Classification of Cancer Patients with Penalized Robust Nonconvex Loss Functions

Zhu Wang

UT Health San Antonio

wangz1@uthscsa.edu

March 16, 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=rhojvaiwkcsaihq&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 300 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 <- "active"
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"
### In cross-validation, we set the number of cores for
### parallel computing by
n.cores <- 5
### 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. By default, the solution path is computed backward as `direction="bwd"` in `nclreg` function. To select optimal penalization tuning parameters, we run five-fold cross-validation averaging classification errors. In cross-validation, we set the number of cores for parallel computing by `n.cores`=5. The classification errors and number of selected variables are summarized and these results can be plotted if `plot.it=TRUE`. This script also contains results with penalized

logistic regression using `glmreg`. The script took less than 6 hours to execute on Intel® CPU @ 2.10GHz × with 5 cores on a single processor.

```
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
```

3

```
        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)
          set.seed(1000 + ii)
          cvm1 <- cv.nclreg(x = dtr, y = ytr, nfolds = 5,
            n.cores = n.cores, s = s[k], lambda = dat.m1$lambda[1:nlam],
            rfamily = type[k], penalty = penalty[j],
            gamma = gam, type = tuning, plot.it = FALSE,
            type.init = dat.m1$type.init, mstop.init = dat.m1$mstop.init,
            nu.init = dat.m1$nu.init, type.path = type.path)
          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%%nrun == 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.      0.0337  0.0337   0.0506   0.0393
## 1st Qu.   0.0674  0.0674   0.0730   0.0843
## Median    0.0730  0.0787   0.0843   0.1011
## Mean      0.0743  0.0773   0.0919   0.1151
## 3rd Qu.   0.0843  0.0843   0.1011   0.1404
## Max.      0.1067  0.1067   0.2247   0.2528
## sd        0.0136  0.0133   0.0296   0.0475
## 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      13.5     8.5      6.0      8.5
## Mean        18.6    14.8     13.8     12.1
## 3rd Qu.     32.2    27.2     26.2     18.2
## Max.        58.0    51.0     50.0     51.0
## sd          19.3    15.3     14.9     12.0
## 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     3.0      3.0      4.0
## Median       4.0     5.0      6.0      7.0
## Mean         4.0     4.7      5.9      7.4
## 3rd Qu.      5.0     6.0      8.0      9.0
## Max.        10.0    14.0     17.0     22.0
## sd           2.1     2.6      3.5      4.4
```

```
## CV based misclassification error
##            cont-0% cont-5% cont-10% cont-15%
## Min.       0.0337  0.0449   0.0562   0.0506
## 1st Qu.    0.0730  0.0730   0.0843   0.0955
## Median     0.0815  0.0843   0.0955   0.1124
## Mean       0.0821  0.0857   0.1034   0.1301
## 3rd Qu.    0.0899  0.0955   0.1124   0.1517
## Max.       0.1180  0.1236   0.2697   0.4101
## sd         0.0144  0.0151   0.0345   0.0570
## 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.        4.0     3.0      3.0      4.0
## Median         5.0     5.0      6.0      7.0
## Mean           5.2     5.5      6.2      7.9
## 3rd Qu.        7.0     7.0      9.0     11.0
## Max.          11.0    14.0     15.0     20.0
## sd             2.0     2.9      3.5      5.0
##
## family Closs, s=0.9
## penalty= SCAD
## gamma= 3.7
## best misclassification error
##            cont-0% cont-5% cont-10% cont-15%
## Min.       0.0337  0.0337   0.0506   0.0393
## 1st Qu.    0.0674  0.0674   0.0716   0.0730
## Median     0.0787  0.0787   0.0843   0.1067
## Mean       0.0770  0.0807   0.1053   0.1364
## 3rd Qu.    0.0843  0.0843   0.1067   0.1868
## Max.       0.1573  0.2809   0.2978   0.3483
## sd         0.0166  0.0291   0.0560   0.0750
## 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        20.5    25.5     24.0      1.0
## Mean          22.4    21.8     20.7     17.3
## 3rd Qu.       44.0    40.0     39.2     36.2
## Max.          58.0    53.0     51.0     51.0
## sd            21.8    19.3     18.6     18.5
## number of variables selected with best error
##            cont-0% cont-5% cont-10% cont-15%
## Min.             1     1.0      1.0      1.0
## 1st Qu.          1     1.0      1.0      1.0
## Median           1     1.0      1.0      1.0
## Mean             1     1.1      1.1      1.3
## 3rd Qu.          1     1.0      1.0      1.0
## Max.             1     6.0      7.0     12.0
## sd               0     0.5      0.7      1.5
```

```
## CV based misclassification error
##           cont-0% cont-5% cont-10% cont-15%
## Min.      0.0393  0.0337   0.0506   0.0449
## 1st Qu.   0.0716  0.0730   0.0787   0.0772
## Median    0.0815  0.0843   0.0843   0.1067
## Mean      0.0812  0.0856   0.1110   0.1437
## 3rd Qu.   0.0899  0.0899   0.1081   0.1966
## Max.      0.1629  0.2809   0.3315   0.3652
## sd        0.0166  0.0294   0.0583   0.0800
## number of variables selected by CV
##           cont-0% cont-5% cont-10% cont-15%
## Min.         1     1.0      1.0      1.0
## 1st Qu.      1     1.0      1.0      1.0
## Median       1     1.0      1.0      1.0
## Mean         1     1.1      1.1      1.3
## 3rd Qu.      1     1.0      1.0      1.0
## Max.         1     6.0      7.0     12.0
## sd           0     0.5      0.7      1.3
##
## family Closs, s=0.9
## penalty= MCP
## gamma= 12
## best misclassification error
##           cont-0% cont-5% cont-10% cont-15%
## Min.      0.0337  0.0337   0.0506   0.0393
## 1st Qu.   0.0674  0.0674   0.0674   0.0787
## Median    0.0787  0.0787   0.0843   0.0983
## Mean      0.0780  0.0792   0.0983   0.1221
## 3rd Qu.   0.0899  0.0899   0.1025   0.1629
## Max.      0.1573  0.1685   0.2978   0.2865
## sd        0.0171  0.0183   0.0459   0.0588
## 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     3.5      2.0      2.0
## Mean        17.6    15.6     15.6     11.1
## 3rd Qu.     36.2    32.2     31.0     21.2
## Max.        59.0    57.0     52.0     49.0
## sd          21.9    18.7     17.2     14.6
## 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         1.3     1.6      1.9      3.2
## 3rd Qu.      1.0     2.0      2.0      4.0
## Max.         4.0     8.0     11.0     13.0
## sd           0.6     1.1      1.7      3.0
```

```
## CV based misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.     0.0337  0.0337   0.0562   0.0506
## 1st Qu.  0.0730  0.0730   0.0787   0.0843
## Median   0.0843  0.0843   0.0899   0.1124
## Mean     0.0820  0.0841   0.1070   0.1314
## 3rd Qu.  0.0899  0.0899   0.1124   0.1699
## Max.     0.1573  0.1685   0.2978   0.3202
## sd       0.0169  0.0209   0.0492   0.0619
## 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     1.6      1.9      3.1
## 3rd Qu.      1.0     2.0      2.0      4.0
## Max.         5.0     6.0      8.0     13.0
## sd           0.8     1.0      1.5      2.8
##
## family Gloss, s=1.01
## penalty= LASSO
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.     0.0337  0.0393   0.0506   0.0393
## 1st Qu.  0.0674  0.0674   0.0730   0.0843
## Median   0.0730  0.0787   0.0843   0.1124
## Mean     0.0753  0.0769   0.0926   0.1225
## 3rd Qu.  0.0843  0.0843   0.1067   0.1531
## Max.     0.1067  0.1067   0.2135   0.2697
## sd       0.0137  0.0140   0.0308   0.0482
## 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      1.8      3.0
## Median      17.5    17.0     20.5     18.0
## Mean        25.7    26.4     25.6     22.0
## 3rd Qu.     49.2    50.5     47.0     36.2
## Max.        84.0    79.0     73.0     71.0
## sd          27.2    25.5     23.6     20.9
## 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     2.8      3.0      4.0
## Median       4.0     5.0      6.0      8.0
## Mean         4.2     4.8      6.4      9.4
## 3rd Qu.      6.0     7.0      8.0     13.0
## Max.        11.0    13.0     21.0     27.0
## sd           2.2     2.8      4.3      6.3
## CV based misclassification error
```

8

```
##            cont-0% cont-5% cont-10% cont-15%
## Min.       0.0393  0.0449   0.0562   0.0562
## 1st Qu.    0.0730  0.0787   0.0843   0.1011
## Median     0.0787  0.0843   0.1011   0.1376
## Mean       0.0825  0.0870   0.1077   0.1433
## 3rd Qu.    0.0899  0.0955   0.1194   0.1742
## Max.       0.1236  0.1236   0.2416   0.3202
## sd         0.0145  0.0153   0.0350   0.0516
## number of variables selected by CV
##            cont-0% cont-5% cont-10% cont-15%
## Min.          1.0     1.0      2.0      1.0
## 1st Qu.       4.0     4.0      5.0      6.0
## Median        5.0     6.0      7.0     11.0
## Mean          5.5     6.4      7.8     11.3
## 3rd Qu.       7.0     8.2     11.0     16.0
## Max.         10.0    16.0     21.0     25.0
## sd            2.1     3.2      3.8      6.1
##
## family Gloss, s=1.01
## penalty= SCAD
## gamma= 3.7
## best misclassification error
##            cont-0% cont-5% cont-10% cont-15%
## Min.       0.0506  0.0506   0.0506   0.0506
## 1st Qu.    0.0730  0.0730   0.0787   0.0787
## Median     0.0843  0.0843   0.0899   0.1124
## Mean       0.0825  0.0861   0.1074   0.1495
## 3rd Qu.    0.0899  0.0955   0.1011   0.2022
## Max.       0.1461  0.2921   0.3315   0.4551
## sd         0.0158  0.0265   0.0579   0.0897
## 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          4.1     5.5      8.3      7.2
## 3rd Qu.       1.0     2.0      3.0      3.2
## Max.         59.0    79.0     80.0     82.0
## sd           10.3    13.9     17.9     15.9
## number of variables selected with best error
##            cont-0% cont-5% cont-10% cont-15%
## Min.            1     1.0      1.0      1.0
## 1st Qu.         1     1.0      1.0      1.0
## Median          1     1.0      1.0      1.0
## Mean            1     1.1      1.2      1.4
## 3rd Qu.         1     1.0      1.0      1.0
## Max.            1    10.0     15.0     27.0
## sd              0     0.9      1.4      2.6
## CV based misclassification error
```

```
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0562  0.0506   0.0506   0.0506
## 1st Qu.   0.0730  0.0787   0.0787   0.0843
## Median    0.0843  0.0899   0.0899   0.1152
## Mean      0.0852  0.0894   0.1124   0.1560
## 3rd Qu.   0.0955  0.0969   0.1124   0.2079
## Max.      0.1573  0.2921   0.3371   0.4607
## sd        0.0165  0.0268   0.0585   0.0907
## number of variables selected by CV
##          cont-0% cont-5% cont-10% cont-15%
## Min.           1     1.0      1.0      1.0
## 1st Qu.        1     1.0      1.0      1.0
## Median         1     1.0      1.0      1.0
## Mean           1     1.1      1.2      1.5
## 3rd Qu.        1     1.0      1.0      1.0
## Max.           1    11.0     15.0     27.0
## sd             0     1.0      1.4      3.0
##
## family Gloss, s=1.01
## penalty= MCP
## gamma= 12
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0337  0.0393   0.0449   0.0506
## 1st Qu.   0.0674  0.0674   0.0730   0.0843
## Median    0.0787  0.0787   0.0843   0.1236
## Mean      0.0795  0.0788   0.1000   0.1308
## 3rd Qu.   0.0899  0.0899   0.1124   0.1629
## Max.      0.1573  0.1461   0.2865   0.3090
## sd        0.0166  0.0175   0.0436   0.0569
## 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    22.0     18.5     12.0
## Mean        20.0    30.2     25.6     20.5
## 3rd Qu.     37.0    61.5     49.0     36.0
## Max.        83.0    81.0     74.0     73.0
## sd          28.7    29.3     25.4     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.      1.0     1.0      1.0      1.0
## Median       1.0     1.0      1.0      3.0
## Mean         1.1     1.4      2.2      4.8
## 3rd Qu.      1.0     1.0      2.0      8.0
## Max.         3.0     7.0     11.0     16.0
## sd           0.4     1.2      2.4      4.6
## CV based misclassification error
```

```
##          cont-0% cont-5% cont-10% cont-15%
## Min.     0.0449  0.0393   0.0506   0.0562
## 1st Qu.  0.0730  0.0730   0.0787   0.0955
## Median   0.0815  0.0843   0.0899   0.1404
## Mean     0.0825  0.0853   0.1088   0.1465
## 3rd Qu.  0.0899  0.0955   0.1236   0.1868
## Max.     0.1573  0.1517   0.2978   0.3146
## sd       0.0163  0.0199   0.0457   0.0608
## 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      4.0
## Mean         1.1     1.5      2.4      5.2
## 3rd Qu.      1.0     1.0      3.0      8.0
## Max.         3.0     9.0     12.0     16.0
## sd           0.4     1.5      2.5      4.6
##
## family Qloss, s=0.5
## penalty= LASSO
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.     0.0337  0.0393   0.0506   0.0393
## 1st Qu.  0.0674  0.0674   0.0730   0.0843
## Median   0.0730  0.0787   0.0843   0.1124
## Mean     0.0752  0.0766   0.0925   0.1221
## 3rd Qu.  0.0843  0.0843   0.1067   0.1461
## Max.     0.1067  0.1067   0.1966   0.2584
## sd       0.0138  0.0137   0.0299   0.0476
## 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      3.0
## Median      16.5    16.5     17.0     15.0
## Mean        25.1    24.3     24.2     21.3
## 3rd Qu.     46.8    45.0     43.2     35.0
## Max.        82.0    78.0     71.0     69.0
## sd          26.1    24.9     22.9     20.6
## 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     3.0      3.0      4.0
## Median       4.0     5.0      6.0      8.5
## Mean         4.5     5.2      6.7      9.8
## 3rd Qu.      6.0     8.0      9.0     14.0
## Max.        12.0    14.0     21.0     28.0
## sd           2.5     3.0      4.4      6.5
## CV based misclassification error
##          cont-0% cont-5% cont-10% cont-15%
```

```
## Min.       0.0337  0.0449  0.0562  0.0562
## 1st Qu.    0.0730  0.0787  0.0843  0.1011
## Median     0.0843  0.0843  0.1011  0.1348
## Mean       0.0825  0.0870  0.1072  0.1415
## 3rd Qu.    0.0899  0.0955  0.1180  0.1685
## Max.       0.1236  0.1236  0.2303  0.3146
## sd         0.0139  0.0153  0.0332  0.0507
## number of variables selected by CV
##          cont-0% cont-5% cont-10% cont-15%
## Min.         1.0     1.0      2.0      1.0
## 1st Qu.      4.0     4.0      5.0      6.0
## Median       6.0     6.0      7.0     10.5
## Mean         5.8     6.7      7.9     11.3
## 3rd Qu.      7.0     9.0     10.0     16.0
## Max.        12.0    16.0     22.0     26.0
## sd           2.3     3.3      4.2      6.2
##
## family Qloss, s=0.5
## penalty= SCAD
## gamma= 3.7
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0506  0.0506  0.0506  0.0506
## 1st Qu.   0.0730  0.0730  0.0787  0.0787
## Median    0.0843  0.0843  0.0899  0.1067
## Mean      0.0823  0.0860  0.1074  0.1467
## 3rd Qu.   0.0899  0.0955  0.1067  0.2022
## Max.      0.1461  0.2921  0.3315  0.4551
## sd        0.0157  0.0265  0.0567  0.0886
## 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         5.2     6.2     10.5      7.7
## 3rd Qu.      2.0     2.2      4.2      4.0
## Max.        91.0    83.0     94.0     74.0
## sd          14.6    15.4     21.0     15.4
## number of variables selected with best error
##          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      1.4      1.1
## 3rd Qu.       1       1      1.0      1.0
## Max.          1       1     15.0      2.0
## sd            0       0      2.2      0.3
## CV based misclassification error
##          cont-0% cont-5% cont-10% cont-15%
```

```
## Min.      0.0562  0.0506  0.0506  0.0562
## 1st Qu.   0.0730  0.0787  0.0787  0.0843
## Median    0.0843  0.0871  0.0927  0.1124
## Mean      0.0858  0.0895  0.1128  0.1524
## 3rd Qu.   0.0955  0.1011  0.1138  0.2037
## Max.      0.1629  0.2921  0.3371  0.4551
## sd        0.0164  0.0268  0.0575  0.0901
## 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      1.4      1.2
## 3rd Qu.       1       1      1.0      1.0
## Max.          1       1     16.0     16.0
## sd            0       0      2.1      1.5
##
## family Qloss, s=0.5
## penalty= MCP
## gamma= 12
## best misclassification error
##         cont-0% cont-5% cont-10% cont-15%
## Min.      0.0337  0.0337  0.0449  0.0506
## 1st Qu.   0.0674  0.0674  0.0730  0.0843
## Median    0.0787  0.0787  0.0843  0.1180
## Mean      0.0791  0.0785  0.0992  0.1293
## 3rd Qu.   0.0899  0.0899  0.1081  0.1643
## Max.      0.1573  0.1404  0.2978  0.2978
## sd        0.0166  0.0161  0.0434  0.0567
## 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    17.0     12.5     11.0
## Mean       20.2    27.3     24.3     20.1
## 3rd Qu.    33.5    56.2     48.0     34.5
## Max.       82.0    78.0     74.0     71.0
## sd         28.1    28.2     25.2     21.5
## 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      3.0
## Mean        1.3     1.5      2.4      5.0
## 3rd Qu.     1.0     1.0      2.0      9.0
## Max.        4.0     6.0     13.0     17.0
## sd          0.7     1.2      2.6      4.7
## CV based misclassification error
##         cont-0% cont-5% cont-10% cont-15%
```

```
## Min.       0.0393  0.0393  0.0449  0.0562
## 1st Qu.    0.0730  0.0730  0.0787  0.0941
## Median     0.0815  0.0843  0.0899  0.1376
## Mean       0.0825  0.0848  0.1078  0.1448
## 3rd Qu.    0.0899  0.0899  0.1236  0.1812
## Max.       0.1629  0.1517  0.2978  0.3596
## sd         0.0168  0.0188  0.0447  0.0622
## 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      4.5
## Mean         1.3     1.7      2.8      5.5
## 3rd Qu.      1.0     2.0      3.0      9.0
## Max.         4.0    10.0     14.0     17.0
## sd           0.6     1.6      3.0      4.5
##
## family Logistic
## penalty= LASSO
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0337  0.0393  0.0506  0.0449
## 1st Qu.   0.0674  0.0716  0.0730  0.0997
## Median    0.0730  0.0787  0.0955  0.1236
## Mean      0.0737  0.0806  0.1028  0.1324
## 3rd Qu.   0.0801  0.0899  0.1194  0.1573
## Max.      0.1067  0.1292  0.2135  0.2640
## sd        0.0138  0.0165  0.0351  0.0473
## which lambda has best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.         2.0     6.0      2.0      6.0
## 1st Qu.      2.0    20.8     21.8     23.0
## Median      43.5    35.5     35.0     45.5
## Mean        42.5    38.6     40.1     47.4
## 3rd Qu.     70.0    56.0     59.0     70.2
## Max.        99.0    98.0     95.0    100.0
## sd          33.2    23.7     25.8     27.6
## 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     1.8      2.0      4.0
## Median       3.0     3.0      5.0      8.0
## Mean         3.6     4.1      6.6     10.8
## 3rd Qu.      5.0     6.0      8.0     18.0
## Max.        13.0    17.0     29.0     33.0
## sd           2.8     3.4      6.0      8.4
## CV based misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0337  0.0506  0.0674  0.0955
```

```
## 1st Qu.  0.0787  0.0941  0.1124  0.1348
## Median   0.0871  0.1067  0.1348  0.1629
## Mean     0.0864  0.1065  0.1378  0.1695
## 3rd Qu.  0.0955  0.1236  0.1587  0.1868
## Max.     0.1236  0.1742  0.2360  0.6404
## sd       0.0157  0.0218  0.0358  0.0635
## number of variables selected by CV
##          cont-0% cont-5% cont-10% cont-15%
## Min.         3.0     7.0      7.0      1.0
## 1st Qu.      9.0    15.0     16.8     15.0
## Median      12.0    18.0     21.0     20.0
## Mean        11.7    17.4     20.1     19.7
## 3rd Qu.     14.0    21.0     24.0     25.2
## Max.        19.0    26.0     34.0     35.0
## sd           3.6     4.3      6.2      7.9
##
## family Logistic
## penalty= SCAD
## gamma= 3.7
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0337  0.0393  0.0506   0.0449
## 1st Qu.   0.0674  0.0716  0.0730   0.0997
## Median    0.0730  0.0787  0.0955   0.1236
## Mean      0.0742  0.0806  0.1026   0.1322
## 3rd Qu.   0.0843  0.0899  0.1194   0.1573
## Max.      0.1067  0.1292  0.2135   0.2640
## sd        0.0141  0.0165  0.0351   0.0472
## which lambda has best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.         2.0     6.0      2.0      6.0
## 1st Qu.      2.0    20.8     21.8     23.0
## Median      37.0    35.5     35.0     45.5
## Mean        39.8    38.6     40.6     47.6
## 3rd Qu.     69.2    56.0     59.2     71.0
## Max.       100.0    96.0     95.0    100.0
## sd          32.8    23.9     25.9     27.8
## 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     1.8      2.0      4.0
## Median       2.0     3.0      5.0      8.0
## Mean         3.0     3.9      6.6     10.8
## 3rd Qu.      4.0     6.0      8.0     18.0
## Max.        10.0    17.0     29.0     33.0
## sd           2.3     3.2      5.8      8.4
## CV based misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0337  0.0506  0.0674   0.0955
```
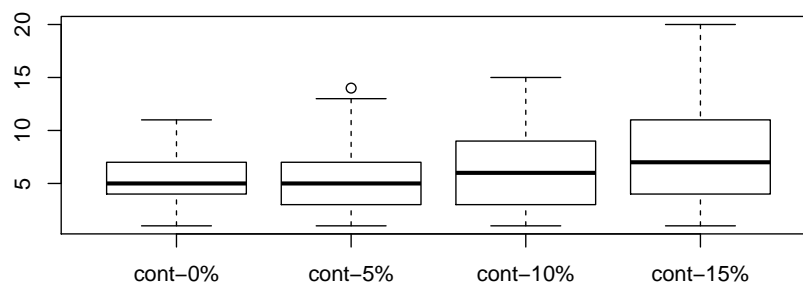
```
## 1st Qu.   0.0787  0.0899  0.1124  0.1348
## Median    0.0843  0.1067  0.1348  0.1601
## Mean      0.0854  0.1051  0.1379  0.1692
## 3rd Qu.   0.0955  0.1236  0.1587  0.1868
## Max.      0.1236  0.1742  0.2416  0.6404
## sd        0.0153  0.0225  0.0358  0.0638
## number of variables selected by CV
##          cont-0% cont-5% cont-10% cont-15%
## Min.         1.0     4.0      3.0      1.0
## 1st Qu.      6.0    12.8     16.0     15.0
## Median       9.0    17.0     20.0     20.0
## Mean         9.0    16.1     19.7     19.6
## 3rd Qu.     12.0    20.2     24.0     25.2
## Max.        18.0    26.0     34.0     35.0
## sd           4.2     5.4      6.5      7.8
##
## family Logistic
## penalty= MCP
## gamma= 12
## best misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0337  0.0337  0.0506  0.0449
## 1st Qu.   0.0618  0.0674  0.0772  0.0955
## Median    0.0758  0.0787  0.0955  0.1292
## Mean      0.0747  0.0803  0.1034  0.1324
## 3rd Qu.   0.0843  0.0899  0.1236  0.1587
## Max.      0.1067  0.1517  0.2247  0.2753
## sd        0.0140  0.0178  0.0371  0.0487
## which lambda has best error
##          cont-0% cont-5% cont-10% cont-15%
## Min.         2.0     5.0      2.0      6.0
## 1st Qu.      2.0    19.8     24.0     23.0
## Median      54.0    32.0     38.0     46.5
## Mean        49.1    37.3     43.6     48.6
## 3rd Qu.     79.0    51.2     67.2     72.0
## Max.        99.0    99.0     99.0    100.0
## sd          35.6    22.8     26.4     27.9
## 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     2.0      4.0      8.0
## Mean         2.9     3.0      5.8      9.2
## 3rd Qu.      4.0     3.2      8.0     14.0
## Max.        11.0    15.0     23.0     28.0
## sd           2.2     2.9      5.0      7.1
## CV based misclassification error
##          cont-0% cont-5% cont-10% cont-15%
## Min.      0.0393  0.0562  0.0618  0.0843
```

```
## 1st Qu.  0.0730  0.0899  0.1124  0.1348
## Median   0.0843  0.1067  0.1404  0.1657
## Mean     0.0847  0.1057  0.1406  0.1744
## 3rd Qu.  0.0955  0.1194  0.1629  0.1966
## Max.     0.1348  0.1742  0.2584  0.6404
## sd       0.0158  0.0224  0.0387  0.0652
## number of variables selected by CV
##         cont-0% cont-5% cont-10% cont-15%
## Min.       1.0     5.0      4.0      1.0
## 1st Qu.    6.0    10.0     11.8     10.8
## Median     8.0    13.0     16.0     17.0
## Mean       8.4    12.9     15.9     15.6
## 3rd Qu.   11.0    15.0     20.0     21.0
## Max.      15.0    21.0     31.0     31.0
## sd         3.0     4.0      5.6      6.9
print(proc.time() - ptm)
##      user    system   elapsed
## 16889.363  2139.966 20017.909
```

Misclassification error
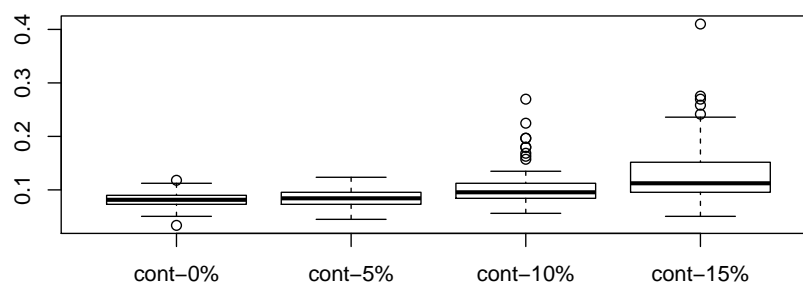
Closs−LASSO

No. variables

Closs−LASSO

17

Misclassification error

Closs−SCAD

No. variables
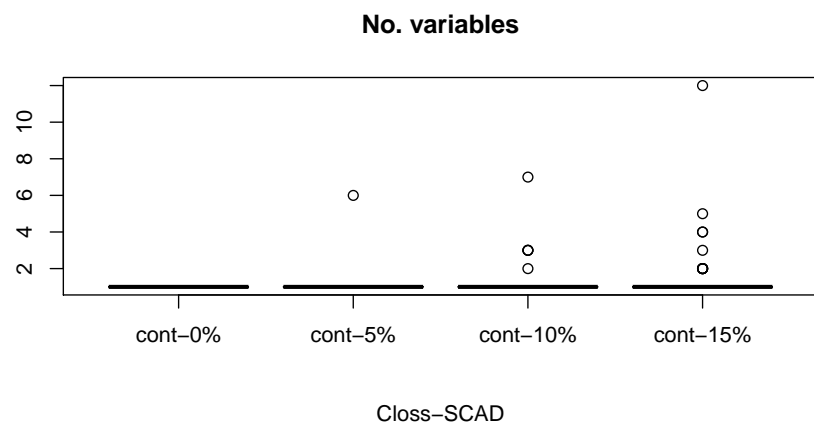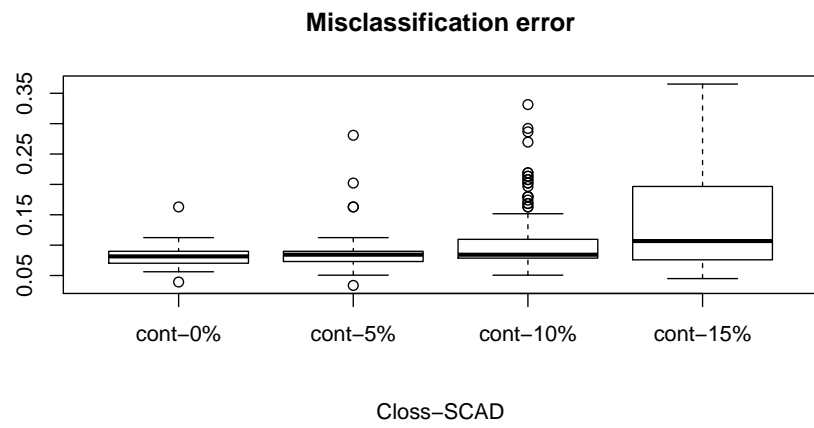
Closs−SCAD

Misclassification error

Closs−MCP

No. variables
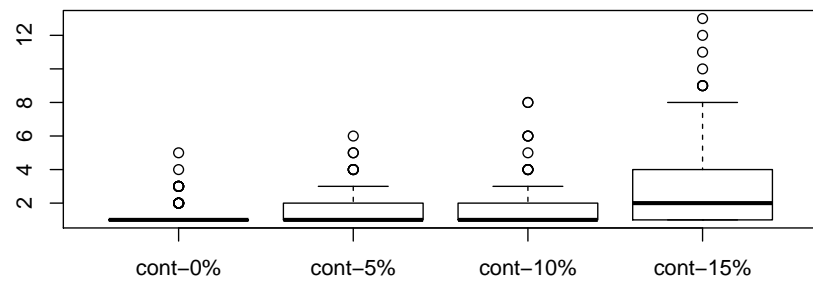
Closs−MCP
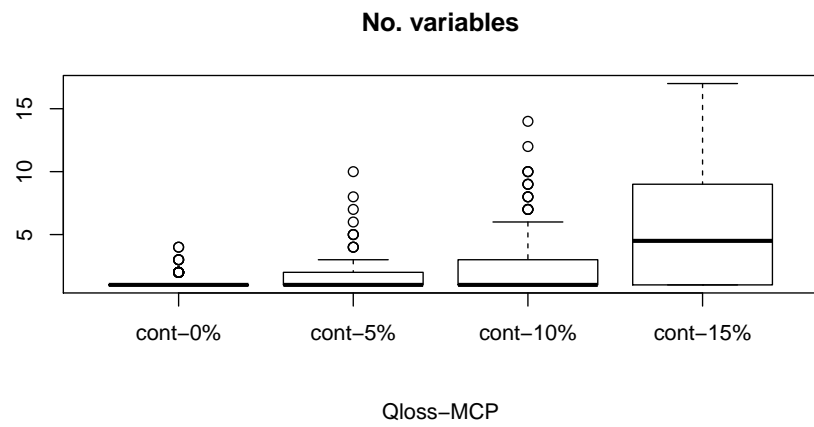
Misclassification error
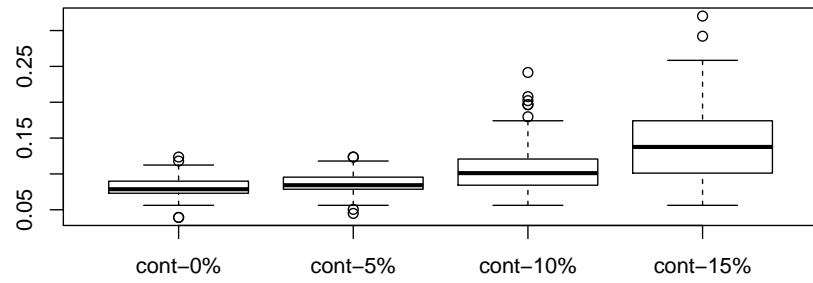
Gloss−LASSO

No. variables
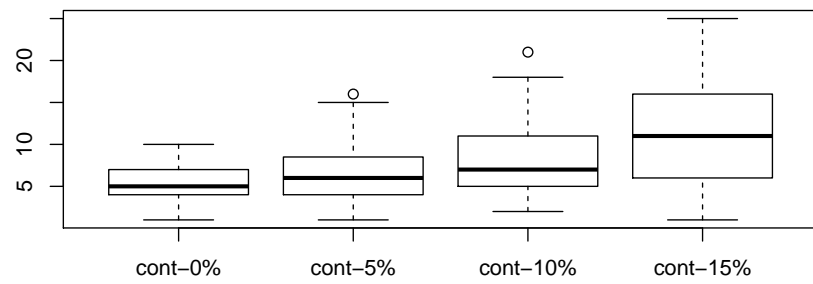
Gloss−LASSO

Misclassification error
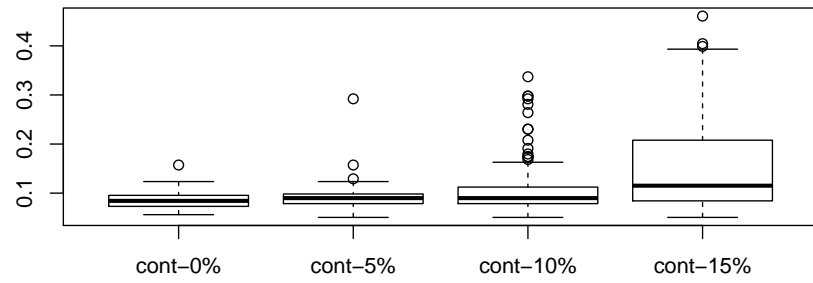
Gloss−SCAD

No. variables

Gloss−SCAD

21
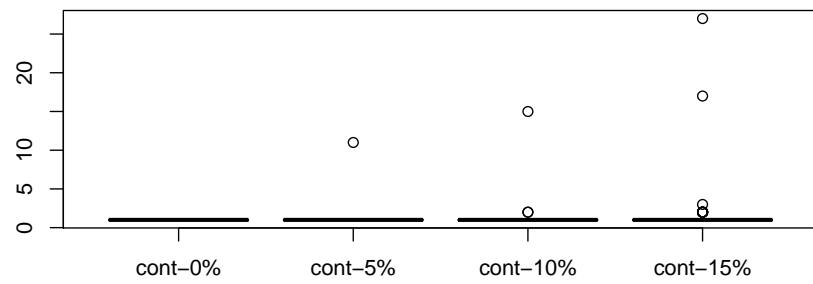
Misclassification error
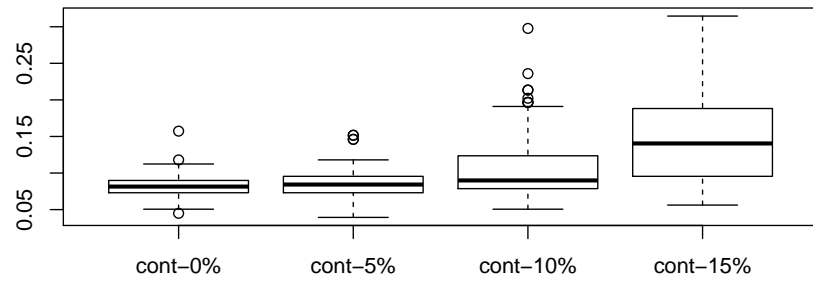
Gloss−MCP

No. variables
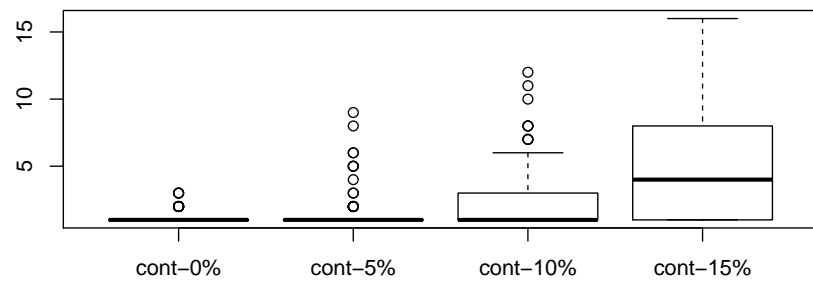
Gloss−MCP

Misclassification error
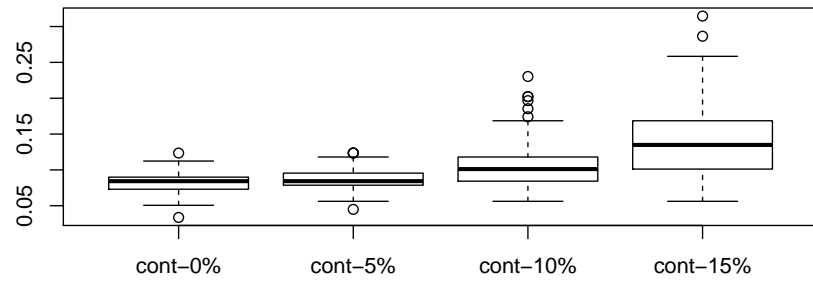
Qloss−LASSO

No. variables

Qloss−LASSO

23

Misclassification error
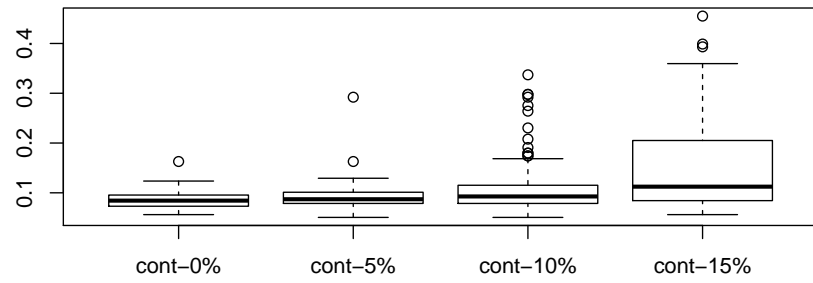
Qloss−SCAD

No. variables

Qloss−SCAD

Misclassification error

Qloss−MCP

No. variables

Qloss−MCP

25
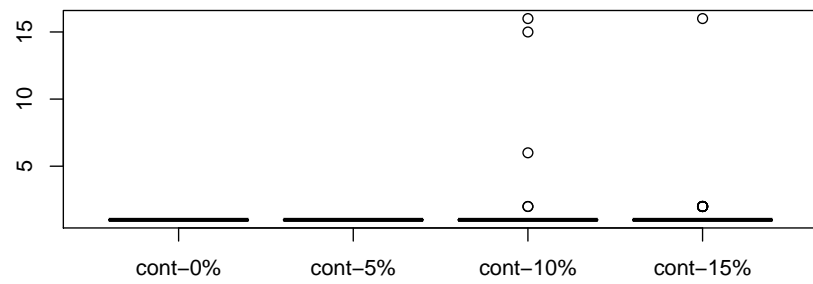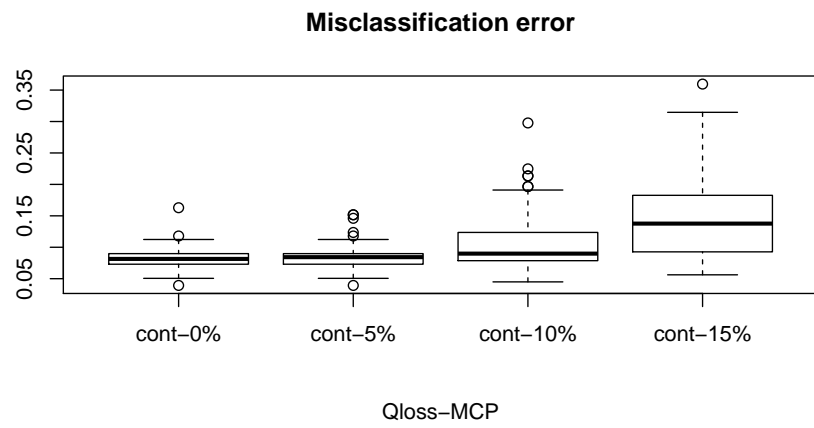
Misclassification error
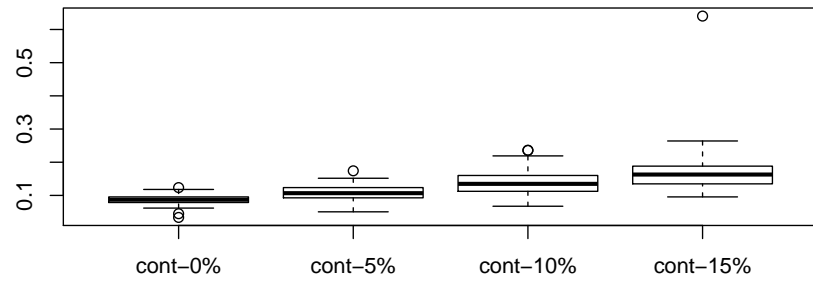
Logistic−LASSO

No. variables

Logistic−LASSO

26
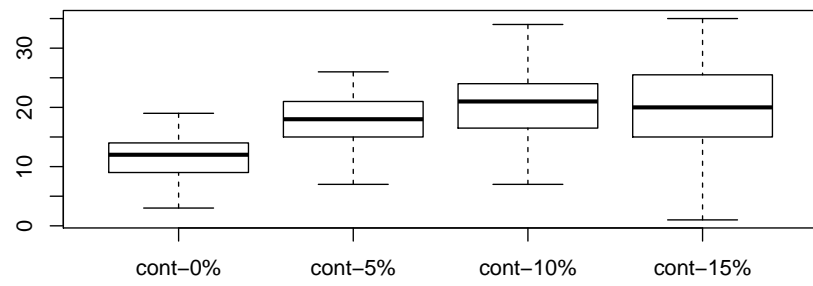
Misclassification error
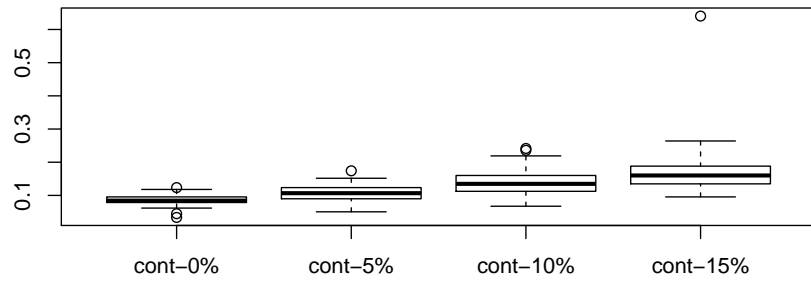
Logistic−SCAD

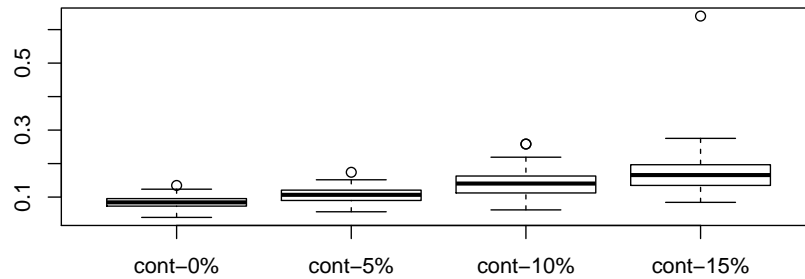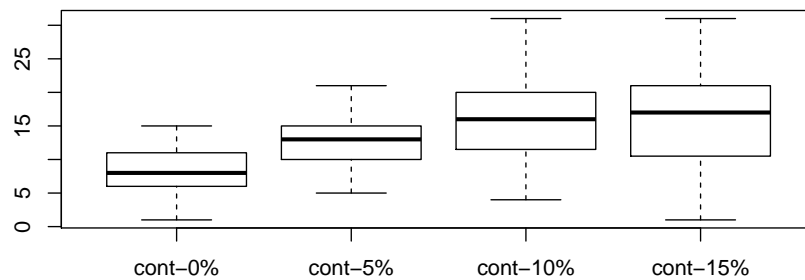**No. variables**

Logistic−SCAD

27
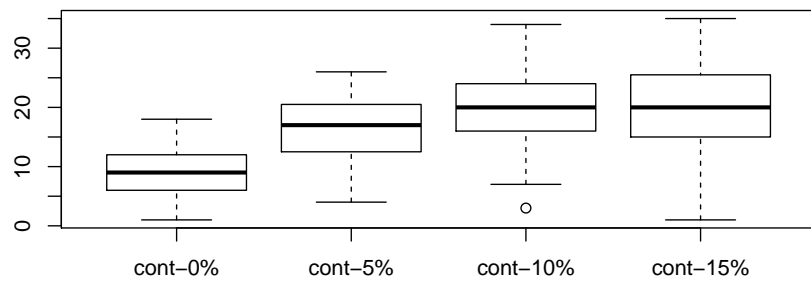
Misclassification error

Logistic–MCP

No. variables

Logistic–MCP

```
sessionInfo()
## R version 3.5.2 (2018-12-20)
## 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_0.3-8  gdata_2.18.0 knitr_1.21
## 
## loaded via a namespace (and not attached):
##  [1] magrittr_1.5      splines_3.5.2     MASS_7.3-51.1
##  [4] doParallel_1.0.14 bst_0.3-16        pscl_1.5.2
##  [7] gbm_2.1.5         lattice_0.20-38   foreach_1.4.4
## [10] stringr_1.4.0     tools_3.5.2       parallel_3.5.2
## [13] grid_3.5.2        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_1.2-15
## [22] gridExtra_2.3     formatR_1.5       codetools_0.2-16
## [25] rpart_4.1-13      evaluate_0.13     stringi_1.3.1
## [28] compiler_3.5.2
```

References

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