Parsimony

Matthew Butner
Federico De Giuli
Yuqi Yang
Xiao Gu

Choosing datasets

- We want low-stress datasets
- UCI is wonderful

Overview

	The predictors we kept		
n=100	**k=0.01**	**k=0.05**	**significant-test**
run1	1,2,3,4	1,2,3	1,2,3
run2	1,2,3,7,9	1,2,3,9	1,2,3,9
run3	1,2,3,4	1,2,3	1,2,3

	The predictors we kept		
n=1000	**k=0.01**	**k=0.05**	**significant-test**
run1	1,2,3	1,2,3	1,2,3,4
run2	1,2,3,4	1,2,3	1,2,3,4
run3	1,2,3	1,2,3	1,2,3

Overview

	The predictors we kept		
n=100	**k=0.01**	**k=0.05**	**significant-test**
run1	1,2,3,4	1,2,3	1,2,3
run2	1,2,3,7,9	1,2,3,9	1,2,3,9
run3	1,2,3,4	1,2,3	1,2,3

	The predictors we kept		
n=1000	**k=0.01**	**k=0.05**	**significant-test**
run1	1,2,3	1,2,3	1,2,3,4
run2	1,2,3,4	1,2,3	1,2,3,4
run3	1,2,3	1,2,3	1,2,3

Overview

	The predictors we kept		
n=10000	**k=0.01**	**k=0.05**	**significant-test**
run1	1,2,3	1,2,3	1,2,3,4,6
run2	1,2,3	1,2,3	1,2,3,4,7
run3	1,2,3	1,2,3	1,2,3,4

	The predictors we kept		
n=100000	**k=0.01**	**k=0.05**	**significant-test**
run1	1,2,3	1,2,3	1,2,3,4
run2	1,2,3	1,2,3	1,2,3,4
run3	1,2,3	1,2,3	1,2,3,4

Overview

	The predictors we kept		
n=10000	k=0.01	k=0.05	significant-test
run1	1,2,3	1,2,3	1,2,3,4,6
run2	1,2,3	1,2,3	1,2,3,4,7
run3	1,2,3	1,2,3	1,2,3,4

	The predictors we kept		
n=100000	k=0.01	k=0.05	significant-test
run1	1,2,3	1,2,3	1,2,3,4
run2	1,2,3	1,2,3	1,2,3,4
run3	1,2,3	1,2,3	1,2,3,4

Overview

	MPG (n=392)		Ion (n=351)		MAGIC (n=19020)		Robot (n=5456)	
	Adj. R^2	p	Adj. R^2	p	Adj. R^2	p	Adj. R^2	p)
Initial	0.707	4	0.70	34	0.702	10	0.40	24
k = 0.05	0.693	1	0.55	4	0.668	3	0.36	3
k = 0.01	0.706	2	0.67	14	0.700	5	0.39	6
Sig. at 0.05	0.706	2	0.51	13	0.702	10	0.40	19

Table 1: Regression datasets

Overview

	MPG (n=392)		Ion (n=351)		MAGIC (n=19020)		Robot (n=5456)	
	Adj. R^2	p	Adj. R^2	p	Adj. R^2	p	Adj. R^2	p)
Initial	0.707	4	0.70	34	0.702	10	0.40	24
k = 0.05	0.693	1	0.55	4	0.668	3	0.36	3
k = 0.01	0.706	2	0.67	14	0.700	5	0.39	6
Sig. at 0.05	0.706	2	0.51	13	0.702	10	0.40	19

Table 1: Regression datasets

Overview

	MPG (n=392)		Ion (n=351)		MAGIC (n=19020)		Robot (n=5456)	
	Adj. R^2	p	Adj. R^2	p	Adj. R^2	p	Adj. R^2	p)
Initial	0.707	4	0.70	34	0.702	10	0.40	24
k = 0.05	0.693	1	0.55	4	0.668	3	0.36	3
k = 0.01	0.706	2	0.67	14	0.700	5	0.39	6
Sig. at 0.05	0.706	2	0.51	13	0.702	10	0.40	19

Table 1: Regression datasets

Overview

	Pima (n=768)		Cancer (n=569)		MAGIC (n=19020)		Digits (n=5081)	
	AIC	p	AIC	p	AIC	p	AIC	p)
Initial	741.4	8	60.0	30	17418	10	662.4	20
k = 0.05	812.72	1	50.0	24	18049	2	741.6	8
k = 0.01	744.30	4	50.0	24	17504	4	672.6	11
Sig. at 0.05	741.4	8	152.5	5	17589	8	706.0	16

Table 2: Classification datasets

Overview

	Pima (n=768)		Cancer (n=569)		MAGIC (n=19020)		Digits (n=5081)	
	AIC	p	AIC	p	AIC	p	AIC	p)
Initial	741.4	8	60.0	30	17418	10	662.4	20
k = 0.05	812.72	1	50.0	24	18049	2	741.6	8
k = 0.01	744.30	4	50.0	24	17504	4	672.6	11
Sig. at 0.05	741.4	8	152.5	5	17589	8	706.0	16

Table 2: Classification datasets

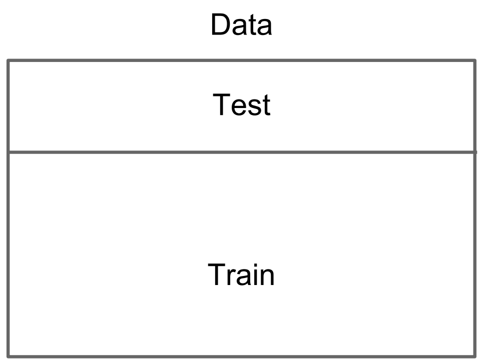
Overview

	Pima (n=768)		Cancer (n=569)		MAGIC (n=19020)		Digits (n=5081)	
	AIC	p	AIC	p	AIC	p	AIC	p)
Initial	741.4	8	60.0	30	17418	10	662.4	20
k = 0.05	812.72	1	50.0	24	18049	2	741.6	8
k = 0.01	744.30	4	50.0	24	17504	4	672.6	11
Sig. at 0.05	741.4	8	152.5	5	17589	8	706.0	16

Table 2: Classification datasets

Cross Validation

- Data
  - Train with some
  - Test against rest



Running Parsimony

AIC

Full model accuracy is 741.4454 .
Removing variable V4 .
Accuracy is now **739.4534** .
Removing variable V5 .
Accuracy is now 739.4617 .
Removing variable V8 .
Accuracy is now 740.5596 .
Removing variable V3 .
Accuracy is now 744.3059 .

Using AIC, when k = 0.01,
we should keep variables 1 2 6 7

Leave One Out

Full model accuracy is 0.7747396 .
Removing variable V4 .
Accuracy is now 0.7747396 .
Removing variable V1 .
Accuracy is now 0.7734375 .
Removing variable V5 .
Accuracy is now 0.7708333 .
Removing variable V3 .
Accuracy is now **0.7760417** .
Removing variable V8 .
Accuracy is now 0.7695312 .
Using leave one out, when k = 0.01,
we should keep variables 2 6 7

Final Results

	AIC		Leave One Out	
	PAC	Variables Kept	PAC	Variables Kept
k = 0.05	813	2	75%	2
k = 0.01	744	1, 2, 6, 7	77%	2, 6, 7

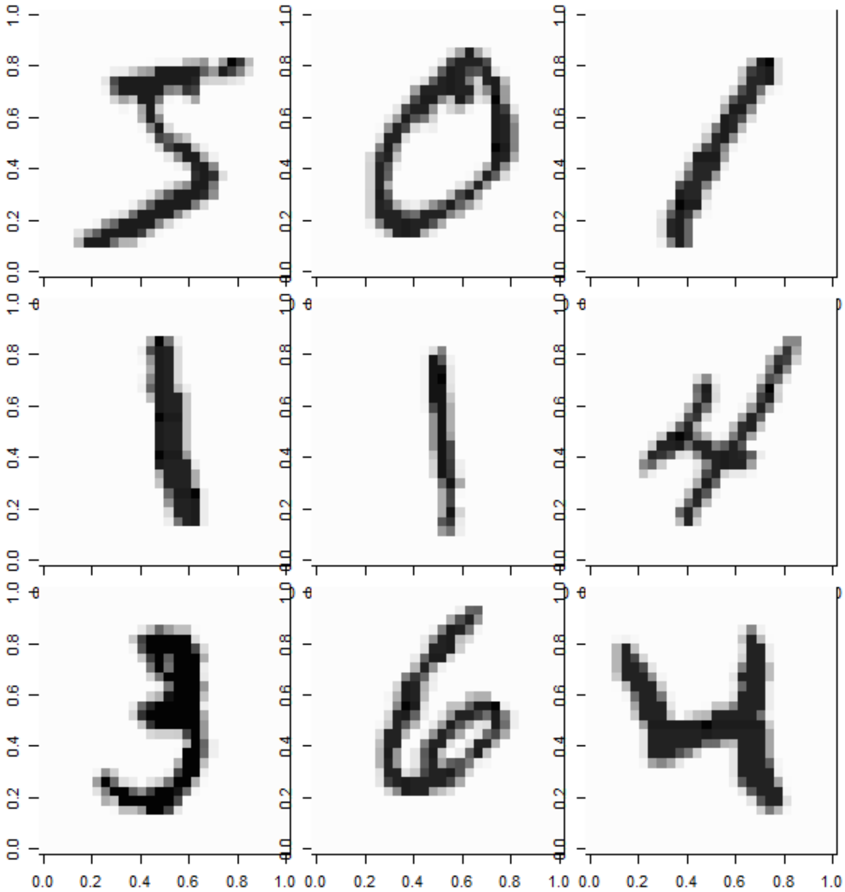
Digits Recoginition

Data discription

In this situation,
we choose a dataset that related to digits recognization.It contains 784 pixels variables. And contain 28000 items which includes all the 0,1,2,3,4,5,6,7,8,9 handwriting digits.

In order to coding 0 and 1, we only choose two type of the digits to fit the GLM model. Here we subset 1s and 8s in both of our traindata(which contains 5081 items) and testdata(which contains 2177 item).

Here are the plots of several items in the data set.



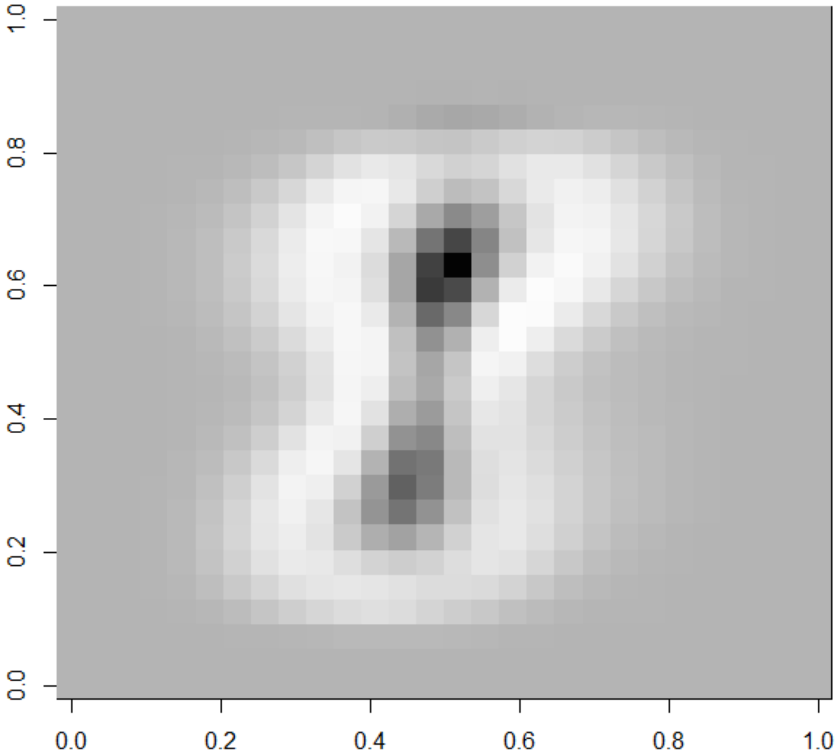
Preprocessing the data

As we can see from the plot of data set, there are many pixels varables are also zero.

So the first things we need to do is to reduce the number of variables which have no help to our prediction So in order to vanish all these things. We first use PCA to reduce the dimension and also eliminate the collinearity.

Then the number of pixels variables we choose is 20 which cover 70 of the variation. The reason we choose 20 is to satisfy the requirement and save the programming time.

The first Principal component



From our parsimony analysis

our output is here when k=0.05

Full model accuracy is 662.3877 .
Removing variable V14 . Accuracy is now 661.4665 .
Removing variable V16 . Accuracy is now 661.1731 .
Removing variable V15 . Accuracy is now 660.652 .
Removing variable V13 . Accuracy is now 660.4225 .
Removing variable V20 . Accuracy is now 660.581 .
Removing variable V18 . Accuracy is now 662.1174 .
Removing variable V9 . Accuracy is now 664.2468 .
Removing variable V17 . Accuracy is now 666.0451 .
Removing variable V8 . Accuracy is now 672.6036 .
Removing variable V11 . Accuracy is now 685.0384 .
Removing variable V3 . Accuracy is now 694.0031 .
Removing variable V19 . Accuracy is now 718.06 .
Removing variable V12 . Accuracy is now 741.5915 .
[1] 1 2 4 5 6 7 10

confusion matrix

True\pred	8	1
8	932	65
1	16	1164

The rate of error is
0. 37

our output is here when k=0.01

Full model accuracy is 662.3877 .
Removing variable PC14 . Accuracy is now 661.4665 .
Removing variable PC16 . Accuracy is now 661.1731 .
Removing variable PC15 . Accuracy is now 660.652 .
Removing variable PC13 . Accuracy is now 660.4225 .
Removing variable PC20 . Accuracy is now 660.581 .
Removing variable PC18 . Accuracy is now 662.1174 .
Removing variable PC9 . Accuracy is now 664.2468 .
Removing variable PC17 . Accuracy is now 666.0451 .
Removing variable PC8 . Accuracy is now 672.6036 .
[1] 1 2 3 4 5 6 7 10 11 12 19

The rate of error is
0.34

confusion matrix

True\pred	8	1
8	939	58
1	17	1163

PC	1	2	3	4	5
p-value	<2e-16	<2e-16	0.000107	0.001438	<2e-16
PC	6	7	8	9	10
p-value	2.53e-10	<2e-16	0.000210	0.005042	2.53e-14
PC	11	12	13	14	15
p-value	0.000150	1.01e-08	0.027292	0.801610	0.297027
PC	16	17	18	19	20
p-value	0.652590	0.297027	0.000111	5.28e-06	0.000535

By using significance test p=0.05, we will remove PC 13, PC14, PC15, PC 16, PC17

Data set Description

The Smaller One

- Auto MPG Data (http://archive.ics.uci.edu/ml/datasets/Auto+MPG)
- Numbers: 392
- Details:

	mpg	displacement	horsepower	weight	acceleration
1	18	307	130	3504	12.0
2	15	350	165	3693	11.5
3	18	318	150	3436	11.0
4	16	304	150	3433	12.0
5	17	302	140	3449	10.5
6	15	429	198	4341	10.0

Data set Description

The larger One

- MAGIC Gamma Telescope Data Set (http://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope)
- Numbers: 19020
- Details:

```
> head(data)
   fLength   dWidth  fSize  fConc fConc1    fAsym  fM3Long fM3Trans fAlpha    fDist class
1  28.7967  16.0021 2.6449 0.3918 0.1982  27.7004  22.0110  -8.2027 40.0920  81.8828     1
2  31.6036  11.7235 2.5185 0.5303 0.3773  26.2722  23.8238  -9.9574  6.3609 205.2610     1
3 162.0520 136.0310 4.0612 0.0374 0.0187 116.7410 -64.8580 -45.2160 76.9600 256.7880     1
4  23.8172   9.5728 2.3385 0.6147 0.3922  27.2107  -6.4633  -7.1513 10.4490 116.7370     1
5  75.1362  30.9205 3.1611 0.3168 0.1832  -5.5277  28.5525  21.8393  4.6480 356.4620     1
6  51.6240  21.1502 2.9085 0.2420 0.1340  50.8761  43.1887   9.8145  3.6130 238.0980     1
> |
```

1. fLength: major axis of ellipse [mm]
2. fWidth: minor axis of ellipse [mm]
3. fSize: 10-log of sum of content of all pixels [in #phot]
4. fConc: ratio of sum of two highest pixels over fSize [ratio]
5. fConc1: ratio of highest pixel over fSize [ratio]
6. fAsym: distance from highest pixel to center, projected onto major axis [mm]
7. fM3Long: 3rd root of third moment along major axis [mm]
8. fM3Trans: 3rd root of third moment along minor axis [mm]
9. fAlpha:angle of major axis with vector to origin [deg]
10. fDist: distance from origin to center of ellipse [mm]
11. class: 1,0 # gamma (signal), hadron (background)

Continuous Y Case

- For Auto_MPG, response Y: mpg

	displacement	horsepower	weight	acceleration
full mode	kept	kept	kept	kept
k=0.01		kept	kept	
k=0.05			kept	
significant test		kept	kept	

Continuous Y Case

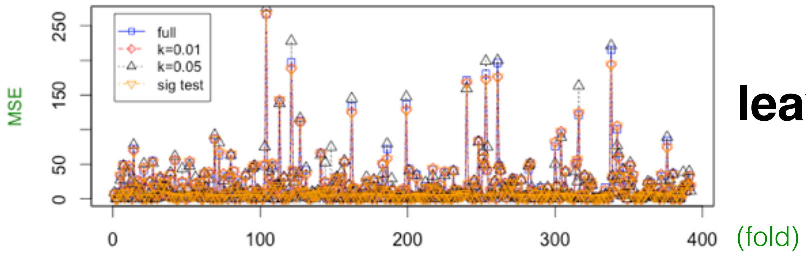
- For Auto_MPG, response Y: mpg

	displacement	horsepower	weight	acceleration	Adj_R^2
full mode	kept	kept	kept	kept	0.7039526
k=0.01		kept	kept		0.7048656
k=0.05			kept		0.6918423
significant test		kept	kept		0.7048656

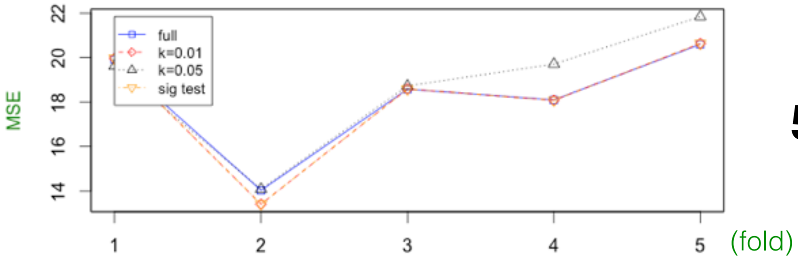
10-fold Cross Validation

number of times variable was chosen					
	displacement	**horsepower**	**weight**	**acceleration**	**E(MSE)**
full mode	10	10	10	10	18.36978
k=0.01	0	10	10	0	18.19453
k=0.05	0	0	10	0	18.95199
significant test	0	10	10	0	18.19453

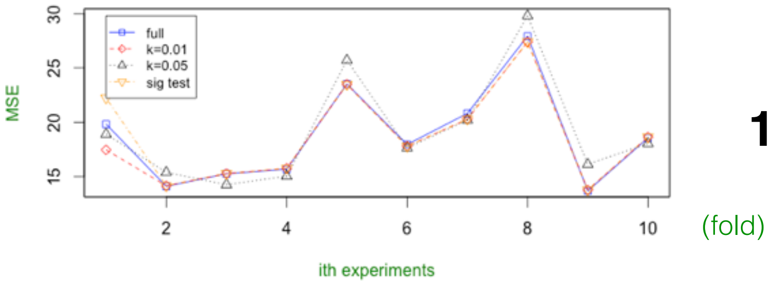
	E(MSE)		
	leave one out	5-fold	10-fold
full mode	18.33959	18.24245	18.72733
k=0.01	18.11295	18.13298	18.39664
k=0.05	18.85161	18.79478	19.10501
significant test	18.11295	18.13298	18.87462



leave one out



5_fold



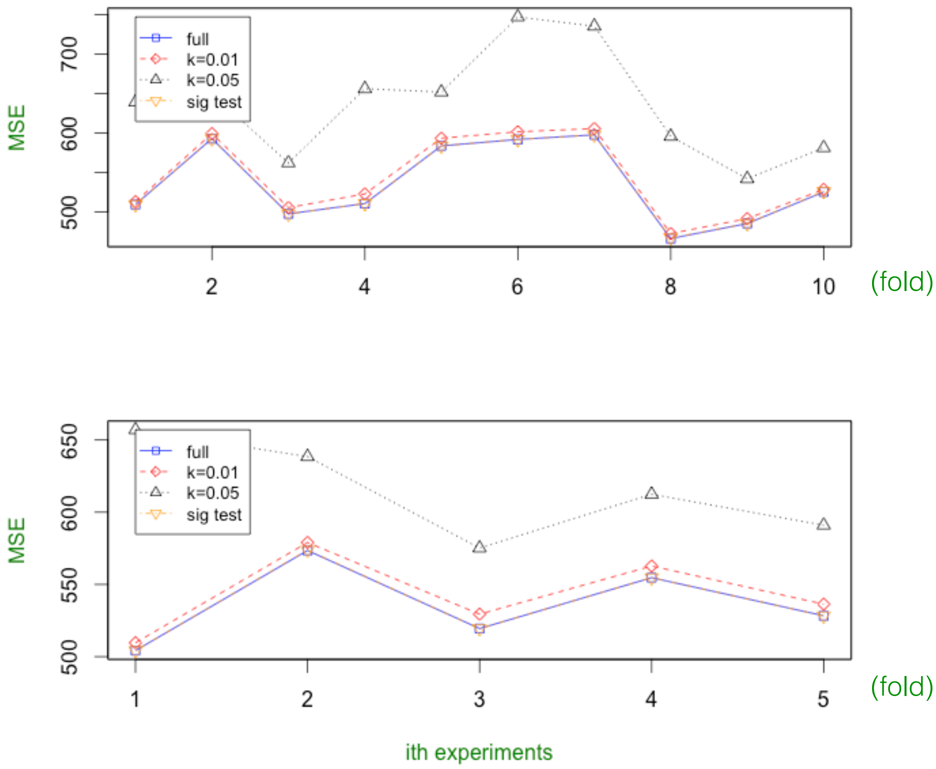
10_fold

Continuous Y Case

- For MAGIC Gamma Telescope, response Y: flength

	number of times variable was chosen										
10-fold	fWidth	fSize	fConc	fConc1	fAsym	fM3Long	fM3Trans	fAlpha	fDist	class	E(MSE)
full mode	10	10	10	10	10	10	10	10	10	10	536.05
k=0.01	10	0	10	0	10	0	0	0	10	10	543.38
k=0.05	10	0	10	0	5	0	0	0	0	0	635.99
significant test	10	10	10	10	10	10	8	10	10	10	536.22

	mMSE	
	5-fold	10-fold
full mode	536.01	536.05
k=0.01	543.42	543.38
k=0.05	614.82	635.99
significant test	536.08	536.22



full
k=0.01
k=0.05
sig test
MSE
(fold)
ith experiments

The validation method was repeated ten times,each time using a different random split of the observations into a training set and testing set.

this plot illustrates the variability in the estimated test MSE that results from this approach