Stepwise Regression and Selection by Hypothesis Testing

1 Theoretical Part: Partial Correlations and Residual Sum of Squares

Look at **paragraph 3.2.1** of the e-book and, with paper and pencil, prove that optimization problem (1) (i.e. min RSS at a given step) and (2) (i.e. max partial correlations) are equivalent:

$$\min_{j \in P \setminus S} \quad \left\| \mathbf{y} - \mathbf{X}^{\star}\hat{\beta}^{\star} - \mathbf{x}_j \beta_j \right\|_2^2 \tag{1}$$

$$\max_{j \in P \setminus S} \quad | \, \mathbf{x}_j^T (\mathbf{y} - \mathbf{X}^{\star}\hat{\beta}^{\star}) \, | \tag{2}$$

where:

(a) $P = \{1, .., p\}$ is the set of all available predictors.

(b) $S = \{a \in P : | \, S \, | = q\}$ is the solution set at the current step of the procedure.

(c) $\mathbf{X}^{\star}$ is the $n \; x \; q$ matrix of predictors selected at previous steps.

(d) $\hat{\beta}^{\star}$ is the OLS estimator of the model $\mathbf{y} = \mathbf{X}^{\star}\beta^{\star} + \boldsymbol{\epsilon}$.

(e) $\mathbf{x}_j$ is one of the $p - q$ predictors left at the current step.

Assume also that all the predictors have unitary norm (i.e. $\mathbf{x}_l^T \mathbf{x}_l = 1 \; \forall \; l \in P$).

(hint: Start working on (1), define the residual $\mathbf{e}^{\star}$ *and expand the norm).*

2 Simulations: Selection by Hypothesis Testing

In this exercise we would like to familiarize with the R package *selectiveInference* and to employ the ForwardStop rule presented in the course (see **paragraph 3.2.2**).

First of all we retrieve the simulation setting used in *Practical 3* with some changes:

1. Generate from a MVN (multivariate normal) a matrix $\mathbf{X_{n*p}}$ with $n = 1000$ and $p = 10$. You can choose the location vector as you wish but set the scale matrix as the identity. Fix also the seed equal to 11 (i.e. set.seed(11)).

2. Choose the generating vector $\boldsymbol{\beta} = [3 \; 1.5 \; 0 \; 2 \; rep(0, 6)]$.

3. Generate $\hat{\mathbf{y}}$ thanks to the relation $\mathbf{y} = \mathbf{X_{n*p}}\ \boldsymbol{\beta} + \boldsymbol{\epsilon}$ where ϵ_i is a standard normal. Suppose for simplicity that the errors are uncorrelated.

Now, after having read the documentation of the R package *selectiveInference* and installed it, perform the following steps:

(a) Use the functions *fs()*, *fsInf()* and *forwardStop()* to do a stepwise regression based on partial correlations and a model selection phase with the ForwardStop rule on your generated data. Try different values for the type one error: how does the choice of α impact the model selection technique?

(b) Given the order of variables produced by *fs()*, use AIC and BIC criteria for model selection to retrieve your final model (*Hint: you do not need to program them, use an existing function of the selectiveInference package*).

(c) Calculate how many models are needed for an exhaustive search in this simulation setting. Use your previous results obtained in *Practical 3* to understand the computational time gained by stepwise regression with respect to exhaustive search. Use the package *tictoc* for this comparison.

(d) (*Optional*) Change the simulation setting outlined above to an high dimensional one: generate from a MVN (multivariate normal) a matrix $\mathbf{X_{n*p}}$ with $n = 100$ and $p = 150$. Evaluate the performance of the ForwardStop rule in this high dimensional setting (i.e. by replicating the model selection task 100 times) thanks to the usual three specific criteria: the proportion of times the correct model is selected (*Exact*), the proportion of times the selected model contains the correct one (*Correct*) and the average number of selected regressors (*Average* $\sharp$). What do you observe? What is the role of α in this case?

3 Real Data Application

Load the *Zambia* dataset, perform a stepwise regression based on partial correlations and a model selection phase based on the ForwardStop rule, AIC and BIC. For simplicity work only on the continuous covariates (i.e. avoiding factors). What do you observe? Which characteristics can you infer regarding ForwardStop rule, AIC and BIC given your results?