36-402 Homework 6

Eu Jing Chua
eujingc

February 25, 2019

```
## Nonparametric Kernel Methods for Mixed Datatypes (version 0.60-9)
## [vignette("np_faq",package="np") provides answers to frequently asked questions]
## [vignette("np",package="np") an overview]
## [vignette("entropy_np",package="np") an overview of entropy-based methods]

## Loading required package: nlme

## This is mgcv 1.8-26. For overview type 'help("mgcv-package")'.
```

Question 1

Q1 a)

Table 1: In-sample MSE

	x
Power-law Model	0.05392
NP Regression	0.05116

Q1 b)

Table 2: Probability of observed MSE difference under Power-law Model

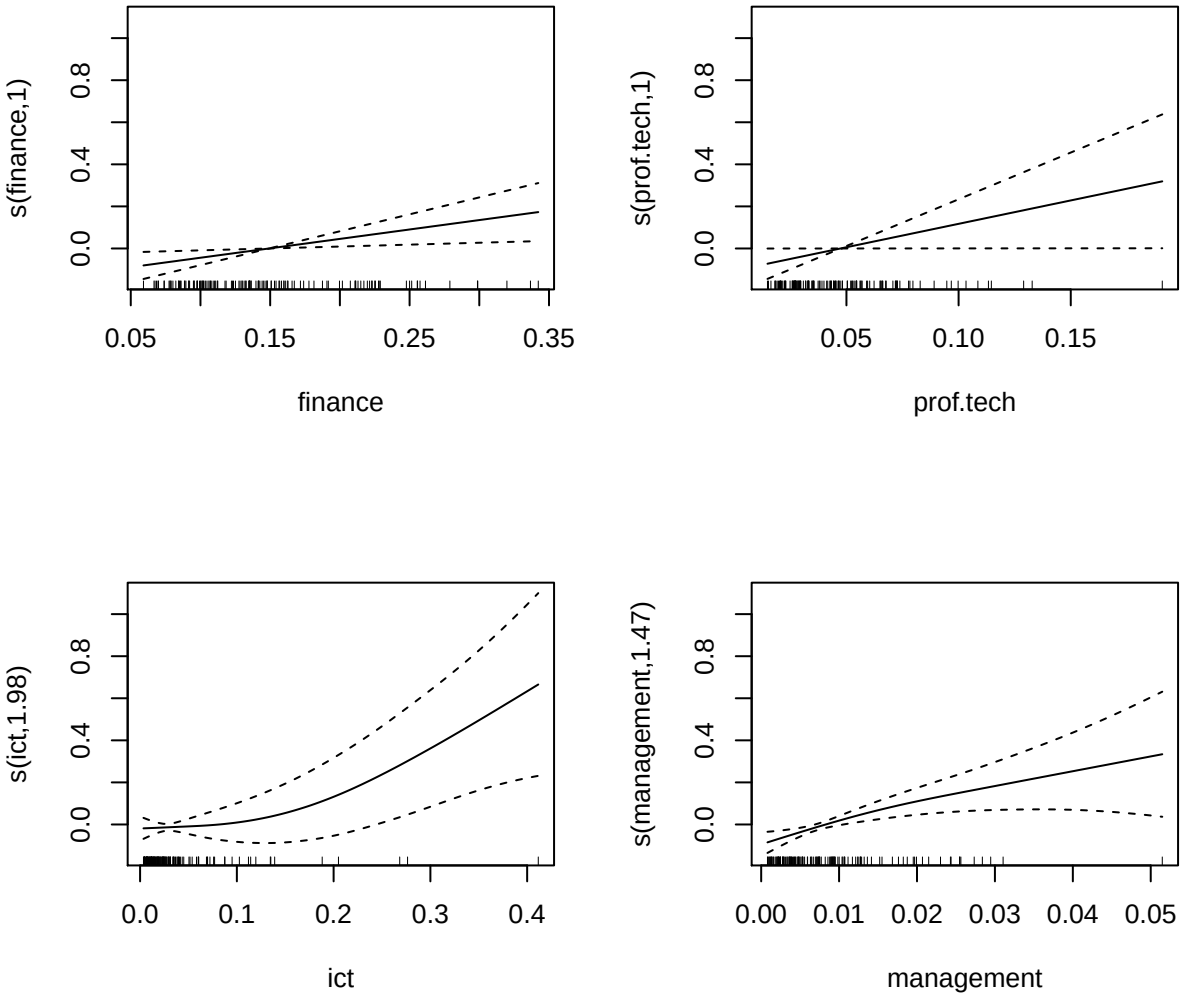
x
0.075

Q1 c)

Assuming we are conducting this hypothesis test at 5% significance, then we conclude that we do not reject that the true model follows a power law, as we have insufficient evidence to reject that.

Question 2

Q2 a)

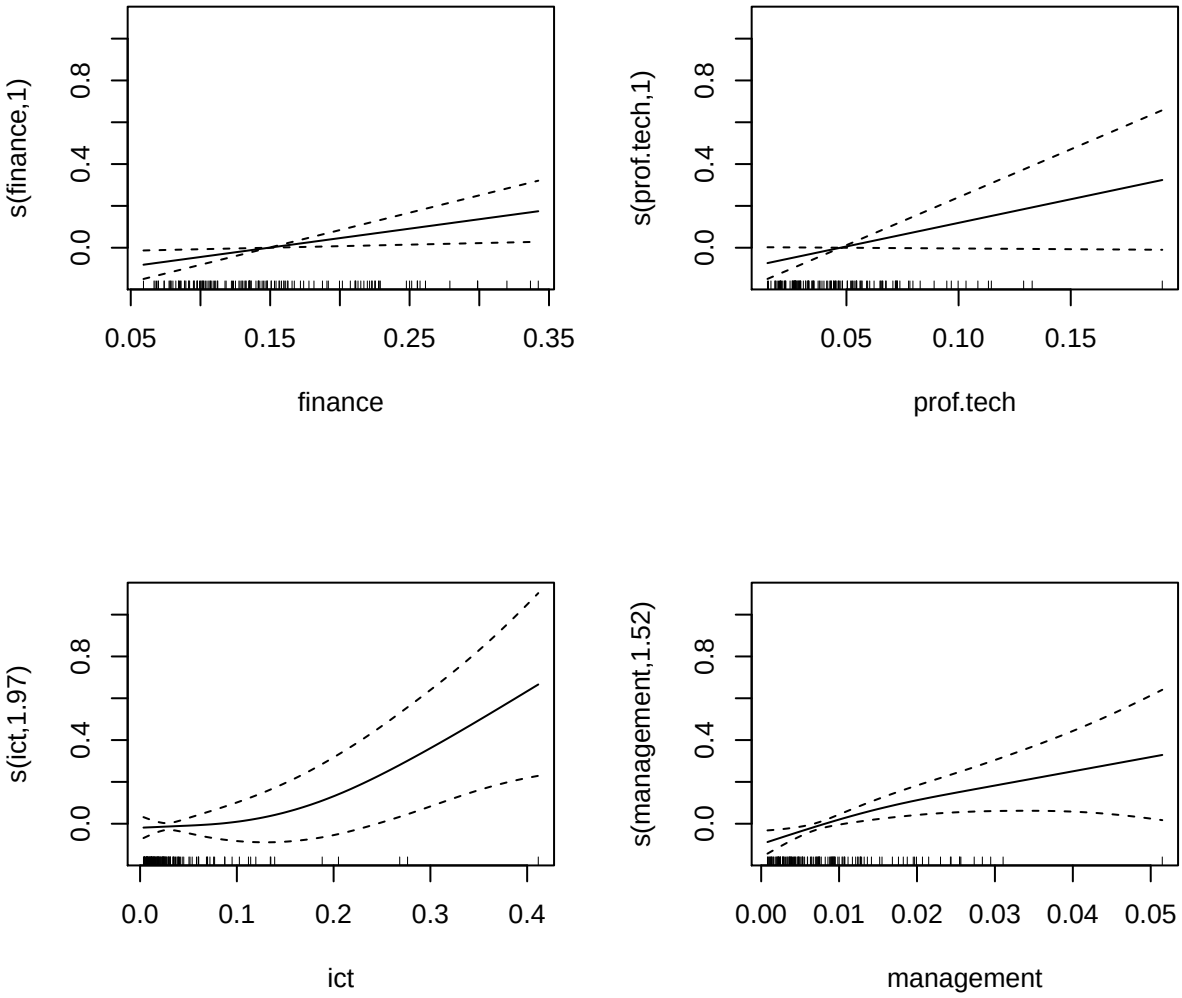


The partial responses of finance, prof.tech, and management all look quite linear with positive slopes. However, the partial response of ict does not look linear but is concave up and increasing.

Q2 b)

Table 3: Coefficient

	x
log(pop)	-0.00266



It seems that the partial response functions did not change much with the additional term added.

Q2 c)

For the bootstrap, we use case-resampling of the original data to generate new samples to produce an estimate of the following:

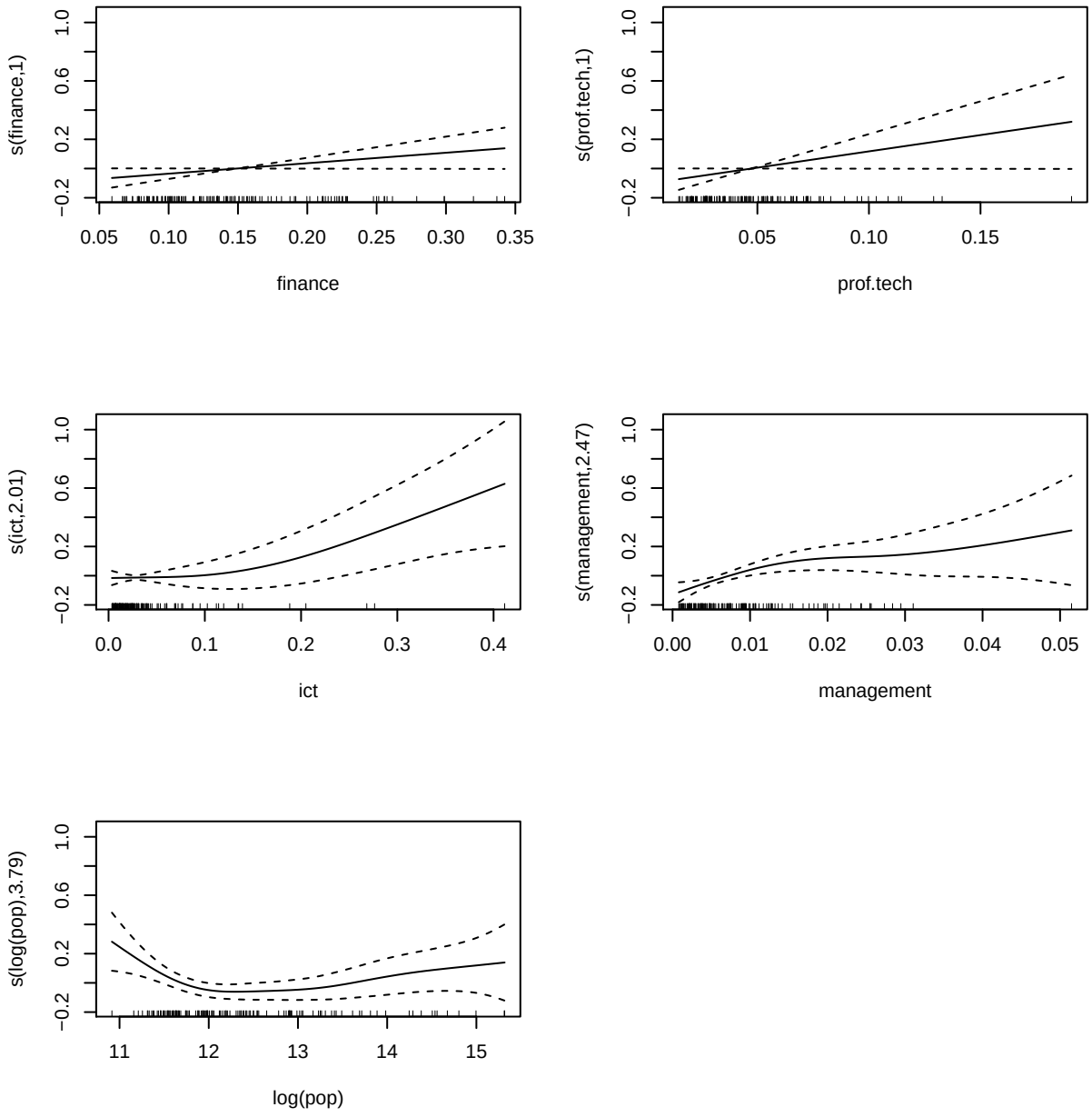
Table 4: 95% C.I. for Coefficient

	lower	upper
log(pop)	-0.03811	0.10015

Q2 d)

The CI from before does include 0. We can conclude that power-law scaling with population size does not seem to be significant in the presence of the other smoothed predictors. Also, this also shows that the power-law model may not be strictly supra-linear scaling, as the confidence interval does not only have positive values in it.

Q2 e)



The partial response functions of finance, prof.tech and ict does not seem to have changed much. However, the function for management is now more non-linear, generally increasing while being concave down then concave up.

The partial response function of log(pop) is non-linear, being roughly concave up where it decreases up till around 13 log population size, before increasing again.

Q2 f)

It seems that population size is a weak but real determinant of a city's per-capita output. In the model of 2 d), we see that in the presence of the other smoothed predictors, the coefficient for the linear term of

log(pop) is non-significant. However, when we model additive model with a smoothed log(pop) term as in 2 e), we see that it has a partial response function that seems significantly non-zero for certain portions such as lower values of log population from 11 to 11.5, as seen from the 2 standard error band around it.

The idea of strictly supra-linear scaling is also not supported, as the coefficient of log(pop) is not strictly positive. Hence, it is weak but still acts as a determinant of a city's per-capita output.

Question 3

Table 5: Probability of observed MSE difference under linear model

x
0.12

Question 4

Q4 a)

$$\mu_j(x_j) = \mathbb{E}\left[Y \mid X_j = x_j\right] \tag{1}$$

$$= \alpha + \sum_{k=1}^{p} \mathbb{E}\left[f_k(X_k) \mid X_j = x_j\right] + 0 \tag{2}$$

$$= \alpha + f_j(x_j) + \sum_{k \neq j}^{p} \mathbb{E}\left[f_k(X_k) \mid X_j = x_j\right] \tag{3}$$

Q4 b) Since X_k is independent of X_j for $k \neq j$, we know that $\mathbb{E}\left[f_k(X_k) \mid X_j = x_j\right] = \mathbb{E}\left[f_k(X_k)\right] = 0$.

$$\mu_j(x_j) = \alpha + f_j(x_j) + \sum_{k \neq j}^{p} \mathbb{E}\left[f_k(X_k) \mid X_j = x_j\right] \tag{4}$$

$$= \alpha + f_j(x_j) \tag{5}$$

$$\mu_j(x_j) - \alpha = f_j(x_j) \tag{6}$$

Q4 c) Consider a model where $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, where $X_1 = X_2$.
If 4b) applied in this case, we would find that $\mu_1(x_1) = \alpha + \beta_1 x_1$.
However, since $X_1 = X_2$, $\mathbb{E}\left[\beta_2 X_2 \mid X_1 = x_1\right] = \beta_2 x_1$, so it should be that $\mu_1(x_1) = \alpha + \beta_1 x_1 + \beta_2 x_1$.

Hence, 4b) does not hold if some X_k is statistically dependent of X_j.

Question 5

Q5 a)

$$\hat{\beta}_{RR} = \underset{\beta}{\operatorname{argmin}} \left[\left[\frac{1}{n} \sum_{i=1}^{n} (y_i - x_i \cdot \beta)^2 \right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right] \tag{7}$$

$$= \underset{\beta}{\operatorname{argmin}} \left[n^{-1} (\mathbf{y} - \mathbf{x}\beta)^T (\mathbf{y} - \mathbf{x}\beta) + \lambda \beta^T \beta \right] \tag{8}$$

Q5 b)

$$\hat{\beta}_{RR} = \underset{\beta}{\operatorname{argmin}} \left[n^{-1}(\mathbf{y} - \mathbf{x}\beta)^T(\mathbf{y} - \mathbf{x}\beta) + \lambda\beta^T\beta \right] \tag{9}$$

$$0 = \frac{\partial}{\partial\beta} n^{-1}(\mathbf{y} - \mathbf{x}\beta)^T(\mathbf{y} - \mathbf{x}\beta) + \lambda\beta^T\beta \tag{10}$$

$$= -n^{-1}2\mathbf{x}^T(\mathbf{y} - \mathbf{x}\beta) + 2\lambda\beta \tag{11}$$

$$n^{-1}\mathbf{x}^T\mathbf{y} = n^{-1}\mathbf{x}^T\mathbf{x}\beta + \lambda\beta \tag{12}$$

$$\mathbf{x}^T\mathbf{y} = \mathbf{x}^T\mathbf{x}\beta + n\lambda\mathbf{I}\beta \tag{13}$$

$$\mathbf{x}^T\mathbf{y} = (\mathbf{x}^T\mathbf{x} + n\lambda\mathbf{I})\beta \tag{14}$$

$$\implies \hat{\beta}_{RR} = (\mathbf{x}^T\mathbf{x} + n\lambda\mathbf{I})^{-1}\mathbf{x}^T\mathbf{y} \tag{15}$$

Q5 c)

As $\lambda \to 0$, $\hat{\beta}_{RR} \to (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} = \hat{\beta}_{OLS}$, we reduce to linear regression by OLS.

We know that $\mathbf{x}^T\mathbf{x}$ is symmetric and positive semi-definite, hence $\mathbf{x}^T\mathbf{x} + n\lambda\mathbf{I}$ is positive definite for $\lambda > 0$. Given that $\mathbf{x}^T\mathbf{x}$ has eigenvalues $\lambda_i \geq 0$, $i = 1 \dots p$, the positive definite matrix $\mathbf{x}^T\mathbf{x} + n\lambda\mathbf{I}$ has eigenvalues $\lambda_i + \lambda > 0$.

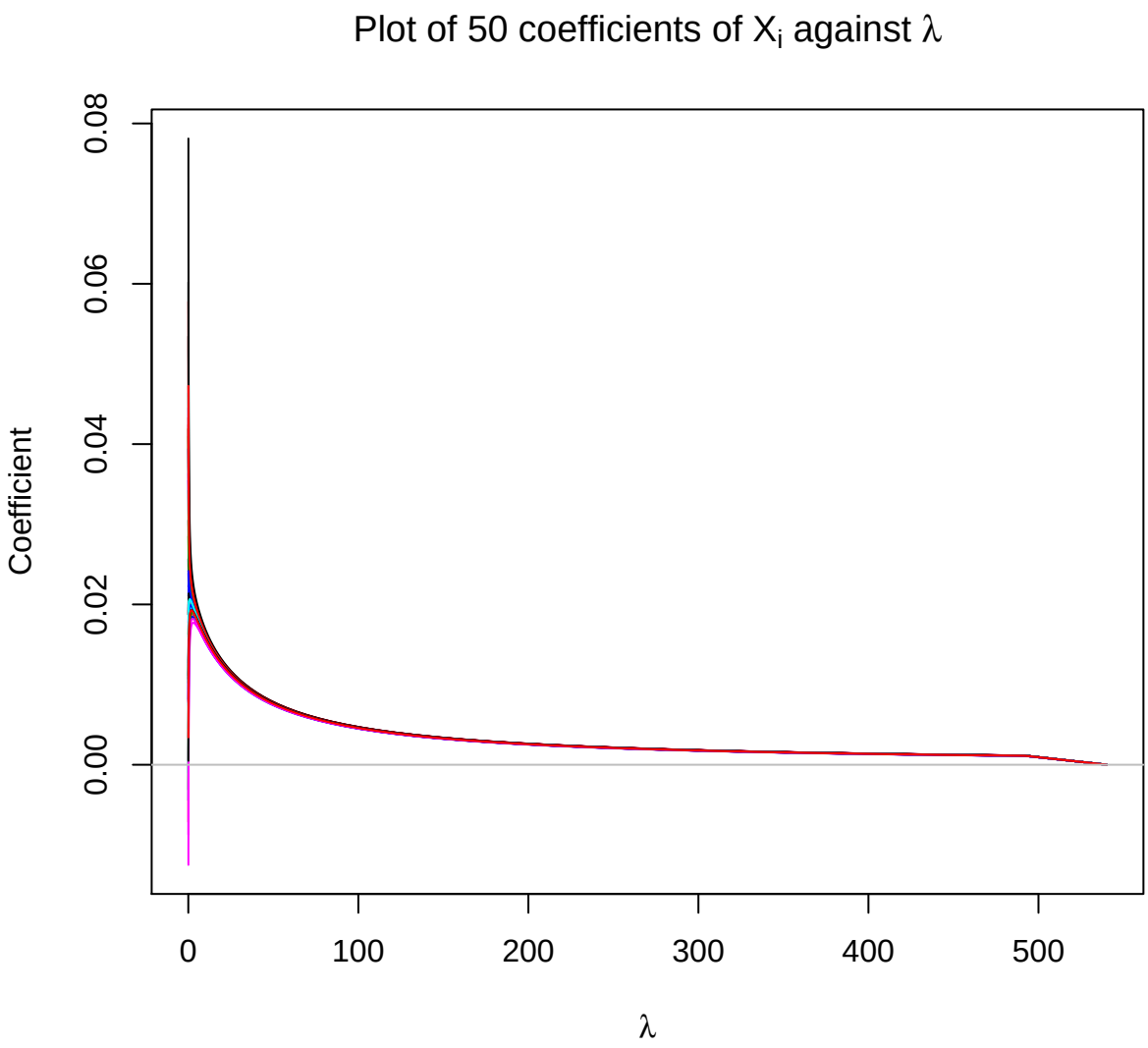
As $\lambda \to \infty$, the eigenvalues $\lambda_i + \lambda \to \infty$. Then the inverse of the positive definite matrix has eigenvalues $\frac{1}{\lambda_i + \lambda} \to 0$, so $(\mathbf{x}^T\mathbf{x} + n\lambda\mathbf{I})^{-1} \to \mathbf{0}$. Thus, this causes $\hat{\beta}_{RR} \to \mathbf{0}$.

Q5 d)

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-16
```



Ridge regression is a shrinkage estimator as increasing the penalty λ causes the estimated coefficients of the model to "shrink" towards 0.

Q5 e)

Table 6: Out-sample MSE

	x
Ridge Regression	0.04948
OLS	0.72071

As seen from the MSEs above, the out-of-sample performance of the Ridge Regression is better than the OLS Regression, having a much lower MSE.