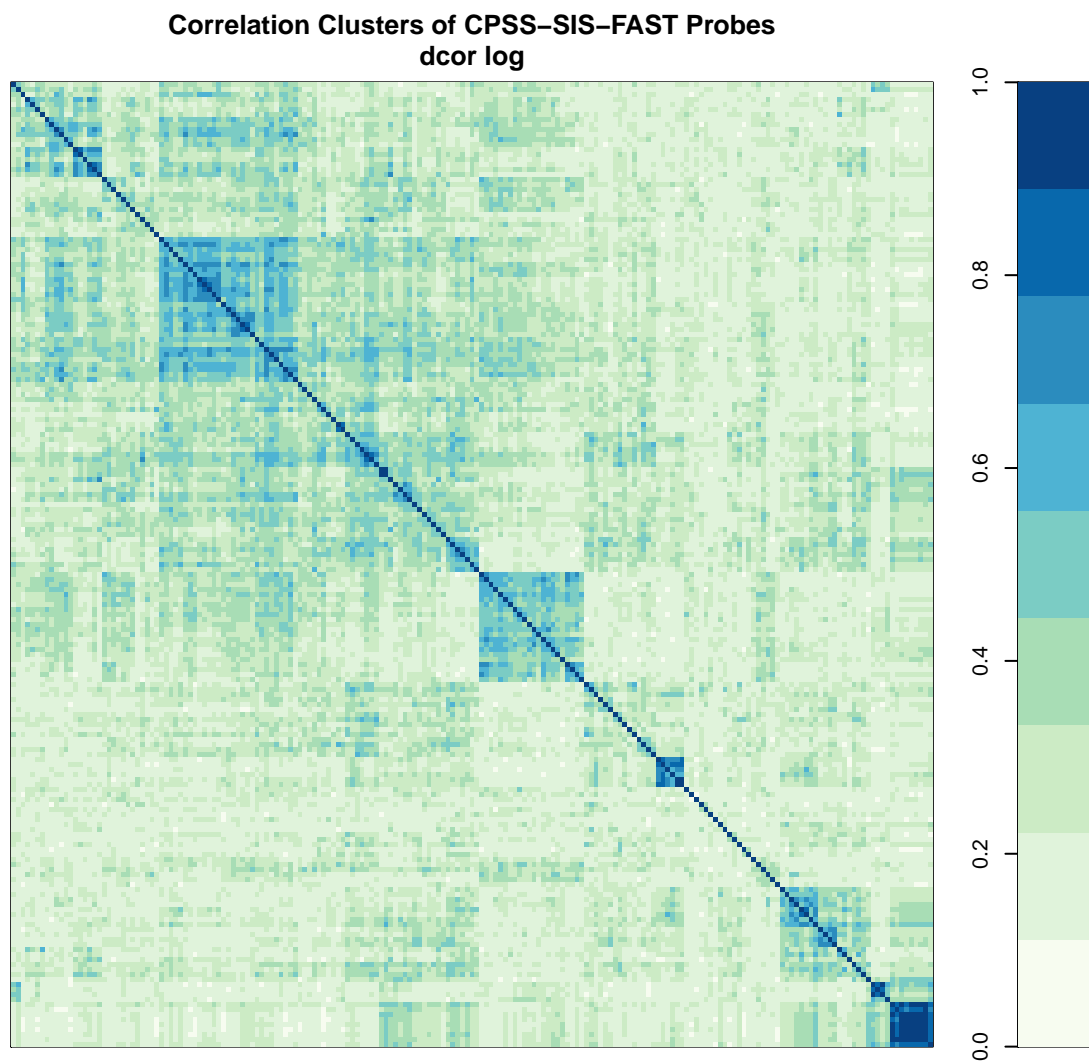
```
corPlot(x.sel.kcor, main = "Correlation Clusters of CPSS-SIS-FAST Probes\nKendall log",
        useRaster = FALSE)
```



```
corPlot(abs(x.sel.kcor), zlim = c(0, 1), pal = "GnBu", main = "Correlation Clusters of CPSS-SIS-FAST Probes",
        useRaster = FALSE)
```

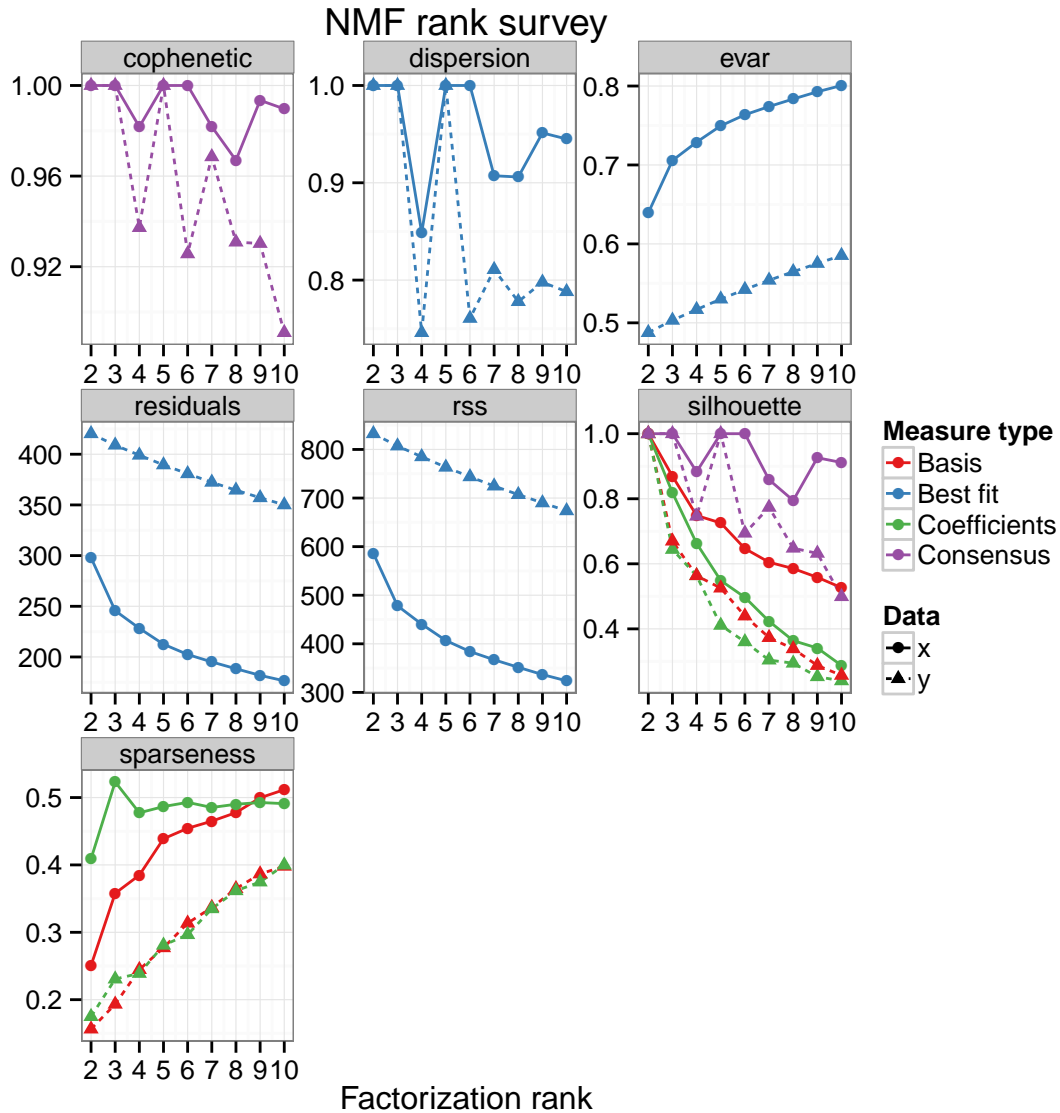



```
corPlot(x.sel.dcor, zlim = c(0, 1), pal = "GnBu", main = "Correlation Clusters of CPSS-SIS-FAST Probes",
        useRaster = FALSE)
```



6 Factorization

```
plot(nmf.runs.rank, nmf.runs.rank.random[[1]])
```

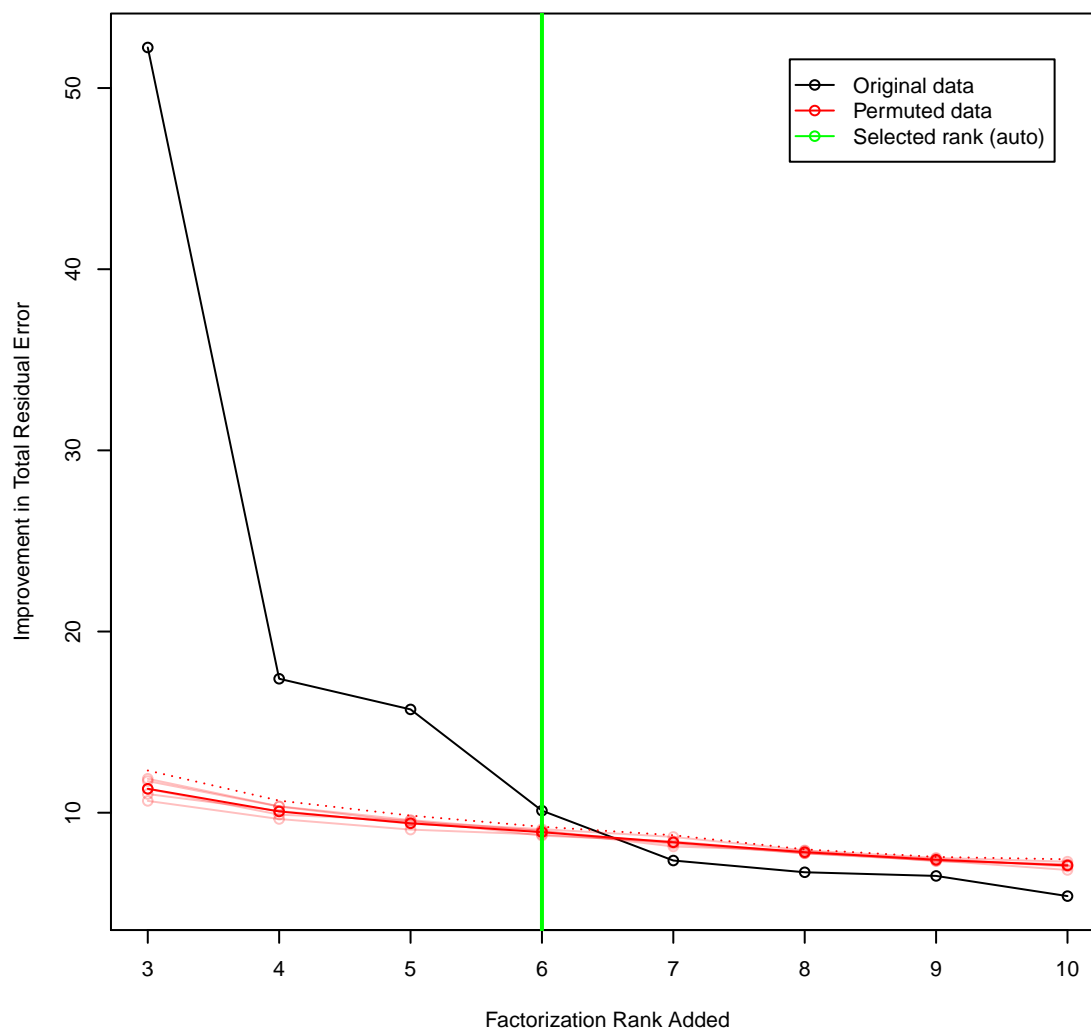


```
# for (i in nmf.runs.rankLfit) { consensusmap(i) }
plot(nmf.rankrange[-1], -temp.orig_resids.delta, type = "o", col = "black",
     pch = 21, ylim = range(-c(temp.orig_resids.delta, temp.perm_resids.delta.mean)),
     xlab = "Factorization Rank Added", ylab = "Improvement in Total Residual Error")
lines(nmf.rankrange[-1], -temp.perm_resids.delta.mean, col = "red", type = "o",
      pch = 21, lwd = 1)
for (i in 1:ncol(temp.perm_resids)) {
  lines(nmf.rankrange[-1], -temp.perm_resids.delta[, i], type = "o", col = rgb(1,
    0, 0, 0.25))
}
lines(nmf.rankrange[-1], -temp.perm_resids.delta.threshold, col = "red", lty = "dotted")
if (nmf.rank.wasauto == TRUE) {
  temp.col = "green"
} else {
  temp.col = "blue"
}
abline(v = nmf.rank, col = temp.col, lwd = 2)
```

```

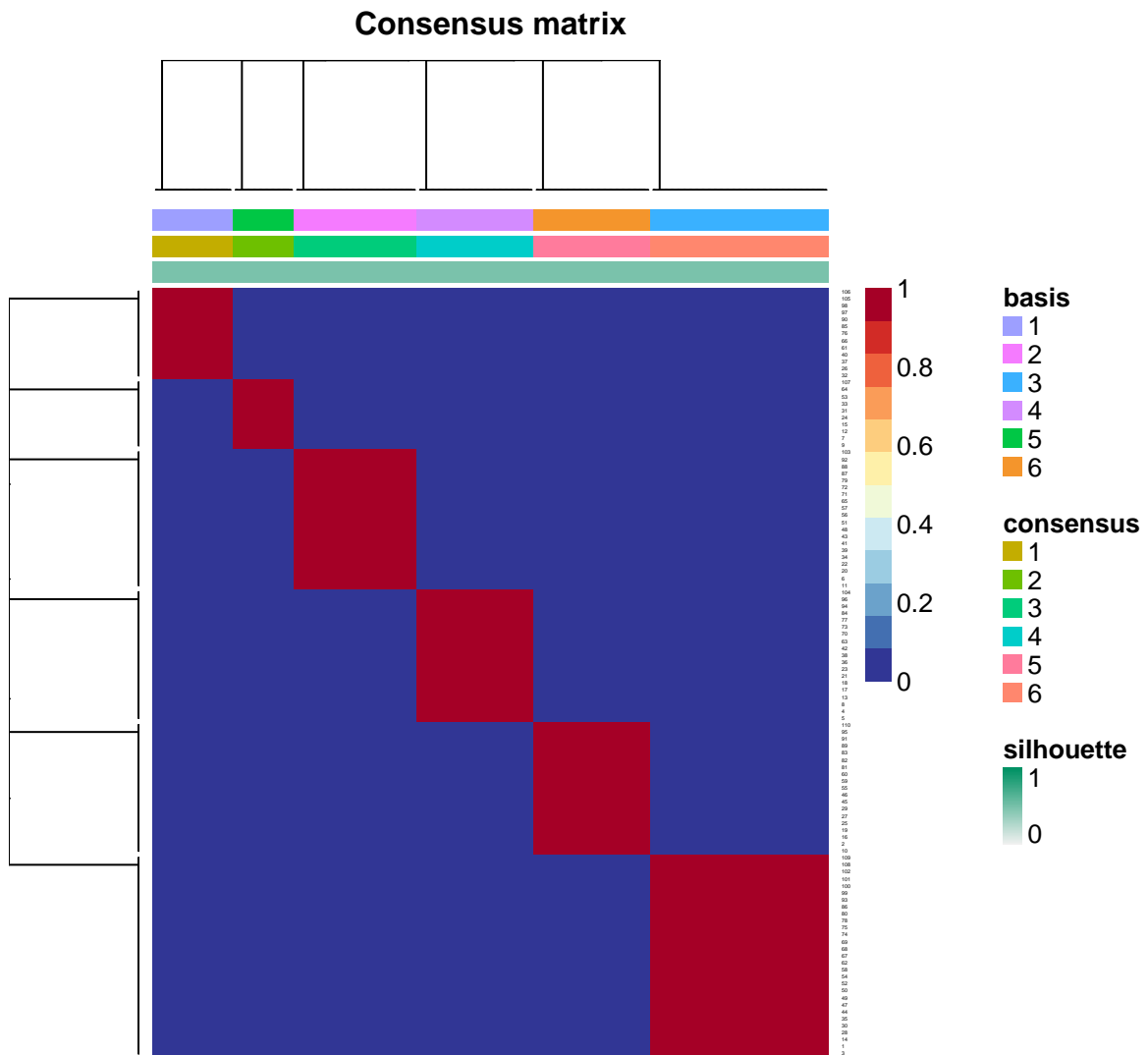
legend("topright", legend = c("Original data", "Permuted data", sprintf("Selected rank (%s)",
  ifelse(nmf.rank.wasauto == TRUE, "auto", "fixed"))), col = c("black", "red",
temp.col), lty = "solid", pch = 21, inset = 0.05)

```

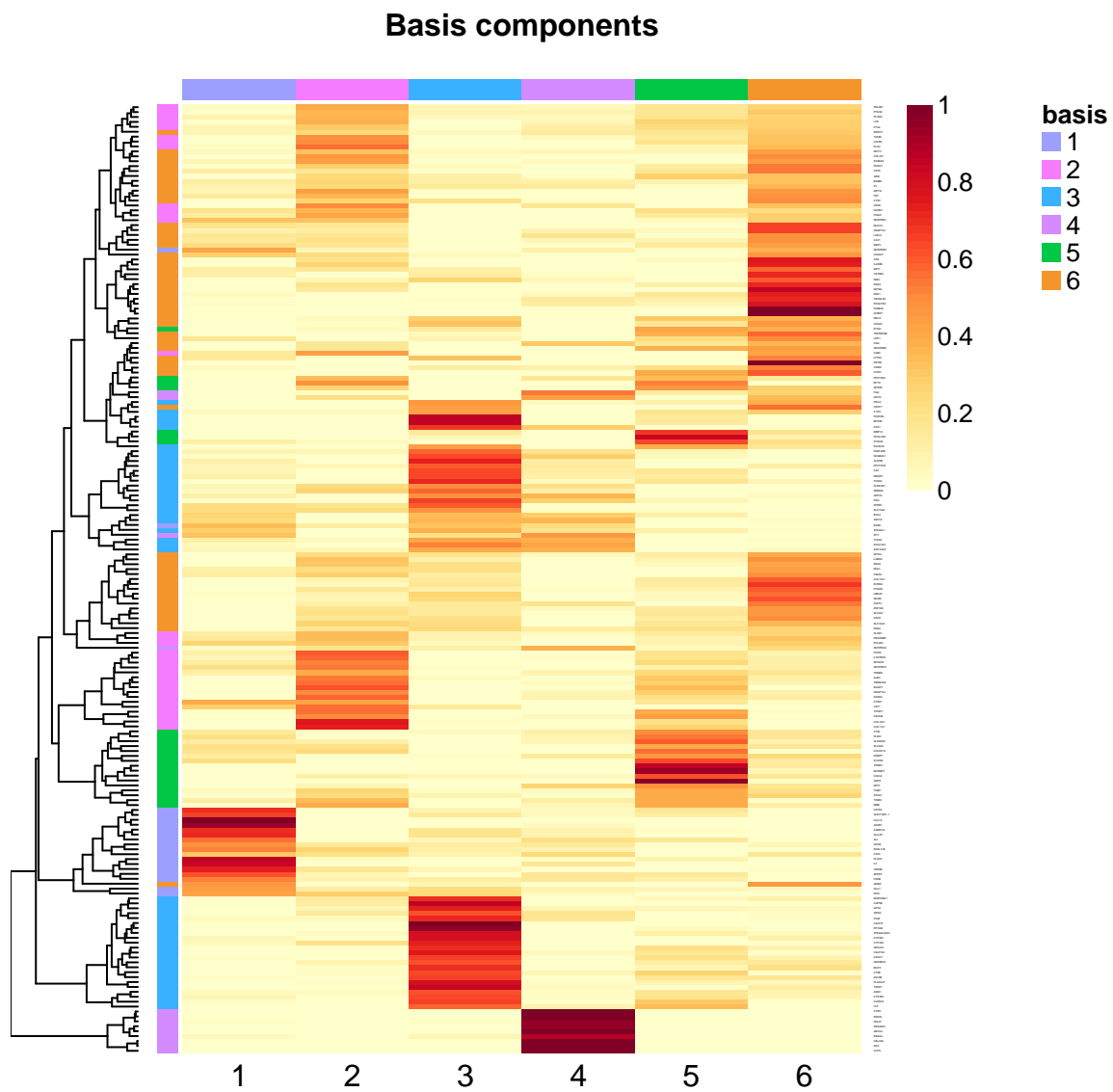


6.1 Fit

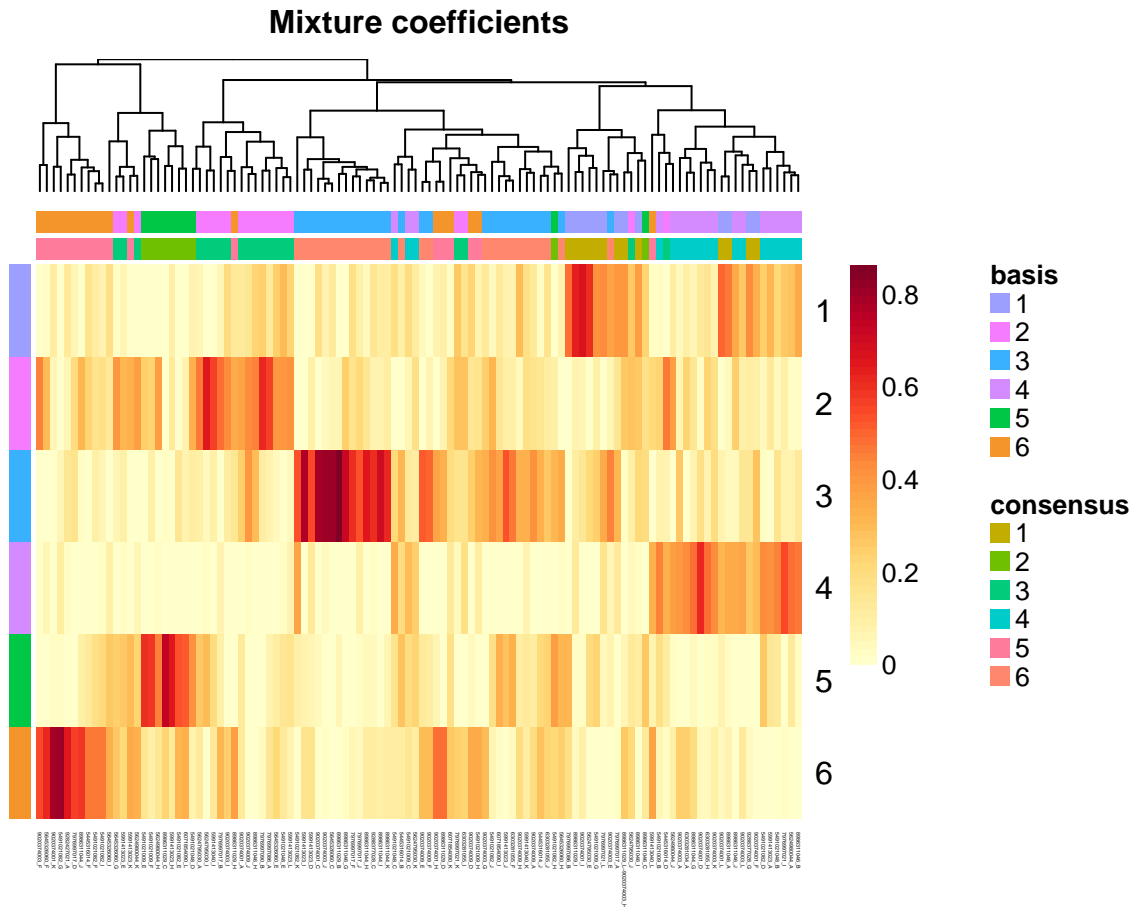
```
consensusmap(nmf.final)
```



```
basimap(nmf.final)
```



```
coefmap(nmf.final)
```



```

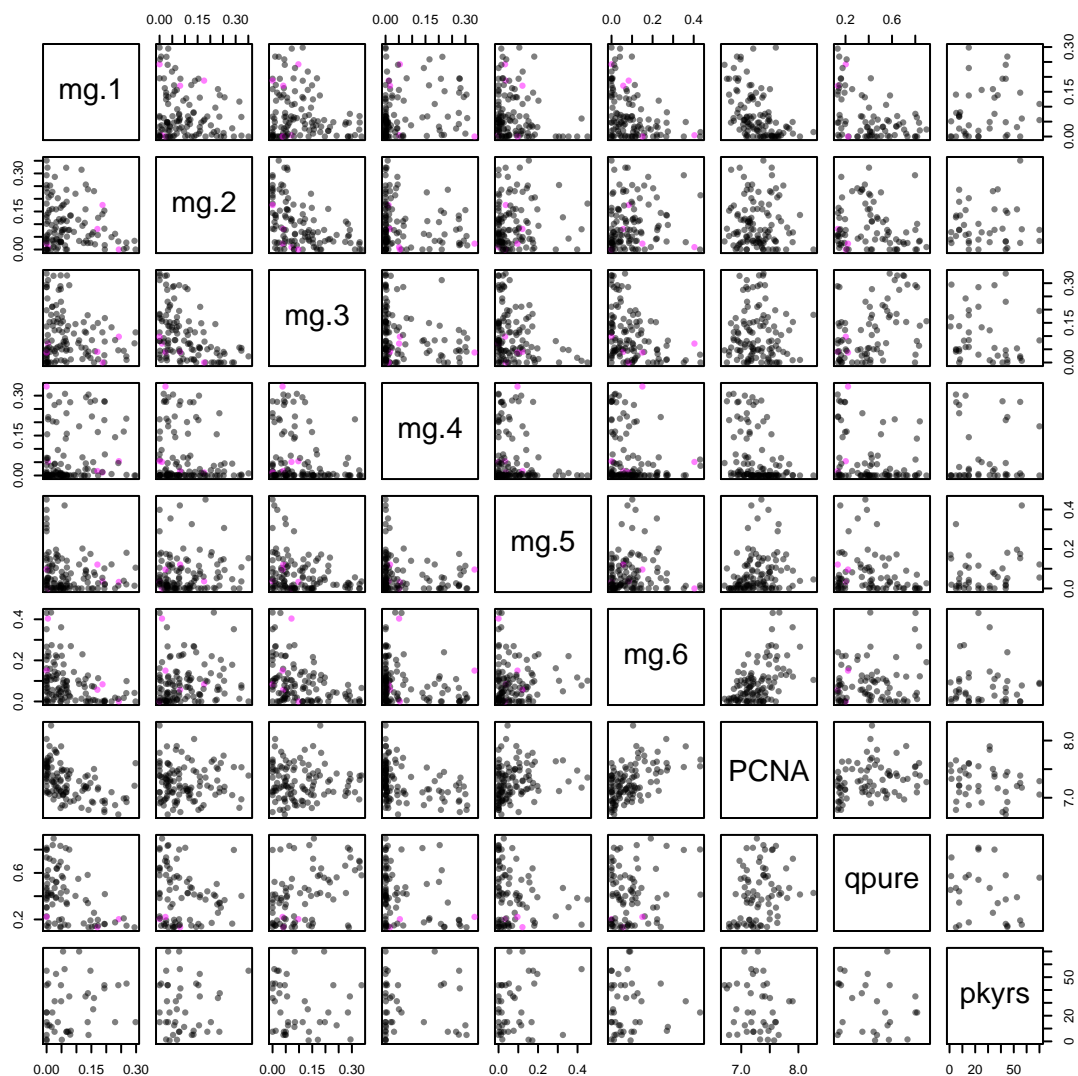
coefs.diag_dsd = apply(xlin.diag_dsd.sel, 2, function(xcol) nnls(basis(nmf.final),
xcol)$x)
coefs.diag_rec = apply(xlin.diag_rec.sel, 2, function(xcol) nnls(basis(nmf.final),
xcol)$x)
coefs.recr_dsd = apply(xlin.recr_dsd.sel, 2, function(xcol) nnls(basis(nmf.final),
xcol)$x)
coefs.pdac_au = apply(xlin.pdac_au.sel, 2, function(xcol) nnls(basis(nmf.final),
xcol)$x)

```

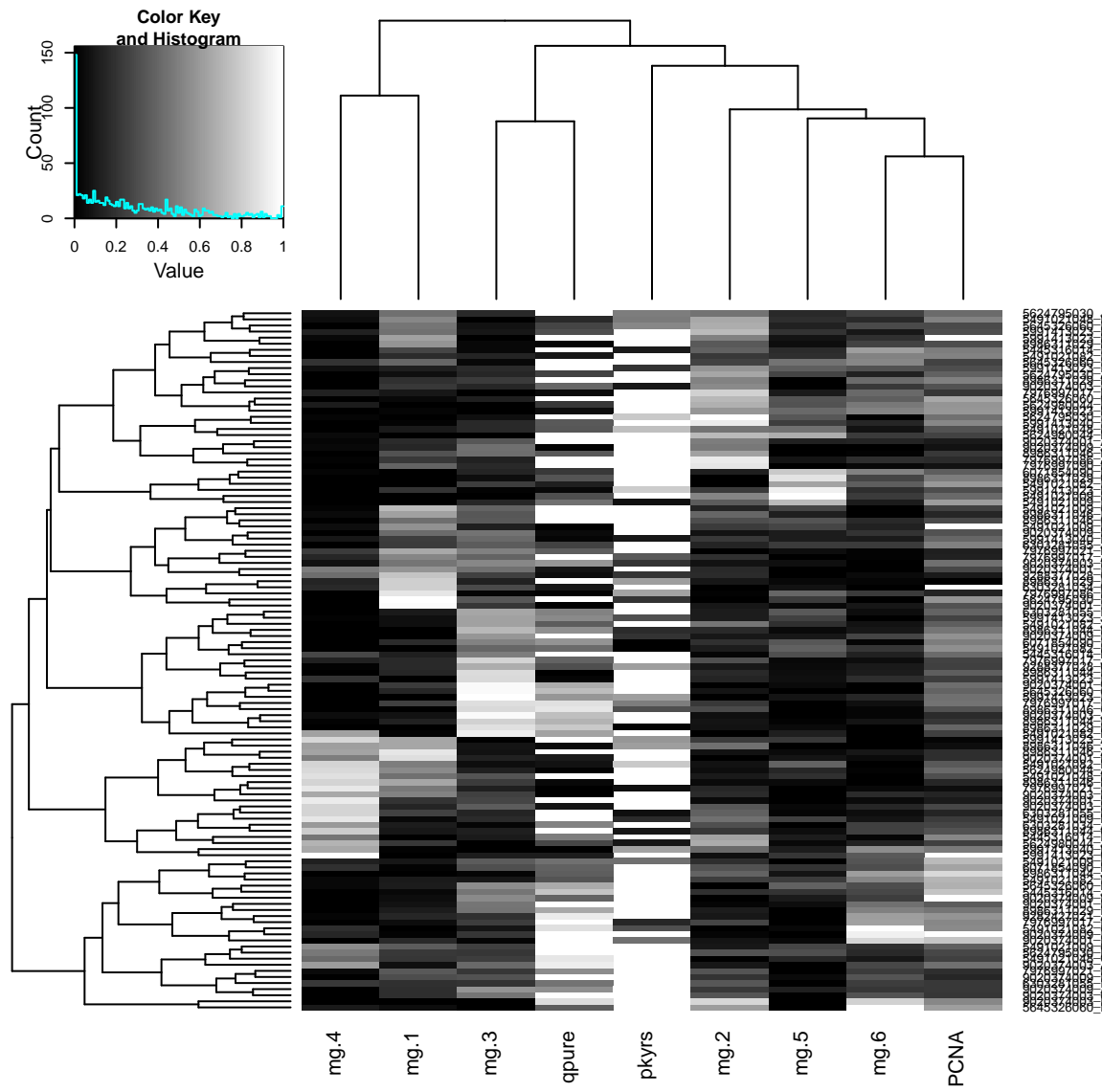
```

temp.pred.pairs = t(rbind(coefs.pdac_au, metapcna.scores[colnames(coefs.pdac_au)]))
colnames(temp.pred.pairs) = paste("mg", 1:ncol(temp.pred.pairs), sep = ".")
colnames(temp.pred.pairs)[ncol(temp.pred.pairs)] = "PCNA"
temp.pred.pairs = cbind(temp.pred.pairs, qpure = samp.pdac_au$Purity.qpure,
pkys = cpvs.pdac_au$History.Smoking.PackYears)
pairs(temp.pred.pairs, pch = 16, cex = 1, col = ifelse(rownames(temp.pred.pairs) %in%
colnames(xlin.diag_dsd.sel), rgb(0, 0, 0, 0.5), rgb(1, 0, 1, 0.5)))

```



```
temp.pred.pairs.rescaled = t((t(temp.pred.pairs) - apply(temp.pred.pairs, 2,
  min, na.rm = TRUE)) / (apply(temp.pred.pairs, 2, function(x) diff(range(x,
  na.rm = TRUE))))))
heatmap.2(temp.pred.pairs.rescaled, trace = "none", scale = "none", col = grey(seq(0,
  1, 0.01)))
```

6.2 Prediction on training set

```
nmf.final.cpv.pvals = data.frame(surv.diag_rec.p = apply(coefs.diag_rec, 1,
  function(xc) pchisq(2 * diff(coxph(y.diag_rec ~ xc)$loglik), df = 1, lower.tail = FALSE)),
  surv.diag_rec.c = apply(coefs.diag_rec, 1, function(xc) coef(coxph(y.diag_rec ~
    xc))), surv.diag_dsd.p = apply(coefs.diag_dsd, 1, function(xc) pchisq(2 *
    diff(coxph(y.diag_dsd ~ xc)$loglik), df = 1, lower.tail = FALSE)), surv.diag_dsd.c = apply(coefs
    1, function(xc) coef(coxph(y.diag_dsd ~ xc))), surv.recr_dsd.p = apply(coefs.recr_dsd,
    1, function(xc) pchisq(2 * diff(coxph(y.recr_dsd ~ xc)$loglik), df = 1,
    lower.tail = FALSE)), surv.recr_dsd.c = apply(coefs.recr_dsd, 1,
    function(xc) coef(coxph(y.recr_dsd ~ xc))), pure.p = apply(coefs.pdac_au,
    1, function(xc) cor.test(samps.pdac_au$purity_qpure, xc, method = "kendall")$p.value),
    pure.s = apply(coefs.pdac_au, 1, function(xc) cor.test(samps.pdac_au$purity_qpure,
    xc, method = "kendall")$statistic))
temp.pvals = as.matrix(nmf.final.cpv.pvals[, grepl("\\.p$", colnames(nmf.final.cpv.pvals))])
```

```
temp.pvals.FWER = matrix(p.adjust(as.vector(temp.pvals), "holm"), nrow = nrow(temp.pvals))
colnames(temp.pvals.FWER) = paste(colnames(temp.pvals), "Holm", sep = ".")
temp.pvals.BY = matrix(p.adjust(as.vector(temp.pvals), "BY"), nrow = nrow(temp.pvals))
colnames(temp.pvals.BY) = paste(colnames(temp.pvals), "BY", sep = ".")
nmf.final.cpv.pvals = cbind(nmf.final.cpv.pvals, temp.pvals.FWER, temp.pvals.BY)
nmf.final.cpv.pvals = nmf.final.cpv.pvals[, order(colnames(nmf.final.cpv.pvals))]
```

	pure.p	pure.p.BY	pure.p.Holm	pure.s	surv.diag_dsd.c	surv.diag_dsd.p	surv.diag_dsd.p.BY	surv.diag_dsd.p.Holm	surv.diag_rec.c	surv.diag_rec.p	surv.diag_rec.p.BY	surv.diag_rec.p.Holm	surv.recr_dsd.c	surv.recr_dsd.p	surv.recr_dsd.p.BY	surv.recr_dsd.p.Holm
1	0.00	0.01	0.01	-3.42	-4.56	0.00	0.04	0.06	-3.78	0.01	0.07	0.11	-2.36	0.16	0.75	0.96
2	0.05	0.30	0.50	-1.96	3.10	0.02	0.15	0.25	1.99	0.11	0.55	0.77	2.62	0.06	0.35	0.56
3	0.00	0.00	0.00	3.91	-5.46	0.00	0.00	0.00	-2.64	0.03	0.19	0.33	-4.66	0.00	0.04	0.04
4	0.03	0.19	0.33	-2.19	-0.95	0.44	1.00	1.00	-1.68	0.16	0.75	0.96	0.56	0.68	1.00	1.00
5	0.07	0.38	0.57	-1.81	2.84	0.02	0.12	0.21	2.73	0.01	0.12	0.20	1.37	0.29	1.00	1.00
6	0.78	1.00	1.00	0.28	6.61	0.00	0.00	0.00	5.49	0.00	0.00	0.00	3.97	0.00	0.04	0.06

Table 1: Resubstitution prediction, all tests

	pure.p.Holm	pure.s	surv.diag_dsd.c	surv.diag_dsd.p.Holm	surv.diag_rec.c	surv.diag_rec.p.Holm	surv.recr_dsd.c	surv.recr_dsd.p.Holm
1	0.01	-3.42	-4.56	0.06	-3.78	0.11	-2.36	0.96
2	0.50	-1.96	3.10	0.25	1.99	0.77	2.62	0.56
3	0.00	3.91	-5.46	0.00	-2.64	0.33	-4.66	0.04
4	0.33	-2.19	-0.95	1.00	-1.68	0.96	0.56	1.00
5	0.57	-1.81	2.84	0.21	2.73	0.20	1.37	1.00
6	1.00	0.28	6.61	0.00	5.49	0.00	3.97	0.06

Table 2: Resubstitution prediction, Holm MTC only

MTC P-values

```
glmnet.coef.1se

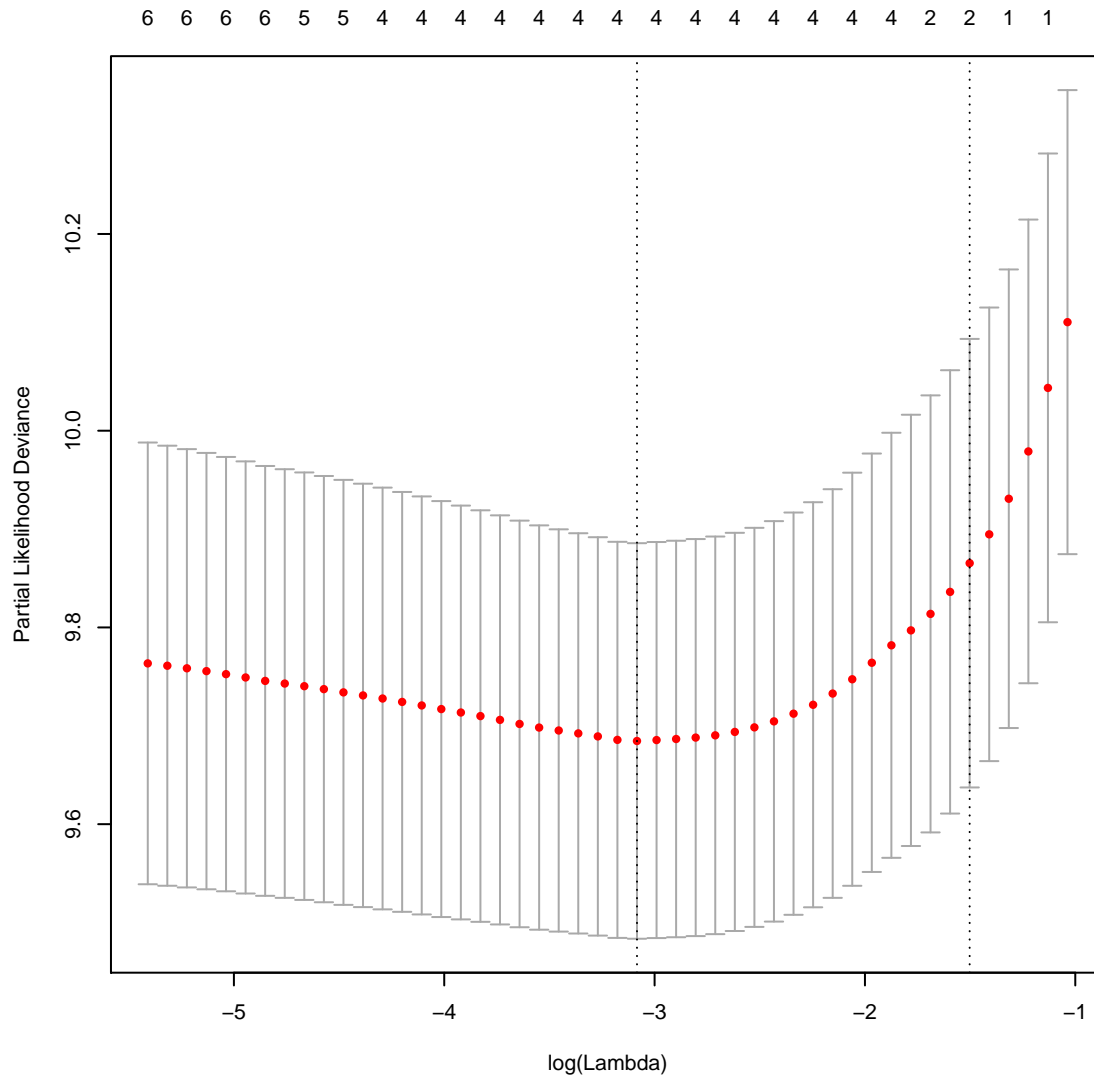
## 6 x 1 sparse Matrix of class "dgCMatrix"
##           1
## mg.1      .
## mg.2      .
## mg.3 -0.1635
## mg.4      .
## mg.5      .
## mg.6  3.0808

glmnet.coef.min

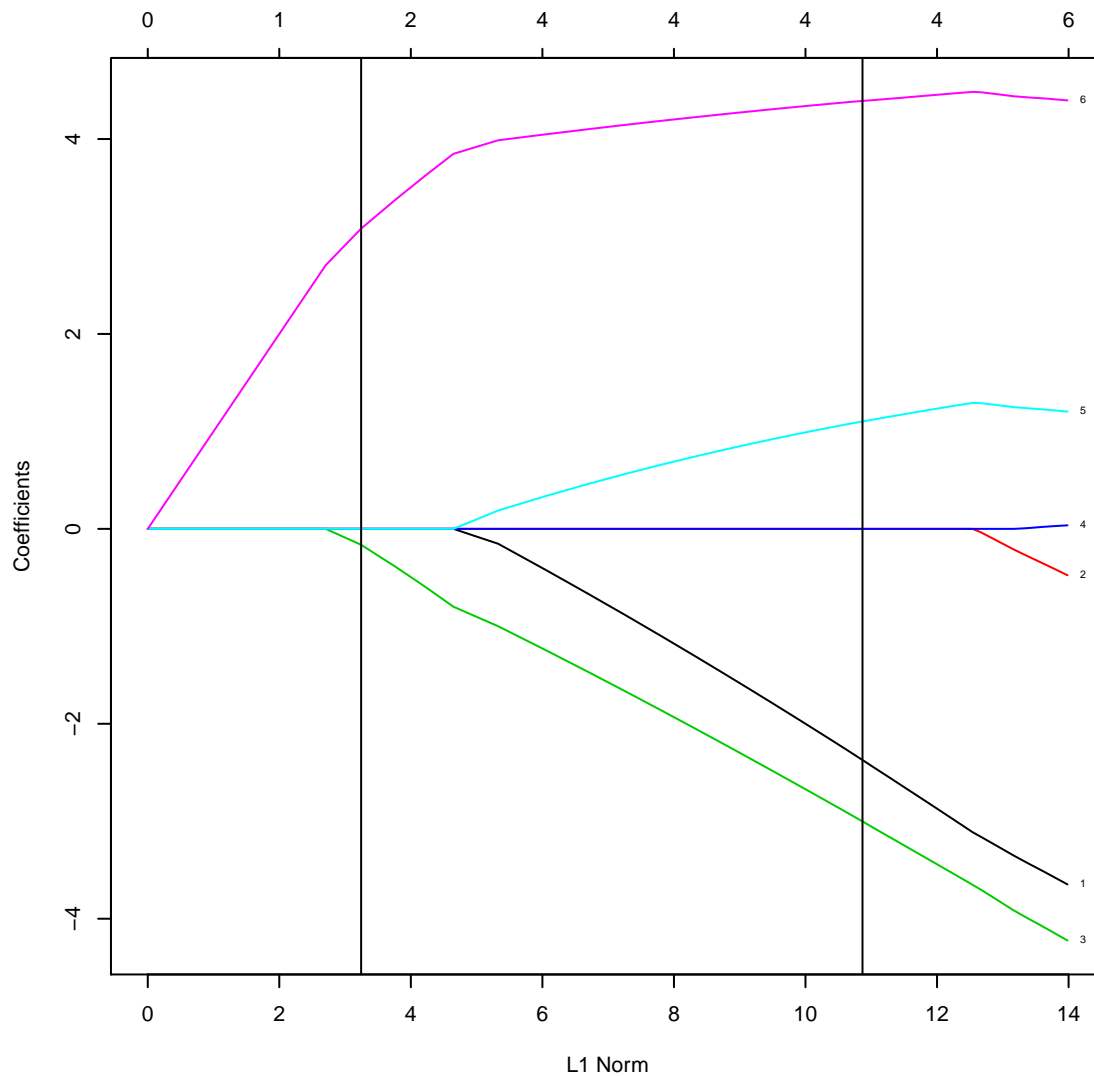
## 6 x 1 sparse Matrix of class "dgCMatrix"
##           1
## mg.1 -2.372
## mg.2      .
## mg.3 -3.002
## mg.4      .
## mg.5  1.102
## mg.6  4.391
```

```
plot(glmnet.fit.cv)
```

LASSO



```
plot(glmnet.fit.cv$glmnet.fit, label = TRUE)
abline(v = sum(abs(glmnet.coef.1se)))
abline(v = sum(abs(glmnet.coef.min)))
```



```
fit.cpv.sig.withoutsmoking.resub = coxph(Surv(Surv.EventTimeFromDiag.DSDeath,
  Surv.Event.DSDeath) ~ Patient.Gender + pspline(History.Diagnosis.AgeAtYears.MedCent,
  3) + Path.Grade.Coarse + Path.TumourLocation.Coarse + pspline(Path.TumourSizeMm.MedCent,
  3) + Path.Invasion.PN + Path.Invasion.VS + Path.Nodes.Regional.Involved.Fraction +
  Staging.pN + Treat.Surgery.ExcisionStatus.Coarse + t(coefs.diag_dsd), data = cpvs.diag_dsd)
anova(fit.cpv.withoutsmoking, fit.cpv.sig.withoutsmoking.resub)

## Analysis of Deviance Table
## Cox model: response is Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath)
## Model 1: ~ Patient.Gender + pspline(History.Diagnosis.AgeAtYears.MedCent, 3) + Path.Grade.Coarse + Path.TumourLocation.Coarse + Path.TumourSizeMm.MedCent + Path.Invasion.PN + Path.Invasion.VS + Path.Nodes.Regional.Involved.Fraction + Staging.pN + Treat.Surgery.ExcisionStatus.Coarse
## Model 2: ~ Patient.Gender + pspline(History.Diagnosis.AgeAtYears.MedCent, 3) + Path.Grade.Coarse + Path.TumourLocation.Coarse + Path.TumourSizeMm.MedCent + Path.Invasion.PN + Path.Invasion.VS + Path.Nodes.Regional.Involved.Fraction + Staging.pN + Treat.Surgery.ExcisionStatus.Coarse + t(coefs.diag_dsd)
## loglik Chisq Df P(>|Chi|)
## 1 -247
## 2 -237 20.9 6 0.0019
```

```

fit.cpv.sig.withsmoking.resub = coxph(Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath) ~
  Patient.Gender + pspline(History.Diagnosis.AgeAtYears.MedCent, 3) + Path.Grade.Coarse +
  Path.TumourLocation.Coarse + pspline(Path.TumourSizeMm.MedCent, 3) +
  Path.Invasion.PN + Path.Invasion.VS + Path.Nodes.Regional.Involved.Fraction +
  Staging.pN + History.Smoking.PackYears + Treat.Surgery.ExcisionStatus.Coarse +
  t(coefs.diag_dsd), data = cpvs.diag_dsd)
anova(fit.cpv.withsmoking, fit.cpv.sig.withsmoking.resub)

## Analysis of Deviance Table
## Cox model: response is Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath)
## Model 1: ~ Patient.Gender + pspline(History.Diagnosis.AgeAtYears.MedCent, 3) + Path.Grade.Coarse +
## Model 2: ~ Patient.Gender + pspline(History.Diagnosis.AgeAtYears.MedCent, 3) + Path.Grade.Coarse +
## loglik Chisq Df P(>|Chi|)
## 1 -68.3
## 2 -59.2 18.2 6 0.0058

```

6.3 Prediction on 10-fold CV

```

cv_preds = readRDS("../analysis/14_SIS_NMF_CV_results.rds")

```

```

apply(cv_preds[, colnames(xlin.diag_dsd.sel)], 1, function(p1) coxph(y.diag_dsd ~
  p1))

## $bs.best
## Call:
## coxph(formula = y.diag_dsd ~ p1)
##
##
##      coef exp(coef) se(coef)      z      p
## p1 0.244      1.28    0.199 1.23 0.22
##
## Likelihood ratio test=1.5 on 1 df, p=0.22 n= 110, number of events= 70
##
## $bs.average
## Call:
## coxph(formula = y.diag_dsd ~ p1)
##
##
##      coef exp(coef) se(coef)      z      p
## p1 0.43      1.54    0.18 2.39 0.017
##
## Likelihood ratio test=5.54 on 1 df, p=0.0186 n= 110, number of events= 70
##
## $lasso.1se
## Call:
## coxph(formula = y.diag_dsd ~ p1)
##
##
##      coef exp(coef) se(coef)      z      p
## p1 0.606      1.83    0.2 3.03 0.0024
##

```

```
## Likelihood ratio test=8.91 on 1 df, p=0.00284 n= 110, number of events= 70
##
## $lasso.min
## Call:
## coxph(formula = y.diag_dsd ~ p1)
##
##
##      coef exp(coef) se(coef)      z p
## p1 -0.000988    0.999    0.189 -0.00522 1
##
## Likelihood ratio test=0 on 1 df, p=0.996 n= 110, number of events= 70
##
## $adalasso.1se
## Call:
## coxph(formula = y.diag_dsd ~ p1)
##
##
##      coef exp(coef) se(coef)      z p
## p1 0.109    1.12    0.219 0.499 0.62
##
## Likelihood ratio test=0.25 on 1 df, p=0.616 n= 110, number of events= 70
##
## $adalasso.min
## Call:
## coxph(formula = y.diag_dsd ~ p1)
##
##
##      coef exp(coef) se(coef)      z p
## p1 0.351    1.42    0.19 1.84 0.065
##
## Likelihood ratio test=3.34 on 1 df, p=0.0676 n= 110, number of events= 70
```

```
# fit.cpusig.withoutsmoking.cv_BSA =
# coxph(Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath) ~
# Patient.Gender + pspline(History.Diagnosis.AgeAtYears.MedCent, 3) +
# Path.Grade.Coarse + Path.TumourLocation.Coarse +
# pspline(Path.TumourSizeMm.MedCent, 3) + Path.Invasion.PN +
# Path.Invasion.VS + Path.Nodes.Regional.Involved.Fraction + Staging.pN +
# Treat.Surgery.ExcisionStatus.Coarse + cv_preds['bs.average',], data =
# cpvs.diag_dsd)
fit.cpv sig.withoutsmoking.cv_L1SE = coxph(Surv(Surv.EventTimeFromDiag.DSDeath,
Surv.Event.DSDeath) ~ Patient.Gender + pspline(History.Diagnosis.AgeAtYears.MedCent,
3) + Path.Grade.Coarse + Path.TumourLocation.Coarse + pspline(Path.TumourSizeMm.MedCent,
3) + Path.Invasion.PN + Path.Invasion.VS + Path.Nodes.Regional.Involved.Fraction +
Staging.pN + Treat.Surgery.ExcisionStatus.Coarse + metapcna.scores[colnames(cv_preds)] +
cv_preds["lasso.1se", ], data = cpvs.diag_dsd)
# anova(fit.cpv.withoutsmoking, fit.cpusig.withoutsmoking.cv_BSA)
# anova(fit.cpv.withoutsmoking, fit.cpusig.withoutsmoking.cv_L1SE)
anova(fit.cpv sig.withoutsmoking.cv_L1SE)

FALSE Analysis of Deviance Table
FALSE Cox model: response is Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath)
FALSE Terms added sequentially (first to last)
```

```
FALSE
FALSE                                loglik Chisq Df
FALSE NULL                          -270
FALSE Patient.Gender                 -269  1.05  1
FALSE pspline(History.Diagnosis.AgeAtYears.MedCent, 3) -262 15.63 10
FALSE Path.Grade.Coarse               -259  4.82  1
FALSE Path.TumourLocation.Coarse      -258  2.53  1
FALSE pspline(Path.TumourSizeMm.MedCent, 3) -253 10.06 10
FALSE Path.Invasion.PN                -252  0.91  1
FALSE Path.Invasion.VS                -252  1.58  1
FALSE Path.Nodes.Regional.Involved.Fraction -252  0.09  1
FALSE Staging.pN                      -252  0.13  1
FALSE Treat.Surgery.ExcisionStatus.Coarse -247  8.27  1
FALSE metapcna.scores[colnames(cv_preds)] -246  3.44  1
FALSE cv_preds["lasso.1se", ]         -244  3.68  1
FALSE                                Pr(>|Chi|)
FALSE NULL
FALSE Patient.Gender                 0.305
FALSE pspline(History.Diagnosis.AgeAtYears.MedCent, 3) 0.111
FALSE Path.Grade.Coarse               0.028
FALSE Path.TumourLocation.Coarse      0.112
FALSE pspline(Path.TumourSizeMm.MedCent, 3) 0.435
FALSE Path.Invasion.PN                0.339
FALSE Path.Invasion.VS                0.208
FALSE Path.Nodes.Regional.Involved.Fraction 0.763
FALSE Staging.pN                      0.722
FALSE Treat.Surgery.ExcisionStatus.Coarse 0.004
FALSE metapcna.scores[colnames(cv_preds)] 0.064
FALSE cv_preds["lasso.1se", ]         0.055
```

```
fit.cpv.sig.withsmoking.cv_L1SE = coxph(Surv(Surv.EventTimeFromDiag.DSDeath,
Surv.Event.DSDeath) ~ Patient.Gender + pspline(History.Diagnosis.AgeAtYears.MedCent,
3) + Path.Grade.Coarse + Path.TumourLocation.Coarse + pspline(Path.TumourSizeMm.MedCent,
3) + Path.Invasion.PN + Path.Invasion.VS + Path.Nodes.Regional.Involved.Fraction +
Staging.pN + History.Smoking.PackYears + Treat.Surgery.ExcisionStatus.Coarse +
metapcna.scores[colnames(cv_preds)] + cv_preds["lasso.1se", ], data = cpvs.diag_dsd)
# anova(fit.cpv.withoutsmoking, fit.cpv.sig.withoutsmoking.cv_BSA)
# anova(fit.cpv.withsmoking, fit.cpv.sig.withsmoking.cv_L1SE)
anova(fit.cpv.sig.withsmoking.cv_L1SE)
```

```
FALSE Analysis of Deviance Table
FALSE Cox model: response is Surv(Surv.EventTimeFromDiag.DSDeath, Surv.Event.DSDeath)
FALSE Terms added sequentially (first to last)
FALSE
FALSE                                loglik Chisq Df
FALSE NULL                          -93.6
FALSE Patient.Gender                 -93.6  0.09  1
FALSE pspline(History.Diagnosis.AgeAtYears.MedCent, 3) -86.5 14.18 10
FALSE Path.Grade.Coarse               -84.8  3.26  1
FALSE Path.TumourLocation.Coarse      -84.5  0.77  1
FALSE pspline(Path.TumourSizeMm.MedCent, 3) -82.3  4.34 10
FALSE Path.Invasion.PN                -79.5  5.64  1
FALSE Path.Invasion.VS                -78.6  1.81  1
```

FALSE Path.Nodes.Regional.Involved.Fraction	-78.4	0.24	1
FALSE Staging.pN	-76.5	3.97	1
FALSE History.Smoking.PackYears	-72.7	7.60	1
FALSE Treat.Surgery.ExcisionStatus.Coarse	-68.3	8.77	1
FALSE metapcna.scores[colnames(cv_preds)]	-64.3	7.97	1
FALSE cv_preds["lasso.1se",]	-62.1	4.45	1
FALSE	Pr(> Chi)		
FALSE NULL			
FALSE Patient.Gender	0.7607		
FALSE pspline(History.Diagnosis.AgeAtYears.MedCent, 3)	0.1651		
FALSE Path.Grade.Coarse	0.0710		
FALSE Path.TumourLocation.Coarse	0.3803		
FALSE pspline(Path.TumourSizeMm.MedCent, 3)	0.9309		
FALSE Path.Invasion.PN	0.0176		
FALSE Path.Invasion.VS	0.1782		
FALSE Path.Nodes.Regional.Involved.Fraction	0.6207		
FALSE Staging.pN	0.0463		
FALSE History.Smoking.PackYears	0.0058		
FALSE Treat.Surgery.ExcisionStatus.Coarse	0.0031		
FALSE metapcna.scores[colnames(cv_preds)]	0.0048		
FALSE cv_preds["lasso.1se",]	0.0349		

6.4 Prediction on validation sets

```
load("../data/15_validation.rda")
```

```
val.basis = basis(nmf.final)
rownames(GSE21501.lingex) = GSE21501.feats$Gene.symbol
rownames(GSE28735.lingex) = GSE28735.feats$Gene.symbol
GSE21501.lingex.for_basis = GSE21501.lingex[match(rownames(val.basis), rownames(GSE21501.lingex)),
]
GSE28735.lingex.for_basis = GSE28735.lingex[match(rownames(val.basis), rownames(GSE28735.lingex)),
]
GSE21501.lingex.for_basis[is.na(GSE21501.lingex.for_basis)] = 0
GSE28735.lingex.for_basis[is.na(GSE28735.lingex.for_basis)] = 0

GSE21501.coefs = apply(GSE21501.lingex.for_basis, 2, function(xcol) nnls(val.basis,
xcol)$x)
GSE28735.coefs = apply(GSE28735.lingex.for_basis, 2, function(xcol) nnls(val.basis,
xcol)$x)
```

```
apply(GSE21501.coefs, 1, function(xc) coxph(Surv(time, event) ~ xc, data = GSE21501.samp))

## [[1]]
## Call:
## coxph(formula = Surv(time, event) ~ xc, data = GSE21501.samp)
##
##
##      coef exp(coef) se(coef)      z      p
## xc -5.31   0.00495    3.44 -1.54 0.12
```



```
##
## Likelihood ratio test=2.57 on 1 df, p=0.109 n= 102, number of events= 66
##
## [[2]]
## Call:
## coxph(formula = Surv(time, event) ~ xc, data = GSE21501.samp)
##
##
##      coef exp(coef) se(coef)      z    p
## xc 2.17      8.79      2.6 0.836 0.4
##
## Likelihood ratio test=0.68 on 1 df, p=0.41 n= 102, number of events= 66
##
## [[3]]
## Call:
## coxph(formula = Surv(time, event) ~ xc, data = GSE21501.samp)
##
##
##      coef exp(coef) se(coef)      z    p
## xc -1.09      0.337      2.73 -0.399 0.69
##
## Likelihood ratio test=0.16 on 1 df, p=0.688 n= 102, number of events= 66
##
## [[4]]
## Call:
## coxph(formula = Surv(time, event) ~ xc, data = GSE21501.samp)
##
##
##      coef exp(coef) se(coef)      z    p
## xc -0.226      0.798      2.47 -0.0914 0.93
##
## Likelihood ratio test=0.01 on 1 df, p=0.927 n= 102, number of events= 66
##
## [[5]]
## Call:
## coxph(formula = Surv(time, event) ~ xc, data = GSE21501.samp)
##
##
##      coef exp(coef) se(coef)      z    p
## xc 2.17      8.75      3.04 0.713 0.48
##
## Likelihood ratio test=0.49 on 1 df, p=0.486 n= 102, number of events= 66
##
## [[6]]
## Call:
## coxph(formula = Surv(time, event) ~ xc, data = GSE21501.samp)
##
##
##      coef exp(coef) se(coef)      z    p
## xc 2.57      13.1      2.08 1.24 0.22
##
## Likelihood ratio test=1.45 on 1 df, p=0.229 n= 102, number of events= 66
##
apply(GSE21501.coefs, 1, function(xc) coxph(Surv(time, event) ~ tstage + nstage +
```

```

xc, data = GSE21501.samp))

## [[1]]
## Call:
## coxph(formula = Surv(time, event) ~ tstage + nstage + xc, data = GSE21501.samp)
##
##
##          coef exp(coef) se(coef)      z      p
## tstage -0.166   0.84664   0.281 -0.593 0.550
## nstage  0.653   1.92135   0.315  2.074 0.038
## xc      -5.237   0.00532   3.545 -1.477 0.140
##
## Likelihood ratio test=7.13 on 3 df, p=0.0678 n= 97, number of events= 63
## (5 observations deleted due to missingness)
##
## [[2]]
## Call:
## coxph(formula = Surv(time, event) ~ tstage + nstage + xc, data = GSE21501.samp)
##
##
##          coef exp(coef) se(coef)      z      p
## tstage -0.161   0.851   0.287 -0.560 0.580
## nstage  0.643   1.903   0.316  2.039 0.041
## xc       1.025   2.788   2.658  0.386 0.700
##
## Likelihood ratio test=4.94 on 3 df, p=0.176 n= 97, number of events= 63
## (5 observations deleted due to missingness)
##
## [[3]]
## Call:
## coxph(formula = Surv(time, event) ~ tstage + nstage + xc, data = GSE21501.samp)
##
##
##          coef exp(coef) se(coef)      z      p
## tstage -0.153   0.858   0.286 -0.535 0.590
## nstage  0.659   1.933   0.314  2.097 0.036
## xc      -1.379   0.252   2.796 -0.493 0.620
##
## Likelihood ratio test=5.04 on 3 df, p=0.169 n= 97, number of events= 63
## (5 observations deleted due to missingness)
##
## [[4]]
## Call:
## coxph(formula = Surv(time, event) ~ tstage + nstage + xc, data = GSE21501.samp)
##
##
##          coef exp(coef) se(coef)      z      p
## tstage -0.158   0.854   0.285 -0.554 0.580
## nstage  0.678   1.970   0.319  2.126 0.034
## xc      -1.041   0.353   2.463 -0.423 0.670
##
## Likelihood ratio test=4.98 on 3 df, p=0.173 n= 97, number of events= 63
## (5 observations deleted due to missingness)
##

```

```
## [[5]]
## Call:
## coxph(formula = Surv(time, event) ~ tstage + nstage + xc, data = GSE21501.samp)
##
##              coef exp(coef) se(coef)      z      p
## tstage -0.179      0.836    0.289 -0.621 0.530
## nstage  0.641      1.898    0.315  2.033 0.042
## xc      1.878      6.543    3.313  0.567 0.570
##
## Likelihood ratio test=5.11 on 3 df, p=0.164 n= 97, number of events= 63
## (5 observations deleted due to missingness)
##
## [[6]]
## Call:
## coxph(formula = Surv(time, event) ~ tstage + nstage + xc, data = GSE21501.samp)
##
##              coef exp(coef) se(coef)      z      p
## tstage -0.110      0.896    0.284 -0.388 0.700
## nstage  0.657      1.928    0.316  2.077 0.038
## xc      2.510     12.310    2.164  1.160 0.250
##
## Likelihood ratio test=6.07 on 3 df, p=0.108 n= 97, number of events= 63
## (5 observations deleted due to missingness)

apply(GSE21501.coefs, 1, function(xc) anova(coxph(Surv(time, event) ~ tstage +
  nstage + xc, data = GSE21501.samp)))

## [[1]]
## Analysis of Deviance Table
## Cox model: response is Surv(time, event)
## Terms added sequentially (first to last)
##
##      loglik Chisq Df Pr(>|Chi|)
## NULL      -242
## tstage   -242  0.01  1      0.928
## nstage   -239  4.79  1      0.029
## xc       -238  2.34  1      0.126
##
## [[2]]
## Analysis of Deviance Table
## Cox model: response is Surv(time, event)
## Terms added sequentially (first to last)
##
##      loglik Chisq Df Pr(>|Chi|)
## NULL      -242
## tstage   -242  0.01  1      0.928
## nstage   -239  4.79  1      0.029
## xc       -239  0.15  1      0.702
##
## [[3]]
## Analysis of Deviance Table
## Cox model: response is Surv(time, event)
```

```

## Terms added sequentially (first to last)
##
##      loglik Chisq Df Pr(>|Chi|)
## NULL      -242
## tstage    -242  0.01  1      0.928
## nstage    -239  4.79  1      0.029
## xc        -239  0.25  1      0.619
##
## [[4]]
## Analysis of Deviance Table
## Cox model: response is Surv(time, event)
## Terms added sequentially (first to last)
##
##      loglik Chisq Df Pr(>|Chi|)
## NULL      -242
## tstage    -242  0.01  1      0.928
## nstage    -239  4.79  1      0.029
## xc        -239  0.18  1      0.668
##
## [[5]]
## Analysis of Deviance Table
## Cox model: response is Surv(time, event)
## Terms added sequentially (first to last)
##
##      loglik Chisq Df Pr(>|Chi|)
## NULL      -242
## tstage    -242  0.01  1      0.928
## nstage    -239  4.79  1      0.029
## xc        -239  0.31  1      0.577
##
## [[6]]
## Analysis of Deviance Table
## Cox model: response is Surv(time, event)
## Terms added sequentially (first to last)
##
##      loglik Chisq Df Pr(>|Chi|)
## NULL      -242
## tstage    -242  0.01  1      0.928
## nstage    -239  4.79  1      0.029
## xc        -239  1.27  1      0.259

apply(GSE28735.coefs, 1, function(xc) coxph(Surv(time, event) ~ xc, data = GSE28735.samp))

## [[1]]
## Call:
## coxph(formula = Surv(time, event) ~ xc, data = GSE28735.samp)
##
##
##      coef exp(coef) se(coef)      z      p
## xc -4.5      0.0111      3.74 -1.2 0.23
##
## Likelihood ratio test=1.55 on 1 df, p=0.213 n= 42, number of events= 29
##
## [[2]]

```

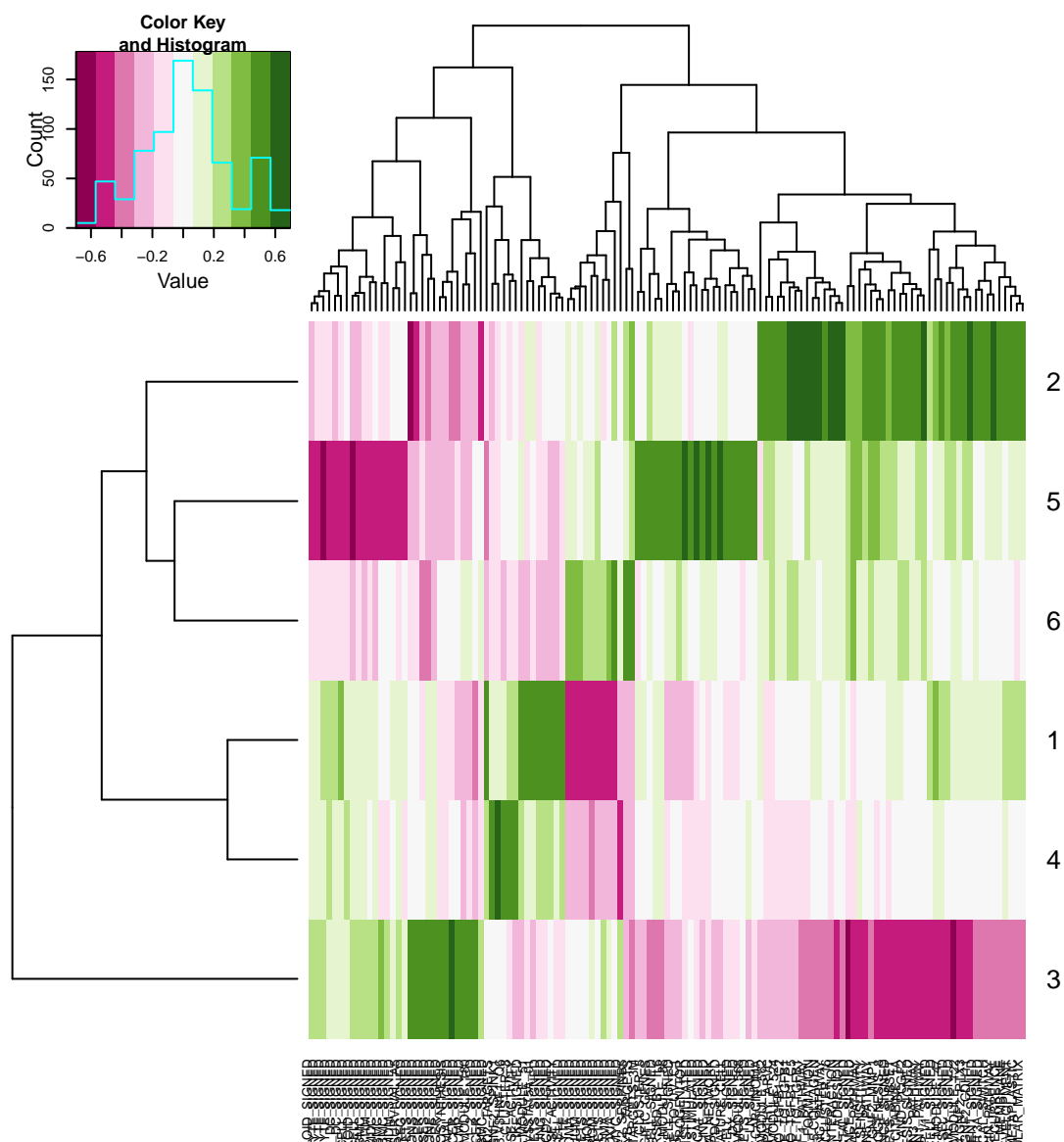
```

## Call:
## coxph(formula = Surv(time, event) ~ xc, data = GSE28735.samp)
##
##
##      coef exp(coef) se(coef)      z      p
## xc 3.46      31.7      2.53 1.36 0.17
##
## Likelihood ratio test=1.63 on 1 df, p=0.201 n= 42, number of events= 29
##
## [[3]]
## Call:
## coxph(formula = Surv(time, event) ~ xc, data = GSE28735.samp)
##
##
##      coef exp(coef) se(coef)      z      p
## xc -7.79  0.000415      3.22 -2.42 0.015
##
## Likelihood ratio test=6.32 on 1 df, p=0.0119 n= 42, number of events= 29
##
## [[4]]
## Call:
## coxph(formula = Surv(time, event) ~ xc, data = GSE28735.samp)
##
##
##      coef exp(coef) se(coef)      z      p
## xc 0.72      2.05      2.81 0.256 0.8
##
## Likelihood ratio test=0.06 on 1 df, p=0.801 n= 42, number of events= 29
##
## [[5]]
## Call:
## coxph(formula = Surv(time, event) ~ xc, data = GSE28735.samp)
##
##
##      coef exp(coef) se(coef)      z      p
## xc -0.469      0.625      2.61 -0.18 0.86
##
## Likelihood ratio test=0.03 on 1 df, p=0.857 n= 42, number of events= 29
##
## [[6]]
## Call:
## coxph(formula = Surv(time, event) ~ xc, data = GSE28735.samp)
##
##
##      coef exp(coef) se(coef)      z      p
## xc 5.73      308      2.32 2.47 0.014
##
## Likelihood ratio test=5.68 on 1 df, p=0.0171 n= 42, number of events= 29

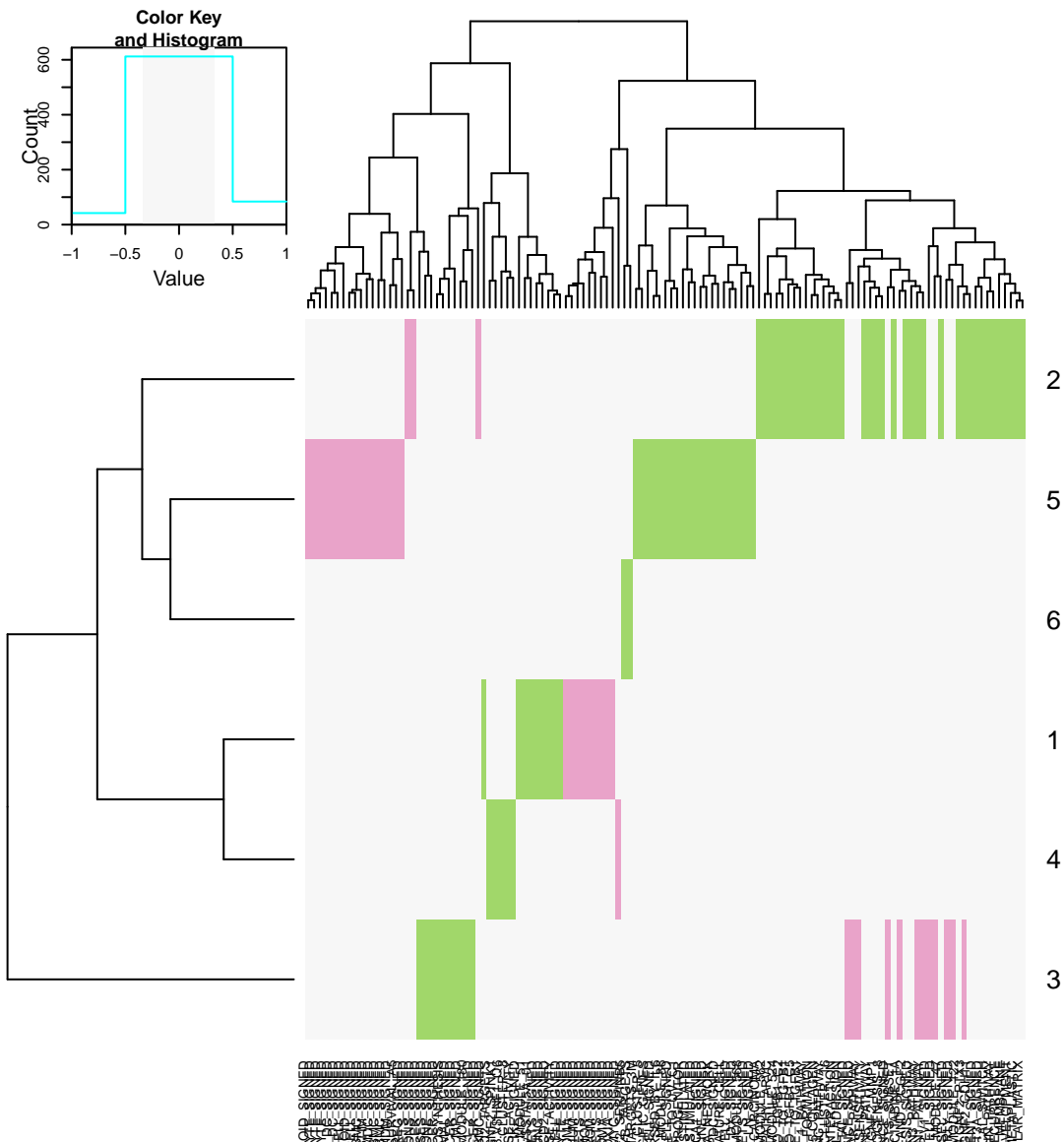
```

6.5 MSigDB score correlation thresholding

```
temp.sel_cols = apply(abs(nmf.final.msigdb.corr) >= sig.corr.threshold, 2, any)
heatmap.2(nmf.final.msigdb.corr[, temp.sel_cols], trace = "none", scale = "none",
  useRaster = TRUE, col = brewer.pal(11, "PiYG"), symbreaks = TRUE)
```



```
heatmap.2(nmf.final.msigdb.corr[, temp.sel_cols], trace = "none", scale = "none",
  useRaster = TRUE, col = brewer.pal(3, "PiYG"), breaks = c(-1, -sig.corr.threshold,
  sig.corr.threshold, 1))
```



```
temp.sig_id = colnames(nmf.final.msigdb.corr)
temp.sig_class = gsub("\\\\.*", "", temp.sig_id)
temp.nsigs = length(temp.sig_id)
temp.nmeta = nrow(nmf.final.msigdb.corr)
tables = lapply(1:temp.nmeta, function(metagene_i) {
  tapply(1:temp.nsigs, temp.sig_class, function(sig_class_is) {
    all_cors = nmf.final.msigdb.corr[, sig_class_is]
    this_cors = all_cors[metagene_i, ]
    this_ids = temp.sig_id[sig_class_is]

    all_sig_cors = abs(all_cors) >= sig.corr.threshold
    this_sig_cors = all_sig_cors[metagene_i, ]

    sigs_to_report = which(this_sig_cors)

    if (length(sigs_to_report) == 0) {
      table = data.frame(GeneSet = c(), Correlation = c(), Metagenes = c())
    }
  })
})
```

```

    } else {
      table = data.frame(GeneSet = this_ids[sigs_to_report], Correlation = this_cors[sigs_to_report],
        Metagenes = apply(all_cors[, sigs_to_report, drop = FALSE],
          2, function(cors) {
            sel = abs(cors) >= sig.corr.threshold
            # A positive number implies that positive GSVA signal is associated with
            # worse prognosis
            paste(which(sel) * sign(cors[which(sel)]) * sign(nmf.final.cpv.pvals$surv.diag_dsd),
              collapse = ",")
          })
      table = table[order(-(table$Correlation)), ]
      rownames(table) <- NULL
    }
    table
  }, simplify = FALSE)
})
tables

## [[1]]
## [[1]]$c1
## data frame with 0 columns and 0 rows
##
## [[1]]$c2
##
##          GeneSet Correlation Metagenes
## 1      c2.KATSANOUELAVL1_TARGETS_SIGNED      0.5096      -1
## 2      c2.ZHAN_MULTIPLE_MYELOMA_CD2_SIGNED      0.5086      -1
## 3      c2.SMID_BREAST_CANCER_NORMAL_LIKE_SIGNED      0.5080      -1
## 4      c2.GREENBAUM_E2A_TARGETS_SIGNED     -0.5009       1
## 5 c2.MARTORIATI_MDM4_TARGETS_NEUROEPITHELIUM_SIGNED     -0.5012       1
## 6      c2.YU_MYC_TARGETS_SIGNED     -0.5029       1
## 7      c2.SABATES_COLORECTAL_ADENOMA_SIGNED     -0.5036       1
## 8      c2.WINTER_HYPOXIA_SIGNED     -0.5241       1
## 9      c2.WEST_ADRENOCORTICAL_TUMOR_SIGNED     -0.5416       1
## 10     c2.SHIPP_DLBCL_VS_FOLLICULAR_LYMPHOMA_SIGNED     -0.5500       1
## 11     c2.LEE_EARLY_T_LYMPHOCYTE_SIGNED     -0.5517       1
## 12     c2.HAHTOLA_SEZARY_SYNDROM_SIGNED     -0.5641       1
##
## [[1]]$c3
##          GeneSet Correlation Metagenes
## 1 c3.V$STAT5A_01      0.5234      -1
##
## [[1]]$c4
##          GeneSet Correlation Metagenes
## 1 c4.MODULE_51      0.5399      -1
##
## [[1]]$c5
##          GeneSet Correlation Metagenes
## 1 c5.PHOSPHORIC_DIESTER_HYDROLASE_ACTIVITY      0.5113      -1
##
## [[1]]$c6
## data frame with 0 columns and 0 rows
##
## [[1]]$c7
##
##          GeneSet Correlation

```



```

## 1      c7.GSE20715_OH_VS_48H_OZONE_TLR4_KO_LUNG_SIGNED      0.5160
## 2 c7.GSE22886_IGM_MEMORY_BCELL_VS_BLOOD_PLASMA_CELL_SIGNED      0.5019
## 3      c7.GSE34205_HEALTHY_VS_RSV_INF_INFANT_PBMCSIGNED      0.5002
## Metagenes
## 1      -1
## 2      -1
## 3      -1
##
##
## [[2]]
## [[2]]$c1
## data frame with 0 columns and 0 rows
##
## [[2]]$c2
##
## 1      c2.REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION/c2.REACTOME_COLLAGEN_FORMA
## 2      c2.PID_SYNDECAN_1_PAT
## 3      c2.VERRECCHIA_DELAYED_RESPONSE_TO_T
## 4      c2.PID_INTEGRIN1_PAT
## 5      c2.PID_AVB3_INTEGRIN_PAT
## 6 c2.FARMER_BREAST_CANCER_CLUSTER_5/c2.ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE/c4.GNF2_
## 7      c2.YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER
## 8      c2.KEGG_ECM_RECEPTOR_INTERAC
## 9      c2.VERRECCHIA_RESPONSE_TO_TGFB
## 10     c2.VERRECCHIA_EARLY_RESPONSE_TO_T
## 11     c2.KEGG_FOCAL_ADHE
## 12     c2.MAHADEVAN_GIST_MORPHOLOGICAL_SW
## 13     c2.CAIRO_LIVER_DEVELOPMENT_SI
## 14     c2.PID_INTEGRIN3_PAT
## 15     c2.KEGG_BASAL_CELL_CARCI
## 16     c2.BURTON_ADIPOGENES
## 17     c2.VERRECCHIA_RESPONSE_TO_TGFB
## 18     c2.CROMER_TUMORIGENESIS_SI
## 19     c2.WIEDERSCHAIN_TARGETS_OF_BMI1_AND_F
## 20     c2.ROZANOV_MMP14_TARGETS_SU
## 21     c2.PID_WNT_SIGNALING_PAT
## 22     c2.KEGG_ARRHYTHMOGENIC_RIGHT_VENTRICULAR_CARDIOMYOPATHY
## 23     c2.LABBE_TARGETS_OF_TGFB1_AND_WNT3A_SI
## 24     c2.LIEN_BREAST_CARCINOMA_METAPLA
## 25     c2.PID_INTEGRIN5_PAT
## 26     c2.LINDGREN_BLADDER_CANCER_HIGH_RECURREN
## 27     c2.POTTI_TOPOTECAN_SENSITI
## 28     c2.MIYAGAWA_TARGETS_OF_EWSR1_ETS_FUSIONS_SI
## 29     c2.BERTUCCI_MEDULLARY_VS_DUCTAL_BREAST_CANCER_SI
## 30     c2.PASINI_SUZ12_TARGETS_SI
## Correlation Metagenes
## 1      0.6490      2
## 2      0.6355      2
## 3      0.6178      2
## 4      0.6067      2,-3
## 5      0.6020      2
## 6      0.5990      2
## 7      0.5963      2
## 8      0.5953      2

```

```

## 9      0.5849      2
## 10     0.5829      2
## 11     0.5758      2
## 12     0.5587      2
## 13     0.5429      2
## 14     0.5409      2,-3
## 15     0.5396      2
## 16     0.5346      2
## 17     0.5312      2
## 18     0.5258      2
## 19     0.5242      2
## 20     0.5228      2
## 21     0.5171      2
## 22     0.5114      2
## 23     0.5081      2
## 24     0.5077      2
## 25     0.5074      2
## 26     0.5047      2
## 27     0.5017      2
## 28    -0.5087     -2
## 29    -0.5436     -2
## 30    -0.5916     -2
##
## [[2]]$c3
## data frame with 0 columns and 0 rows
##
## [[2]]$c4
##
##      GeneSet Correlation Metagenes
## 1      c4.GNF2_PTX3      0.5533      2,-3
## 2      c4.MODULE_122      0.5369      2
## 3      c4.GNF2_MMP1      0.5366      2
## 4      c4.MODULE_562      0.5178      2
## 5 c4.MODULE_419/c4.MODULE_524      0.5128      2
## 6      c4.MODULE_47      0.5003      2
##
## [[2]]$c5
##
##      GeneSet
## 1      c5.COLLAGEN
## 2 c5.PROTEINACEOUS_EXTRACELLULAR_MATRIX/c5.EXTRACELLULAR_MATRIX_PART/c5.EXTRACELLULAR_MATRIX
## 3      c5.BASEMENT_MEMBRANE
## 4      c5.SKELETAL_DEVELOPMENT
##      Correlation Metagenes
## 1      0.6496      2
## 2      0.5336      2
## 3      0.5148      2
## 4      0.5101      2
##
## [[2]]$c6
## data frame with 0 columns and 0 rows
##
## [[2]]$c7
## data frame with 0 columns and 0 rows
##
##

```

```
## [[3]]
## [[3]]$c1
## data frame with 0 columns and 0 rows
##
## [[3]]$c2
##                                     GeneSet Correlation
## 1  c2.CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_SIGNED      0.5882
## 2                c2.REACTOME_GLYCEROPHOSPHOLIPID_BIOSYNTHESIS    0.5269
## 3                c2.SMID_BREAST_CANCER_RELAPSE_IN_BONE_SIGNED    0.5215
## 4                        c2.LIU_PROSTATE_CANCER_SIGNED           0.5202
## 5                c2.WAMUNYOKOLI_OVARIAN_CANCER_GRADES_1_2_SIGNED  0.5178
## 6  c2.WANG_BARRETTS_ESOPHAGUS_AND_ESOPHAGUS_CANCER_SIGNED      0.5175
## 7                c2.WAMUNYOKOLI_OVARIAN_CANCER_LMP_SIGNED        0.5165
## 8                        c2.WALLACE_PROSTATE_CANCER_SIGNED       0.5155
## 9                c2.DOANE_BREAST_CANCER_CLASSES_SIGNED          0.5111
## 10               c2.WOO_LIVER_CANCER_RECURRENCE_SIGNED          -0.5000
## 11                        c2.PID_UPA_UPAR_PATHWAY               -0.5011
## 12               c2.SUZUKI_RESPONSE_TO_TSA_AND_DECITABINE_1A    -0.5141
## 13               c2.HUANG_DASATINIB_RESISTANCE_SIGNED           -0.5145
## 14                        c2.LIM_MAMMARY_STEM_CELL_SIGNED       -0.5175
## 15                        c2.PID_INTEGRIN3_PATHWAY               -0.5175
## 16               c2.ROY_WOUND_BLOOD_VESSEL_SIGNED              -0.5235
## 17                        c2.PID_INTEGRIN1_PATHWAY              -0.5248
## 18  c2.LIEN_BREAST_CARCINOMA_METAPLASTIC_VS_DUCTAL_SIGNED      -0.6110
## 19  c2.VECCHI_GASTRIC_CANCER_ADVANCED_VS_EARLY_SIGNED          -0.6217
##      Metagenes
## 1          -3
## 2          -3
## 3          -3
## 4          -3
## 5          -3
## 6          -3
## 7          -3
## 8          -3
## 9          -3
## 10         3
## 11         3
## 12         3
## 13         3
## 14         3
## 15        -2,3
## 16         3
## 17        -2,3
## 18         3
## 19         3
##
## [[3]]$c3
## data frame with 0 columns and 0 rows
##
## [[3]]$c4
##                                     GeneSet Correlation Metagenes
## 1  c4.MODULE_139/c4.MODULE_180      0.5195          -3
## 2                c4.GNF2_PTX3      -0.5155        -2,3
```

```

##
## [[3]]$c5
## data frame with 0 columns and 0 rows
##
## [[3]]$c6
##           GeneSet Correlation Metagenes
## 1 c6.LEF1_UP.V1_SIGNED      -0.5597      3
##
## [[3]]$c7
## data frame with 0 columns and 0 rows
##
##
## [[4]]
## [[4]]$c1
## data frame with 0 columns and 0 rows
##
## [[4]]$c2
##           GeneSet Correlation Metagenes
## 1          c2.BERGER_MBD2_TARGETS      0.5646      -4
## 2 c2.TERAMOTO_OPN_TARGETS_CLUSTER_8      0.5274      -4
## 3      c2.LEE_LIVER_CANCER_MYC_SIGNED     -0.5203      4
##
## [[4]]$c3
##           GeneSet Correlation Metagenes
## 1 c3.V$HNF1_Q6      0.5124      -4
##
## [[4]]$c4
##           GeneSet Correlation Metagenes
## 1 c4.GNF2_SERPINI2/c4.GNF2_SPINK1      0.6959      -4
##
## [[4]]$c5
##           GeneSet Correlation Metagenes
## 1 c5.CARBOXYPEPTIDASE_ACTIVITY      0.5342      -4
##
## [[4]]$c6
## data frame with 0 columns and 0 rows
##
## [[4]]$c7
## data frame with 0 columns and 0 rows
##
##
## [[5]]
## [[5]]$c1
## data frame with 0 columns and 0 rows
##
## [[5]]$c2
##           GeneSet
## 1          c2.IVANOVA_HEMATOPOIESIS_LATE_PROGENITOR
## 2          c2.MARSON_BOUND_BY_FOXP3_STIMULATED
## 3          c2.SESTO_RESPONSE_TO_UV_C1
## 4          c2.NAKAYAMA_SOFT_TISSUE_TUMORS_PCA1_SIGNED
## 5          c2.IVANOVA_HEMATOPOIESIS_MATURE_CELL
## 6 c2.MARTINEZ_RB1_AND_TP53_TARGETS_SIGNED/c2.MARTINEZ_TP53_TARGETS_SIGNED
## 7          c2.KAMIKUBO_MYELOID_CEBPA_NETWORK

```

```

## 8 c2.MARSON_BOUND_BY_FOXP3_UNSTIMULATED
## 9 c2.VALK_AML_CLUSTER_5
## 10 c2.RAGHAVACHARI_PLATELET_SPECIFIC_GENES
## 11 c2.LIAN_LIPA_TARGETS_6M/c2.LIAN_LIPA_TARGETS_3M
## 12 c2.BROCKE_APOPTOSIS_REVERSED_BY_IL6
## 13 c2.SHEDDEN_LUNG_CANCER_GOOD_SURVIVAL_A5
## Correlation Metagenes
## 1 0.6114 5
## 2 0.5798 5
## 3 0.5491 5
## 4 0.5413 5
## 5 0.5410 5
## 6 0.5304 5
## 7 0.5280 5
## 8 0.5154 5
## 9 0.5134 5
## 10 0.5124 5
## 11 0.5005 5
## 12 0.5001 5
## 13 -0.5080 -5
##
## [[5]]$c3
## data frame with 0 columns and 0 rows
##
## [[5]]$c4
## GeneSet Correlation Metagenes
## 1 c4.MODULE_86 0.5240 5
## 2 c4.MODULE_491/c4.MODULE_568 0.5063 5
##
## [[5]]$c5
## data frame with 0 columns and 0 rows
##
## [[5]]$c6
## data frame with 0 columns and 0 rows
##
## [[5]]$c7
## GeneSet
## 1 c7.GSE29618_MONOCYTE_VS_PDC_SIGNED
## 2 c7.GSE29618_MONOCYTE_VS_PDC_DAY7_FLU_VACCINE_SIGNED
## 3 c7.GSE3982_MAC_VS_NKCELL_SIGNED
## 4 c7.GSE15767_MED_VS_SCS_MAC_LN_SIGNED
## 5 c7.GSE3982_BASOPHIL_VS_CENT_MEMORY_CD4_TCELL_SIGNED
## 6 c7.GSE11057_PBMC_VS_MEM_CD4_TCELL_SIGNED
## 7 c7.GSE9006_HEALTHY_VS_TYPE_1_DIABETES_PBMC_4MONTH_POST_DX_SIGNED
## 8 c7.GSE11057_NAIVE_CD4_VS_PBMC_CD4_TCELL_SIGNED
## 9 c7.GSE3982_DC_VS_MAC_LPS_STIM_SIGNED
## 10 c7.GSE17721_POLYIC_VS_GARDIQUIMOD_1H_BMDM_SIGNED
## 11 c7.GSE29618_PDC_VS_MDC_SIGNED
## 12 c7.GSE6269_HEALTHY_VS_STREP_AUREUS_INF_PBMC_SIGNED
## 13 c7.GSE20366_CD103_KLRG1_DP_VS_DN_TREG_SIGNED
## 14 c7.GSE22886_NAIVE_CD8_TCELL_VS_DC_SIGNED
## 15 c7.GSE22886_NAIVE_CD8_TCELL_VS_MONOCYTE_SIGNED
## 16 c7.GSE11057_CD4_CENT_MEM_VS_PBMC_SIGNED
## 17 c7.GSE22886_NAIVE_CD4_TCELL_VS_DC_SIGNED

```

```

## 18          c7.GSE11057_NAIVE_VS_MEMORY_CD4_TCELL_SIGNED
## 19          c7.GSE11057_CD4_EFF_MEM_VS_PBMV_SIGNED
## 20          c7.GSE10325_BCELL_VS_MYELOID_SIGNED
## 21          c7.GSE22886_NAIVE_TCELL_VS_MONOCYTE_SIGNED
## 22          c7.GSE10325_LUPUS_CD4_TCELL_VS_LUPUS_MYELOID_SIGNED
## 23          c7.GSE22886_NAIVE_CD4_TCELL_VS_MONOCYTE_SIGNED
##      Correlation Metagenes
## 1      0.5760      5
## 2      0.5712      5
## 3      0.5573      5
## 4      0.5502      5
## 5      0.5352      5
## 6      0.5314      5
## 7      0.5209      5
## 8     -0.5042     -5
## 9     -0.5042     -5
## 10     -0.5076     -5
## 11     -0.5086     -5
## 12     -0.5086     -5
## 13     -0.5233     -5
## 14     -0.5267     -5
## 15     -0.5274     -5
## 16     -0.5352     -5
## 17     -0.5355     -5
## 18     -0.5379     -5
## 19     -0.5420     -5
## 20     -0.5519     -5
## 21     -0.5610     -5
## 22     -0.5699     -5
## 23     -0.5825     -5
##
##
## [[6]]
## [[6]]$c1
## data frame with 0 columns and 0 rows
##
## [[6]]$c2
##      GeneSet Correlation Metagenes
## 1 c2.LEI_MYB_TARGETS      0.509      6
##
## [[6]]$c3
## data frame with 0 columns and 0 rows
##
## [[6]]$c4
##      GeneSet Correlation Metagenes
## 1 c4.GNF2_CDH3/c4.GNF2_SERPINB5      0.5532      6
##
## [[6]]$c5
## data frame with 0 columns and 0 rows
##
## [[6]]$c6
## data frame with 0 columns and 0 rows
##
## [[6]]$c7

```

```
## data frame with 0 columns and 0 rows
```

7 Session information

```
session_info

## R version 3.1.1 (2014-07-10)
## Platform: x86_64-unknown-linux-gnu (64-bit)
##
## locale:
##   [1] LC_CTYPE=en_AU.UTF-8      LC_NUMERIC=C
##   [3] LC_TIME=en_AU.UTF-8      LC_COLLATE=en_AU.UTF-8
##   [5] LC_MONETARY=en_AU.UTF-8  LC_MESSAGES=en_AU.UTF-8
##   [7] LC_PAPER=en_AU.UTF-8     LC_NAME=en_AU.UTF-8
##   [9] LC_ADDRESS=en_AU.UTF-8   LC_TELEPHONE=en_AU.UTF-8
##  [11] LC_MEASUREMENT=en_AU.UTF-8 LC_IDENTIFICATION=en_AU.UTF-8
##
## attached base packages:
## [1] splines    parallel  methods    stats      graphics  grDevices  utils
## [8] datasets  base
##
## other attached packages:
## [1] doParallel_1.0.8    iterators_1.0.7    foreach_1.4.2
## [4] ahaz_1.14           survival_2.37-7    NMF_0.20.5
## [7] Biobase_2.26.0      BiocGenerics_0.12.1 cluster_1.15.3
## [10] rngtools_1.2.4      pkgmaker_0.22      registry_0.2
## [13] energy_1.6.2        glmnet_1.9-8       Matrix_1.1-4
## [16] glmulti_1.0.7       rJava_0.9-6
##
## loaded via a namespace (and not attached):
## [1] boot_1.3-13         codetools_0.2-9    colorspace_1.2-4
## [4] compiler_3.1.1      digest_0.6.4       ggplot2_1.0.0
## [7] grid_3.1.1          gridBase_0.4-7     gtable_0.1.2
## [10] lattice_0.20-29     MASS_7.3-35        munsell_0.4.2
## [13] plyr_1.8.1          proto_0.3-10       RColorBrewer_1.0-5
## [16] Rcpp_0.11.3         reshape2_1.4       scales_0.2.4
## [19] stringr_0.6.2       tools_3.1.1        xtable_1.7-4

sessionInfo()

## R version 3.1.1 (2014-07-10)
## Platform: x86_64-unknown-linux-gnu (64-bit)
##
## locale:
##   [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
##   [3] LC_TIME=en_US.UTF-8      LC_COLLATE=en_US.UTF-8
##   [5] LC_MONETARY=en_US.UTF-8  LC_MESSAGES=en_US.UTF-8
##   [7] LC_PAPER=en_US.UTF-8     LC_NAME=en_US.UTF-8
##   [9] LC_ADDRESS=en_US.UTF-8   LC_TELEPHONE=en_US.UTF-8
##  [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=en_US.UTF-8
##
## attached base packages:
```

```
## [1] parallel  methods  splines  stats      graphics  grDevices  utils
## [8] datasets  base
##
## other attached packages:
## [1] stargazer_5.1      xtable_1.7-4      gplots_2.14.2
## [4] RColorBrewer_1.0-5  glmnet_1.9-8      Matrix_1.1-4
## [7] glmulti_1.0.7      rJava_0.9-6      nnls_1.4
## [10] NMF_0.20.5         synchronicity_1.1.4 bigmemory_4.4.6
## [13] BH_1.54.0-5        bigmemory.sri_0.1.3 Biobase_2.26.0
## [16] BiocGenerics_0.12.1 cluster_1.15.3    rngtools_1.2.4
## [19] pkgmaker_0.22      registry_0.2      energy_1.6.2
## [22] survival_2.37-7    knitr_1.8
##
## loaded via a namespace (and not attached):
## [1] bitops_1.0-6      boot_1.3-13      caTools_1.17.1
## [4] codetools_0.2-9   colorspace_1.2-4 digest_0.6.4
## [7] doParallel_1.0.8  evaluate_0.5.5   foreach_1.4.2
## [10] formatR_1.0       gdata_2.13.3     ggplot2_1.0.0
## [13] grid_3.1.1        gridBase_0.4-7   gtable_0.1.2
## [16] gtools_3.4.1      highr_0.4        iterators_1.0.7
## [19] KernSmooth_2.23-13 labeling_0.3      lattice_0.20-29
## [22] MASS_7.3-35       munsell_0.4.2    plyr_1.8.1
## [25] proto_0.3-10      Rcpp_0.11.3      reshape2_1.4
## [28] scales_0.2.4      stringr_0.6.2    tools_3.1.1
```