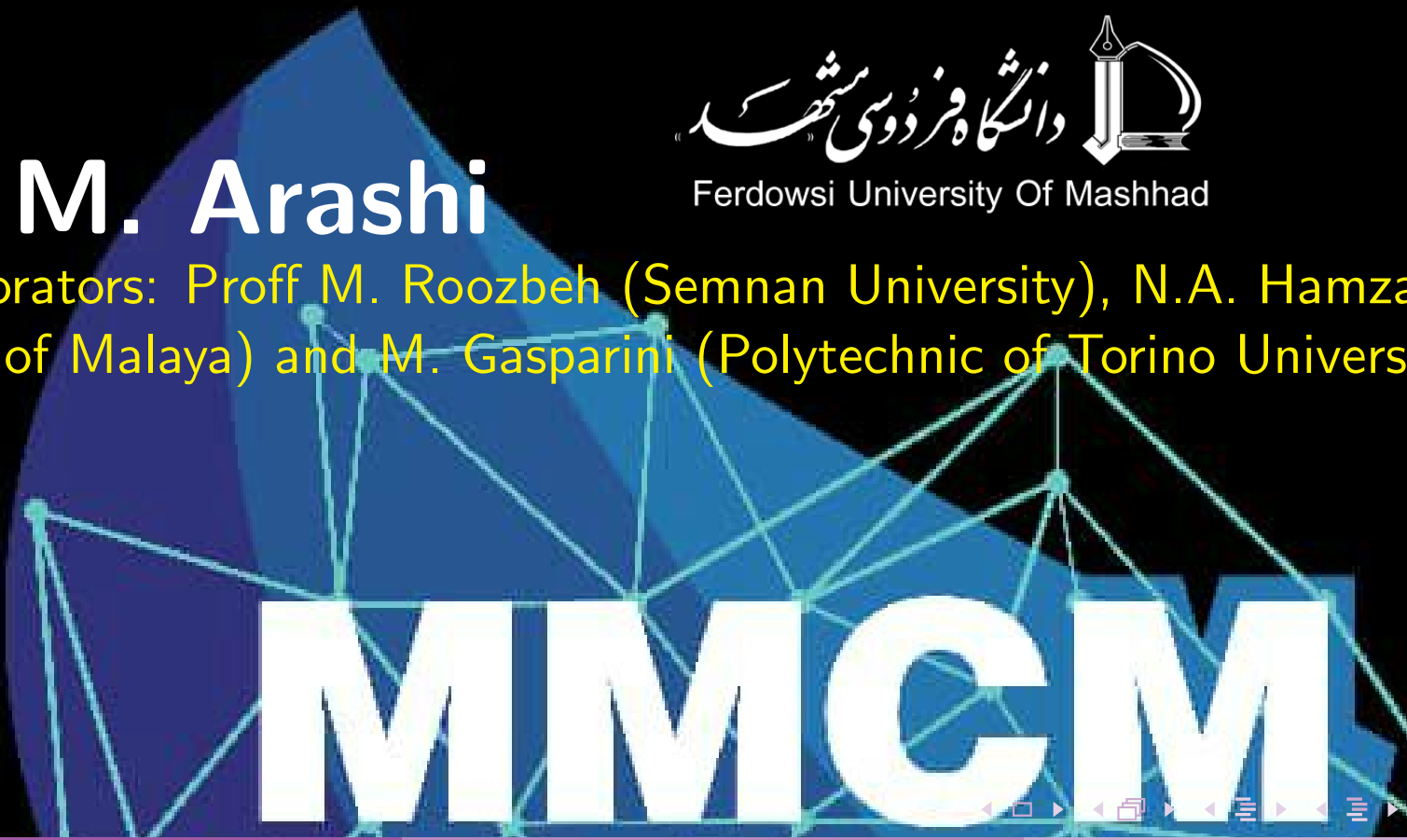
Contraction Estimation in High-dimensional Data Analysis

M. Arashi

دانشگاه فردوسی مشهد
Ferdowsi University Of Mashhad

Collaborators: Proff M. Roozbeh (Semnan University), N.A. Hamzah (University of Malaya) and M. Gasparini (Polytechnic of Torino University)



Outline

What is HD problem?

With the rapid advancement in data acquisition technologies and data processing capabilities, the data analytics environment has changed drastically over the last decades and we have "large p small n problems"

1. image analysis
2. microarray analysis
3. text analysis
4. disease classification
5. image analysis
6. pattern recognition

What is HD problem? cont.

Let $\rho = \frac{p}{n}$. Then $\rho \in (1, \infty)$ accounts for high dimensions.
However p can be smaller than n, but still very large.
To better understand low/high-dimensional regime, let $p = p_n$ denote the growth of the number of features as n grows. Then examples such as $p_n = \left[\frac{n}{2}\right]$, $p_n = \left[4.5n^{\frac{1}{4}}\right]$, $p_n = \left[\frac{n}{b\log(n)}\right]$ where $b > 1$ are used for $p_n < n$ and $\log(p_n) = O_p\left(n^b\right)$ where $0 < b < 1$ refers to $p_n > n$.

What is HD problem? cont.

Table: Examples of low/high-dimensional regime (n, p_n).

	low-dimensional				high-dimensional	
n \ p_n	$\frac{n}{2}$	$4.5n^{\frac{1}{4}}$	$\frac{n}{1.5\log(n)}$	$\frac{n}{2\log(n)}$	$e^{n^{0.4}}$	$e^{n^{0.5}}$
25	13	10	5	4	27	148
50	25	12	9	6	70	1177
100	50	14	14	11	244	22026
250	125	18	30	23	2238	7358659
500	250	21	54	40	21338	5141855148

Riboflavin Data cont.

Dataset of riboflavin (vitamin B2) production by Bacillus subtilis contains $n = 71$ observations of $p = 4088$ predictors (gene expressions) and a one-dimensional response (riboflavin production).

Riboflavin Data cont.

Use R!

```
library(hdi)
data(riboflavin)
str(riboflavin)
```

```
'data.frame':        71 obs. of  2 variables:
$ y: num  -6.64 -6.95 -7.93 -8.29 -7.31 ...
$ x: 'AsIs' num [1:71, 1:4088] 8.49 7.64 8.09 7.89 6.81 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : chr [1:71] "b_Fbat107PT24.CEL" "b_Fbat107PT30.CEL" "b
.. ..$ : chr [1:4088] "AADK_at" "AAPA_at" "ABFA_at" "ABH_at" .
```

Riboflavin Data cont.

```
Format

y Log-transformed riboflavin production rate
(original name: q_RIBFLV).
x (Co-)variables measuring the logarithm of the
expression level of 4088 genes.
```

There is a single real valued response variable which is the logarithm of the riboflavin production rate and $p = 4088$ explanatory variables measuring the logarithm of the expression level of 4088 genes.

Riboflavin Data cont.

Research question: How can we predict the production rate based on the genes' expression level?

Use the model

$$y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \epsilon_i,$$

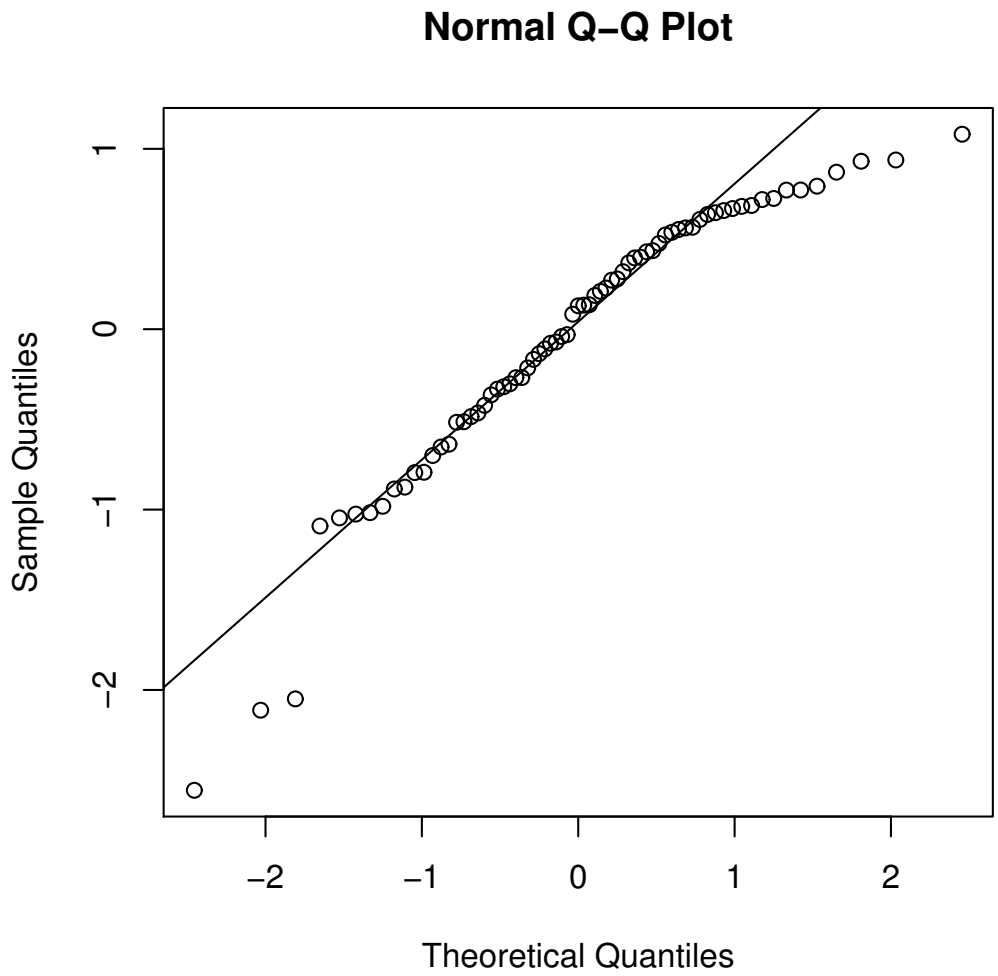
where $\boldsymbol{\beta}$ is the vector of regression coefficients and ϵ_i is the i^{th} error component.

The OLS fails:::$\hat{\boldsymbol{\beta}} = \cancel{(\mathbf{X}^\top \mathbf{X})^{-1}} \mathbf{X}^\top \mathbf{y}$

Use ridge: $\hat{\boldsymbol{\beta}}(k) = (\mathbf{X}^\top \mathbf{X} + k\mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$

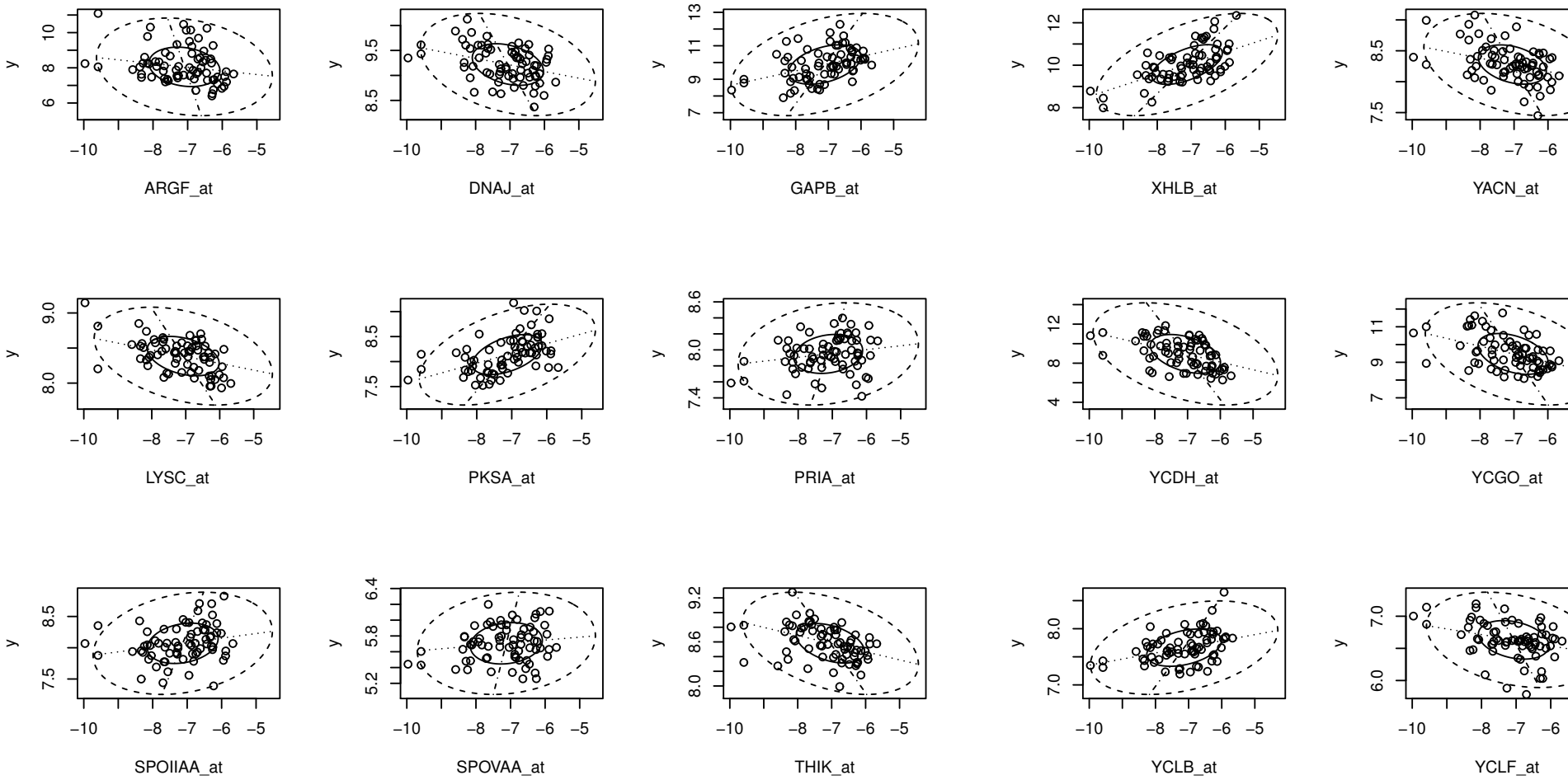
Riboflavin Data cont.

The following figure shows the normal Q–Q plot based on the ridge estimation for the riboflavin production data set.



Riboflavin Data cont.

Also, the bivariate boxplot for selected genes.



Riboflavin Data cont.

This diagram is based on calculating robust measures of location, scale, and correlation; it consists essentially a pair of concentric ellipses, one of which (the hinge) includes 50% of the data and the other (called the fence) delineates potentially troublesome outliers. In addition, robust regression lines of both response on predictor and vice versa are shown, with their intersection showing the bivariate location estimator. The acute (large) angle between the regression lines will be small (large) for a large (small) absolute value of correlations.

Our visualizations clearly revealed that the data contains some outliers.

We need a robust procedure for estimation of parameter and give the prediction model in high-dimension.

Rank-estimation

Rank-based estimators are highly efficient and robust procedures to outliers in the response space.
In short, rank regression is a simple technique which consists of replacing the data with their corresponding ranks. Rank regression and related inferential methods are useful in situations where

1. the relation between the response and covariate variables is nonlinear and monotonic and a simple and practical nonlinear form is of interest rather than polynomial, spline, kernel and/or other forms
2. there are outliers present in the study and we need a nonparametric robust procedure
3. the mere presence of so many important input variables makes it difficult to think in terms to find an appropriate parametric nonlinear model

Rank-estimation cont.

Consider the setting where observed data are realizations of $\{(\mathbf{X}_i, y_i)\}_{i=1}^n$ with p-dimensional covariates $\mathbf{X}_i \in \mathbf{R}^p$ and univariate continuous response variables $y_i \in \mathbf{R}$. A simple regression model has form

$$y_i = \mathbf{X}_i^\top \beta + \epsilon_i, \tag{2.1}$$

where β is the vector of regression coefficients and ϵ_i is the i^{th} error component.

Rank-estimator

Under some regularity conditions, the rank-estimate of $\boldsymbol{\beta}$ is given by

$$\begin{aligned} \hat{\boldsymbol{\beta}}_\psi &= \arg\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_\psi \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \hat{\mathbf{y}}_\psi, \end{aligned} \tag{2.2}$$

where $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$ and $\hat{\mathbf{y}}_\psi$ is the minimizer of dispersion function $D_\psi(\boldsymbol{\eta}) = \|\mathbf{y} - \boldsymbol{\eta}\|_\psi$ over $\boldsymbol{\eta} \in \mathcal{C}(\mathbf{X})$, where $\mathcal{C}(\mathbf{X})$ is the column space spanned by the columns of $\mathbf{X}$. Thus, $\hat{\boldsymbol{\beta}}_\psi$ is the solution to the rank-normal equations $\mathbf{X}^\top \mathbf{a}(R(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})) = \mathbf{0}$ and $D_\psi(\mathbf{X}\hat{\boldsymbol{\beta}}_\psi) = \|\mathbf{y} - \hat{\mathbf{y}}_\psi\|_\psi$.

Robust methods

Now, using a similar approach in formulating the ridge estimator, we use the following rank ridge regression estimator

$$\hat{\boldsymbol{\beta}}_{\psi}(k) \quad = \quad \mathbf{C}_n(k)^{-1}\mathbf{X}^{\top}\hat{\mathbf{y}}_{\psi}, \quad \mathbf{C}_n(k) = \left(\frac{1}{n}\mathbf{X}^{\top}\mathbf{X} + k\mathbf{I}_p\right) \qquad (2.3)$$

where $k > 0$ is the ridge parameter.

Robust contraction

In order to improve upon the rank ridge regression estimator, we define the Stein-type shrinkage estimator (SSE) as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\psi}^{(S)}(k,d) &= \left(1-\frac{d}{R_n(k)}\right)\hat{\boldsymbol{\beta}}_{\psi}(k) \\ &= \hat{\boldsymbol{\beta}}_{\psi}(k) - dR_n(k)^{-1}\hat{\boldsymbol{\beta}}_{\psi}(k), \quad d>0.\end{aligned}$$

The SSE shrinks the coefficients towards the origin using the weight function $R_n(k)$, where

$$R_n(k) = \sigma_a^{-2}\mathbf{a}^{\top}(R(\mathbf{y}))\mathbf{X}\mathbf{C}_n(k)^{-1}[\mathbf{X}(k)]^{-1}\mathbf{C}_n(k)^{-1}\mathbf{X}^{\top}\mathbf{a}(R(\mathbf{y}))$$

and $\mathbf{X}(k)$ is an invertible matrix given by

$$\mathbf{X}(k) = \mathbf{C}_n(k)^{-1} - \left[k\mathbf{C}_n(k)^{-2}\right],$$

$\sigma_a^2 = \frac{1}{n-1}\sum_{j=1}^{n} a^2(j) \doteq 1$, and $k > 0$. The amount of shrinkage is controlled by the shrinkage coefficient d.

Theorem

$\hat{\boldsymbol{\beta}}_{\psi}^{(S)}(k,d)$ is a shrinkage estimator under l_q-norm under some regularity conditions as stated below

(i): *Under the set of local alternatives $\mathcal{K}_n : \boldsymbol{\beta} = n^{-\frac{1}{2}}\boldsymbol{\delta}$, with $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_p)^\top$, $\delta_i \neq 0$, $i = 1, \cdots, p$, we have $\|\hat{\boldsymbol{\beta}}_{\psi}^{(S)}(k,d)\|^q < \left\|\hat{\boldsymbol{\beta}}_{\psi}(k)\right\|^q$.*

(ii) *For $k > n/2$, $d > 0$, we have*

$$\|\hat{\boldsymbol{\beta}}_{\psi}^{(S)}(k,d)\|^q < \left\| \left(1 - \frac{d}{R_n(k)}\right) \hat{\mathbf{y}}_{\psi} \right\|^q .$$

(iii) *Assume $\lambda_i = o(n)$, $i = 1, \cdots, n$. For $k > \sup_{1 \leq i \leq n} \lambda_i$, (ii) holds in limit.*

Why two tuning parameters?

$$\hat{\boldsymbol{\beta}}_{\psi}^{(S)}(k,d) = \hat{\boldsymbol{\beta}}_{\psi}(k) - dR_n(k)^{-1}\hat{\boldsymbol{\beta}}_{\psi}(k), \quad d > 0.$$

The proposed SSE may be criticized since it depends on the two tuning parameters k and d and it may come to mind why we need an estimator with two tuning parameters, when we have the rank ridge regression estimator.

1. Apparently, as $d \to 0$, $\hat{\boldsymbol{\beta}}_{\psi}^{(S)}(k,d) \to \hat{\boldsymbol{\beta}}_{\psi}(k)$ and thus for small values d the gain in estimation is just the information provided by the robust ridge parameter, even if $\boldsymbol{\beta} = \mathbf{0}$. Thus, even if we agree that the rank ridge regression estimator shrink the coefficients to zero, the information provided by the weight $R_n(k)$, controlled by d in the SSE, is useful.
2. The ridge estimator does not select variables, thus, we can not estimate the zero vector using the rank ridge regression estimator, however, the shrinkage coefficient d maybe obtained such that for a given k, $d = R_n(k)$ and the resulting shrinkage estimator becomes equal to zero. This might be a rare event, but theoretically sounds.
3. The last but not the least, for the set of local alternatives $\mathcal{K}_n$, as in Theorem 1, the proposed SSE shrinks more than the rank ridge regression estimator. Thus in order to have robust shrinkage estimator, the SSE with two tuning parameters is preferred.

Tuning selection

The GCV function is then defined as

$$\mathrm{GCV}\left(\hat{\boldsymbol{\beta}}_{\psi}^{(S)}(k,d)\right) = \frac{\frac{1}{n}\left\|\left(\mathbf{I}_n - \mathbf{L}(k,d)\right)\boldsymbol{y}\right\|^2}{\left(1-\mu_1(k,d)\right)^2},$$

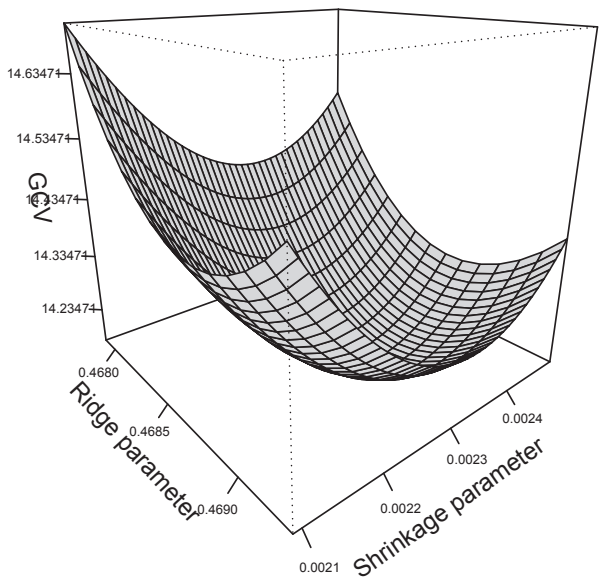
where $\mu_1(k,d) = \frac{1}{n}\mathrm{tr}\mathbf{L}(k,d)$, with

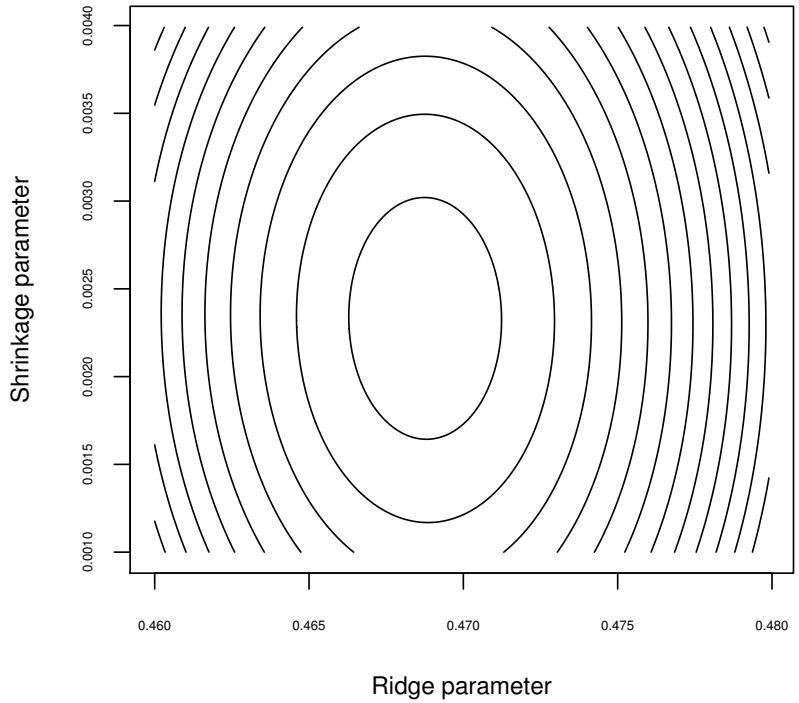
$$\mathbf{L}(k,d) = \left(1 - \frac{d}{R_n(k)}\right) 2\hat{\tau}_{\psi}\mathbf{X}\mathbf{C}_n(k)^{-1}\mathbf{X}^{\top}.$$

The latter expression is termed as the hat matrix of $\mathbf{y}$, and $\hat{\tau}_{\psi}$ is a consistent estimator.

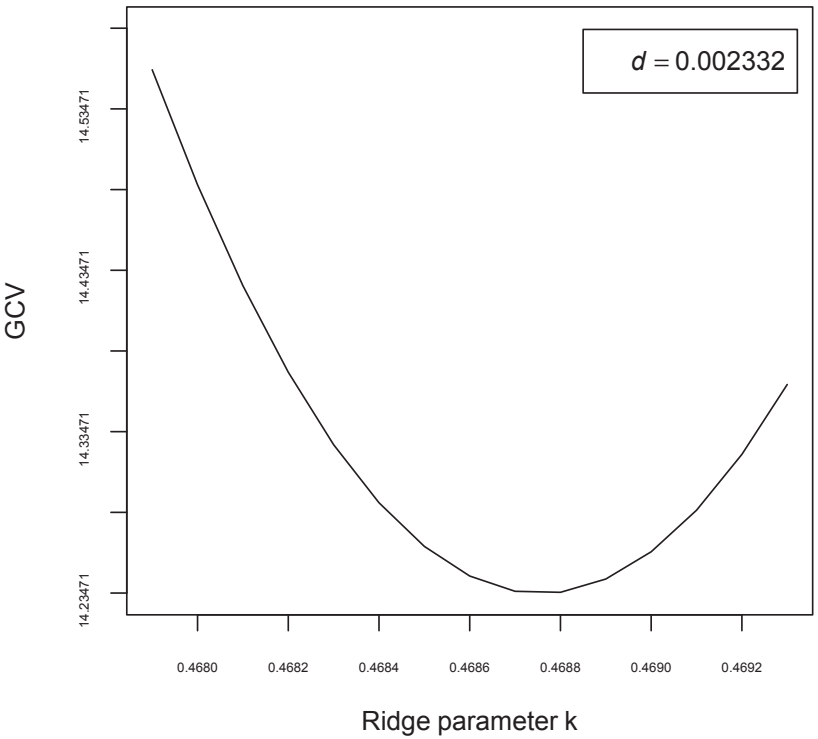
Back to real data: Does the GCV work?

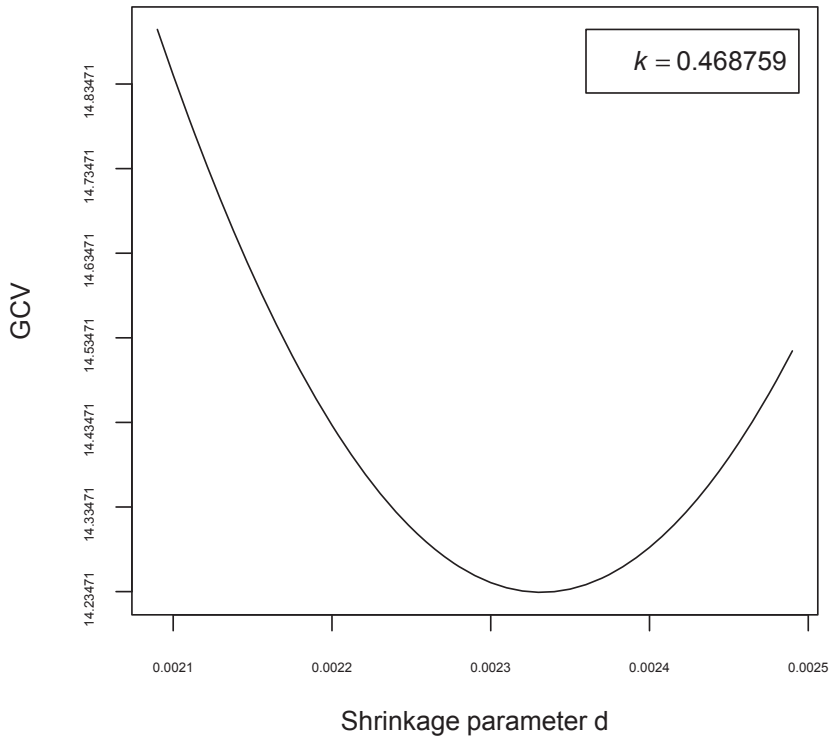
3D diagram of the GCV of $\hat{\boldsymbol{\beta}}_{\psi}^{(S)}(k, d)$ versus k and d for the riboflavin production data set:





2D slices of the GCV of $\hat{\beta}_{\psi}^{(S)}(k, d)$ versus k and d for the riboflavin production data set:





The 2D (3D) diagrams of the GCV are convex functions (surfaces) and hence they have a global minimum. This guarantees the existence of optimum values of k and d which minimize the GCV's. The minimum of $\mathrm{GCV}\big(\hat{\boldsymbol{\beta}}_{\psi}^{(S)}(k,d)\big)$ approximately occurs at $k_{opt} = 0.468759$ and $d_{opt} = 0.002332$.

To measure the prediction accuracy of proposed estimators, the leave-one-out cross-validation (CV) criterion was used, which is defined by

$$\mathrm{CV}(\hat{\boldsymbol{\beta}}) = \frac{1}{\mathrm{n}} \sum_{\mathrm{s}=1}^{\mathrm{n}} \left(\mathbf{y}_{(-\mathrm{s})} - \mathbf{X}_{(-\mathrm{s})}^{\top} \hat{\boldsymbol{\beta}}_{(-\mathrm{s})} \right)^2,$$

where $\hat{\boldsymbol{\beta}}_{(-s)}$ is obtained by replacing $\boldsymbol{X}$ and $\boldsymbol{y}$ with $\boldsymbol{X}_{(-s)} = \left(\tilde{x}_{jk(-s)} \right)$, $1 \leq k \leq n$, $1 \leq j \leq p$, $\boldsymbol{y}_{(-s)} = \left(\tilde{y}_{1(-s)}, \ldots, \tilde{y}_{n(-s)} \right)^{\top}$, $\tilde{x}_{lk(-s)} = x_{lk} - \sum_{j \neq s}^{n} W_{nj}(t_s) x_{lj}$, $\tilde{y}_{k(-s)} = y_k - \sum_{j \neq s}^{n} W_{nj}(t_s) y_j$. Here $\boldsymbol{y}_{(-s)}$ is the predicted value of response variable where sth observation left out of the estimation of the $\boldsymbol{\beta}$.

In the following table, a GOF criterion R-squared is calculated for comparing the proposed estimators using

$$\mathrm{R}^2(\hat{\boldsymbol{\beta}}) = 1 - \frac{\mathrm{SSE}(\hat{\boldsymbol{\beta}})}{\mathrm{S}_{yy}},$$

where $\mathrm{SSE}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ for $\hat{y}_i = \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}$ and $\mathrm{S}_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

Estimator	$\hat{\boldsymbol{\beta}}(k)$	$\hat{\boldsymbol{\beta}}_\psi(k)$	$\hat{\boldsymbol{\beta}}_\psi^{(S)}(k,d)$
CV	13.10023	9.88070	**8.01023**
min(GCV)	22.88144	17.00815	**14.23133**
R^2	0.707080	0.759485	**0.798853**

Take Home Message

We saw $\hat{\beta}_{\psi}^{(S)}(k, d)$ performs better than the ridge regression estimator, since it offers smaller GCV and bigger R-squared values in the presence of multicollinearity and outliers.

Moreover, because of the existence of outliers in the data set, it can be seen that R-squared's of robust type estimators are more acceptable than the R-squared of non-robust type estimator.

For high-dimensional data analysis, contraction estimation may help obtaining more efficient estimation strategies for prediction purposes.

Thank you for your attention and patience