AI Strategy and Digital transformation
2. Regularization

Piotr Wójcik
University of Warsaw (Poland)
pwojcik@wne.uw.edu.pl

January 2025

Regularization – motivation

- if the actual relation between the target and explanatory variables is close to linear, linear regression will have **low bias**
- if the number of observations (n) is significantly greater than the number of variables (p), the linear regression result will also have **small variance**, so it will give good results also on the test sample
- if, however, n is not much bigger than p, the linear regression variance will increase and there may be a problem of **overfitting** and hence – weaker forecasts on the test sample
- when $p > n$ linear regression does not have a unique solution, the model variance grows to infinity – linear regression **can not** be applied

Regularization – motivation – cont'd

- imposing additional restrictions on the estimated β coefficients, one can significantly **reduce the variance of the model**, at the cost of some **increase in the bias of the model**
- this can lead to a significant improvement in forecasts from the model and also allow the use of linear regression even when $p > n$
- often many variables included in the regression model are irrelevant – they do not influence the studied phenomenon
- leaving them in the model causes unnecessary **increase in its complexity**
- removing these variables – setting their parameters to 0 – will result in a model which is **easier to interpret**

Regularization – motivation – cont'd

- Let's assume that we need to **explain a person's weight** based on the observation of people ine the room
- We would do a fairly decent job just by saying that **taller people are heavier**
- Then we would probably say that **men are on average heavier than women**, and so on.
- At some point we would run out of **sensible rules**.
- This would make us create rules that apply to small subgroups of individuals or even single observations.
- This would **lead to overfitting**
- If adding new rules is costly, there is a **trade-off between cost of a new rule and its explanatory power** (e.g. increasing goodness of fit)
- Depending on the way in which this cost is introduced to the loss function it can lead to a decrease of a parameter or even its elimination.

Methods for variables selecting

- we already previously discussed before some methods of automatic variable selection (stepwise) or their initial filtering
- the alternative is to use the so-called **regularization** (also called **shrinkage**), which, depending on the variant, might also be a method of selecting variables – imposing restrictions on some parameters of the model – equating them to 0
- in this case, a full model with p variables is estimated with an additional constraint, which causes the estimated parameters to be closer to 0 (*shrinking* to 0) or some of them even be equal to 0

Methods of regularization – ridge regression and LASSO

- **intuitively** – the simpler the model, the lower the risk of overfitting
- so simplifying the model, **even at the cost of some additional bias**, may result in better forecasts on the test sample
- in regularization, the simplification of the model consists in **reducing the value of some parameters in direction to 0** (*shrinking*)
- the two most-known **methods of regularization**, imposing restrictions on parameters that bring them closer to 0, are the **ridge regression** and **LASSO** (*Least Absolute Shrinkage Selector Operator*)

Regularization

- in the OLS method we look for such parameters β, that **minimize the sum of the squared errors** of the model:

$$\min_{\hat{\beta}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{\hat{\beta}} RSS$$

- in the case of regularization, we also want the parameters to be **as small as possible** (nearest zero)

Regularization – *ridge regression*

- in the ridge regression the above formula is extended by an additional element:

$$\min_{\hat{\beta}}\left[\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right]=\min_{\hat{\beta}}\left[\sum_{i=1}^{n}e_i^2+\lambda\sum_{j=1}^{p}\beta_j^2\right]$$

- where $\lambda \geq 0$ is a parameter that **requires tuning**
- just like in OLS, we are looking for parameters β, which give a **good fit to the data** – low RSS (the first element of the formula)
- at the same time the second element imposes a penalty for too large values of parameters $\lambda\sum_{j=1}^{p}\beta_j^2$ (*shrinkage penalty*) – the higher λ the stricter the "punishment"
- **NOTE!** It is worth noting that "penalty" **does not include a constant term** – β_0 from the model

Regularization – *ridge regression* – cont'd

- adding a **penalty** in the optimization results in searching for parameters that **fit the data well**, but **are as small as possible** (nearest 0)
- parameters at less important variables will not necessarily be equal to 0, but their impact on the model **will be limited** (closer to zero values of β)
- for $\lambda = 0$ the model simplifies to regular linear regression (OLS)
- different values of λ will result in **different model parameters** β
- finding the optimal value of the parameter λ is is found with the use of **cross-validation**

Regularization – *ridge regression* – standardization

- in a standard linear regression, parameter estimates are not sensitive to changing the variable scale
- multiplying the X_j variable by the constant c will multiply its parameter by $1/c$
- in other words, regardless of the resizing of j-th variable, the product of the variable and the parameter $X_j\beta_j$ will remain unchanged
- in turn, in ridge regression, due to additional constraints, changing the variable scale may cause a disproportionate change in the value of the estimated parameter and the product of $X_j\beta$
- the product $X_j\beta_j$ can change even if the scale of **other explanatory variables** is changed
- therefore, it is recommended to use regularized regression on **standardized** variables – reduced to **common scale** (it is enough to divide each variable by its standard deviation)

Regularization – *ridge regression* – disadvantages

- the main disadvantage of ridge regression, as the method of model selection, is leaving all p variables in the model, although with reduced parameter values – unless $\lambda = \infty$
- it does not need to be a problem in prediction – the accuracy of forecasts may be high, but it may make the model difficult to interpret – concluding which variables are the most important
- the solution to this problem is to use a different "penalty" formula for too large parameter values

Regularization – *LASSO*

- in **LASSO** regression the linear regression formula is expanded in the following way:

$$\min_{\hat{\beta}}[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|] = \min_{\hat{\beta}}[RSS + \lambda \sum_{j=1}^{p} |\beta_j|]$$

- similarly to ridge regression, LASSO shrinks the values of β parameters to zero
- another way of punishing causes that for a sufficiently large but finite value of the parameter λ some parameters β will take the value 0
- so the LASSO method can be considered as the **variable selection method** in the model
- models obtained from LASSO regression are usually easier to interpret than the result of ridge regression

Regularization – alternative representation

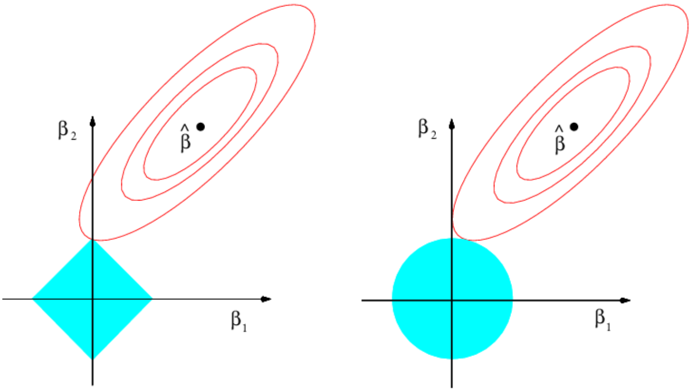
The optimization performed in the ridge and LASSO regressions can alternatively be shown as:

- **ridge**: $\underset{\hat{\beta}}{min} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ subject to $\sum_{j=1}^{p} \beta_j^2 \leq s$
- **LASSO**: $\underset{\hat{\beta}}{min} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ subject to $\sum_{j=1}^{p} |\beta_j| \leq s$
- in other words, for each value of λ there is a number s that will give identical estimation results

Regularization – alternative presentation – cont'd

- for example for $p = 2$ LASSO coefficients have the lowest RSS value among all combinations of parameters lying within the **diamond** specified by $|\beta_1| + |\beta_2| \leq s$.
- similarly in the ridge regression coefficients have the lowest RSS from all the combinations within the **circle** described by: $\beta_1^2 + \beta_2^2 \leq s$
- we can think of this additional limitation as a *budget* in which the β parameters of the model must fit
- when s is very large (λ is small), the budget is not very restrictive, so the parameters can be large
- you can interpret ridge and LASSO regressions as a **computationally efficient** alternative to **choosing the best subset of variables**

Regularization – graphical representation



Source: James et al (2017), p. 222

Regularization – graphical representation – comment

- the OLS solution is marked as $\hat{\beta}$, while the **blue diamond and circle** indicate the restrictions imposed by the ridge and LASSO regression respectively
- for a sufficiently large s areas describing the *budget* restriction will include $\hat{\beta}$ – then regression and LASSO will give the same result as OLS
- however, in the above figure, the OLS solution lies outside the set of available options
- ellipses with centroids in $\hat{\beta}$ represent coutours with **equal values** of RSS

Regularization – graphical representation – comment

- the further away from the OLS solution, the higher the RSS
- LASSO and ridge solution are points on ellipses tangent to the *budget* limitation
- because the ridge regression has a spherical restriction, the point of contact will NOT be on any of the axes
- in the case of LASSO – on the contrary – due to the shape of the constraint, the tangent point will usually be on one axis
- then (at least) one of the coefficients will be equal to 0

Which to use?

- none of the two discussed methods dominates in each case
- it can be expected that **LASSO** will work better in a situation where a relatively **small number of predictors have significant coefficients**, and the remaining predictors have coefficients that are very small or equal to zero
- **ridge regression** will work better when the dependent variable is a function of many predictors and all have coefficients of **comparable value**
- however, the number of predictors affecting the dependent variable is never known a priori for real data
- that is why it is always worth comparing both methods and verifying which is better eg using cross validation

Elastic net

- a combination of both constraints can be used – this is called **elastic net**
- in this case there is one more parameter needed (α) and the additional restriction for OLS becomes:

$$\underset{\hat{\beta}}{min}[RSS + \lambda \sum_{j=1}^{p} \left(\alpha|\beta_j| + (1-\alpha)|\beta_j|^2\right)]$$

- ridge and LASSO are special cases of such an elastic net – for $\alpha = 1$ one obtains pure LASSO, while pure ridge regression is obtained for $\alpha = 0$
- elastic net in linear models tends to **group correlated variables** – their parameters are kept at similar level

regularization – practical exercises in python

Thank you for your attention