The Annals of Statistics
2014, Vol. 42, No. 2, 478–482
DOI: 10.1214/13-AOS1175B
Main article DOI: 10.1214/13-AOS1175

DISCUSSION: "A SIGNIFICANCE TEST FOR THE LASSO"

BY T. TONY CAI[1] AND MING YUAN[2]

University of Pennsylvania and University of Wisconsin–Madison

We congratulate the authors for an interesting article and an innovative proposal to testing the significance of the predictor variables selected by the Lasso. There is much material for thought and exploration. Research on high-dimensional regression has been very active in recent years, but most of the efforts have so far focused on estimation. Despite the popularity of the Lasso as a variable selection technique, the problem of making valid inference for a model chosen by the Lasso is largely unsettled. The current paper pinpoints some of the challenges in making valid inference in the high-dimensional setting and presents a thought-provoking approach to address them.

Following the notation used in the paper, let A be the model selected at the kth step of either the Lasso or forward stepwise regression and j be the index of the variable to be added in the next step. This paper considers the problem of testing the null hypothesis that the underlying model corresponding to the true regression coefficient vector β^* is nested in the current selected model, that is,

$$H_0 : \operatorname{supp}(\beta^*) \subseteq A.$$

As pointed out in the paper, a classical approach to testing two fixed nested models A and $A \cup \{j\}$ is the chi-squared test, which is based on the test statistic

$$R_j = (\mathrm{RSS}_A - \mathrm{RSS}_{A\cup\{j\}})/\sigma^2$$

and compares it to the quantile of the χ_1^2 distribution. The test fails, as noted, when applying to the forward stepwise regression or the Lasso in a vanilla fashion because it fails to account for the fact that neither A nor $\{j\}$ is fixed. The randomness of A can be addressed using a conditional argument as suggested by the authors. The effect of the way that the new index j is selected is more subtle. The seemingly lack of a remedy to this problem motives the authors to focus on the Lasso and to propose the so-called covariance test statistic

$$\begin{aligned} T_k &= \bigl(\langle y, X\hat{\beta}(\lambda_{k+1})\rangle - \langle y, X_A\tilde{\beta}_A(\lambda_{k+1})\rangle\bigr)/\sigma^2 \\ &= R_j - \lambda_{k+1}\bigl(\langle s_{A\cup\{j\}}, \hat{\beta}^{\mathrm{LS}}_{A\cup\{j\}}\rangle - \langle s_A, \hat{\beta}^{\mathrm{LS}}_A\rangle\bigr)/\sigma^2, \end{aligned} \tag{1}$$

Received December 2013.

[1] Supported in part by NSF Grant DMS-12-08982 and NIH Grant R01 CA127334.

[2] Supported in part by NSF Career Award DMS-08-46234 and FRG Grant DMS-12-65202.