Calibrating the Lasso with AIC, BIC and CV

Leo Davy   Martin Gjorgjevski

ENS Lyon
M2 Advanced Mathematics

January 2022

Model selection 1

Given

$$Y = X\beta^* + \epsilon$$

we want to find β^* with the assumptions:

- n observations
- p parameters
- $\epsilon \sim N(0, \sigma^2 I)$
- p_0 non-zero coefficients (each equal to 1)

Many estimators (models) can be defined : Least squares, maximum likelihood estimator, Cross-validation, Lasso, Ridge...

(Idealistic) goal: Find the best estimator (model)

Difficulties of selecting an estimator

Once we have: $\hat{\beta}_{LS}, \hat{\beta}_{MLE}, \hat{\beta}_{CV}, \hat{\beta}_{Lasso}, \cdots$ which one is the "right" one ?

Depending on the goal and on the context (the properties of the data, β^* or ϵ), the answer will change.

Goals and contexts

Properties of the estimator: over/under-fitting, bias-variance tradeoff, generalization error,...
Properties of the data: dimensionality (n/p), sparsity (p_0/p), noise,...

Model selection and criterions

Given a class of models $\mathcal{M}$, instead of finding the correct β^* we want to find the "best" one $\hat{\beta}$ among the estimates $(\beta_m)_{m\in\mathcal{M}}$ that we possess.

Use of a criterion

A criterion C is a function which assigns to each estimate $\hat{\beta}_m$ a number $C(Y, m)$.
We define the best estimate for a given criterion C as:

$$\hat{\beta}_C = argmin_(\hat{\beta}_m)_{\mathcal{M}} C(Y, \hat{\beta}_m)$$

Again, for a class of criterions $\mathcal{C}$ we obtain a set of estimates $(\hat{\beta}_c)_{c\in\mathcal{C}}$ and we want to select the "right" one.

Example : the Elastic Net

We can define a class of models $\mathcal{M}$ indexed by $(\alpha, \lambda) \in [0, 1] \times \mathbb{R}^+$ as

$$\mathcal{C} = (EN(\alpha, \lambda))_{(\alpha,\lambda)}$$

where

$$EN(\alpha, \lambda)(Y, \beta) := ||Y - X\beta||_2^2 + \lambda((1 - \alpha)||\beta||_2 + \alpha||\beta||_1).$$

This gives us a (big) family of estimates

$$(\hat{\beta}_c)_{c \in \mathcal{C}} = (\hat{\beta}_{\alpha,\lambda})_{\alpha,\lambda}$$

and we would like to define a criterion to extract a good estimate from these.

Our setting

We assume that β^* is sparse (with p_0 non-zero coefficients) so it is natural to consider the Lasso problem:

$$Lasso(\lambda)(Y, \beta) = EN(\alpha = 1, \lambda) = ||Y - X\beta||_2^2 + \lambda||\beta||_1.$$

What is the influence of λ ?

Heuristic

- $\lambda << 1$: Lasso is similar to least squares (not sparse, most precise)
- $\lambda >> 1$: Lasso gives the 0 solution (the sparsest, not precise)

There should have a λ which trades between complexity (dimension of the model) and the precision of the model.

Criterions to calibrate the Lasso

In order to perform model selection, we will consider several criterions which have been introduced with different goals in mind:

- Akaike Information Criterion (AIC, Akaike-1971): Minimizing KL-divergence
- Bayesian Information Criterion (BIC, Schwarz-1978): Bayesian model
- Cross-validation (CV): Minimizing generalization error and optimizing prediction performance

AIC

In the Gaussian setting, it coincides with Mallow's Cp criterion:

- $\hat{\lambda}_{AIC} = argmin_{m \in \mathcal{M}} ||Y - X\beta_{\lambda_m}||^2 + 2\sigma^2 ||\beta_{\lambda_m}||_0$
- obtained from minimizing an unbiased estimator $\hat{r}_m$ for the risk $r_m = E||X\beta_{\lambda_m} - X\beta^*||^2$
- overfitting occurs because it doesn't take into account the variance of $\hat{r}_m$

BIC

BIC is similar to AIC, in that it is a penalized risk estimator:

- $\hat{\lambda}_{BIC} = argmin_{m \in \mathcal{M}} ||Y - X\beta_{\lambda_m}||^2 + \log(n)\sigma^2 ||\beta_{\lambda_m}||_0$
- Derived from a Bayesian approach, after putting uniform prior on the models
- Penalizes model complexity more heavily then AIC, yet it can still overfit

CV

Cross validation is performed by spliting the data into a training set and a testing set.

- Compute residual sum of squares (RSS) between test data and prediction on the training set
- Cycle through the data, picking a different train and test set each time, then compute the average RSS error
- CV selects the λ with smallest average RSS error

Simulations

The simulations involve 3 steps: First step: For a list $\Lambda = (\lambda_1, \cdots, \lambda_B)$, we compute $AIC(\hat{\beta}_\lambda)_{\lambda \in \Lambda}, BIC(\hat{\beta}_\lambda)_{\lambda \in \Lambda}, CV(\hat{\beta}_\lambda)_{\lambda \in \Lambda}$.

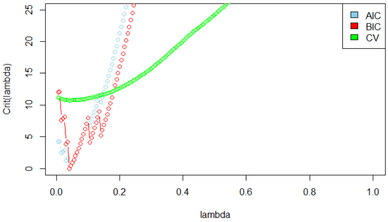

Figure: Values of criterions depending on lambda (n=100, p=10)

Second step: For each criterion $C \in \{AIC, BIC, CV\}$ we define the best estimator as $\hat{\theta}_C = argmin_{(\hat{\theta}_\lambda)_\Lambda} C(\hat{\theta}_\lambda)$

Comparing the estimates obtained

Third step: We want to compare the performance of each criterion. Several choices (of measures) can be made.
For the estimation of parameters:

$$||\hat{\beta} - \beta^*||.$$

However, β^* might not be in the models considered, so instead of β^* we can use $\hat{\beta}_{opt} = \hat{\beta}_{\lambda_{opt}}$ where

$$\lambda_{opt} = argmin_\lambda ||\hat{\beta}_\lambda - \beta^*||.$$

This way $\hat{\beta}_{opt}$ is the closest parameters we can obtain.

Prediction error

If we want to be more focused on prediction, we can consider one of:

$$||Y - X\hat{\beta}|| \quad \text{prediction error}$$
$$||X\beta^* - X\hat{\beta}|| \quad \text{true prediction error}$$
$$||X\hat{\beta}_{opt} - X\hat{\beta}|| \quad \text{best prediction error}$$

In the following, n is fixed $n = 100$. Code has been implemented[1] in R and Python.
When fixed, $\sigma^2 = 1$ and $p_0 = 5$

[1] Figures have been obtained on R.

Influence of dimensionality (fixed n, σ^2, varying p)

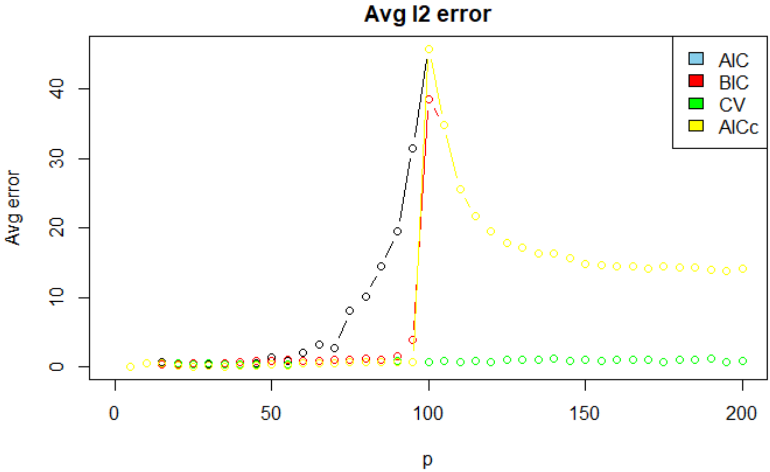

Figure: Average quadratic error for varying p

Influence of noise low-dim (fixed n, p, varying σ^2)

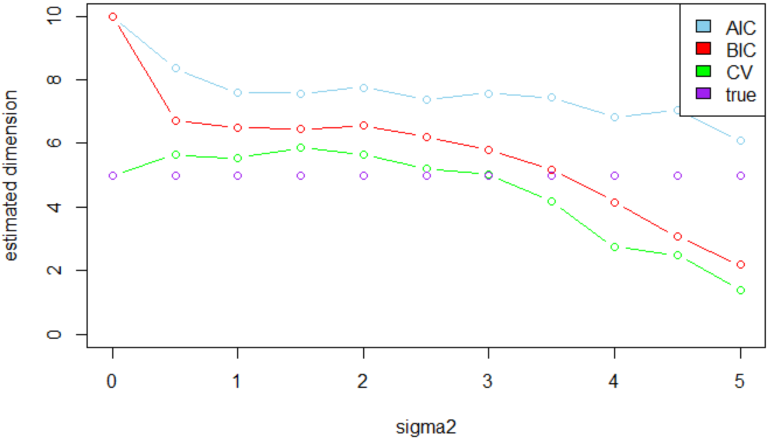

Figure: Influence of noise for $p = 10$

Influence of noise high-dim (fixed n, p, varying σ^2)

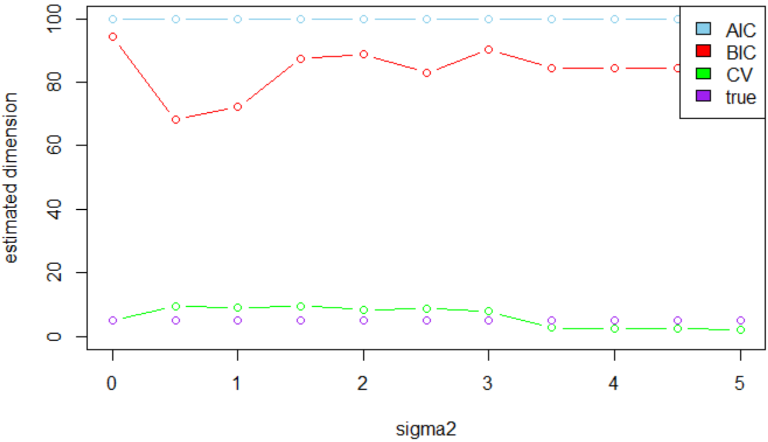

Figure: Influence of noise for $p = 100$

Influence of sparsity (fixed n, p, σ^2, varying p_0)

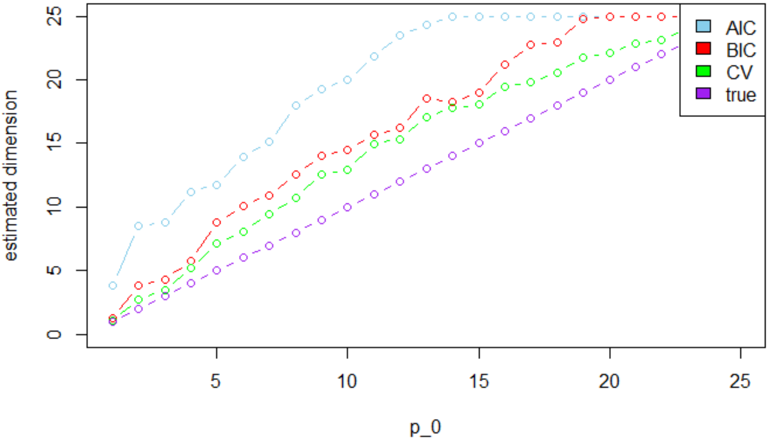

Figure: Influence of sparsity $p = 25$

The big picture (fixed n, σ^2, varying p, p_0)

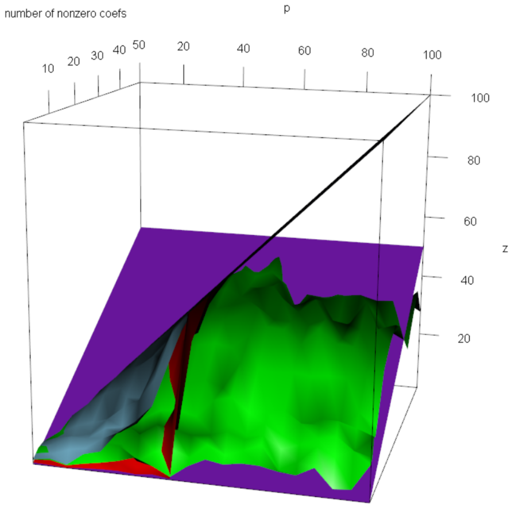

Figure: Average dimension estimated by each criterion. AIC: Blue, BIC: Red, CV: Green, true: Purple

The big picture (fixed n, σ^2, varying p, σ^2)

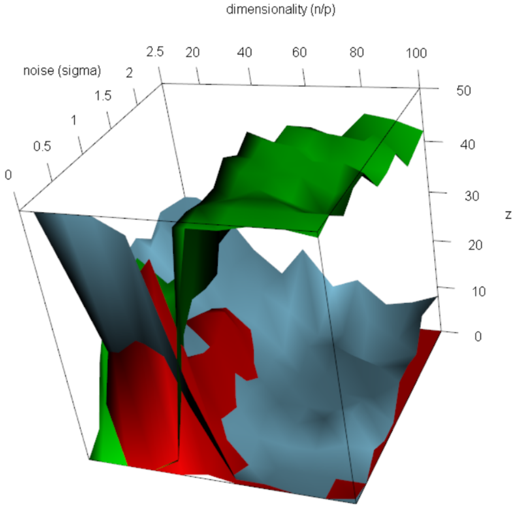

Figure: Criterion closest to λ_{opt}. AIC: Blue, BIC: Red, CV: Green (higher is better)

The big picture (fixed n, σ^2, varying p, σ^2)

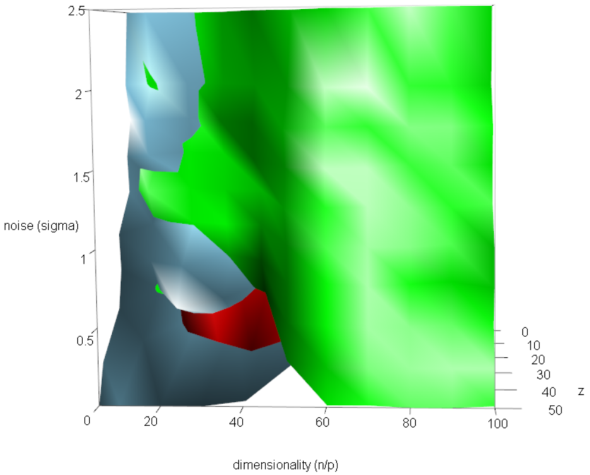

Figure: Criterion closest to λ_{opt}. AIC: Blue, BIC: Red, CV: Green (shown is best)

The big picture (fixed n, σ^2, varying p, σ^2)

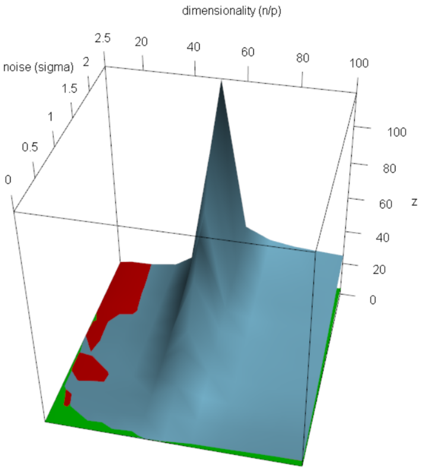

Figure: Quadratic error for $n = 50$. AIC: Blue, BIC: Red, CV: Green (lower is better)

The big picture (fixed n, σ^2, varying p, σ^2)

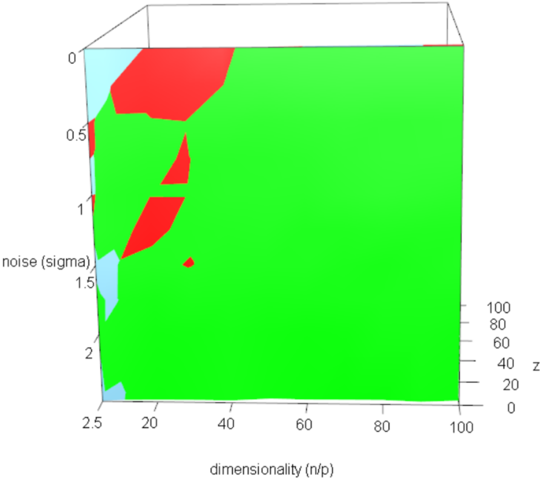

Figure: Quadratic error for $n = 50$. AIC: Blue, BIC: Red, CV: Green (shown is best)

The big picture (fixed n, σ^2, varying p, p_0)

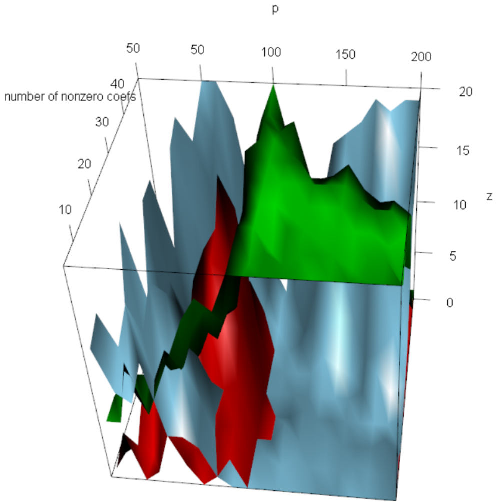

Figure: Criterion closest to λ_{opt}. AIC: Blue, BIC: Red, CV: Green (higher is better, $n = 100$)

The big picture (fixed n, σ^2, varying p, p_0)

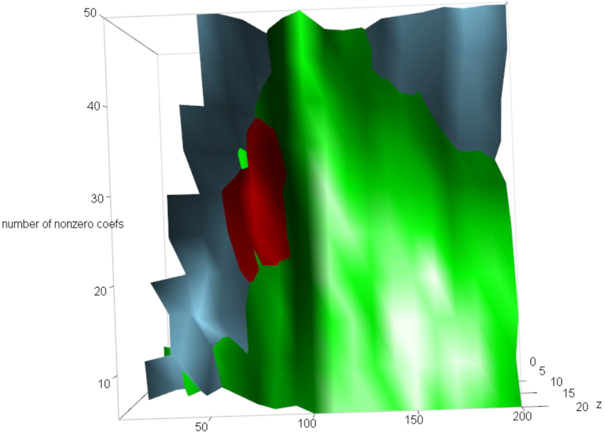

Figure: Criterion closest to λ_{opt}. AIC: Blue, BIC: Red, CV: Green (shown is best, $n = 100$)

The big picture (fixed n, σ^2, varying p, p_0)

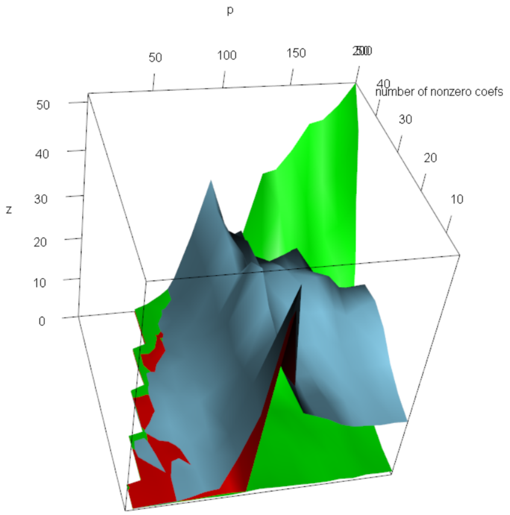

Figure: Quadratic error. AIC: Blue, BIC: Red, CV: Green (lower is better, $n = 100$)

The big picture (fixed n, σ^2, varying p, p_0)

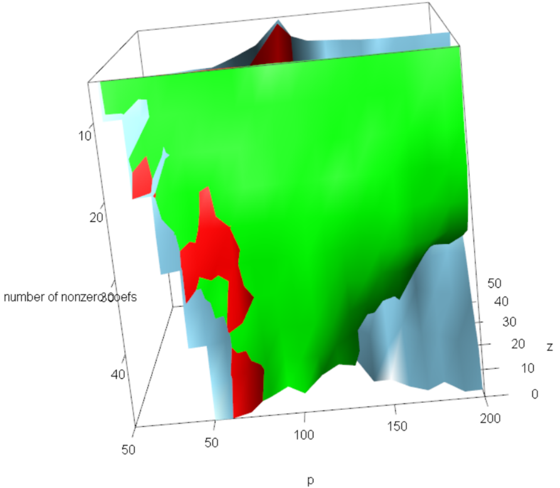

Figure: Quadratic error. AIC: Blue, BIC: Red, CV: Green (shown is best, $n = 100$)

Conclusions

- Depending on the context some estimators will perform better than others
- In most contexts, Cross-Validation performs better than others
- In low dimensions, AIC and BIC are good alternatives to CV
- AIC and BIC are easier and faster to implement than CV (at least for gaussian case)
- In practice, selecting a criterion for a model, amounts to an assumption on the true data distribution