

Classification of Cancer Patients with Penalized Robust Nonconvex Loss Functions

Zhu Wang
UT Health San Antonio
wangz1@uthscsa.edu

May 18, 2019

This document presents analysis for the MAQC-II project, human breast cancer data set with penalized classification algorithms developed in Wang (2016) and implemented in R package `mpath`.

Dataset comes from the MicroArray Quality Control (MAQC) II project and includes 278 breast cancer samples with 164 estrogen receptor (ER) positive cases. The data files `GSE20194_series_matrix.txt.gz` and `GSE20194_MDACC_Sample_Info.xls` can be downloaded from <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=rhojvaihkcshq&acc=GSE20194>. After reading the data, some unused variables are removed. From 22283 genes, the dataset is pre-screened to obtain 3000 genes with the largest absolute values of the two-sample t-statistics. The 3000 genes are standardized.

```
# The data files below were downloaded on June 1, 2016
require("gdata")
library("mpath")
bc <- t(read.delim("GSE20194_series_matrix.txt.gz", sep = "",
  header = FALSE, skip = 80))
colnames(bc) <- bc[1, ]
bc <- bc[-1, -c(1, 2)]
### The last column is empty with variable name
### !series_matrix_table_end, thus omitted
bc <- bc[, -22284]
mode(bc) <- "numeric" ### convert character to numeric
dat1 <- read.xls("GSE20194_MDACC_Sample_Info.xls", sheet = 1,
  header = TRUE)
y <- dat1$characteristics..ER_status
y <- ifelse(y == "P", 1, -1)
table(y)
## y
## -1 1
## 114 164
res <- rep(NA, dim(bc)[2])
for (i in 1:dim(bc)[2]) res[i] <- abs(t.test(bc[, i] ~ y)$statistic)
### find 3000 largest absolute value of t-statistic
```

```

tmp <- order(res, decreasing = TRUE)[1:3000]
dat <- bc[, tmp]
### standardize variables
dat <- scale(dat)

```

Set up configuration parameters.

```

### number of replicates
nrun <- 100
### penalty type
penalty <- c("enet", "snet", "mnet")
### Smallest value for lambda, as a fraction of lambda.max, the
### smallest value for which all coefficients are zero except
### the intercept
ratio <- 0.25
type.path <- "nonactive"
nlam <- ifelse(type.path != "onestep", 30, 100)
### The training data is contaminated by randomly switching
### response variable labels at varying pre-specified
### proportions
per <- c(0, 0.05, 0.1, 0.15)
### what quantity is minimized for tuning parameter selection
tuning <- "error"
### robust nonconvex loss function, rfamily type and logistic
type <- c("closs", "gloss", "qloss", "binomial")
### and corresponding labels
type1 <- c("Closs", "Gloss", "Qloss", "Logistic")
### and corresponding tuning parameter
s <- c(0.9, 1.01, 0.5)
mstop <- 50
plot.it <- TRUE

```

The training data contains randomly selected 50 samples with positive estrogen receptor status and 50 samples with negative estrogen receptor status, and the rest were designated as the test data. The training data is contaminated by randomly switching response variable labels at varying pre-specified proportions `per`=0, 0.05, 0.1, 0.15. This process is repeated `nrun`=100 times. Robust non-convex loss functions include C-loss, G-loss and Q-loss, each with penalty LASSO, SCAD and MCP. The initial values are derived using the boosting package `bst` with `mstop`=50 and `nu` provided below depending on loss function `type`. For SCAD and MCP penalty, a penalty tuning parameter `gam` is provided below. To select optimal penalization tuning parameters, we run five-fold cross-validation averaging classification errors. The classification errors and number of selected variables are tabularized and plotted with `plot.it=TRUE`. The script uses `type.path="nonactive"`. Alternatively, setting `type.path="active"`, computing time can be reduced by more than 50 percent with slightly inferior results in some cases. The test errors based on cross-validation tuning parameters between `type.path="nonactive"` and `type.path="active"` have ranges (0, 0.0002), (-0.0007, 0.0002), (-0.0021, 0.0002) and (- 0.0035, -0.0003)

for `per=0, 0.05, 0.1, 0.15`, respectively. Finally, this script also contains results with penalized logistic regression using `glmreg`.

```
summary7 <- function(x) c(summary(x), sd = sd(x))
ptm <- proc.time()
for (k in (1:4)) {
  ### k controls family argument rfamily type (see above)
  if (type[k] == "gloss")
    nu <- 0.1 else nu <- 0.01
  for (j in (1:3)) {
    ### j controls argument penalty type (see above)
    gam <- ifelse(penalty[j] == "snet", 3.7, 12)
    err.m1 <- nvar.m1 <- errbest.m1 <- lambest.m1 <- matrix(NA,
      ncol = 4, nrow = nrun)
    nvarbest.m1 <- mstopcv.m1 <- matrix(NA, ncol = 4, nrow = nrun)
    colnames(err.m1) <- c("cont-0%", "cont-5%", "cont-10%",
      "cont-15%")
    colnames(mstopcv.m1) <- colnames(nvarbest.m1) <- colnames(err.m1)
    colnames(nvar.m1) <- colnames(err.m1)
    colnames(errbest.m1) <- colnames(err.m1)
    colnames(lambest.m1) <- colnames(err.m1)
    for (ii in 1:nrun) {
      set.seed(1000 + ii)
      trid <- c(sample(which(y == 1))[1:50], sample(which(y ==
        -1))[1:50])
      dtr <- dat[trid, ]
      dte <- dat[-trid, ]
      ytrold <- y[trid]
      yte <- y[-trid]
      ### number of patients/no. variables in training and test data
      dim(dtr)
      dim(dte)
      ### randomly contaminate data
      ntr <- length(trid)
      set.seed(1000 + ii)
      con <- sample(ntr)
      for (i in (1:4)) {
        ### i controls how many percentage of data contaminated, see
        ### argument per above
        ytr <- ytrold
        percon <- per[i]
        ### randomly flip labels of the samples in training set
        ### according to pre-defined contamination level
        if (percon > 0) {
          ji <- con[1:(percon * ntr)]
          ytr[ji] <- -ytrold[ji]
        }
        ### fit a model with nclreg for nonconvex loss or glmreg for
        ### logistic loss, and use cross-validation to select best
        ### penalization parameter
      }
    }
  }
}
```

```

if (type[k] %in% c("closs", "gloss", "qloss")) {
  dat.m1 <- nclreg(x = dtr, y = ytr, s = s[k],
    iter = 100, rfamily = type[k], penalty = penalty[j],
    lambda.min.ratio = ratio, gamma = gam, mstop.init = mstop,
    nu.init = nu, type.path = type.path, decreasing = FALSE,
    type.init = "bst")
  lambda <- dat.m1$lambda[1:nlam]
  set.seed(1000 + ii)
  cvm1 <- cv.nclreg(x = dtr, y = ytr, nfolds = 5,
    n.cores = n.cores, s = s[k], lambda = lambda,
    rfamily = type[k], penalty = penalty[j],
    gamma = gam, type = tuning, plot.it = FALSE,
    type.init = dat.m1$type.init, mstop.init = dat.m1$mstop.init,
    nu.init = dat.m1$nu.init, type.path = type.path,
    decreasing = dat.m1$decreasing)
  err1 <- predict(dat.m1, newdata = dte, newy = yte,
    type = "error")
} else {
  dat.m1 <- glmreg(x = dtr, y = (ytr + 1)/2,
    family = type[k], penalty = penalty[j], lambda.min.ratio = ratio,
    gamma = gam)
  set.seed(1000 + ii)
  cvm1 <- cv.glmreg(x = dtr, y = (ytr + 1)/2,
    nfolds = 5, n.cores = n.cores, lambda = dat.m1$lambda,
    family = type[k], penalty = penalty[j], gamma = gam,
    plot.it = FALSE)
  err1 <- apply((yte > -1) != predict(dat.m1,
    newx = dte, type = "class"), 2, mean)
}
optmstop <- cvm1$lambda.which
err.m1[ii, i] <- err1[optmstop]
nvar.m1[ii, i] <- length(predict(dat.m1, which = optmstop,
  type = "nonzero"))
errbest.m1[ii, i] <- min(err1, na.rm = TRUE)
lambest.m1[ii, i] <- which.min(err1)
nvarbest.m1[ii, i] <- length(predict(dat.m1,
  which = which.min(err1), type = "nonzero"))
}
if (ii%nrnrun == 0) {
  if (type[k] %in% c("closs", "gloss", "qloss"))
    cat(paste("\nfamily ", type1[k], ", s=", s[k],
      sep = ""), "\n") else cat(paste("\nfamily ", type1[k], sep = ""),
    "\n")
  pentype <- switch(penalty[j], enet = "LASSO",
    mnet = "MCP", snet = "SCAD")
  cat("penalty=", pentype, "\n")
  if (penalty[j] %in% c("snet", "mnet"))
    cat("gamma=", gam, "\n")
  cat("best misclassification error\n")
}

```

```

print(round(apply(errbest.m1, 2, summary7), 4))
cat("which lambda has best error\n")
print(round(apply(lambest.m1, 2, summary7), 1))
cat("number of variables selected with best error\n")
print(round(apply(nvarbest.m1, 2, summary7),
  1))
cat("CV based misclassification error\n")
print(round(apply(err.m1, 2, summary7), 4))
cat("number of variables selected by CV\n")
print(round(apply(nvar.m1, 2, summary7), 1))
if (plot.it) {
  par(mfrow = c(2, 1))
  boxplot(err.m1, main = "Misclassification error",
    subset = "", sub = paste(type1[k], "-", pentype,
      sep = ""))
  boxplot(nvar.m1, main = "No. variables", subset = "",
    sub = paste(type1[k], "-", pentype, sep = ""))
}
}
}
}
}
##
## family Closs, s=0.9
## penalty= LASSO
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0506  0.0506   0.0449   0.0393
## 1st Qu.    0.0674  0.0674   0.0730   0.0730
## Median    0.0730  0.0730   0.0787   0.0843
## Mean      0.0746  0.0778   0.0844   0.1010
## 3rd Qu.    0.0843  0.0843   0.0899   0.1138
## Max.      0.0955  0.1461   0.1798   0.3483
## sd        0.0116  0.0152   0.0230   0.0445
## which lambda has best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1.0    1.0    1.0    1.0
## 1st Qu.      1.0    1.0    1.0    1.0
## Median     13.0   13.5    7.0   11.0
## Mean       18.0   17.9   11.6   14.2
## 3rd Qu.    30.2   33.2   17.2   24.2
## Max.       58.0   50.0   44.0   46.0
## sd        18.3   16.4   12.8   13.5
## number of variables selected with best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1.0    1.0    1.0    1.0
## 1st Qu.      2.0    2.0    3.0    3.0
## Median      3.5    4.0    5.0    5.0
## Mean       3.8    4.3    5.6    6.1

```

```

## 3rd Qu.      5.0      6.0      7.2      9.0
## Max.        12.0     12.0     17.0     16.0
## sd          2.1      2.6      3.0      3.5
## CV based misclassification error
##           cont-0% cont-5% cont-10% cont-15%
## Min.       0.0506  0.0506  0.0562  0.0618
## 1st Qu.    0.0730  0.0730  0.0772  0.0899
## Median     0.0787  0.0843  0.0899  0.1011
## Mean       0.0820  0.0860  0.0928  0.1174
## 3rd Qu.    0.0899  0.0955  0.1011  0.1348
## Max.       0.1348  0.1573  0.1854  0.3483
## sd         0.0147  0.0169  0.0256  0.0467
## number of variables selected by CV
##           cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1.0
## 1st Qu.      3.8      3.0      4.0      4.0
## Median       5.0      5.0      6.0      7.0
## Mean         5.2      5.1      6.0      7.2
## 3rd Qu.      7.0      7.0      8.0     10.0
## Max.        12.0     12.0     16.0     17.0
## sd          2.3      2.6      3.1      4.2
##
## family Closs, s=0.9
## penalty= SCAD
## gamma= 3.7
## best misclassification error
##           cont-0% cont-5% cont-10% cont-15%
## Min.       0.0449  0.0449  0.0506  0.0506
## 1st Qu.    0.0674  0.0674  0.0674  0.0730
## Median     0.0787  0.0730  0.0787  0.0843
## Mean       0.0762  0.0794  0.0866  0.1103
## 3rd Qu.    0.0843  0.0843  0.0899  0.1348
## Max.       0.1011  0.2640  0.2753  0.2809
## sd         0.0121  0.0235  0.0340  0.0573
## which lambda has best error
##           cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1.0
## 1st Qu.      1.0      1.0      1.0      1.0
## Median      22.5     23.5      2.5      1.0
## Mean        22.8     20.4     17.8     16.6
## 3rd Qu.     42.8     40.0     35.0     36.2
## Max.        58.0     50.0     50.0     48.0
## sd          21.7     19.2     18.1     18.8
## number of variables selected with best error
##           cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1.0
## 1st Qu.      1.0      1.0      1.0      1.0
## Median       1.0      1.0      1.0      1.0
## Mean         1.0      1.0      1.1      1.2

```

```

## 3rd Qu.      1.0      1.0      1.0      1.0
## Max.         2.0      2.0      2.0      7.0
## sd           0.2      0.1      0.2      1.0
## CV based misclassification error
##           cont-0% cont-5% cont-10% cont-15%
## Min.        0.0506 0.0449 0.0562 0.0506
## 1st Qu.     0.0730 0.0730 0.0730 0.0787
## Median      0.0787 0.0787 0.0843 0.0899
## Mean        0.0805 0.0839 0.0910 0.1152
## 3rd Qu.     0.0899 0.0899 0.0955 0.1404
## Max.        0.1067 0.2753 0.2753 0.3090
## sd          0.0115 0.0261 0.0358 0.0605
## number of variables selected by CV
##           cont-0% cont-5% cont-10% cont-15%
## Min.         1       1       1.0     1.0
## 1st Qu.       1       1       1.0     1.0
## Median        1       1       1.0     1.0
## Mean          1       1       1.0     1.2
## 3rd Qu.       1       1       1.0     1.0
## Max.          1       1       2.0     7.0
## sd            0       0       0.1     0.8
##
## family Closs, s=0.9
## penalty= MCP
## gamma= 12
## best misclassification error
##           cont-0% cont-5% cont-10% cont-15%
## Min.        0.0449 0.0449 0.0506 0.0506
## 1st Qu.     0.0674 0.0674 0.0674 0.0730
## Median      0.0787 0.0730 0.0787 0.0843
## Mean        0.0772 0.0804 0.0862 0.1055
## 3rd Qu.     0.0843 0.0899 0.0899 0.1292
## Max.        0.1124 0.2191 0.2303 0.2584
## sd          0.0127 0.0245 0.0311 0.0503
## which lambda has best error
##           cont-0% cont-5% cont-10% cont-15%
## Min.         1.0     1.0     1.0     1.0
## 1st Qu.       1.0     1.0     1.0     1.0
## Median        1.0    11.5     1.0     3.0
## Mean         17.9    18.8    11.7    12.1
## 3rd Qu.       35.5    35.5    21.0    25.0
## Max.         60.0    55.0    52.0    49.0
## sd           21.6    19.2    15.6    15.3
## number of variables selected with best error
##           cont-0% cont-5% cont-10% cont-15%
## Min.         1.0     1.0     1.0     1.0
## 1st Qu.       1.0     1.0     1.0     1.0
## Median        1.0     1.0     1.0     2.0
## Mean          1.4     1.5     1.8     2.6

```

```

## 3rd Qu.      2.0      2.0      2.0      4.0
## Max.        7.0      6.0      7.0     10.0
## sd          0.9      0.9      1.2      2.2
## CV based misclassification error
##           cont-0% cont-5% cont-10% cont-15%
## Min.       0.0449  0.0449  0.0562  0.0562
## 1st Qu.    0.0730  0.0730  0.0730  0.0787
## Median     0.0815  0.0843  0.0843  0.0899
## Mean       0.0817  0.0860  0.0923  0.1145
## 3rd Qu.    0.0899  0.0899  0.0955  0.1362
## Max.       0.1124  0.2191  0.2416  0.2584
## sd         0.0128  0.0243  0.0334  0.0522
## number of variables selected by CV
##           cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1.0
## 1st Qu.      1.0      1.0      1.0      1.0
## Median       1.0      1.0      1.0      2.0
## Mean         1.5      1.5      1.8      2.7
## 3rd Qu.      2.0      2.0      2.0      4.0
## Max.         7.0      7.0      8.0     11.0
## sd           1.0      1.1      1.4      2.4
##
## family Gloss, s=1.01
## penalty= LASSO
## best misclassification error
##           cont-0% cont-5% cont-10% cont-15%
## Min.       0.0449  0.0506  0.0449  0.0562
## 1st Qu.    0.0674  0.0674  0.0674  0.0730
## Median     0.0730  0.0730  0.0787  0.0899
## Mean       0.0752  0.0770  0.0836  0.1038
## 3rd Qu.    0.0843  0.0843  0.0899  0.1194
## Max.       0.0955  0.1236  0.1685  0.3146
## sd         0.0120  0.0139  0.0223  0.0410
## which lambda has best error
##           cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1.0
## 1st Qu.      1.0      1.0      3.8      7.8
## Median     15.0     20.0     18.0     25.0
## Mean       25.9     27.5     22.8     27.4
## 3rd Qu.    52.0     50.0     38.0     44.0
## Max.       83.0     76.0     78.0     77.0
## sd         26.9     25.6     21.3     22.0
## number of variables selected with best error
##           cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1.0
## 1st Qu.      2.0      2.0      3.0      4.0
## Median       3.5      4.0      6.0      7.0
## Mean         4.0      4.8      6.0      8.0
## 3rd Qu.      5.0      7.0      8.0     11.0

```



```

## Max.      10.0    17.0    21.0    21.0
## sd        2.4     3.3     3.6     4.9
## CV based misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0506 0.0506  0.0618  0.0674
## 1st Qu.    0.0730 0.0787  0.0787  0.0899
## Median    0.0843 0.0843  0.0899  0.1180
## Mean      0.0833 0.0858  0.0975  0.1266
## 3rd Qu.    0.0955 0.0913  0.1067  0.1573
## Max.      0.1236 0.1629  0.2303  0.3146
## sd        0.0138 0.0177  0.0292  0.0454
## number of variables selected by CV
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1.0     1.0     1.0     2.0
## 1st Qu.      4.0     4.0     4.0     6.0
## Median       5.0     5.5     6.0     9.5
## Mean        5.2     6.0     7.6    10.6
## 3rd Qu.      6.2     8.0    10.0    15.0
## Max.       11.0    15.0    22.0    24.0
## sd          2.2     2.9     4.6     5.7
##
## family Gloss, s=1.01
## penalty= SCAD
## gamma= 3.7
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0562  0.0562  0.0618
## 1st Qu.    0.0730 0.0730  0.0730  0.0772
## Median    0.0787 0.0843  0.0843  0.0955
## Mean      0.0813 0.0860  0.0944  0.1271
## 3rd Qu.    0.0899 0.0899  0.0955  0.1601
## Max.      0.1124 0.2640  0.3427  0.3371
## sd        0.0125 0.0253  0.0425  0.0721
## which lambda has best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1.0     1.0     1.0     1.0
## 1st Qu.      1.0     1.0     1.0     1.0
## Median       1.0     1.0     1.0     1.0
## Mean        1.9     2.3     3.0     6.8
## 3rd Qu.      1.2     2.0     2.0     2.0
## Max.       34.0    88.0    81.0    95.0
## sd          3.7     8.7     9.8    17.4
## number of variables selected with best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1      1      1      1.0
## 1st Qu.      1      1      1      1.0
## Median       1      1      1      1.0
## Mean        1      1      1      1.6
## 3rd Qu.      1      1      1      1.0

```

```

## Max.          1          1          1      21.0
## sd            0          0          0       2.5
## CV based misclassification error
##           cont-0% cont-5% cont-10% cont-15%
## Min.        0.0562 0.0562 0.0562 0.0674
## 1st Qu.      0.0730 0.0787 0.0787 0.0829
## Median       0.0787 0.0843 0.0843 0.1011
## Mean         0.0836 0.0888 0.0972 0.1335
## 3rd Qu.      0.0955 0.0955 0.1011 0.1699
## Max.         0.1124 0.2697 0.3427 0.3539
## sd           0.0128 0.0251 0.0432 0.0724
## number of variables selected by CV
##           cont-0% cont-5% cont-10% cont-15%
## Min.          1          1          1       1.0
## 1st Qu.        1          1          1       1.0
## Median         1          1          1       1.0
## Mean           1          1          1       2.3
## 3rd Qu.        1          1          1       1.0
## Max.           1          1          1      26.0
## sd             0          0          0       4.9
##
## family Gloss, s=1.01
## penalty= MCP
## gamma= 12
## best misclassification error
##           cont-0% cont-5% cont-10% cont-15%
## Min.        0.0449 0.0506 0.0506 0.0506
## 1st Qu.      0.0674 0.0674 0.0674 0.0730
## Median       0.0787 0.0787 0.0787 0.0843
## Mean         0.0784 0.0803 0.0839 0.1029
## 3rd Qu.      0.0899 0.0899 0.0899 0.1292
## Max.         0.1011 0.2022 0.2135 0.3483
## sd           0.0122 0.0195 0.0244 0.0498
## which lambda has best error
##           cont-0% cont-5% cont-10% cont-15%
## Min.          1.0       1.0       1.0       1.0
## 1st Qu.        1.0       1.0       1.0       5.8
## Median         1.0       1.5      10.0      21.0
## Mean          21.0      23.1      19.8      27.4
## 3rd Qu.        45.8      48.2      35.5      49.5
## Max.           84.0      77.0      68.0      72.0
## sd            29.2      27.4      22.0      23.5
## number of variables selected with best error
##           cont-0% cont-5% cont-10% cont-15%
## Min.          1.0       1.0       1.0       1.0
## 1st Qu.        1.0       1.0       1.0       1.0
## Median         1.0       1.0       1.0       1.0
## Mean           1.1       1.4       1.6       2.7
## 3rd Qu.        1.0       1.0       2.0       3.0

```

```

## Max.      3.0      9.0      9.0     13.0
## sd        0.4      1.3      1.3      2.8
## CV based misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0506  0.0506   0.0562   0.0618
## 1st Qu.    0.0730  0.0730   0.0730   0.0843
## Median     0.0787  0.0787   0.0843   0.1067
## Mean       0.0820  0.0849   0.0933   0.1304
## 3rd Qu.    0.0899  0.0899   0.1011   0.1742
## Max.       0.1292  0.2022   0.2640   0.3764
## sd         0.0135  0.0213   0.0341   0.0587
## number of variables selected by CV
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1.0
## 1st Qu.      1.0      1.0      1.0      1.0
## Median       1.0      1.0      1.0      2.0
## Mean         1.1      1.5      2.1      4.7
## 3rd Qu.      1.0      1.0      2.0      8.0
## Max.         7.0      9.0     17.0     18.0
## sd           0.7      1.4      2.8      5.0
##
## family Qloss, s=0.5
## penalty= LASSO
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0449  0.0506   0.0449   0.0562
## 1st Qu.    0.0674  0.0674   0.0674   0.0730
## Median     0.0730  0.0730   0.0787   0.0899
## Mean       0.0747  0.0768   0.0838   0.1027
## 3rd Qu.    0.0843  0.0843   0.0899   0.1180
## Max.       0.0955  0.1180   0.1742   0.3090
## sd         0.0117  0.0136   0.0227   0.0407
## which lambda has best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1.0
## 1st Qu.      1.0      1.0      3.0      8.0
## Median     15.5     20.5     18.0     25.0
## Mean       25.7     26.0     22.9     26.9
## 3rd Qu.    51.0     44.2     36.2     43.2
## Max.       81.0     74.0     74.0     74.0
## sd        27.1     24.5     21.2     21.0
## number of variables selected with best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1.0
## 1st Qu.      2.0      2.0      3.0      4.0
## Median       4.0      5.0      6.0      7.0
## Mean         4.3      5.0      6.2      8.0
## 3rd Qu.      6.0      7.0      8.0     11.0
## Max.       11.0     17.0     16.0     20.0

```

```

## sd          2.5      3.4      3.6      4.8
## CV based misclassification error
##           cont-0% cont-5% cont-10% cont-15%
## Min.      0.0506 0.0506  0.0618  0.0674
## 1st Qu.    0.0730 0.0772  0.0843  0.0899
## Median    0.0815 0.0843  0.0899  0.1180
## Mean      0.0822 0.0851  0.0971  0.1246
## 3rd Qu.    0.0899 0.0899  0.1025  0.1475
## Max.      0.1236 0.1517  0.2191  0.3090
## sd        0.0140 0.0162  0.0280  0.0447
## number of variables selected by CV
##           cont-0% cont-5% cont-10% cont-15%
## Min.        1.0     2.0     1.0     2.0
## 1st Qu.      4.0     4.0     4.0     6.0
## Median      5.0     6.0     6.5     9.0
## Mean       5.8     6.2     7.9    10.5
## 3rd Qu.     7.0     8.2    10.0    14.0
## Max.      11.0    14.0    23.0    27.0
## sd         2.3     2.8     5.1     5.9
##
## family Qloss, s=0.5
## penalty= SCAD
## gamma= 3.7
## best misclassification error
##           cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0562  0.0562  0.0618
## 1st Qu.    0.0730 0.0730  0.0730  0.0787
## Median    0.0787 0.0843  0.0843  0.0955
## Mean      0.0810 0.0856  0.0938  0.1253
## 3rd Qu.    0.0899 0.0899  0.0955  0.1559
## Max.      0.1124 0.2640  0.3427  0.3708
## sd        0.0126 0.0251  0.0425  0.0697
## which lambda has best error
##           cont-0% cont-5% cont-10% cont-15%
## Min.        1.0     1.0     1.0     1.0
## 1st Qu.      1.0     1.0     1.0     1.0
## Median      1.0     1.0     1.0     1.0
## Mean       2.0     1.9     1.9     3.3
## 3rd Qu.     2.0     2.0     2.0     2.0
## Max.      33.0    15.0    11.0    54.0
## sd         3.5     2.0     1.9     8.0
## number of variables selected with best error
##           cont-0% cont-5% cont-10% cont-15%
## Min.        1      1      1      1.0
## 1st Qu.      1      1      1      1.0
## Median      1      1      1      1.0
## Mean       1      1      1      1.5
## 3rd Qu.     1      1      1      1.0
## Max.        1      1      1     20.0

```

```

## sd          0          0          0          2.4
## CV based misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0562  0.0562  0.0674
## 1st Qu.    0.0730 0.0787  0.0772  0.0843
## Median     0.0787 0.0843  0.0843  0.0983
## Mean       0.0833 0.0893  0.0971  0.1309
## 3rd Qu.    0.0913 0.0955  0.1011  0.1685
## Max.       0.1124 0.2753  0.3427  0.3764
## sd         0.0129 0.0256  0.0429  0.0705
## number of variables selected by CV
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1         1         1         1.0
## 1st Qu.      1         1         1         1.0
## Median       1         1         1         1.0
## Mean         1         1         1         2.0
## 3rd Qu.      1         1         1         1.0
## Max.         1         1         1        26.0
## sd           0         0         0         4.1
##
## family Qloss, s=0.5
## penalty= MCP
## gamma= 12
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0449 0.0506  0.0506  0.0506
## 1st Qu.    0.0674 0.0674  0.0674  0.0730
## Median     0.0787 0.0787  0.0787  0.0843
## Mean       0.0781 0.0798  0.0848  0.1013
## 3rd Qu.    0.0899 0.0899  0.0899  0.1236
## Max.       0.1067 0.1966  0.2079  0.3258
## sd         0.0122 0.0193  0.0266  0.0480
## which lambda has best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1.0
## 1st Qu.      1.0      1.0      1.0      2.5
## Median       1.0      2.5      6.0     21.0
## Mean        21.9     22.5     18.8     26.4
## 3rd Qu.     51.2     49.0     32.2     46.5
## Max.        82.0     77.0     75.0     69.0
## sd          29.5     26.7     22.7     22.8
## number of variables selected with best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1
## 1st Qu.      1.0      1.0      1.0      1
## Median       1.0      1.0      1.0      2
## Mean         1.2      1.4      2.0      3
## 3rd Qu.      1.0      1.0      2.0      4
## Max.         2.0      9.0     17.0     15

```

```

## sd          0.4      1.2      2.1      3
## CV based misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0506 0.0562 0.0618
## 1st Qu.    0.0730 0.0730 0.0772 0.0843
## Median     0.0787 0.0787 0.0843 0.1011
## Mean       0.0819 0.0842 0.0930 0.1270
## 3rd Qu.    0.0899 0.0899 0.0969 0.1643
## Max.       0.1292 0.2022 0.2640 0.3876
## sd         0.0132 0.0192 0.0339 0.0583
## number of variables selected by CV
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1.0
## 1st Qu.      1.0      1.0      1.0      1.0
## Median       1.0      1.0      1.0      2.0
## Mean         1.3      1.5      2.2      4.8
## 3rd Qu.      1.0      2.0      2.0      9.0
## Max.         7.0      9.0     17.0     17.0
## sd           0.8      1.2      2.8      4.8
##
## family Logistic
## penalty= LASSO
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0449 0.0449 0.0562 0.0562
## 1st Qu.    0.0660 0.0674 0.0730 0.0843
## Median     0.0730 0.0787 0.0843 0.1124
## Mean       0.0729 0.0811 0.0939 0.1176
## 3rd Qu.    0.0787 0.0899 0.1067 0.1461
## Max.       0.1011 0.1685 0.1854 0.3034
## sd         0.0114 0.0200 0.0297 0.0413
## which lambda has best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.        2.0      6.0      2.0      5.0
## 1st Qu.      2.0     21.0     19.0     19.0
## Median     51.5     34.0     32.5     32.0
## Mean       45.4     40.7     37.3     39.1
## 3rd Qu.     73.2     55.0     54.5     55.5
## Max.       99.0    100.0    100.0    100.0
## sd         33.2     24.3     24.4     25.8
## number of variables selected with best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.        1.0      1.0      1.0      1.0
## 1st Qu.      1.0      1.0      2.0      3.0
## Median       3.0      3.0      3.0      6.0
## Mean         3.8      4.0      5.7      8.9
## 3rd Qu.      5.0      5.0      7.0     11.5
## Max.        15.0     16.0     29.0     32.0
## sd           3.2      3.5      5.8      8.1

```

```

## CV based misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0618  0.0618  0.0843
## 1st Qu.    0.0772 0.0885  0.1011  0.1292
## Median     0.0843 0.1011  0.1236  0.1573
## Mean       0.0851 0.1058  0.1290  0.1567
## 3rd Qu.    0.0955 0.1180  0.1461  0.1812
## Max.       0.1292 0.2079  0.2191  0.3371
## sd         0.0143 0.0262  0.0356  0.0406
## number of variables selected by CV
##          cont-0% cont-5% cont-10% cont-15%
## Min.         4.0    4.0    3.0    4.0
## 1st Qu.       9.0   13.0   13.0   13.0
## Median      11.5   16.0   19.0   20.0
## Mean        11.7   16.5   18.6   19.4
## 3rd Qu.     15.0   20.0   24.0   26.0
## Max.        20.0   29.0   33.0   33.0
## sd          3.6    5.2    7.3    7.7
##
## family Logistic
## penalty= SCAD
## gamma= 3.7
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0449 0.0449  0.0562  0.0562
## 1st Qu.    0.0674 0.0674  0.0730  0.0843
## Median     0.0730 0.0787  0.0843  0.1124
## Mean       0.0737 0.0811  0.0940  0.1176
## 3rd Qu.    0.0843 0.0899  0.1067  0.1461
## Max.       0.1011 0.1685  0.1854  0.3034
## sd         0.0115 0.0200  0.0297  0.0413
## which lambda has best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.         2.0    6.0    2.0    5.0
## 1st Qu.       2.0   21.0   19.0   19.0
## Median      49.5   34.0   32.0   32.0
## Mean       44.3   39.7   36.7   39.1
## 3rd Qu.     73.8   53.2   52.5   55.5
## Max.      100.0   98.0  100.0  100.0
## sd        34.4   23.4   24.2   25.8
## number of variables selected with best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.         1.0    1.0    1.0    1.0
## 1st Qu.       1.0    1.0    2.0    3.0
## Median       2.5    3.0    3.0    6.0
## Mean        3.4    3.8    5.5    8.9
## 3rd Qu.      5.0    4.0    7.0   11.5
## Max.       15.0   16.0   29.0   32.0
## sd          2.8    3.4    5.6    8.1

```

```

## CV based misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0618  0.0618  0.0843
## 1st Qu.    0.0772 0.0843  0.1011  0.1292
## Median     0.0843 0.1011  0.1236  0.1573
## Mean       0.0846 0.1051  0.1289  0.1566
## 3rd Qu.    0.0899 0.1138  0.1461  0.1812
## Max.       0.1292 0.2079  0.2191  0.3371
## sd         0.0127 0.0264  0.0359  0.0409
## number of variables selected by CV
##          cont-0% cont-5% cont-10% cont-15%
## Min.         2.0    3.0    3.0    4.0
## 1st Qu.       6.0   12.0   12.8   13.0
## Median       8.0   15.0   19.0   20.0
## Mean        9.1   15.3   18.5   19.1
## 3rd Qu.     12.0   20.0   24.0   25.0
## Max.       20.0   29.0   33.0   33.0
## sd         4.3    5.8    7.3    7.6
##
## family Logistic
## penalty= MCP
## gamma= 12
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0449 0.0449  0.0506  0.0562
## 1st Qu.    0.0674 0.0674  0.0730  0.0843
## Median     0.0730 0.0787  0.0843  0.1152
## Mean       0.0743 0.0798  0.0936  0.1184
## 3rd Qu.    0.0843 0.0899  0.1011  0.1461
## Max.       0.1011 0.1685  0.1966  0.2921
## sd         0.0120 0.0199  0.0314  0.0418
## which lambda has best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.         2.0    6.0    1.0    5.0
## 1st Qu.       2.0   24.5   20.0   22.5
## Median      55.5   41.0   38.0   34.0
## Mean       50.0   45.3   38.7   41.4
## 3rd Qu.     84.0   63.2   53.8   53.0
## Max.       98.0   97.0   99.0  100.0
## sd        35.4   24.7   24.6   26.1
## number of variables selected with best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.         1.0    1.0    1.0    1.0
## 1st Qu.       1.0    1.0    2.0    3.0
## Median       2.0    3.0    3.0    5.0
## Mean        3.0    3.6    5.0    8.1
## 3rd Qu.      4.0    5.0    7.0   11.0
## Max.       10.0   16.0   17.0   28.0
## sd         2.2    3.1    4.3    7.1

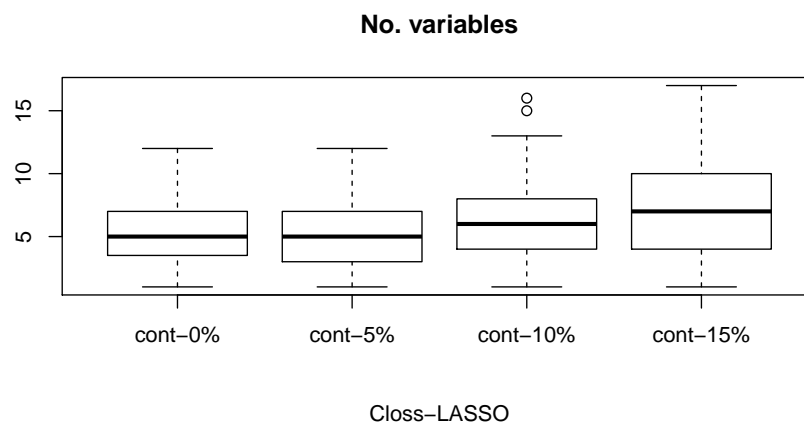
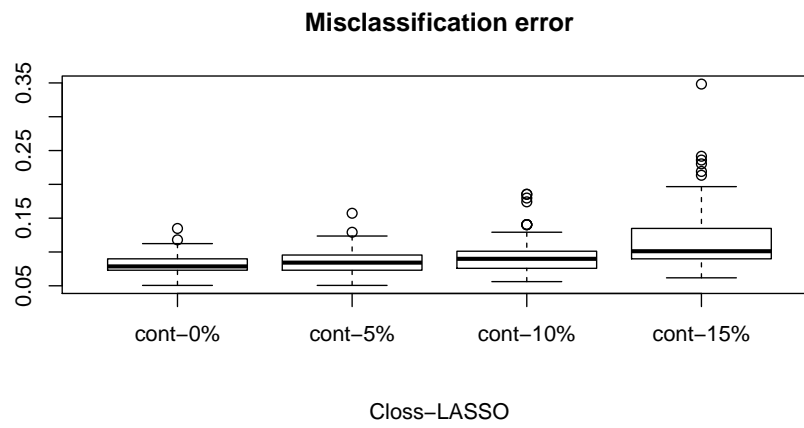
```

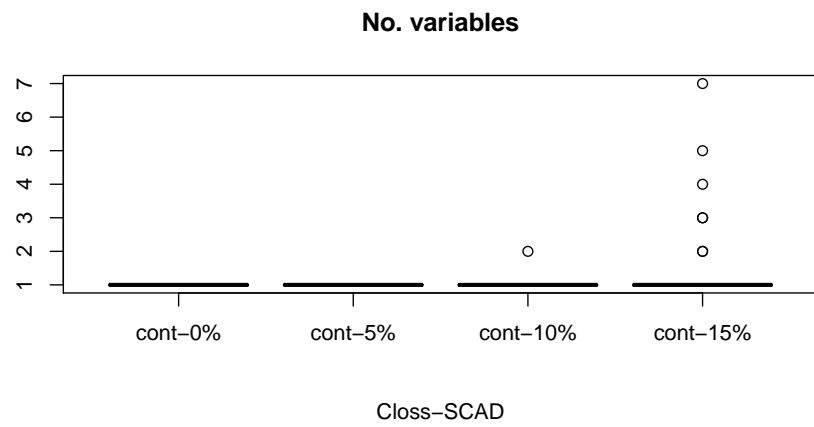
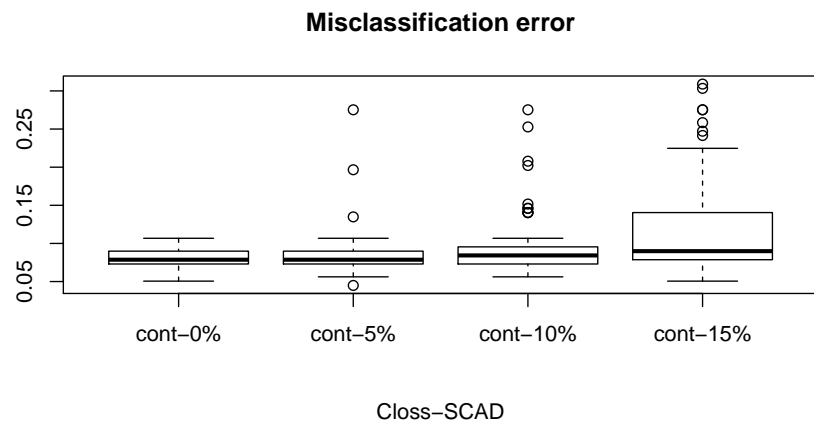


```

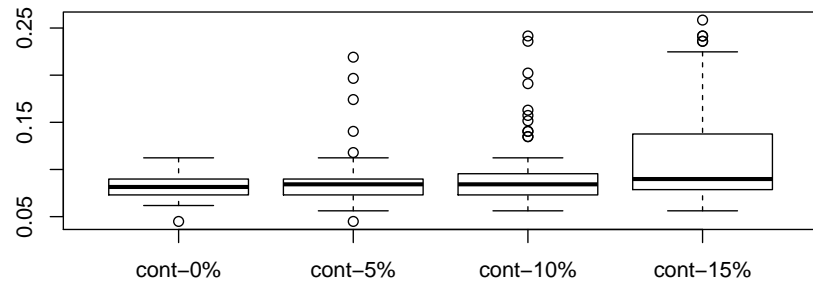
## CV based misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562 0.0618  0.0562  0.0730
## 1st Qu.    0.0730 0.0843  0.1011  0.1292
## Median     0.0843 0.0955  0.1208  0.1629
## Mean       0.0844 0.1032  0.1289  0.1608
## 3rd Qu.    0.0955 0.1124  0.1517  0.1854
## Max.       0.1292 0.2360  0.2360  0.3315
## sd         0.0132 0.0268  0.0397  0.0443
## number of variables selected by CV
##          cont-0% cont-5% cont-10% cont-15%
## Min.        2.0    2.0    1.0    2.0
## 1st Qu.      6.0    9.0    8.0   10.0
## Median       8.0   12.0   15.0   15.0
## Mean        8.4   11.9   14.1   15.1
## 3rd Qu.     11.0   15.0   20.0   21.0
## Max.        17.0   23.0   27.0   29.0
## sd          3.3    4.5    6.7    6.7
print(proc.time() - ptm)
##          user      system   elapsed
## 57645.816    14.694 59073.200

```



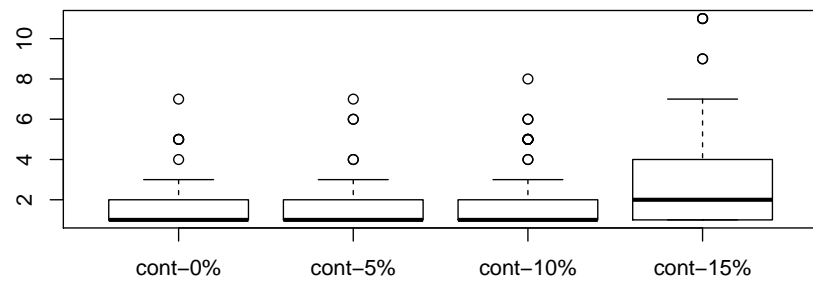


Misclassification error



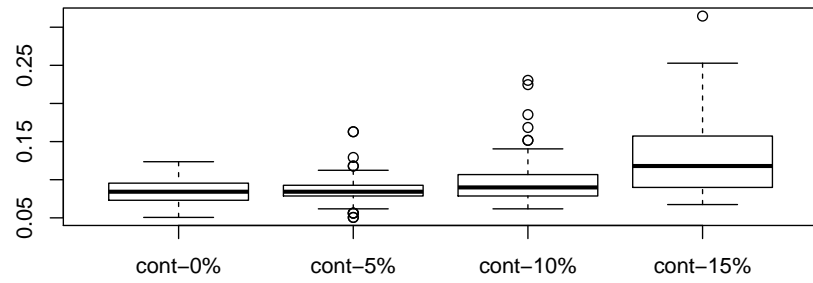
Closs-MCP

No. variables



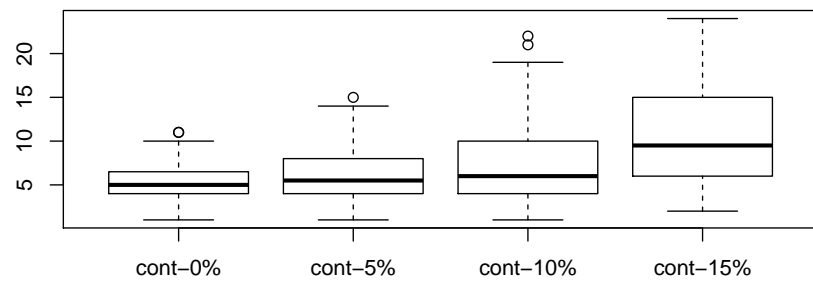
Closs-MCP

Misclassification error

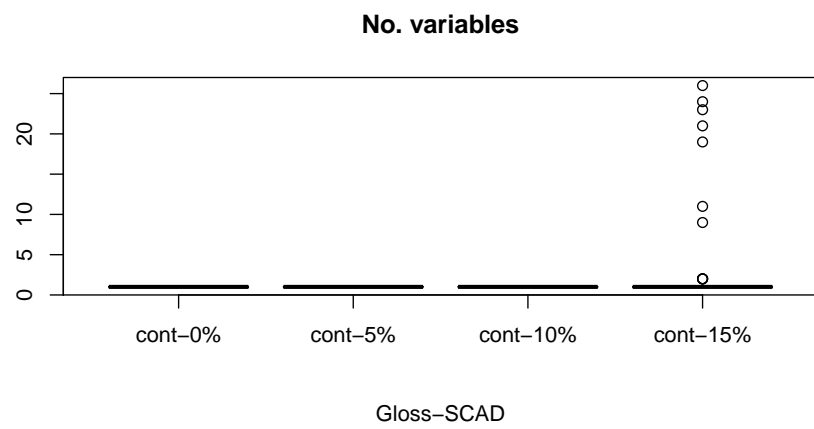
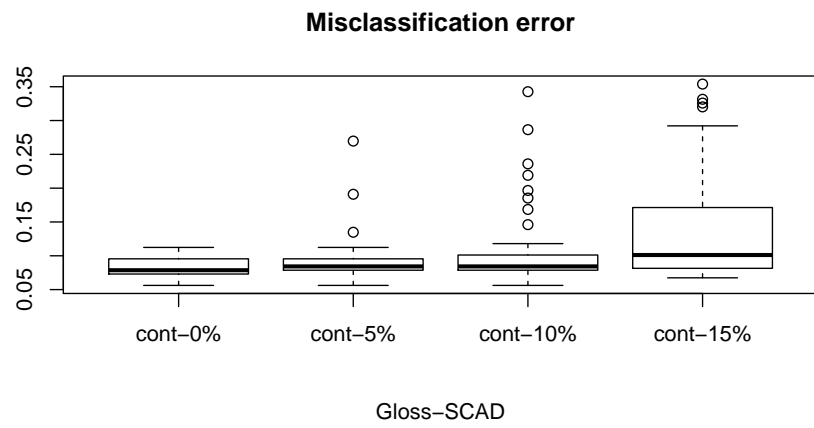


Gloss-LASSO

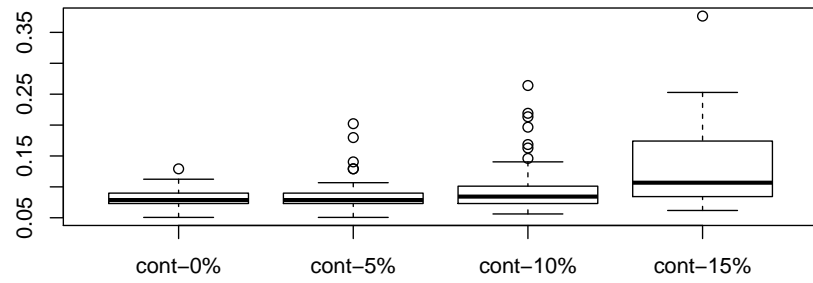
No. variables



Gloss-LASSO

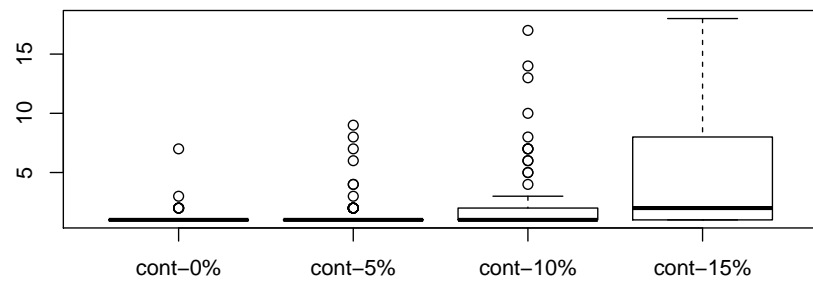


Misclassification error



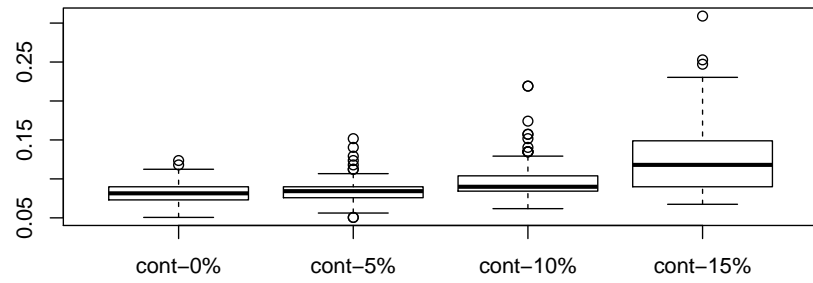
Gloss-MCP

No. variables



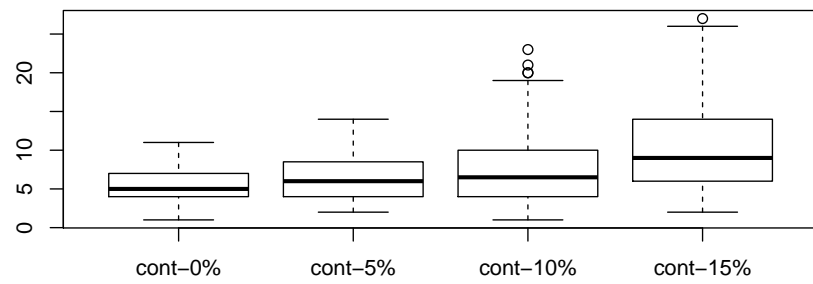
Gloss-MCP

Misclassification error



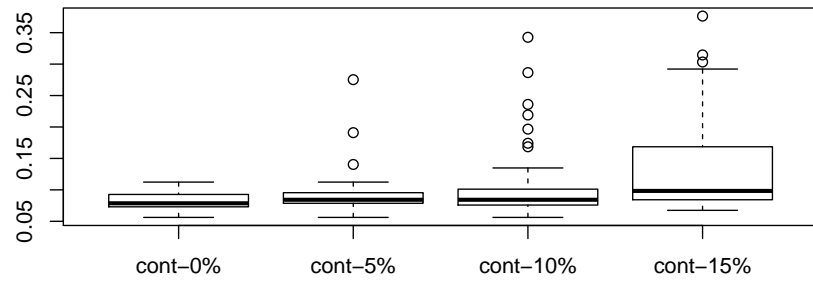
Qloss-LASSO

No. variables



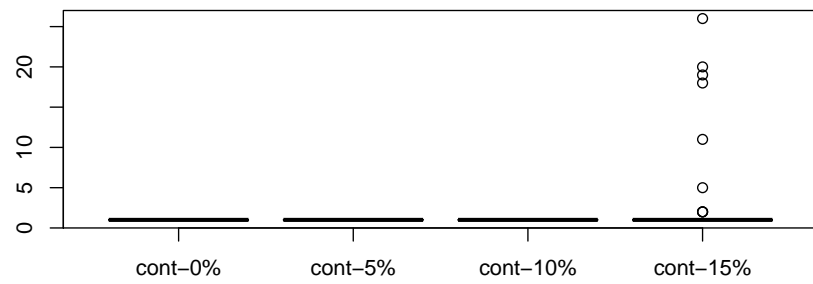
Qloss-LASSO

Misclassification error



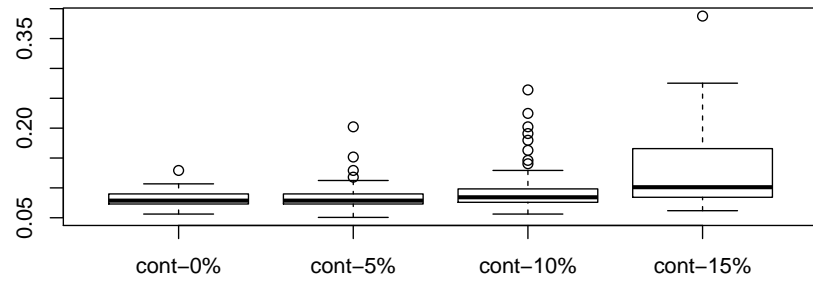
Qloss-SCAD

No. variables



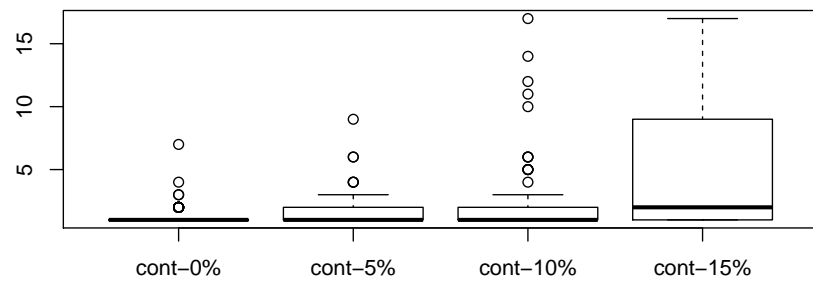
Qloss-SCAD

Misclassification error



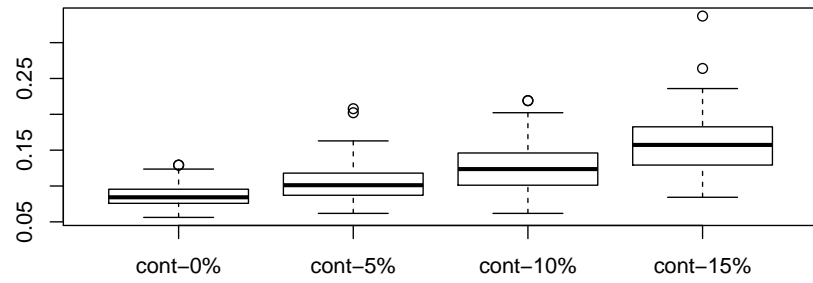
Qloss-MCP

No. variables



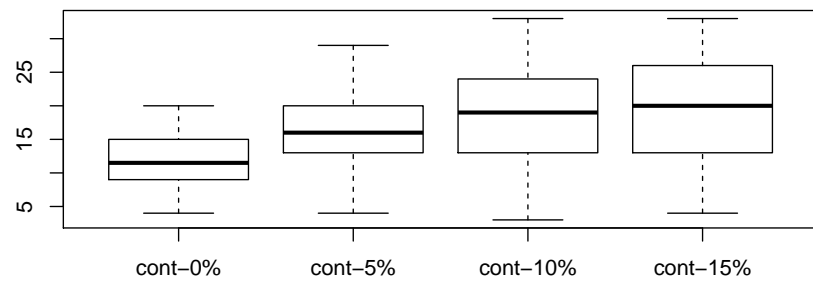
Qloss-MCP

Misclassification error



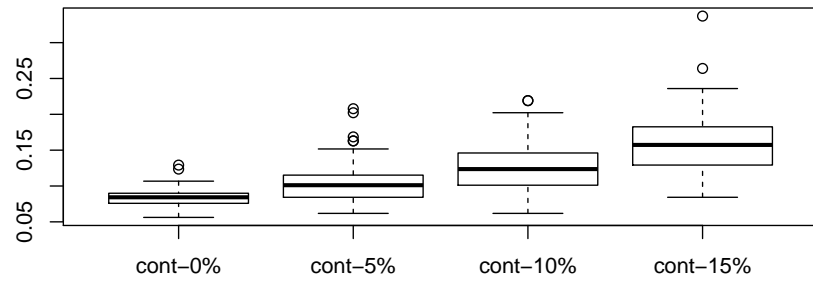
Logistic-LASSO

No. variables



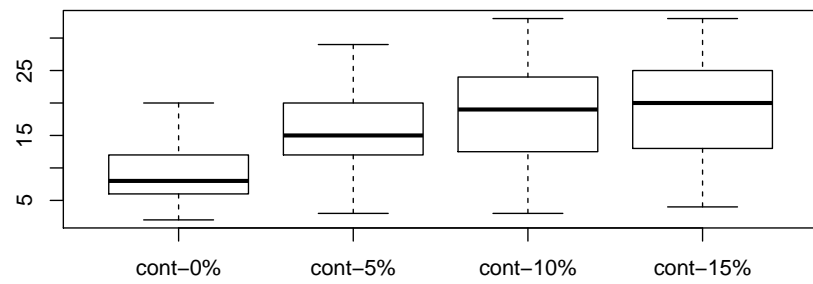
Logistic-LASSO

Misclassification error



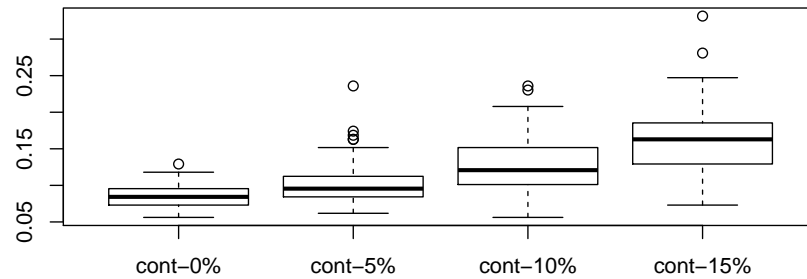
Logistic-SCAD

No. variables



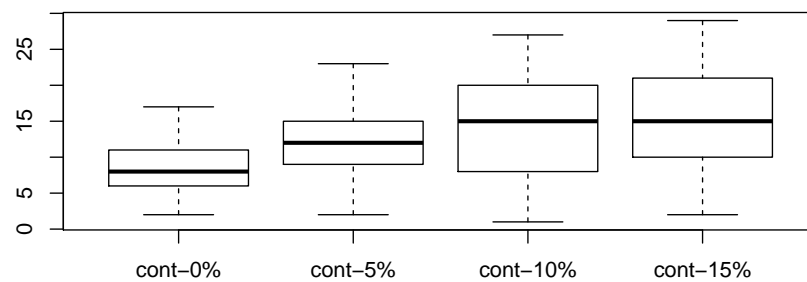
Logistic-SCAD

Misclassification error



Logistic-MCP

No. variables



Logistic-MCP

```
sessionInfo()
## R version 3.6.0 (2019-04-26)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.2 LTS
##
## Matrix products: default
## BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
##
## 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=C
##  [9] LC_ADDRESS=C              LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets
## [6] methods    base
```

```
##
## other attached packages:
## [1] mpath_0.3-13 gdata_2.18.0 knitr_1.21
##
## loaded via a namespace (and not attached):
## [1] magrittr_1.5      splines_3.6.0     MASS_7.3-51.1
## [4] doParallel_1.0.14 bst_0.3-18        pscl_1.5.2
## [7] gbm_2.1.5         lattice_0.20-38   foreach_1.4.4
## [10] stringr_1.4.0     tools_3.6.0       parallel_3.6.0
## [13] grid_3.6.0        glmnet_2.0-16     gtable_0.2.0
## [16] xfun_0.4          iterators_1.0.10  gtools_3.8.1
## [19] survival_2.43-3   numDeriv_2016.8-1 Matrix_1.2-17
## [22] gridExtra_2.3     formatR_1.5       codetools_0.2-16
## [25] rpart_4.1-15      evaluate_0.13     stringi_1.3.1
## [28] compiler_3.6.0
```

References

Zhu Wang. Quadratic majorization for robust nonconvex loss with applications to variable selection. 2016. manuscript.