

# Penalized regression inference regarding variable selection in high dimensions: presentation of selected methods implemented in R

Marta Karas

Contact: <http://statsox.github.io>

**Association between an outcome variable and predictors.** To assess the association between an outcome  $y \in \mathbb{R}^n$  and a set of predictors  $x_j \in \mathbb{R}^n, j = 1, \dots, p$ , one might consider the model:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

where  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$ ,  $\beta \in \mathbb{R}^p$  is vector of coefficients, and  $\epsilon \in \mathbb{R}^n$  is a vector of errors with mean zero and constant variance. If the number of variables  $p$  is much smaller than  $n$ , we could perform a formal statistical test for whether an element of  $\beta$  is zero using classical methods, such as likelihood ratio or Wald test. However, **in the high-dimensional setting, when the number of variables  $p$  is large, these tests have low power, or are undefined.**

**Penalized regression techniques.** In the case where  $p$  is large, penalized regression techniques such as Ridge and Lasso can be employed to obtain  $\beta$  estimates:

$$\hat{\beta}_\lambda = \arg \min_{b \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}b\|_2^2 + \lambda J(b) \right\},$$

where  $J(b) = \|b\|_1$  for Ridge and  $J(b) = \frac{1}{2} \|b\|_2$  for Lasso. However, **Lasso and Ridge yield biased estimators of  $\beta$ , thus these procedures do not provide  $p$ -values or confidence intervals.**

**Methods. Penalized regression inference.** Here, we present examples of usage of a few selected methods available in R:

- `lassoscore {lassoscore}`: **Score test based on penalized regression.** Performs penalized regression of an outcome on all but a single feature, and test for correlation of the residuals with the held-out feature; applied on each feature in turn.
- `hdi {hdi}`: **Multi sample-splitting.** Splits the sample into two equal halves,  $I_1$  and  $I_2$ . First half  $I_1$  is used for variable selection (with the use of Lasso) and the second half  $I_2$ , with the reduced set of selected variables (from  $I_1$ ), is used for "classical" statistical inference in terms of  $p$ -values. Repeats this  $B$  times and aggregate obtained  $p$ -values.
- `grace.test {Grace}`: **Grace test.** Proposes how to overcome that Ridge is a biased estimator of  $\beta$  and its estimation bias is negligible only if the Ridge tuning parameter  $\lambda$  is close to zero. To construct a test statistic for the null hypothesis  $H_0 : \beta_j^* = 0$  for some  $j \in \{1, \dots, p\}$ , it adjusts for the potential estimation bias by using a stochastic bound derived from an initial estimator. Since with this adjustment the tuning parameter  $\lambda$  needs not be very small, coefficient estimation and corresponding  $p$ -values for penalized regression might be obtained.

**Methods. Assessing the inference results.** In regression settings, False Discovery Rate (FDR) is often used to describe the proportion of false "discoveries" (whose coefficients in the true *full model* are zero). However, when applying FDR to settings with the presence of correlated predictors, more than one variable is likely to be capturing the same underlying signal. **Then, "classical" FDR suffers from unintuitive and potentially undesirable behavior.**

- Here, we use **False Variable Rate (FVR)** measure ([x]), which considers a variable to be an interesting selection if it captures signal that has not been explained by any other variable in the selected model. Mathematically, for a selected variables set  $A \subseteq \{1, \dots, p\}$ , we project the mean  $\mathbf{X}\beta$  from the *full model* onto subset of predictors  $\mathbf{X}_A$  to obtain a projected mean  $\mathbf{X}\beta^{(A)}$ . We define a selected variable to be a false selection if it has a zero coefficient in this projected mean vector.

Genes	A*	B <sub>1</sub> *	B <sub>2</sub>	B <sub>3</sub>	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	D
Selected Sets	{A, B <sub>2</sub> , C <sub>1</sub> }				{A, B <sub>1</sub> , B <sub>2</sub> }			
Projected	✓	✓	✗		✓	✓	✗	

Figure 1. **False Variable Rate (FVR)** criterion illustration. Variables are denoted as correct selections if they are capturing unique signal among the selected variables. Thus  $B_2$  is correctly selected in the first set. However,  $B_2$  is considered a false selection in the second set because it adds no information beyond  $B_1$ . Figure & caption source: X.

**Code example.** Assume we are given data matrix  $X_{100 \times 300} \sim N(0, \Sigma)$ , true signal  $\beta$  and observed response variable  $Y \sim N(X\beta, 1^2)$ .

