A model selection criterion for LASSO

Ho Yin HO*

Hong Kong University of Science and Technology hyhoaa@ust.hk

Man-Yu WONG

Hong Kong University of Science and Technology mamywong@ust.hk

LASSO (*Tibshirani*, 1996) is a shrinkage and selection method for linear regression model through minimizing the residual sum of squares plus a penalty term, which is equal to a product of the L_1 norm of the vector of estimated coefficients and a tuning parameter, λ. It is a popular model selection method since it assigns zero value to unimportant factors and hence provides an interpretable model. Model selection in LASSO is usually done by cross-validation. Later, a modified Least Angle Regression (LARS) algorithm was used to compute the solution path for LASSO, which is the set of LASSO solution for all nonnegative value of λ (*Efron*, *Hastie*, *Johnstone* and *Tibshirani*, 2004). Selection criteria using the solution path provided by Least Angle Regression (LARS) algorithm, such as C_p, AIC and BIC, were studied (*Zou*, *Hastie* and *Tibshirani*, 2007). In our work, we propose an alternative selection criterion, which is also based on the solution path. We consider λ as a random variable. It is noted that a large value of λ is needed for LASSO to assign a zero value to a factor with important effect. Thus, λ can be used to test how important a factor is. We identify the connection between LASSO and the traditional ordinary least square regression (OLS). We use this relationship to derive the distribution of λ under Gaussian noise and the corresponding hypothesis test statistic as our selection criterion. The tuning parameter λ can be expressed as an order statistic of a set of non-negative random variables. The exact distribution of λ would be difficult to obtain. Instead, we concentrate on the derivation of the critical value by approximation. The selection criterion derived from this project will be applied to some systematic supersaturated designs which the number of observations of such design is less than the number of factors and also satisfy the "nearly orthogonal" condition (*Booth* and *Cox*, 1962).

Key Words: Hypothesis testing, Least angle regression (LARS), Ordinary least square regression (OLS), Order statistics, Systematic supersaturated factorial designs