# report_3

Kamil Zaborniak

## Contents

# Multiple regression properties

In order to analyze later designs, we will first recall the properties of the trace of a symmetric matrix, consider the matrix $X^TX$, where $X$ is a real matrix of sizes $n \times p$. Then we will repeat the matrix selection criteria and solve simple tasks related to them. Finally, we will recall how LASSO works.

1. The trace of the symmetric real matrix

   Let $X$ be a real symmetric matrix of size $n \times n$. Let $\lambda_1, \ldots, \lambda_n$ and $e_1, \ldots, e_n$ be the eigenvalues and eigenvectors of matrix $X$, respectively, where the vector $e_i$ corresponds to the eigenvalue of $\lambda_i$. From previous reports, we know that for any symmetric matrix we can write as $X = P\Lambda P^T$, where the matrix $P = [e_1, \ldots, e_n]$ and $\Lambda$ is a diagonal matrix whose diagonal is the vector $[\lambda_1, \ldots, \lambda_n]$. Then the trace of $X$:

   $$tr(X) = tr\left(P\Lambda P^T\right) = tr\left(P^TP\Lambda\right) = tr(\Lambda) = \sum_{i=1}^{n} \lambda_i.$$

2. Properties of $X^TX$, where $X$ is the real matrix of the dimension $n \times p$.

   Let's suppose vector $a \in \mathbb{R}^p \setminus \{0\}$, then:

   $$a^T\left(X^TX\right)a = (Xa)^T(Xa) = \|Xa\|^2 \geq 0.$$

   This is the proof that $X^TX$ is semipositive definite. Let $\lambda_i$ and $e_i$, where $i \in \{1, \ldots, n\}$ be the eigenvalue and the corresponding eigenvector of the matrix $X^TX$. We know that $\forall_{i \in \{1,\ldots,n\}} X^TXe_i = \lambda_i e_i$. Hence:

   $$\begin{aligned} e_i^TX^TXe_i &\geq 0 \\ e_i^T\lambda_i e_i &\geq 0 \\ \lambda_i &\geq 0 \end{aligned}$$

   What proves that eigenvalues of $X^TX$ are larger or equal to zero.

   When $p > n$, then at least one eigenvalue of $X^TX$ is equal to 0. The rank of $X^TX$:

   $$rank\left(X^TX\right) \leq \min\left(rank\left(X^T\right), rank\left(X\right)\right) \leq n.$$

   Hence $det(X^TX) = 0 = \prod_{i=1}^{p} \lambda_i$ ($X^TX$ is symmetric), what is means that at least one of eigenvalues $\lambda_i$ must be equal to 0.

3. Model selection criteria

   Below we will consider three known criteria for selecting models:

   - Akaike information criterion (AIC) - is the general model selection criteria. It minimizes $RSS + 2k\sigma^2$, where $k$ is the number of estimated model parameters.
   - Bayesian Information Criterion (BIC) - is used to approximate the posterior probability of a given model. We use when the number of data ($n$) is greater than 8.it minimizes $RSS + \sigma^2 k \log n$.
   - Risk Information Criterion (RIC) - it minimizes $RSS + \sigma^2 \cdot 2k \log p$. We use this criterion when the number of model variables ($p$) is large.

   Let's assume that our data contains 10 variables. We fit 10 regression models including the first variable, the first two variables, etc. The residual sums of squares for these 10 consecutive models are equal to $(1731, 730, 49, 38.9, 32, 29, 28.5, 27.8, 27.6, 26.6)$. The sample size is equal to 100. Assuming that the standard deviation of the error term is known: $\sigma = 1$. The 6th. model will be selected by AIC, and 5th model will be selected by BIC and RIC.

   Assuming the orthogonal design ($X^TX = I$) and $n = p = 10000$ and none of the variables is really important (i.e. $p_0 = p$). The probability of a type I error according to AIC is:

   $$P_{AIC}\left(X_i \text{ is selected } |\hat{\beta}_i = 0\right) = 2(1 - \Phi(\sqrt{2})),$$

hence the expected number of false discoveries is equal to

$$0.15729 \cdot 10000 \approx 1573.$$

The probability of a type I error according to BIC is:

$$P_{BIC}\left(X_i \text{ is selected } |\hat{\beta}_i = 0\right) = 2(1 - \Phi(\sqrt{\log n})).$$

The expected number of false discoveries is equal to:

$$0.002406 \cdot 10000 \approx 24.$$

The probability of a type I error according to RIC is:

$$P_{RIC}\left(X_i \text{ is selected } |\hat{\beta}_i = 0\right) = 2(1 - \Phi(\sqrt{2\log p})).$$

The expected number of false discoveries is equal to:

$$0.000017712 \cdot 10000 \approx 0.$$

4. Ridge regression properties under the orthogonal design.

Ridge regression is used in the context of fitting linear regression models in which the number of design matrix variables is greater than the number of data ($X_{n\times p}$, p>n). It was proposed to reduce the variance of least squares estimators. The idea behind ridge regression estimators is similar to James-Stein estimators. It introduces bias, but the variance of the estimator will be smaller, which makes the mean square error smaller. Ridge regression estimators significantly improve least squares estimators.

Assume a design matrix $X$ such that $X^TX = I$, then

$$\hat{\beta}_{RR} = (X^TX + \gamma I)^{-1}X^TY = \frac{1}{1+\gamma}X^TY = \frac{1}{1+\gamma}\hat{\beta}_{LS},$$

where $Y$ is the regressand.

The formula of the bias of the ridge regression estimators under the orthogonal design matrix has form:

$$b\left(\hat{\beta}_{RR}\right) = \mathbb{E}\left[\left\|\hat{\beta}_{RR} - \beta\right\|\right] = \mathbb{E}\left[\left\|\hat{\beta}_{RR} - \beta\right\|\right] = \mathbb{E}\left[\left\|\frac{1}{1+\gamma}\hat{\beta}_{LS} - \beta\right\|\right] = -\frac{\gamma}{1+\gamma}\beta.$$

Knowing that the variance of the least squares estimate of $\hat{\beta}_{LS}$ is $\sigma^2 I$, the variance of $\hat{\beta}_{RR}$ is equal to:

$$Var\left[\hat{\beta}_{RR}\right] = Var\left[\frac{1}{1+\gamma}\hat{\beta}_{LS}\right] = \frac{\sigma^2}{(1+\gamma)^2}I.$$

Mean square error of $\hat{\beta}_{RR}$ is equal to:

$$MSE\left(\hat{\beta}_{RR}\right) = b\left(\hat{\beta}_{RR}\right)^2 + Var\left[\hat{\beta}_{RR}\right] = \frac{\gamma^2}{(1+\gamma)^2}\|\beta\| + \frac{p\sigma^2}{(1+\gamma)^2}.$$

Ridge regression estimator is better than least square estimator when $\|\beta\|^2 \leq p\sigma^2$ or $\gamma < \frac{2p\sigma^2}{\|\beta\|^2 - p\sigma^2}$ with $\|\beta\|^2 > p\sigma^2$.

Let's consider the following situation. For a given data set with 40 explanatory variables the residual sums of squares from the least squares method and the ridge regression are equal to: 4.5 and 11.6, respectively. For the ridge regression the trace of $X(X^TX+\gamma I)^{-1}X^T$ is equal to 32. Prediction errors are equal to:

$$PE_{RR} = 11.6 + 2\sigma^2 \cdot 32 = 11.6 + 64\sigma^2,$$
$$PE_{LS} = 4.5 + 2\sigma^2 \cdot 40 = 4.5 + 80\sigma^2.$$

When $\sigma^2 > \frac{71}{160}$, then the ridge regression is better.

5. Properties of LASSO under the orthogonal design.

   Least absolute shrinkage and selection operator. LASSO resets the ith. coordinate $\hat{\beta}_{LS}$ if $|\hat{\beta}_{LS,i}| < \lambda$. Usually (in our case) $\lambda = \sigma\Phi(1-\alpha/(2p)) \approx \sigma\sqrt{2\log p}$, where $\alpha$ is the significance level.

   The expected value of false discoveries is equal to:

$$p \cdot P\left(|\hat{\beta}_{LS,i}| > \lambda | \beta_i = 0\right) = p \cdot P\left(\frac{|\hat{\beta}_{LS,i}|}{\sigma} > \frac{\lambda}{\sigma} | \beta_i = 0\right) = p \cdot 2\left(1 - \Phi\left(\frac{\lambda}{\sigma}\right)\right) = p \cdot 2\left(1 - (1 - \frac{\alpha}{2p})\right) = \alpha.$$

   The power of LASSO is equal to:

$$P\left(|\hat{\beta}_{LS,i}| > \lambda | \beta_i \neq 0\right) = P\left(\frac{\hat{\beta}_{LS,i} - \beta_i}{\sigma} < \frac{-\lambda - \beta_i}{\sigma} | \beta_i \neq 0\right) + P\left(\frac{\hat{\beta}_{LS,i} - \beta_i}{\sigma} > \frac{\lambda - \beta_i}{\sigma} | \beta_i \neq 0\right) = \Phi\left(\frac{-\lambda - \beta_i}{\sigma}\right) + 1 - \Phi\left(\frac{\lambda - \beta_i}{\sigma}\right).$$

## James-Stein estimators

To consider James-Stein estimators, let's analyze the data contained in the *Lab3.Rdata* dataset. This set contains expressions of 300 genes for 210 individuals.

First, we scale the data. We subtract the mean of this vector from the vector of each column of the $xx$ matrix, then divide it by its standard deviation and add the mean to it. This action is intended to preserve the mean of the column vector while equalizing its standard deviation to 1. Knowing that the average expression of each gene is 10, we subtract 10 from each column vector.

Based on the 'standardized' data obtained, we proceed to the next activities. Based on the first five records, we determine the maximum likelihood estimator and both James-Stein estimators (shrunk towards zero and towards common mean). To determine them, we will use $\sigma^2 = 0.2$ because $\hat{\mu}_{MLE} \sim N(\beta, 0.2I)$. The maximum likelihood estimator of the vector of average gene expressions determined on the basis of the first 5 records is simply the average vector of the first five records. The James-Stein estimator shrunk towards to zero is given by:

$$\hat{\mu}_{JSc} = c_{JS}\hat{\mu}_{MLE},$$

where $c_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\|\hat{\mu}_{MLE}\|^2}\right)$. The James-Stein estimator shrunk towards to common mean is given by:

$$\hat{\mu}_{JSd} = (1-d)\hat{\mu}_{MLE} + d\overline{\hat{\mu}_{MLE}},$$

where $d = \frac{p-3}{p-1}\frac{\sigma^2}{Var[\mu_{MLE}]}$. To evaluate the estimators, we will assume that the actual vector of average gene expressions is the vector of averages of the remaining 205 individuals.

The graph of the determined estimators is as follows.

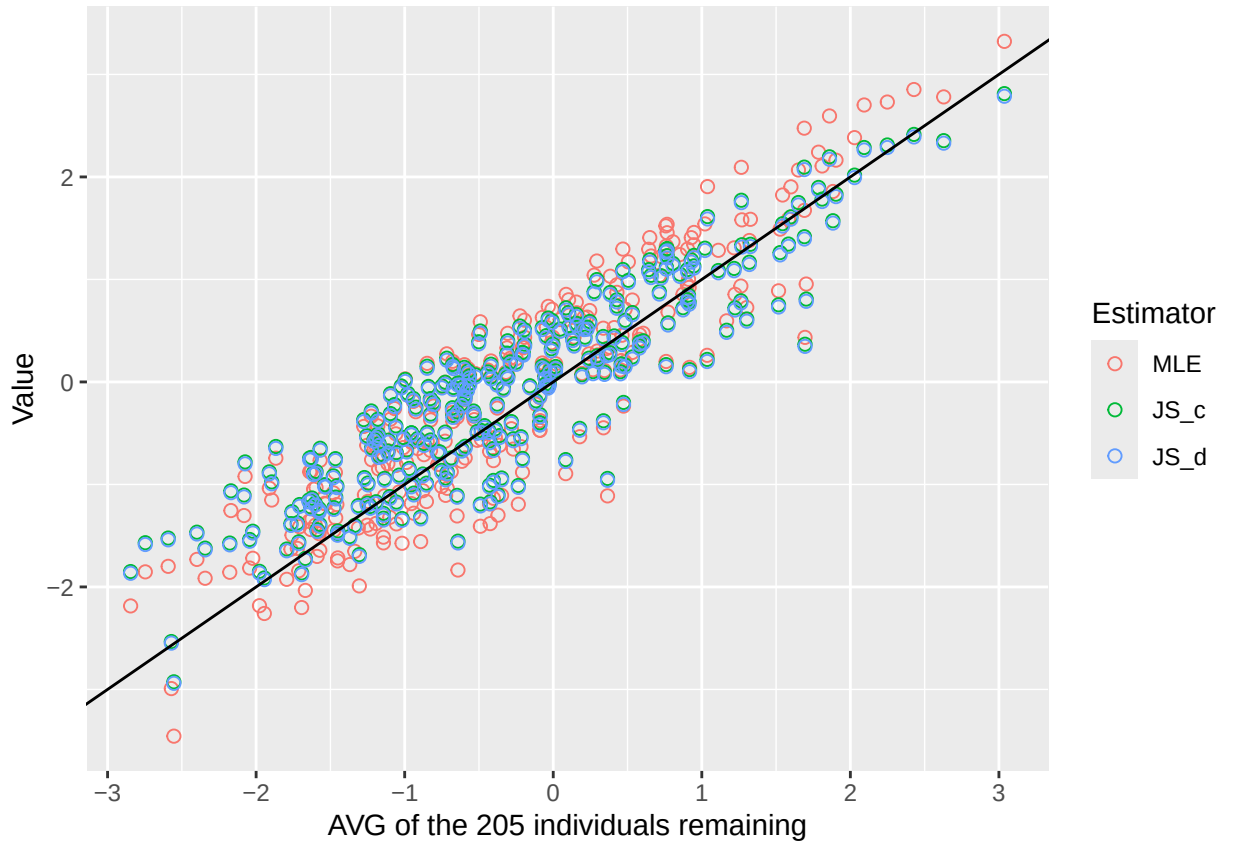


Figure 1: Scatter plot of the estimators versus average gene expressions.

We see that the James-Stein estimators lie very close to each other and lie closer to the $x = y$ line than the maximum likelihood estimators. The table below presents the squared errors of the estimators, where $SE(\hat{\mu}) = \|\hat{\mu} - \mu\|^2$ and $\mu$ is the vector of means of the remaining 205 records.

| | $\hat{\beta}_{MLE}$ | $\hat{\beta}_{JSc}$ | $\hat{\beta}_{JSd}$ |
|---|---|---|---|
| SE | 83.61966 | 77.66406 | 75.14637 |

Table 1: Square errors of the estimators.

James-Stein estimators have the least squared error, but the estimator shrunk towards common mean has the smallest SE. The above results and the graph confirm the theory that James-Stein estimators are better than the maximum likelihood estimator when the estimated parameter is dimension larger than 2.

# Prediction Error in Multiple Regression

## Simulation

In order to consider prediction error estimates, we will perform a simulation. In a single step, we will generate a $X$ matrix of size $1000 \times 950$, whose elements will come from the $N(0, 0.001)$ distribution. We will generate the vector of the response variable according to the model $Y = X\beta + \epsilon$, where $\beta = (3, 3, 3, 3, 3, 0, ..., 0)^T$ and $\epsilon \sim N(0, I)$. We will build models based on $k$ first variables, where $k \in \{2, 5, 10, 100, 500, 950\}$.

For each model we will estimate $\beta$ using the least squares method and determine the $RSS = \|Y - X\hat{\beta}_{LS}\|^2$ and the expected value of the prediction error

$$PE = \mathbb{E}\|X(\beta - \hat{\beta})\| + n\sigma^2 = \|X(\beta - \hat{\beta})\| + 1000.$$

We will determine the PE estimates assuming that $\sigma$ is known:

$$\hat{PE}_1 = RSS + 2 \cdot 1 \cdot k,$$

and assuming that $\sigma$ is unknown:

$$\hat{PE}_2 = RSS + 2 \cdot \frac{RSS}{n-k} \cdot k.$$

We will determine the last PE estimator based on leave-one-out cross-validation:

$$\hat{PE}_3 = \sum_{i=1}^{n} \left( \frac{Y_i - \hat{Y}_i}{1 - M_{ii}} \right)^2,$$

where $M = X(X^T X)^{-1} X^T$.

After repeating a single simulation step 30 times for each $k$ value, we will average the actual prediction error value and the PE estimator values, and plot the differences between the actual values and the estimators.

## Results

| k | PE | $\hat{PE}_1$ | $\hat{PE}_2$ | $\hat{PE}_3$ |
|---|---|---|---|---|
| 2 | 1001.960 | 1013.021 | 1013.065 | 1013.134 |
| 5 | 1004.906 | 1011.597 | 1011.663 | 1011.699 |
| 10 | 1011.0847 | 1005.9952 | 1005.9143 | 1005.9260 |
| 100 | 1102.1281 | 1098.9009 | 1098.6567 | 1110.5355 |
| 500 | 1493.3838 | 1503.5889 | 1510.7666 | 2017.4134 |
| 950 | 1960.0718 | 1948.0197 | 1872.7715 | 19828.6578 |

Table 2: Real and estimated values of the prediction error.

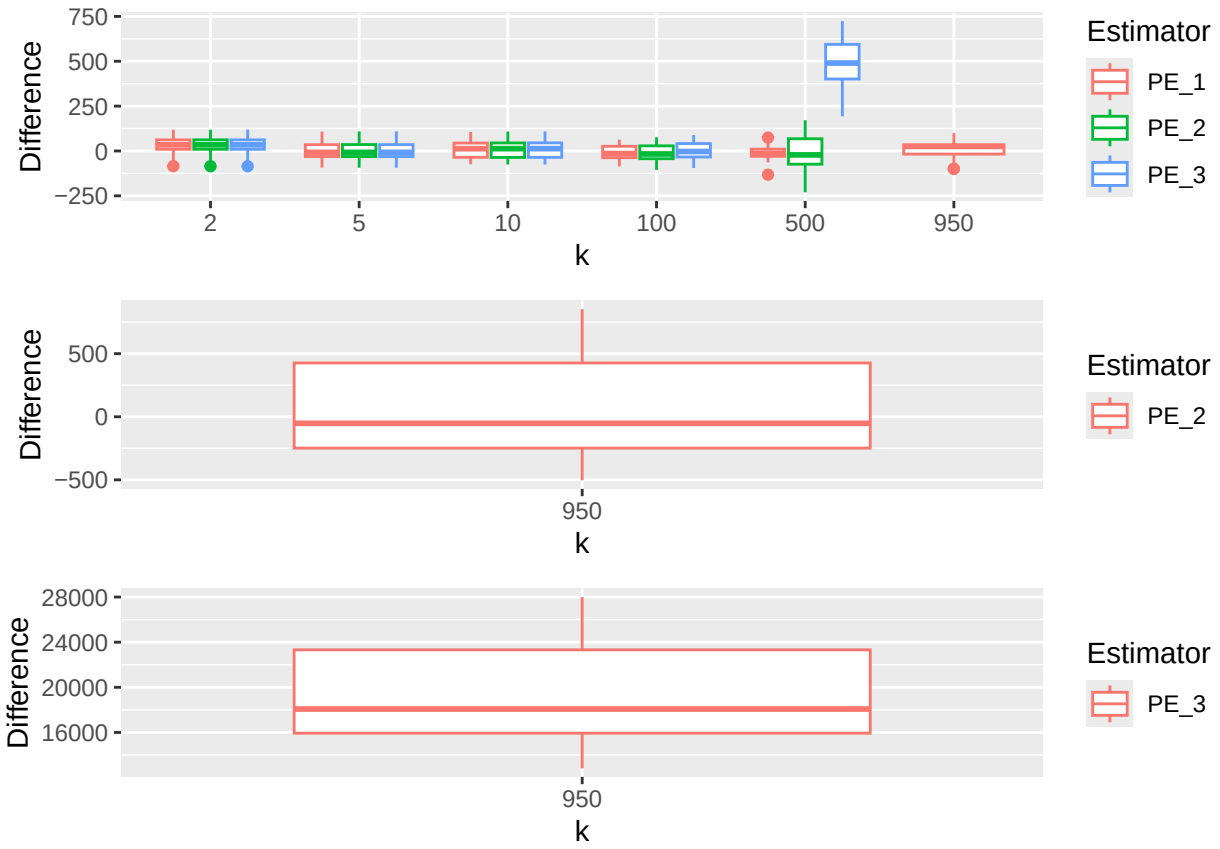


Figure 2: Boxplots of differences betweent real PE and its estimators, depends on k.

Based on above table and plot I would choose model with first five variables. The models based on 2, 5, 10, 100 variables gives similar results. For $k \geq 500$ the best estimator is the first estimator and the worst estimator is the LOO-CV estimator.

# Multiple regression - model selection and regularization

## Simulation

Using simulation, we will generate a $X$ plan matrix of size $1000 \times 950$, whose elements will come from the $N(0, 0.001)$ distribution. After that we will generate the vector of the response variable according to the model:

$$Y = X\beta + \epsilon,$$

where $\beta_1 = \ldots = \beta_20 = 6$, $\beta_{21} = \ldots = \beta_{950} = 0$ and $\epsilon \sim N(0, I)$. We will Analyse this data using:

- mBIC2 criterion,
- Ridge with the tuning parameter selected by cross-validation
- LASSO with the tuning parameter selected by cross-validation
- LASSO with the tuning parameter $\lambda = \Phi^{-1}(1 - \frac{0.1}{2p}$

- SLOPE with the BH sequence of the tuning parameters $\lambda_i = \Phi^{-1}(1 - \frac{0.1i}{2p})$

For each of these methods we will calculate the square estimation errors $\|\hat{\beta} - \beta\|^2$ and $\|X(\hat{\beta} - \beta)\|^2$. In case of LASSO and SLOPE we will consider also estimators obtained by performing the regular least squares fit within the selected model. For all methods apart from ridge we will calculate also the False Discovery Proportion and the True Positive Proportion (Power).

## Results

| Method | SE | $\|X(\hat{\beta} - \beta)\|^2$ | FDP | TPP |
|---|---|---|---|---|
| mBIC2 | 677.365 | 658.449 | 0 | 0.1 |
| Ridge (c-v) | 687.131 | 666.694 | | |
| LASSO (c-v, ".min") | 344.543 | 329.295 | 0.558 | 0.95 |
| LASSO (c-v, "1se") | 596.587 | 597.005 | 0.181 | 0.45 |
| LASSO | 340.048 | 324.429 | 0.577 | 0.95 |
| SLOPE | 393.968 | 363.156 | 0.840 | 0.95 |

Table 3: Result of the simulation.