Shrinkage Estimators of OLS and Linear Models in Machine Learning

Machine Learning & Computational Statistics

Ivailo Dimov

February 12, 2026

Readings

In addition to the lecture notes, the following are required readings:

- Chapter 3 in Friedman, Hastie, and Tibshirani (2017). While you should read 3.3 and 3.6-3.9, a more cursory read of these sections is fine.

You may also enjoy watching these scikit-learn video tutorials (optional):

- PyData 2015, and
- PyData 2016.

Shrinkage Estimators for Linear Regression

Recall: Classical Linear Regression

Under the classical linear regression assumptions, recall that the Gauss-Markov theorem states that our OLS estimator $\boldsymbol{\beta}_{\mathrm{OLS}} \in \mathbb{R}^p$ has the smallest variance amongst all linear unbiased estimators.

This raises the question: Is there a biased linear estimator that is "better" than OLS in some way?

Mean Squared Error and the Bias-Variance Tradeoff I

Let $\widehat{\beta}$ be an estimate of β and $\mathbb{E}[\widehat{\beta}] = \widehat{\beta}^*$. We say that $\widehat{\beta}^* - \beta$ is the *bias*.

The *mean squared error* (MSE) of an estimator is defined as

$$\mathbb{E}[(\widehat{\beta} - \beta)^2]\,. \tag{1}$$

A simple calculation shows that[1]

$$\begin{aligned}
\mathbb{E}[(\widehat{\beta} - \beta)^2] &= \mathbb{E}[(\widehat{\beta} - \widehat{\beta}^* + \widehat{\beta}^* - \beta)^2] \\
&= \mathbb{E}[(\widehat{\beta} - \widehat{\beta}^*)^2] + (\widehat{\beta}^* - \beta)^2 \\
&= \mathrm{var}[\widehat{\beta}] + (\widehat{\beta}^* - \beta)^2
\end{aligned}$$

We call this decomposition the *bias-variance tradeoff*:

- First term: Variance of the estimator.
- Second term: Squared bias of the estimator.

Mean Squared Error and the Bias-Variance Tradeoff II

It follows from the Gauss-Markov theorem that OLS has the smallest MSE amongst all linear unbiased estimators.

But is there a biased linear estimator with a smaller MSE?

[1] For you: Show that the covariance term is zero.

A Simple Shrinkage Estimator I

To make things simple, let us consider the case where our regressors, $\mathbf{X} \in \mathbb{R}^{n \times p}$, are orthogonal i.e. $\mathbf{X}'\mathbf{X} = \mathbf{I}$.

Suppose we replace our OLS estimate $\widehat{\beta}_j$ with something slightly smaller

$$\widetilde{\beta}_j = \frac{1}{1+\lambda}\widehat{\beta}_j \,, \tag{2}$$

where $\lambda \in \mathbb{R}_+$.

We observe that (a) if $\lambda = 0$ we obtain OLS, and (b) if $\lambda \to \infty$ then $\widetilde{\beta}_j \to 0$. This estimator is an example of a *shrinkage estimator*. In this case we are shrinking towards zero.

A Simple Shrinkage Estimator II

Let us compute the MSE of the coefficients. As our regressors are orthogonal, the MSE of this estimator is

$$
\begin{aligned}
\sum_{j=1}^{p} \mathbb{E}\Big(\frac{1}{1+\lambda}\widehat{\beta}_j - \beta_j\Big)^2 &= \Big(\frac{1}{1+\lambda}\Big)^2 \sum_{j=1}^{p} \mathbb{E}(\widehat{\beta}_j - \beta_j)^2 \\
&\quad + \Big(\frac{\lambda}{1+\lambda}\Big)^2 \sum_{j=1}^{p} \beta_j^2 \\
&= p\sigma^2\Big(\frac{1}{1+\lambda}\Big)^2 + \Big(\frac{\lambda}{1+\lambda}\Big)^2 \sum_{j=1}^{p} \beta_j^2 \,. \qquad (3)
\end{aligned}
$$

To minimize this function, we solve the FOC with respect to λ, obtaining the optimal shrinkage[2]

$$
\lambda^* = \frac{p\sigma^2}{\sum_{j=1}^{p} \beta_j^2} \,. \qquad (4)
$$

A Simple Shrinkage Estimator III

While in general we do not know the true OLS coefficients and therefore this estimator cannot be used in practice, this formula provides intuition:

- If the OLS coefficients are large relative to the residual variance, then λ^* should be small (and vice versa).

In principle, with the right choice of shrinkage we can obtain an estimator with a lower MSE. The estimator may be biased. However, the bias-variance tradeoff states that what we pay for in bias we can make up for in decreased variance up to some optimal point.

Next we look at some shrinkage estimators that are based upon this idea.

A Simple Shrinkage Estimator IV

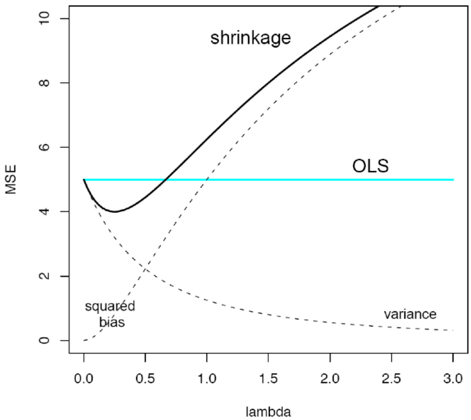

Figure 1: Bias-variance tradeoff of a linear estimator. (Source: Mark Hansen, UCLA, Applied Regression Analysis: Generalized Linear Models.)

[2] For you: Verify that this is indeed the minimum.

The James-Stein Estimator I

Suppose we know the value of σ^2. The expansion

$$\frac{1}{1+\lambda} = 1 - \lambda + O(\lambda^2) \tag{5}$$

inspired James and Stein (1961) to define the following shrinkage estimator

$$\widetilde{\beta}_j = \left(1 - \frac{(p-2)\sigma^2}{\sum_{j=1}^{p} \widehat{\beta}_j^2}\right) \widehat{\beta}_j \tag{6}$$

where $\{\widehat{\beta}_j\}_{j=1}^{p}$ are the estimated OLS coefficients. This estimator is known as the *James-Stein estimator* (JS).

The James-Stein Estimator II

As the regressors are orthogonal, the MSE of the OLS and James-Stein estimators are[3]

$$\mathrm{MSE}_{\mathrm{OLS}} = p\sigma^2 \,, \tag{7}$$

$$\mathrm{MSE}_{\mathrm{JS}} = 2\sigma^2 \,. \tag{8}$$

Note that the MSE of the JS estimator is independent of p! When $p \geqslant 3$ its MSE is therefore smaller than that of OLS.

The James-Stein Estimator III

Remark

What we saw above is referred to as *Stein's paradox* in the statistics literature. It is the phenomenon that when three or more parameters are estimated at the same time, then there are estimators that combine these parameters that have a lower MSE than any estimator that use the parameters separately. This paradox was discovered by Charles Stein of Stanford University. You can read more about Stein's paradox in his original paper (see, Stein (1956)) and by clicking here.

[3] When the regressors are orthogonal, for OLS notice

$$\sum_{j=1}^{p} \mathbb{E}[(\widehat{\beta}_j - \beta_j)^2] = \sum_{j=1}^{p} \mathrm{var}[\widehat{\beta}_j] = p\sigma^2 \, .$$

For you: Verify the MSE for the James-Stein estimator.

The Sclove Estimator I

There are two issues with the James-Stein estimator:

1. The shrinkage factor $1 - \frac{(p-2)\sigma^2}{\sum_{j=1}^{p} \widehat{\beta}_j^2}$ can be negative, and
2. σ^2 is often unknown in practice.

To address the first issue, Sclove (1968) suggested a simple modification: Set the coefficients to zero if you get something negative. In other words, define a new shrinkage estimator via

$$\widetilde{\beta}_j = \left(1 - \frac{(p-2)\sigma^2}{\sum_{j=1}^{p} \widehat{\beta}_j^2}\right)^+ \widehat{\beta}_j \tag{9}$$

where $(x)^+ := \max(x, 0)$.

The Sclove Estimator II

For the second issue, if σ^2 is unknown, Sclove (1968) proposed the modified estimator

$$\widetilde{\beta}_j = \left(1 - c\frac{\widehat{\varepsilon}'\widehat{\varepsilon}}{\sum_{j=1}^{p}\widehat{\beta}_j^2}\right)^+ \widehat{\beta}_j \tag{10}$$

for some value of c and where $\widehat{\varepsilon}$ are the residuals from OLS. This is called the *Sclove* estimator.

The Sclove Estimator III

In the situation where all independent variables are orthogonal, there is a simple interpretation of this estimator.

Recall that the F-test statistic for $H_0 : \beta_1 = \ldots = \beta_j = 0$, is given by

$$F = \frac{\sum_{j=1}^{p} \widehat{\beta}_j^2 / p}{\widehat{\varepsilon}' \widehat{\varepsilon} / (n-p)} \sim F_{p,n-p} \,. \tag{11}$$

Using this fact, we see that the Sclove estimator can be expressed as

$$\widetilde{\beta}_j = \left(1 - \frac{\alpha}{F}\right)^{+} \widehat{\beta}_j \tag{12}$$

where $\alpha = c\frac{p}{n-p}$.

In other words, we set all the coefficients to zero unless $F > \alpha$. Equivalently, we set all the coefficients to zero if we reject an F-test at significance level α.

Decision Theory and Loss Functions

Decision Theory and Loss Functions I

Suppose we have a random input vector $\mathbf{x}$ and a random output variable y.

The joint probability distribution $p(\mathbf{x}, y)$ provides a complete summary of the uncertainty associated with these variables. Determining $p(\mathbf{x}, y)$ from a set of training data is an exercise in statistical inference that in general can be very difficult to solve.

However, in many practical settings we do not need the full joint probability distribution. Rather, we are often interested in "just" predicting the value of y given values of the input $\mathbf{x}$. How to best make this prediction is addressed in *decision theory*.

Decision Theory and Loss Functions II

We seek a function $f : \mathbb{R}^n \to \mathbb{R}$ such that "$f(\mathbf{x}) \approx y$" for each input $\mathbf{x}$. To measure the quality of the resulting prediction, we use the *squared error loss*

$$L_2(y, f(\mathbf{x})) := (y - f(\mathbf{x}))^2 \,. \tag{13}$$

The resulting *expected loss*, $\mathbb{E}[L]$, is given by

$$\begin{aligned} \mathbb{E}[L_2] &:= \iint L_2(y, f(\mathbf{x})) p(\mathbf{x}, y) \, \mathrm{d}\mathbf{x} \, \mathrm{d}y & (14) \\ &= \iint \big(y - f(\mathbf{x})\big)^2 p(\mathbf{x}, y) \, \mathrm{d}\mathbf{x} \, \mathrm{d}y \,. & (15) \end{aligned}$$

The expected loss with a squared error loss function is of course the mean squared error (MSE) we saw before. This time we are evaluating the MSE of the prediction. Above we used MSE to evaluate estimators.

The Best Predictor Under Mean Squared Error

The Best Predictor Under Mean Squared Error I

Proposition

$\mathbb{E}[y \mid \mathbf{x}]$ is the best predictor of y in that it minimizes the mean squared error. In other words,

$$\mathbb{E}[y \mid \mathbf{x}] = \text{argmin}_{f(\mathbf{x})} \mathbb{E}\left[(y - f(\mathbf{x}))^2\right] .$$

The Best Predictor Under Mean Squared Error II

Proof. Let us denote by $f(\mathbf{x})$ any forecast for y. We can write the forecast error as

$$y - f(\mathbf{x}) = (y - \mathbb{E}[y \mid \mathbf{x}]) + (\mathbb{E}[y \mid \mathbf{x}] - f(\mathbf{x})) \,. \tag{16}$$

Therefore, the squared loss is

$$\begin{aligned}(y - f(\mathbf{x}))^2 = (y - \mathbb{E}[y \mid \mathbf{x}])^2 &+ 2(y - \mathbb{E}[y \mid \mathbf{x}])(\mathbb{E}[y \mid \mathbf{x}] - f(\mathbf{x})) \\ &+ (\mathbb{E}[y \mid \mathbf{x}] - f(\mathbf{x}))^2 \,.\end{aligned}$$

Taking expectation of both sides, we obtain

$$\begin{aligned}\mathsf{MSE} &= \mathbb{E}\left[(y - f(\mathbf{x}))^2\right] \\ &= \mathbb{E}\left[(y - \mathbb{E}[y \mid \mathbf{x}])^2\right] + 2\mathbb{E}\left[(y - \mathbb{E}[y \mid \mathbf{x}])(\mathbb{E}[y \mid \mathbf{x}] - f(\mathbf{x}))\right] \\ &\quad + \mathbb{E}\left[(\mathbb{E}[y \mid \mathbf{x}] - f(\mathbf{x}))^2\right] \,. \end{aligned} \tag{17}$$

The Best Predictor Under Mean Squared Error III

We observe that by the law of total expectations the cross-term above can be rewritten as

$$\begin{aligned} 2\mathbb{E}\left[\phi(\mathbf{x})(y - \mathbb{E}[y \mid \mathbf{x}])\right] = & \qquad (18) \\ = 2\mathbb{E}\left[\mathbb{E}\left[\phi(\mathbf{x})(y - \mathbb{E}[y \mid \mathbf{x}]) \mid \mathbf{x}\right]\right] = 0\,, & \qquad (19) \end{aligned}$$

where $\phi(\mathbf{x}) := \mathbb{E}[y \mid \mathbf{x}] - f(\mathbf{x})$. Hence we have

$$\begin{aligned} \mathsf{MSE} &= \mathbb{E}\left[(y - \mathbb{E}[y \mid \mathbf{x}])^2\right] + \mathbb{E}\left[(\mathbb{E}[y \mid \mathbf{x}] - f(\mathbf{x}))^2\right] & (20) \\ &\geq \mathbb{E}\left[(y - \mathbb{E}[y \mid \mathbf{x}])^2\right]\,, & (21) \end{aligned}$$

where equality is obtained for $f(\mathbf{x}) \equiv \mathbb{E}[y \mid \mathbf{x}]$. □

The Best Predictor Under Mean Squared Error IV

Remark

The squared loss is not the only possible choice of loss function for regression problem. There are situations where squared loss can lead to poor results and other approaches are needed. One example is when the conditional distribution $p(y \mid \mathbf{x})$ is multimodal. One generalization of the squared loss is the *Minkowski loss*

$$L_q(y, f(\mathbf{x})) := |y - f(\mathbf{x})|^q \,. \tag{22}$$

The Minkowski loss reduces to the squared loss for $q = 2$ and we showed above that the minimum of $\mathbb{E}\left[L_2\right]$ is given by the conditional mean.

One can show that:
(a) for $q = 1$ the minimum is achieved by some function you will determine in the homework, and
(b) for $q \to 0$ the minimum is achieved by the conditional mode.

The Best Linear Predictor Under Mean Squared Error

The Best Linear Predictor Under Mean Squared Error I

If we restrict ourselves to predictors that are linear functions of $\mathbf{x}$, it is natural to ask: What is the best linear predictor (in a MSE sense) of y based on $\mathbf{x}$?

Consider $\boldsymbol{\beta}^*$ satisfying the orthogonality condition

$$\mathbb{E}\left[\mathbf{x}\left(y - \mathbf{x}'\boldsymbol{\beta}^*\right)\right] = \mathbf{0}\,,$$

which can equivalently be expressed as $\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]\boldsymbol{\beta}^* = \mathbb{E}(\mathbf{x}y)$.

The Best Linear Predictor Under Mean Squared Error II

This is saying that $\boldsymbol{\beta}^*$ is determined such that the forecast error $y - \mathbf{x}'\boldsymbol{\beta}^*$ is orthogonal to $\mathbf{x}$.[4]

Assuming $\mathbb{E}\left(\mathbf{x}\mathbf{x}'\right)$ is nonsingular, the solution to the orthogonality condition is given by

$$\boldsymbol{\beta}^* = \left[\mathbb{E}\left(\mathbf{x}\mathbf{x}'\right)\right]^{-1} \mathbb{E}(\mathbf{x}y)\,.$$

The Best Linear Predictor Under Mean Squared Error III

We define the linear projection of y on $\mathbf{x}$ as

$$\widehat{\mathbb{E}}^*(y \mid \mathbf{x}) := \mathbf{x}'\boldsymbol{\beta}^* \tag{23}$$

where $\boldsymbol{\beta}^*$ is given by the formula above. This is called the *least squares projection* of y onto $\mathbf{x}$.

The Best Linear Predictor Under Mean Squared Error IV

Not surprisingly, we have the following result.

Proposition

The least squares projection $\widehat{\mathbb{E}}^*(y \mid \mathbf{x})$ is the best linear predictor of y in that it minimizes the MSE.

[4] Recall that we showed is satisfied by OLS.

Penalized Regression: The Ridge and Lasso I

As above, let us assume the regressors are orthogonal. Consider the loss function

$$L(\boldsymbol{\beta}) := (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta} \tag{24}$$

for some $\lambda \in \mathbb{R}_+$. Of course, this is the SSR we saw in OLS with an additional term that is proportional to the squared l^2-norm of $\boldsymbol{\beta}$.

Penalized Regression: The Ridge and Lasso II

Minimizing this loss function with respect to $\boldsymbol{\beta}$, we obtain[5]

$$\widetilde{\boldsymbol{\beta}} = \frac{1}{1+\lambda}\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}}\,. \tag{25}$$

We saw this type of result earlier where λ is a parameter that controls the amount of shrinkage.

This example suggests that we can obtain the shrinkage estimators above by modifying the standard OLS loss function. In fact, as we will see, this principle is very general.

[5] This is for you to verify on your own.

Ridge Regression I

Consider the following optimization problem with general regressors (not necessarily orthogonal as in the previous section).

$$\text{min}_{\boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\lambda\boldsymbol{\beta}'\boldsymbol{\beta} \tag{26}$$

is referred to as *ridge regression*. Some additional terminology:

- The term $\|\boldsymbol{\beta}\|_2^2$ is an example of a *regularizer*.
- λ is referred to as the *regularization parameter* and determines how much we want to penalize the flexibility of our model.
- Ridge regression is an example of *penalized linear regression*, also referred to as *regularized linear regression*.

Ridge Regression II

Remark

Regularization is not new. Tikhonov (see, for example, Tikhonov and Arsenin (1977)) began the research on developing stable methods for the numerical solution of inverse problems by using l^2-regularization. For history buffs, Neittaanmäki, Rossi, Majava, Pironneau, Engl, and Tröltzsch (2004) is interesting reading.

The finite-dimensional case was further developed in a statistical setting by Hoerl (1962) and Hoerl and Kennard (1970). Regularization is an important topic in machine learning and you will see it in this and many other courses.

Read about why it is called ridge regression here.

Ridge Regression III

The solution to ridge regression is obtained immediately by "completing the squares" of the loss function, yielding

$$L(\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta}\,. \tag{27}$$

Computing the FOC in our heads, we obtain the solution

$$\boldsymbol{\beta}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}\,. \tag{28}$$

Ridge Regression: Some Remarks I

Remark

- When λ is large enough the problem is non-singular even if $\mathbf{X}'\mathbf{X}$ is singular.
- For each choice of λ we obtain a different solution. Indexing each solution by λ allows us to trace out a path of the solutions (as a function of λ).
- As before, when $\lambda \to 0$ then $\widehat{\boldsymbol{\beta}}_{\text{ridge}} \to \widehat{\boldsymbol{\beta}}_{\text{OLS}}$; and when $\lambda \to \infty$ then $\widehat{\boldsymbol{\beta}}_{\text{ridge}} \to 0$.
- There are different ways to choose the "right" λ. Today it is standard practice to use *cross-validation* (CV).

Ridge Regression: Some Remarks II

Remark

Just like mean-variance optimization (MVO) in classical portfolio theory, ridge regression can be formulated in several equivalent ways. One such equivalent formulation is

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{29}$$

$$\text{s.t. } \boldsymbol{\beta}'\boldsymbol{\beta} \leqslant s \tag{30}$$

for some $s \in \mathbb{R}_+$. This formulation offers additional intuition as to what is happening. With many (highly) correlated variables in a linear regression model, their coefficients can become poorly determined and exhibit high variance. For example, a large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin. By imposing a "size constraint" on the coefficients as above this issues is alleviated.

Ridge Regression: Some Remarks III

Remark (Ridge Regression vs. Bayesian Regression)

Ridge regression can also be derived as the mean of a posterior distribution if we assume $y_i \sim \mathcal{N}(\mathbf{x}_i' \boldsymbol{\beta}, \sigma^2)$ with each parameter β_j having prior distribution $\mathcal{N}(0, \tau^2)$. Then the negative log-posterior density of $\boldsymbol{\beta}$ is exactly $L_{\text{ridge}}(\boldsymbol{\beta})$, with $\lambda = \sigma^2/\tau^2$.

Ridge Regression: Some Remarks IV

Remark (Ridge Regression and Normalization of the Data)

Ridge regression is dependent on the origin chosen for the “response” (the intercept). In their implementation of ridge regression, most numerical software libraries automatically normalize the data to so-called *correlation form* such that

$$\sum_{i=1}^{n} x_{ik} = 0 \tag{31}$$

and

$$\sum_{i=1}^{n} x_{ik}^{2} = 1\,. \tag{32}$$

Lasso Regression I

Consider the loss function

$$L(\boldsymbol{\beta}) := (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 \,, \tag{33}$$

for some $\lambda \in \mathbb{R}_+$ and where we have an l^1-regularizer.

Lasso Regression II

The resulting optimization problem

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) \tag{34}$$

is referred to as *least absolute shrinkage and selection operator regression*, or in short, *lasso regression*.

Lasso Regression III

Unlike ridge regression, in general there is no closed form solution for the lasso. However, when the regressors orthogonal, it can be shown that

$$\widehat{\beta}_{\text{lasso},j} = \widehat{\beta}_{\text{OLS},j} \cdot \max\left(0, 1 - \frac{\lambda}{|\widehat{\beta}_{\text{OLS},j}|}\right) . \tag{35}$$

Lasso Regression IV

Lasso regression is equivalent to

$$\text{min}_{\boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \tag{36}$$

$$\text{s.t. } ||\boldsymbol{\beta}||_1 \leqslant s \tag{37}$$

for some $s \in \mathbb{R}_+$.

Lasso Regression V

Remark (Lasso Regression and Normalization of the Data)

Just like ridge regression, the lasso depends on the origin chosen for the response. Therefore, most numerical software libraries automatically normalize the data to correlation form as above.

A Few Other Common Regularization Techniques for Regression

A Few Other Common Regularization Techniques for Regression I

In Friedman, Hastie, and Tibshirani (2017) you will read about other techniques for regression and how they compare to ridge and lasso regression.

We make a few general remarks about these techniques below.

A Few Other Common Regularization Techniques for Regression II

1. *Elastic net* combines the ridge and lasso penalties and hence uses the following regularizer

$$\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \,. \tag{38}$$

2. *Least Angle Regression* (LAR) is "democratic" version of forward stepwise regression.
   - LAR is closely related to the lasso and provides an extremely efficient algorithm for computing the entire lasso path.
3. *Principal Component Regression* (PCR) is referring to linear regression on orthogonalized regressors (obtained from PCA) and is similar to ridge regression.
   - Note that if the number of regressors is equal to the full rank, then PCR is the same as OLS.

A Few Other Common Regularization Techniques for Regression III

- Otherwise, if the number of regressors is less than the full rank, then PCR is a form of shrinkage that discards PCs corresponding to the smallest eigenvalues.
- Principal components regression is very similar to ridge regression. Ridge regression shrinks the coefficients of the principal components.

4. *Partial Least Squares* (PLS) finds directions that have high variance and have high correlation with the response (cf. principal component regression that find directions that have high variance).

References

- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2017). *The Elements of Statistical Learning*. Second Edition. Springer Series in Statistics Springer, Berlin.
- Hoerl, Arthur E. (1962). “Application of Ridge Analysis to Regression Problems”. In: *Chemical Engineering Progress* 58.3, pp. 54–59.
- Hoerl, Arthur E. and Robert W. Kennard (1970). “Ridge regression: Biased Estimation for Nonorthogonal Problems”. In: *Technometrics* 12.1, pp. 55–67.
- James, William and Charles Stein (1961). “Estimation with Quadratic Loss”. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. 1961, pp. 361–379.
- Neittaanmäki, P., T. Rossi, K. Majava, O. Pironneau, HW Engl, and F. Tröltzsch (2004). “Tikhonov Type Regularization Methods: History and Recent Progress”. In: *European Congress on Computational Methods in Applied Sciences and Engineering*. Ed. by P. P. Neittaanmäki, T. Rossi,