Collinearity and High Dimensionality

Comparing the performance of PCR to Shrinkage Methods in simple simulated examples

Gabriella Stabile, Giuseppe Martinelli and Chiara Cavigli

MASL a.a. 2022/2023

(Multi)Collinearity

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

Case: $\mathbf{X}_1$ =a+b $\mathbf{X}_2$

$\mathbf{X} =$

$$\begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n1} & \dots & x_{nk} \end{pmatrix}$$

- Impossibility to uniquely estimate β_1 and β_2 ($\text{rk}(\mathbf{X}) \neq \text{p} = \text{k+1}$).
- The information that $\mathbf{X}_1$ provides about $\mathbf{Y}$ is redundant in the presence of $\mathbf{X}_2$.
- Multicollinearity is much more likely to occur in high-dimensional settings.

High-Dimensionality

Refers to the case where we have a high number of predictors. Commonly we have that: $p > n$ or $p \approx n$. Linear models often result problematic for the following reasons:

- OLS provides too much flexibility, leading to **overfitting**.
- *Curse of Dimensionality*.
- High variance of the estimated coefficients.

The methods we use to overcome collinearity are usually also helpful in the case of high dimensional settings.

Principal Components Regression

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

We estimate $\boldsymbol{\alpha}$ by OLS:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y} = (\boldsymbol{\Sigma}_m\mathbf{U}_m^T\mathbf{U}_m\boldsymbol{\Sigma}_m)^{-1}\boldsymbol{\Sigma}_m\mathbf{U}_m^T\mathbf{y} = \boldsymbol{\Sigma}_m^{-1}\mathbf{U}_m^T\mathbf{y}$$

The fitted values:

$$\hat{\mathbf{y}}^{PCR} = \mathbf{Z}\hat{\boldsymbol{\alpha}}$$

PCR approach consists in fitting a least squares model regressing $\mathbf{y}$ on $\mathbf{z}_1$,...,$\mathbf{z}_M$ for M$\leq$p (M typically chosen by cross-validation).

This technique is not scale invariant (each $\mathbf{x}_j$, j=1,..,p must be standardized).

It leaves M high-variance directions and discards the rest.

More M is large more the bias decreases while the variance increases.

Approaches and Strategy

We have a number of possible approaches to deal with both high correlation and high dimensionality in the data:

- **Subset Selection** $\rightarrow$ Choosing the best predictors.
- **Shrinkage/Regularization** $\rightarrow$ Acting on the coefficients by shrinking them towards zero.
- **Dimensionality Reduction** $\rightarrow$ They help us synthesize the information contained in multiple predictors

In particular our aim is to compare Regularized estimators and Principal Component Regression in terms of prediction accuracy (MSE), Properties and Computational efficiency.

Data Generating Process

We have simulated four data sets:

- One data set with 200 observations and 10 covariates (n>p)
- One data set with 180 observations and 200 covariates (p>n)
- One data set with 200 observations and 10 covariates but in which only the first eight variables contain all the information about Y (n>p and the design matrix $\mathbf{X}$ has 2 collinear predictors)
- One data set with 150 observations and 200 covariates (p>n with highly correlated predictors).

We consider only continuous numerical variables.

Case n>p

In this case we have n=200 observations and p=10 predictors generated independently.

	OLS	**Ridge**	**Lasso**	**Elastic**	**PCR**
Validation		$\lambda = 0.81$	$\lambda = 0.10$	$\lambda = 0.18$	10
$MSE_{Training}$	0.84	1.51	0.94	0.99	1.99
MSE_{Test}	1.14	2.07	1.31	1.37	0.26

- Regularized estimators tend to choose a small λ.
- PCR uses all 10 components.
- OLS is BLUE. In this case, there would be no reason to use shrinkage methods.

Case n>p (with few perfectly correlated covariates)

In this experiment we generated 2 of the predictors from the previous dataset as linearly dependent on other covariates.

	OLS	**Ridge**	**Lasso**	**Elastic**	**PCR**
Validation		$\lambda = 0.84$	$\lambda = 0.04$	$\lambda = 0.05$	8
$MSE_{Training}$	0.95	0.99	0.97	0.97	22
MSE_{Test}	1.09	1.65	1.08	1.08	0.29

- OLS works, but we don't get estimates $\hat{\beta}$ for the redundant predictor(s).
- Regularized estimators will push the redundant variables towards 0 (for Lasso exactly to 0).
- According to CV, 8 components are enough to perform PCR.

Case p>n

This simulation involves n=180 and a higher number of predictors p=200.

	OLS	**Ridge**	**Lasso**	**Elastic**	**PCR**
Validation		$\lambda = 254$	$\lambda = 25$	$\lambda = 51$	70
$MSE_{Training}$	0	1343	3140	48	325
MSE_{Test}	241221	2460	3002	1497	736

- OLS overfits severely.
- According to the results the lowest MSE was achieved by PCR using 70 components.
- Shrinkage methods chose a much higher value of λ.

Case p>n (with highly correlated predictors)

While still in a high-dimensional setting, we now generate a stronger correlation between the predictors.

	OLS	**Ridge**	**Lasso**	**Elastic**	**PCR**
Validation		$\lambda = 188$	$\lambda = 1.76$	$\lambda = 2.66$	29
$MSE_{Training}$	0	420	143	105	372
MSE_{Test}	13822	706	640	586	506

- PCR chooses a very slim number of components, and once again reported the lowest MSE.
- When higher correlation is present among the predictors, regularized estimators are unaffected, and perform well.

Remarks

- OLS is an optimal solution for cases involving n>p and low correlation between predictors, though it's also possible to adapt the model to overcome collinearity.
- PCR appears to perform better but it must not be concluded that this method should be used alone, as it's common practice to use a combination of PCA and other methods.
- Regularized estimators provide solid solutions to high-dimensional and high-correlation settings, but λ must be chosen wisely.

Computational considerations

The computational complexity of all the regularized estimators (except for Subset Selection) is the same as OLS, which is:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- $O(C^2N)$ to multiply $\mathbf{X}^T$ by $\mathbf{X}$
- $O(CN)$ to multiply $\mathbf{X}^T$ by $\mathbf{Y}$
- $O(C^3)$ to compute the LU (or Cholesky) factorization of $\mathbf{X}^T\mathbf{X}$ and use that to compute the product $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

Regarding PCR, the complexity depends on the computation of the covariance matrix $O(C^2N)$ and its eigen-value decomposition $O(C^3)$. So, the complexity of PCA is $O(C^2N+C^3)$

References

- Fahrmeir L., Kneib T., Lang S., Marx B. (2013), *Regression: Models, Methods and Applications*, Springer
- James G., Witten D., Hastie T., Tibshirani R. (2013), *An Introduction to Statistical Learning, with Applications in R*, Springer
- Hastie T., Tibshirani R., Friedman J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer