Introduction: HyperCube

Chi Po Choi, Amy T. Kim

2015-06-10

Introduction

The R package HyperCube is an implementation of the Hypercube Estimator introduced by (Beran 2014). Hypercube estimator is a richer class of regularized estimators of linear model, which extends penalized least squares estimators with quadratic penalties. The R package HyperCube lets users Hypercube Estimators to fit linear model.

In this document, we first briefly introduce the theoretical background of Hypercube Estimator. Then we demonstrate how to use the R package HyperCube to produce the examples given in (Beran 2014).

Theoretical background

Motivation

Given a data set (y, X) where y is the $n \times 1$ vector of observations, X is a given $n \times p$ design matrix of rank $p \leq n$, we fit the linear model:

$$y = X\beta + \epsilon.$$

with the components of ϵ are independent $\sim (0, \sigma^2)$, and finite fourth moment.

Let $\eta = \mathrm{E}(y) = X\beta$.

The least squares estimator of η is $\hat{\eta}_{LS} = X(X'X)^{-1}X'y$. However, $\hat{\eta}_{LS}$ usually overfits. If the error vector ϵ is Gaussian and $p \geq 3$, then $\hat{\eta}$ is an inadmissible estimator of η under the quadratic risk function $R(\hat{\eta}) = E|\hat{\eta} - \eta|^2$. In other words, the least square estimator of η does not minimized the risk.

We would like to find an estimator $\hat{\eta}$ so that the risk $\mathrm{E}|\hat{\eta} - \eta|^2$ is minimized. We may sacrifice the unbiasedness (least square estimator) to obtain a risk-minimizing estimator of $\hat{\eta}$. It is the so-call bias-variance trade-off. The Hypercube estimator is a class of $\hat{\eta}$ which performs much better than least square estimator in term of minimizing the risk.

Definition of the Hypercube Estimator

Define the hypercube estimator of η to be

$$\hat{\eta}_{\mathrm{H}}(V) = A(V)y \quad \text{with} \quad A(V) = XV(VX'XV + I_p - V^2)^{-1}VX'$$

where V is a symmetric matrix with all eigenvalues $\in [0, 1]$ and A(V) is called the operator. We can compute the V so that the risk $\mathrm{E}|\hat{\eta}_{\mathrm{H}}(V) - \eta|^2$ is minimized. Thus, we obtain an estimator which is better than the least square estimator.

Let $\eta = \mathrm{E}(y) = X\beta$. We would like to find an Hypercube Estimator $\hat{\eta}$,

$$\hat{\eta}_{\mathrm{H}}(V) = XV(VX'XV + I_p - V^2)^{-1}VX'y$$

where V is a symmetric matrix with all eigenvalues $\in [0, 1]$. (That's why it is named Hypercube Estimator.)

Penalized Least Squares Estimator

Let W be any $p \times p$ positive semidefinite matrix, and the associated penalized least squared (PLS) estimators of η

$$\hat{\eta}_{\text{PLS}} = X\hat{\beta}_{\text{PLS}}$$

where

$$\hat{\beta}_{\text{PLS}} = \text{argmin}_{\beta}[|y - X\beta|^2 + \beta' W \beta] = (X'X + W)^{-1}X'y$$

The mapping from $\hat{\beta}_{\text{PLS}}(W)$ to $\hat{\beta}_{\text{PLS}}(W)$ is one-to-one. The matrix $V = (I_P + W)^{-\frac{1}{2}}$ is symmetric with all eigenvalues in $(0, 1]$.

$$\hat{\eta}_{\text{PLS}}(W) = \hat{\eta}_H((I_P + W)^{-\frac{1}{2}})$$

Minimizing the estimated risk over V

The normalized quadratic risk is

$$R(\hat{\eta}_H, \eta, \sigma^2) = p^{-1}E|\hat{\eta}_H(V) - \eta|^2 = p^{-1}tr[\sigma^2 A^2(V) + (I_n - A(V))^2\eta\eta']|.$$

The risk depends on the unknown parameters η and σ^2. We consider the normalized estimated risk

$$\hat{R_H}(V) = p^{-1}[|y - A(V)y|^2 + \{2tr(A(V)) - n\}\hat{\sigma}^2]$$

where $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

The goal of the Hypercube Estimator is to choose V for the Hyercube Estimator to obtain $\hat{\eta}$ with smaller estimated risk.

Demonstration of Examples

Example 1

It is the Example 1 in (Beran 2014).

In the Canadian earnings data considered by (Ullah 1985), we consider a linear model on log(incomes) versus ages.

A model for the data is

$$y = Cm + e$$

Here y is $n \times 1$ vector of observation, m is the $p \times 1$ vector of mean, C is the $n \times p$ data-incidence matrix with elements 0 or 1. The is a special case of linear model in which $X = C, \text{and} \beta = m$

Consider the $(g - 1) \times g$ difference matrix $\Delta(g) = \{\delta_{u,w}\}$ in which $\delta_{u,u} = 1, \delta_{u,u+1} = -1$ for every u and all other entries are zero.

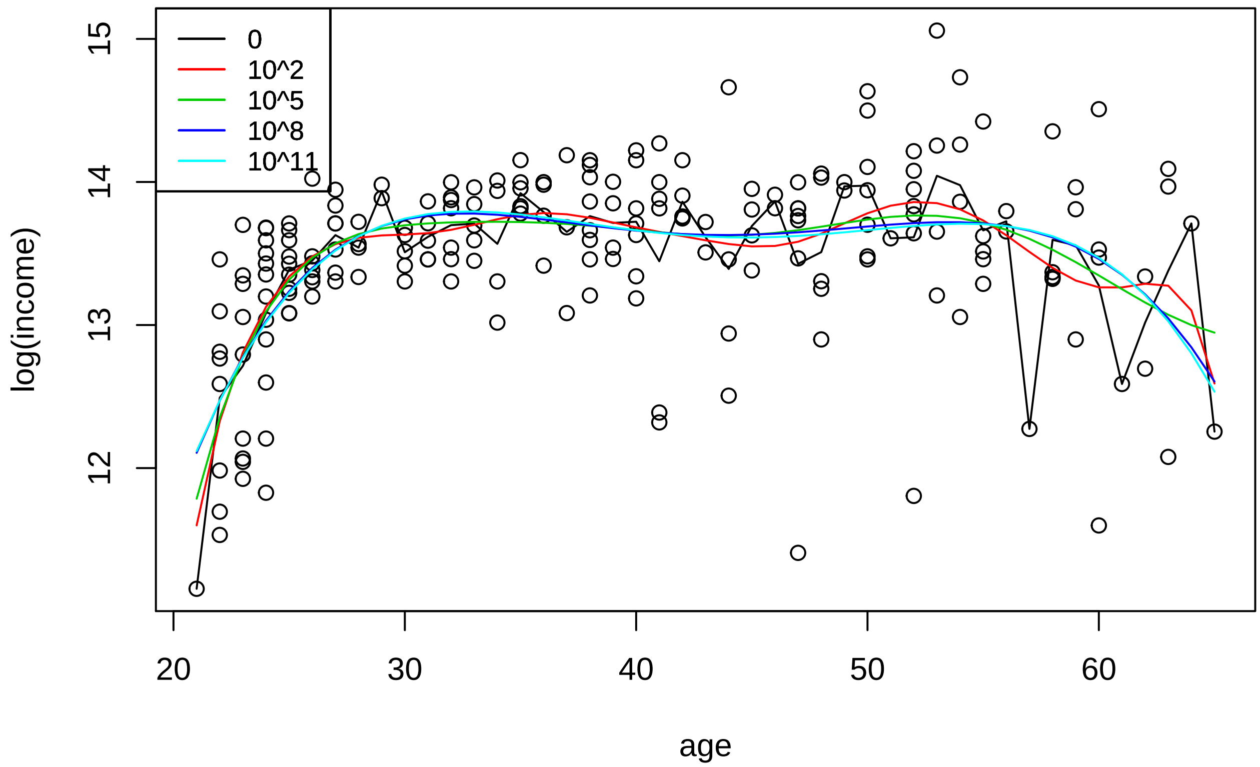
Here, define $D_5 = \Delta(p - 4)\Delta(p - 3)\Delta(p - 2)\Delta(p - 1)\Delta(p)$ with $p = 45$

Let $W(v) = vD_5'D_5$, for every $v \geq 0$

$$m\hat{_{\text{PLS}}}(W(\nu)) = \text{argmin}[|y - Cm|^2 + \nu|D_5 m|^2] = (C'C + W(\nu))^{-1}C'y$$

We use Hypercube estimator to obtain fits to the Canadian earnings data for various values of penalty weight ν.

```r
library(HyperCube)

# The package includes the data set canadian.earnings.
# The age is considered as factor in the data set.
canadian.earnings$age <- factor(canadian.earnings$age)

# Plot the data
plot(as.numeric(as.character(canadian.earnings$age)), canadian.earnings$log.income,
     xlab = "age", ylab = "log(income)")

# The number of ages in the data set, p, in Example 1 in Beran (2014).
p <- length(unique(canadian.earnings[,1]))

# D_5 as in equation (3.10) in Beran (2014)
D <- diffMatrix(p, 5)

# The parametor nu in equation (3.11) in Beran (2014)
nu <- c(0, 10^c(2,5,8,11))

# Plotting Hypercube Estimator fits for varying nu
lcolor <- 1:5
for(k in 1:5) {

  # The matrix W in equation (3.11) in Beran (2014)
  W <- nu[k] * t(D) %*% D

  # Convert W to V, as described in (1.6) in Beran (2014)
  V <- plsW2V(W)

  # Hyperpercube Estimator Fit
  hcmod <- hypercube( log.income ~ age -1, data=canadian.earnings, V)

  # Plot the fits
  lines(as.numeric(levels(canadian.earnings$age)),
        hcmod$coefficients, col = lcolor[k])
  legend("topleft", cex = 0.8,
         legend = c("0", "10^2", "10^5", "10^8", "10^11"),
         lty = rep(1,5), col=1:5)
}
```



Example 2

It is the Example 2 in (Beran 2014).

We consider the rat litter data treated by (Scheffé 1959). Each response recorded is the average weight-gain of a rat litter when the infants in the litter are nursed by a rat foster-mother. Factor 1, with four levels, is the genotype of the foster-mother. Factor 2, with the same levels, is the genotype of the infant litter.

Two-way ANOVA considers competing least squares fits to the rat litter data. Let $u_r = (1/2, 1/2, 1/2, 1/2)'$ for $r = 1, 2$. Set $J_r = u_r u_r'$ and $H_r = I_4 - J - R$. The standard ANOVA projections are

$$P_1 = J_2 \otimes J_1, \qquad P_2 = J_2 \otimes H_1, \qquad P_3 = H_2 \otimes J_1, \qquad P_4 = H_2 \otimes H_1.$$

The $\{P_k\}$ are symmetric, idempotent, mutually orthogonal matrices such that $\sum_{k=1}^4 P_k = I$. Let $d = (d_1, d_2, d_3, d_4) \in [0,1]^4$ and $V(d) = d_1 V_1 + d_2 V_2 + d_3 V_3 + d_4 V_4$. We want to minimize the risk $\hat{\eta}_H(V(d))$ over $d \in [0,1]^4$.

```
library(HyperCube)

# The package includes the data set litter.
# The formula specifying the two-way layout is "weight ~ mother:infant -1".
# hypercubeOptimization computes the optimal d
hcmodopt <- hypercubeOptimization( weight ~ mother:infant -1, data = litter)

# The optimal d
# Same result as stated in Example 2 in Beran (2014)
hcmodopt$projcoef
```

```
## [1] 0.9971043 0.6931901 0.0000000 0.4150027
```

```r
# Compare the estimated risk
summary(hcmodopt$est)
```

```
## Call:
## hypercube.formula(formula = formula, data = data, V = V)
##
## The estimated risk of hypercube estimation: 16.1214335561813
## The estimated risk of least square estimation:  54.2403666666667
```

References

Beran, Rudolf. 2014. "Hypercube Estimators: Penalized Least Squares, Submodel Selection, and Numerical Stability." *Computational Statistics & Data Analysis* 71. Elsevier: 654–66.

Scheffé, H. 1959. "The Analysis of Variance. a Wiley Publication in Mathematical Statistics." Wiley.

Ullah, Aman. 1985. "Specification Analysis of Econometric Models." *Journal of Quantitative Economics* 1: 187–209.