Statistical learning

Assignment 3

Jakub Skalski

October 6, 2024

1. Trace of symmetric real matrix is the sum of its eigenvalues

We can easily show this by first decomposing real symmetric matrix A into its eigenvectors and eigenvalues and then using the circular property of the trace:

$$tr(A) = tr(P\Lambda P^T) = tr(PP^T\Lambda) = tr(\Lambda) = \sum_i \lambda_i$$

2. Properties of $A = X^T X$

2.1 Positive semi-definite

We say that matrix A is positive semi-definite if it satisfies $\mathbf{v}^T A\mathbf{v} \geq 0$ for all vectors $\mathbf{v} \in \mathbb{R}^n$. Consider any vector $\mathbf{v} \in \mathbb{R}^n$, then:

$$\mathbf{v}^T A\mathbf{v} = v^T X^T Xv = (Xv)^T Xv = \|Xv\|^2 \geq 0$$

2.2 Non-negative eigenvalues

Consider any eigenvector $\mathbf{v} \in \mathbb{R}^n$. Since A is positive semi-definite:

$$\mathbf{v}^T A\mathbf{v} = \mathbf{v}^T \lambda\mathbf{v} = \mathbf{v}^T\mathbf{v}\lambda \geq 0$$

Clearly, $\mathbf{v}^T\mathbf{v}$ is strictly positive, therefore $\lambda \geq 0$.

2.3 At least one eigenvalue is zero when p > n

Matrix is singular, if and only if it is not full rank:

$$rank(A) < p$$

First, note that $rank(A)$ is bounded from above by the rank of X^T:

$$rank(A) = rank(X^T X) \leq rank(X^T) \leq n$$

From the original assumptions it follows that $rank(X^T X) \leq n < p$.

3. Model selection

The data consists of 100 observations of 10 variables. We fit 10 regression models, where k-th model includes only the first k variables. 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). Let us consider different criteria for the model selection under the assumption of a standard error term.

3.1 Akaike Information Criterion

Table 1: Values of the criterion $AIC = RSS + 2k\sigma^2$ for each model

k	1	2	3	4	5	6	7	8	9	10
AIC	1733	734	55	46.9	42	41	42.5	43.8	45.6	46.6

3.2 Bayesian Information Criterion

Table 2: Values of the criterion $BIC = RSS + klog(n)\sigma^2$ for each model

k	1	2	3	4	5	6	7	8	9	10
$\approx BIC$	1735	739	63	57	55	57	61	65	69	72

3.3 Risk Inflation Criterion

Table 3: Values of the criterion $RIC = RSS + 2klog(p)\sigma^2$ for each model

k	1	2	3	4	5	6	7	8	9	10
$\approx RIC$	1735	739	63	57	55	57	61	65	69	72

4. False discovery rate

Assuming the orthogonal design ($X^TX = I$) and n = p = 10000 we calculate the expected number of false discoveries of AIC, BIC and RIC for $\hat{\beta}_i \sim \mathcal{N}(0, \sigma^2)$ when $\beta_i = 0$.

4.1 Akaike Information Criterion

AIC selects variables satisfying $\hat{\beta}_i \geq \sqrt{2}\sigma$, meaning the probability of type I error is:

$$P(X_i \text{ selected } | \beta_i = 0) = 2(1 - \Phi(\sqrt{2}) = 0.16$$

The number of false discoveries for this criterion is $0.16 * 10000 = 1600$.

4.2 Bayesian Information Criterion

BIC selects variables satisfying $\hat{\beta}_i \geq \sqrt{\log n}\sigma$, meaning the probability of type I error is:

$$P(X_i \text{ selected } | \beta_i = 0) = 2(1 - \Phi(\sqrt{\log n})) \approx 0.0024$$

The number of false discoveries for this criterion is $0.0024 * 10000 = 24$.

4.3 Risk Inflation Criterion

BIC selects variables satisfying $\hat{\beta}_i \geq \sqrt{2 \log p}\sigma$, meaning the probability of type I error is:

$$P(X_i \text{ selected } | \beta_i = 0) = 2(1 - \Phi(\sqrt{2 \log p})) \approx 0.00002$$

The number of false discoveries for this criterion is $0.00002 * 10000 = 0.2$.

5. Choosing the right criterion

5.1 Akaike Information Criterion

When the primary aim is to select a model that provides the best predictive performance rather than identifying the true underlying model, since it directly coincides with minimizing the prediction error.

5.2 Bayesian Information Criterion

When there is a belief that one of the candidate models is the true model and the sample size is relatively large, making it ideal for explanatory analysis.

5.3 Risk Inflation Criterion

Mostly useful in high-dimensional settings, such as variable selection in regression models with a large number of predictors.

6. Ridge regression formulas

The multiple regression models are used to describe the relationship between a dependent variable and multiple independent variables. In matrix notation, the multiple regression model can be formulated as follows:

$$Y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

The general ridge regression solution can be obtained analytically by solving for the zero gradient of this convex function:

$$\nabla[(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta] = 2X^TX\beta - 2X^TY + 2\lambda\beta = 0$$

The closed form general solution is thus:

$$\hat{\beta} = (X^TX + \lambda I)^{-1}X^TY$$

Under orthogonal design this simplifies to:

$$\hat{\beta} = \Lambda X^TY, \quad \text{where } \Lambda := \frac{1}{1+\lambda}$$

6.1 Bias

First, let us compute the expectation of $\hat{\beta}$:

$$E[\hat{\beta}] = \Lambda X^T E[Y] = \Lambda X^T X\beta = \Lambda\beta$$

Then the ridge regression bias is equal to the following:

$$E[\hat{\beta}] - \beta = \Lambda\beta - \beta = -\frac{\lambda}{1+\lambda}\beta$$

Which, in the case of OLS, where $\lambda = 0$ is zero.

6.2 Variance

The ridge regression estimator variance is the following:

$$V[\hat{\beta}] = V[\Lambda X^TY] = V[\Lambda X^T(X\beta + \epsilon)] = V[\Lambda X^T\epsilon] = \Lambda X^T\sigma^2 X\Lambda = \sigma^2\Lambda^2 = \frac{\sigma^2}{(1+\lambda)^2}$$

Which for OLS is simply σ^2.

6.3 Mean squared error

$$E[(\hat{\beta} - \beta)^T(\hat{\beta} - \beta)] = V[\hat{\beta}] + (E[\hat{\beta}] - \beta)^2 = \frac{\sigma^2}{(1+\lambda)^2} + \lambda^2\Lambda^2\beta^T\beta = \frac{\sigma^2 + \lambda^2\beta^T\beta}{(1+\lambda)^2}$$

which equates to σ^2 for OLS.

7. Prediction error

Suppose that 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 $X(X^TX + yI)^{-1}X^T$ is equal to 32. We will now compute and compare the resulting prediction errors of these two methods.

7.1 Ordinary least squares

$$\hat{PE_o} = RSS + 2\sigma^2 p = 4.5 + 80\sigma^2$$

7.2 Ridge regression

$$\hat{PE_r} = RSS + 2\sigma^2 Tr(M) = 11.6 + 64\sigma^2$$

7.3 Comparison

Assuming $\sigma = 1$, we surmise that PE_o is greater than PE_r.

8. LASSO

8.1 False discovery rate and power

Under orthogonal design, Lasso selects variable X_i when $|\hat{\beta}_i| > \lambda$, where $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2)$, so the false discovery rate equates to $P(X_i \text{ selected } |\beta_i = 0) = 2(1 - \Phi(\frac{\lambda}{\sigma}))$, while the power is $P(X_i \text{ selected } |\beta_i \neq 0) = 2(1 - \Phi(\frac{\lambda - \beta_i}{\sigma}))$.

8.2 Computing adaptive LASSO

Solving adaptive LASSO is actually the same as first solving regular LASSO for some scaled input data and then correcting the produced estimator.

8.2.1 Explanation

Let $\beta_j = \frac{1}{w_j} \cdot \beta_j^w$. Replacing in the minimization problem we obtain:

$$\hat{\beta}^w = \arg\min_{\beta^w} \left\{ \left\| y - X\left(\frac{1}{\mathbf{w}} \cdot \beta^w\right) \right\|_2^2 + \lambda \sum_{j=1}^{p} w_j \left| \frac{1}{w_j} \cdot \beta_j^w \right| \right\}$$

Now, we can simply multiply into the absolute value in the latter sum to get rid of the w_j, so that it becomes:

$$\hat{\beta}^w = \arg\min_{\beta^w} \left\{ \left\| y - X \left(\frac{1}{\mathbf{w}} \cdot \beta^w \right) \right\|_2^2 + \lambda \sum_{j=1}^{p} \left| \beta_j^w \right| \right\}$$

Finally, scaling the input data to account for this change we arrive at the regular LASSO formulation we are familiar with:

$$\hat{\beta}^w = \arg\min_{\beta^w} \left\{ \| y - X^w \beta^w \|_2^2 + \lambda \sum_{j=1}^{p} \left| \beta_j^w \right| \right\}$$

8.2.2 Solution

The approach is known as the LARS algorithm:

1. Define $\mathbf{x}_{i,j}^w = \mathbf{x}_j / w_j, \quad j = 1, 2, \ldots, p.$
2. Solve the lasso problem for the scaled data,
3. Output $\hat{\beta}_j = \hat{\beta}_j^w / w_j, \quad j = 1, 2, \ldots, p.$

Typically, one would simply pass the vector of absolute inverses of the initial weights to some regular LASSO solver (like glmnet, for instance).

8.3 Closed form solution under orthogonal design

Solving LASSO is equivalent to minimizing the following objective function:

$$\frac{1}{2} ||Y - X\beta||^2 + \sum_{i}^{p} \lambda_i |\beta_i|$$

Expanding and rearranging the terms we obtain:

$$-Y^T X\beta + \frac{1}{2} ||\beta||^2 + \sum_{i}^{p} \lambda_i |\beta_i|$$

Recall that solution to the least squares problem is $\hat{\beta} = X^T Y$, and so we can simplify further:

$$\sum_{i}^{p} -\bar{\beta}_i \beta_i + \frac{1}{2} \beta_i^2 + \lambda_i |\beta_i| = \sum_{i}^{p} Z_i$$

Minimizing the above amounts to minimizing each individual Z_i. Taking the subdifferential of such with respect to β and solving for zero gradient we arrive at:

$$\beta_i = \hat{\beta}_i - \lambda_i$$

Which is only feasible for the non-negatives. Adjusting for the general case we obtain the following closed form solution:

$$\beta_i^* = sgn(\hat{\beta}_i)(|\hat{\beta}_i| - \lambda_i)^+$$

8.4 Relation to the ordinary least squares method

One could obtain weights for each variable through OLS along with the estimator and use them for the adaptive lasso. For instance, suppose that the ordinary least squares weight w_1 is $\frac{1}{4}$ and its estimator of β_1 under the orthogonal design is equal to 3. Also, the regular LASSO estimator of this parameter is equal to 2. We can compute the tuning parameter first:

$$2 = sgn(3)(3 - \lambda)^+ \implies \lambda = 1$$

Then, the adaptive LASSO estimator would be:

$$sgn(\hat{\beta}_i)(|\hat{\beta}_i| - \lambda_i w_i) = 3 - \frac{1}{4} = \frac{11}{4}$$

9. Project 1

9.1 James-Stein Estimators

The most common estimation approach is the Maximum Likelihood Estimator (MLE), which often corresponds to the sample mean in many contexts. The sample mean is an unbiased estimator, meaning its expected value is equal to the true parameter value. When estimating the mean of a multivariate normal distribution with three or more dimensions, the sample mean is not the best estimator in terms of mean squared error (Stein paradox). The James-Stein estimator improves upon the sample mean by ”shrinking” it towards a central point (often the origin). This shrinkage reduces the mean squared error of the estimator.

9.1.1 Shrink to zero

$\hat{\mu}_{JS} = \left(1 - \sigma^2 \frac{p-2}{\|\bar{x}\|^2}\right) \bar{x}$

9.1.2 Shrink to the common mean

$\hat{\mu}_{JS} = \mu_0 + \left(1 - \sigma^2 \frac{p-2}{\|\bar{x} - \mu_0\|^2}\right) (\bar{x} - \mu_0)$

9.1.3 Experiment

Here, we test the validity of the previous statements. We compute the James-Stein estimators on a standardized genes data and compare them to the classical maximum

likelihood estimator (MLE). Estimators are computed from the first 5 observations, while the remaining 205 observations are used for validation.

Table 4: Mean squared estimation errors

estimator	mle	zero	mean
MSE	≈ 84	≈ 85	≈ 7

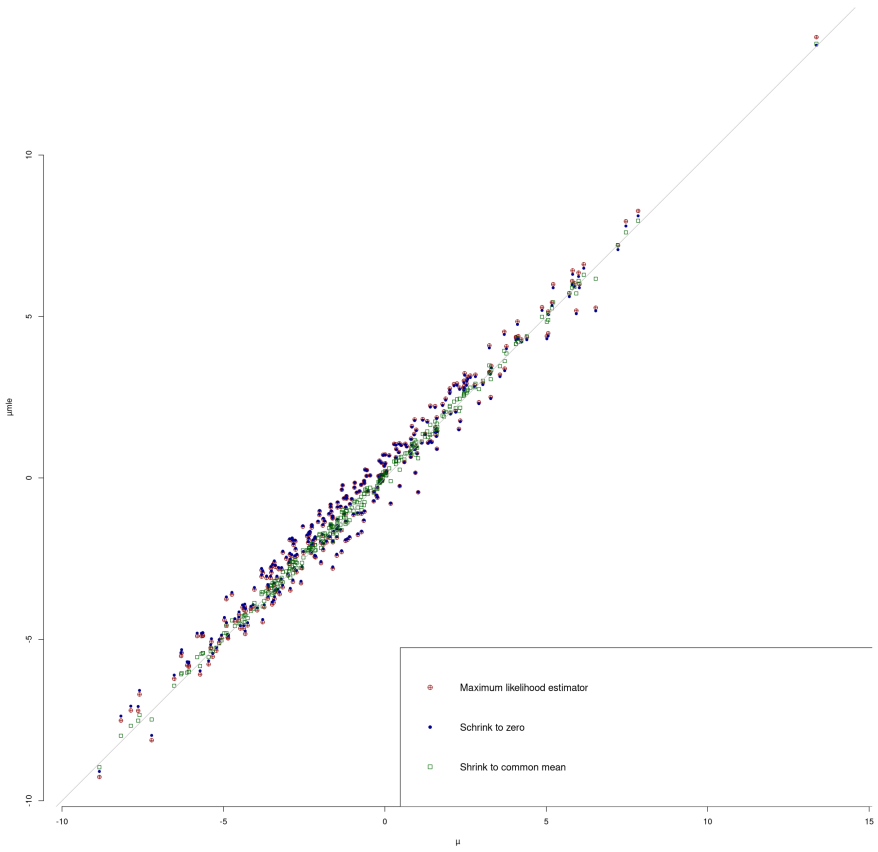

Figure 1: Estimators

9.2 Multiple Regression

Here we look into the ordinary least square method. The generated data is sampled from $N\left(0, \frac{1}{\sqrt{1000}}\right)$. The response variable Y is modeled as $Y = X\beta + \epsilon$, where the first five elements of β are 3, and the rest are 0. The error term ϵ follows $N(0, I)$. We fit the model for $k = 2, 5, 10, 100, 500, 950$ variables.

9.2.1 Prediction Errors

For each model we compute prediction error estimators:

- $PE1 = RSS + 2ps^2$
- $PE2 = RSS + 2p\sigma^2$
- $PE3 = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{1 - M_{ii}} \right)^2$

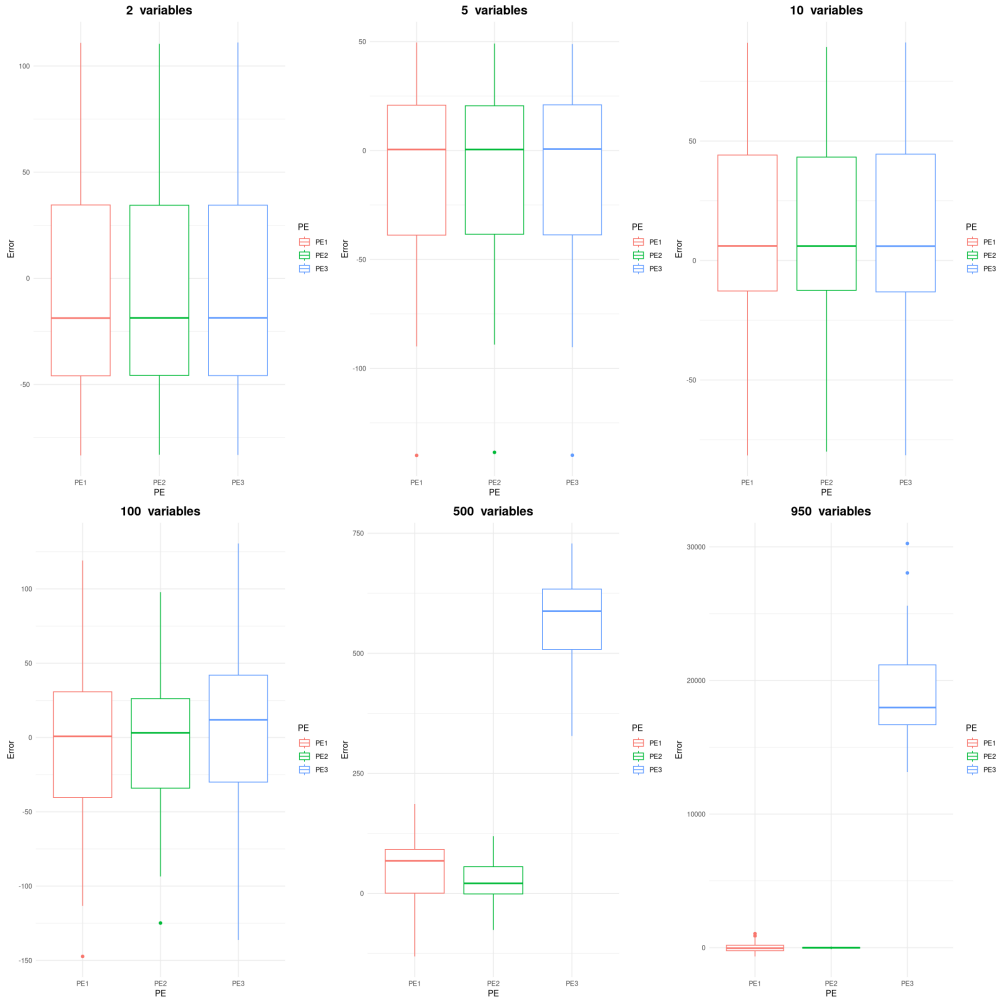

Figure 2: Prediction error residuals

PE2 is clearly the superior estimator (since it uses the real variance) while PE3 (the cross-validation error) appears to struggle for highly dimensional data.

Table 5: Rounded conditional prediction error estimate averages across thirty runs

k	PE	PE1	PE2	PE3
2	1002	1001	1001	1001
5	1005	1006	1006	1006
10	1009	1006	1006	1006
100	1098	1099	1099	1111
500	1497	1501	1500	2006
950	1950	1899	1948	20293

The bivariate model achieves supreme prediction error scores by a very small margin. Using only the prediction error as the metric it would make for the optimal choice.

10. Project 2

The following experiment is designed to investigate the efficacy of various regression techniques, particularly those involving regularization, such as ridge regression, LASSO, and SLOPE. We also compare the results with mBIC2 model selection criterion. For each of the described methods we calculate the square estimation errors $E1 = \|\hat{\beta} - \beta\|^2$, $E2 = \|X(\hat{\beta} - \beta)\|^2$, FDP and TPP.

10.1 Multiple regression methods

10.2 Ordinary least squares estimator

$$\hat{\beta}^{OLS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

Computed with cv.glmnet and SLOPE by setting λ to zero.

10.2.1 Ridge estimator

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Computed with cv.glmnet by passing the $alpha = 0$ parameter.

10.2.2 LASSO estimator

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

When fitting a cross-validated lasso model using cv.glmnet, two lambda values are commonly reported:

- **lambda.min**: The value of lambda that gives the minimum mean cross-validated error. This is often referred to as the ”best” lambda because it directly minimizes the prediction error on the validation set.
- **lambda.1se**: The largest value of lambda for which the mean cross-validated error is within one standard error of the minimum. This lambda value usually results in a sparser model (fewer non-zero coefficients), potentially improving interpretability and generalization by favoring simpler models.

10.3 Results

Table 6: Rounded error averages across ten runs

model	E1	E2	FDP	TPP
mBIC2	22	22	0	1
lasso (min)	99	87	0.81	1
lasso (1se)	126	114	0.53	1
lasso (arg)	1479	652	0.97	1
SLOPE	18493	989	0.97	1
ridge	553	405	-	-
lasso (ols)	126	114	-	-
SLOPE (ols)	720	691	-	-