R Code for 'Nonparametric Notion of Residual and Test for Conditional Independence'

In this article, we discuss the R implementation of the test statistic for the test of conditional indendence of X, Y given Z. In the present version the codes provided in this page can only be used when X and Y are real valued and $Z \in \mathbb{R}^d$ for $d \geq 1$.

In the follwoing, we generate $n = 100$ i.i.d. copies of (X, Y, Z) when $d = 2$. Moreover, we assume that $Z \sim N(0, \sigma_z^2 \mathbf{I}_{5\times 5})$, $X = W + Z_1 + \epsilon$, and $X = W + Z_1 + \epsilon'$, where Z_1 is the first coordinate of Z and ϵ, ϵ', and W are three independent mean zero Gaussian random variables. Moreover, we assume that ϵ, ϵ', and W are independent of Z, and $var(\epsilon) = var(\epsilon') = \sigma_E^2$, and $var(W) = \sigma_W^2$. Note that this is the simulation scenario used in Section 4 of Patra, Sen, and Székely (2015). Note that X and Y are conditionally independent of Z only when $\sigma_W = 0$. In the following, we fixed σ_W to be 0.1.

```
n <-100
d <- 2
sigma.Z <- 0.3
sigma.W <- 0.1
sigma.E <- 0.2
Z <- matrix(rnorm(n*d,0,sigma.Z),nr=n, nc=d)
W <- rnorm(n,0,sigma.W)
eps <- rnorm(n,0,sigma.E)
eps.prime <- rnorm(n,0,sigma.E)
X <- W + Z[,1] +eps
Y <- W + Z[,1] +eps.prime
Data <- cbind(X,Y,Z)
colnames(Data) <- c('X', 'Y', paste('Z.', 1:d,sep=''))
head(Data)
```

```
##                X          Y         Z.1         Z.2
## [1,] -0.15577811 -0.3921237 -0.24730964 -0.06967209
## [2,] -0.04944131  0.0404184 -0.09926920  0.40181352
## [3,] -1.24943962 -1.1810316 -0.95382946 -0.03040068
## [4,]  0.78575367  0.4238296  0.57912533  0.16761585
## [5,]  0.12827329  0.1109989  0.03614325 -0.08641120
## [6,]  0.25492997  0.2665227  0.17681404  0.08911656
```

In the following we calculate the test statistic $\hat{\mathcal{E}}_n$, see (3.7) of Patra, Sen, and Székely (2015).

```
source('Npres_Fucntions.R')
```

```
## Nonparametric Kernel Methods for Mixed Datatypes (version 0.60-2)
## [vignette("np_faq",package="np") provides answers to frequently asked questions]
```

```
test.stat <- npresid.statistics(Data,d)
```

As the limiting behavior of $\hat{\mathcal{E}}_n$ is unknown, in the following we approximate the asymptotic distribution through a model based bootstrap procedure (see Section 3.2.1 of Patra, Sen, and Székely (2015)) and evaluate the p-value of the proposed test. In the following "boot.replic" denotes the number of bootstrap replications. We recommend using a bootstrap replication of size 1000.

```
out <- npresid.boot(Data,d,boot.replic=50)
```

```
## [1] "Starting bootstrap"
## [1] "50 bootstrap samples obtained"
## [1] "At bootstrap iteration 25 of 50"
## [1] "At bootstrap iteration 50 of 50"
```

```
str(out, max.level = 1)
```

```
## List of 7
##  $ statistic           : num 4.3
##  $ p.value             : num 0.78
##  $ method              : chr "Cond Indep test: p-values by inverting F_hat to get bootstrap samples
##  $ bandwidth.method    : chr "least-squares cross-validation, see \"np\" package"
##  $ data.desrip         : chr "dimension of Z is 2, sample size 100, dimension of (X,Y,Z) is 4, boots
##  $ bootstrap.stat.values: num [1:50] 4.03 8.76 3.82 7.75 1.53 ...
##  $ cond.dist.obj       :List of 6
```

Here "cond.dist.obj" is the the list conataining estimators of $F_{X|Z}$, $F_{Y|Z}$, and F_Z evaluated at the data points (denoted by F.x_z, F.y_z, and F.z_hat) and the bandwidth used (denoted by Fbw.x_z, Fbw.y_z, and Fbw.z_z) to evaluate the conditional distribution fucntions. Note that we use the functions available in the "np" package (see Hayfield and Racine (2008)) to compute the optimal bandwiths as well as the estimates of the conditional distribution functions.

```
str(out$cond.dist.obj, max.level = 1)
```

```
## List of 6
##  $ F.x_z  : num [1:100] 0.594 0.466 0.5 0.589 0.556 ...
##  $ F.y_z  : num [1:100] 0.246 0.586 0.5 0.47 0.532 ...
##  $ Fbw.x_z:List of 64
##   ..- attr(*, "class")= chr "condbandwidth"
##  $ Fbw.y_z:List of 64
##   ..- attr(*, "class")= chr "condbandwidth"
##  $ Fbw.z_z:List of 2
##  $ F.z_hat: num [1:100, 1:2] 0.227 0.371 0.005 0.968 0.533 ...
```

The estimated p-value of the test proceudure is given through

```
p.value <- out$p.value
```

The required R file "Npres_Fucntions.R" can be downloaded here. Note the fucntions require the following R-packages: boot, data.table, energy, np, and stats.

References

Hayfield, Tristen, and Jeffrey S. Racine. 2008. "Nonparametric Econometrics: The Np Package." *Journal of Statistical Software* 27 (5). http://www.jstatsoft.org/v27/i05/.

Patra, Rohit K., Bodhisattva Sen, and Gábor Székely. 2015. "A Consistent Bootstrap Procedure for the Maximum Score Estimator." *Statist. Probab. Lett. (to Appear)*. http://arxiv.org/abs/1409.3886.