Post-selection inference for l_1-penalized likelihood models

Presenter: Wei JIANG
June 1st 2017
Machine Learning Journal Club, CMAP

Abstract

According to the article[2], we present a new method for post-selection inference for l_1(lasso)-penalized likelihood models, including generalized regression models. Our approach generalizes the post-selection framework presented in Lee et al. (2013)[1]. The method provides P-values and confidence intervals that are asymptotically valid, conditional on the inherent selection done by the lasso. We present applications of this work to (regularized) logistic regression, Cox's proportional hazards model, and the graphical lasso. We do not provide rigorous proofs here of the claimed results, but rather conceptual and theoretical sketches.

1 Introduction

In many applications of regression model, we start with a large pool of candidate variables, such as genes or demographic features, and does not know a priori which are relevant. This is especially problematic when there are more variables than observations, since then the model is unidentifiable. In such settings, it is tempting to let the data decide which variables to include in the model. For example, one common approach when the number of variables is not too large is to fit a linear model with all variables included, observe which ones are significant at level α, and then refit the linear model with only those variables included. The problem with this is that the p-values can no longer be trusted, since the variables that are selected will tend to be those that are significant. Intuitively, we are "overfitting" to a particular realization of the data.
To solve the problem, we make inference after the selection of model. Here we discuss only the post-selection inference for lasso-penalized likelihood models. We started from a simple Gaussian case then extend it to the generalized lasso regression.

2 Post-selection inference

2.1 Gaussian case

Suppose that we have data (x_i, y_i), $i = 1, 2, \cdots, N$ consisting features $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ and outcomes y_i. Let $X = \{x_{ij}\}$. For the Gaussian case $y \sim N(\mu, I \cdot \sigma^2)$.
We denote the selected model by M with sign vector s_M. Solve the l_1-regularized minimization problem

$$\hat{\beta} = \text{argmin}\frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1$$

then we can define a model selected by Lasso :

$$\{\hat{M} = j : \hat{\beta}_j \neq 0\}$$

The post-selection inference seeks to make inference for some functional $\gamma^T\beta$ given $\hat{M} = M$. For example, γ might be chosen so that $\gamma^T\beta$ is the partial regression coefficient for the jth predictor.
KKT conditions state that $\{\hat{M}, \hat{s}_M\} = \{M, s_M\}$ iif there exists $\hat{\beta_M} \in \mathbb{R}$ and $u_{-M} \in \mathbb{R}^{-M}$

$$X_M^T(X_M\hat{\beta}_M - y) + \lambda s_M = 0$$
$$X_{-M}^T(X_M\hat{\beta}_M - y) + \lambda u_{-M} = 0$$
$$sign(\hat{\beta}_M) = s_M$$
$$||u_{-M}|| \leq 1$$

This allows us to write the set of responses y that yield the same M and s in the polyhedral form

$$\{\hat{M} = M, \hat{s}_M = s_M\} = \{y : Ay \leq b\}$$

with A, b not depending on y, and obtain a statistic based on truncated Gaussian distribution

$$F^{\mathcal{V}^-,\mathcal{V}^+}_{\gamma^T\mu,\sigma^2||\gamma||^2}(\gamma^T y)\Big|\{Ay < b\} \sim U(0,1)$$

where $\mathcal{V}^-, \mathcal{V}^+$ are computable values that are functions of γ, A and b.
If we choose $\gamma = (X^+_M)^T e_j$ where $X^+_M = (X^T_M X_M)^{-1} X^T_M$ then $\gamma^T\mu = \beta^M_j$ and $\gamma^T y$ is the partial least-square estimate for one predictor.
We can easily derive the post-selection p-values and confidence intervals for each predictor.

2.2 Generalized regression model

We consider a generalized regression model with linear predictor $\eta = x^T\beta$ and likelihood $l(\beta)$. To solve the l_1-regularized minimization problem

$$\hat{\beta} = \text{argmin}_\beta - l(\beta) + \lambda||\beta||_1$$

, a common strategy is to express the usual Newton-Raphson update as an iteratively reweighted least square (IRLS). Define

$$W = W(\beta) = -(\frac{\partial^2 l}{\partial\eta\partial\eta^T})\Big|_{\eta=X\beta}$$

and

$$z = z(\beta) = X\beta + W^{-1}(\frac{\partial l}{\partial\eta})\Big|_{\eta=X\beta}$$

IRLS proceeds as follows:

1. initialize $\hat{\beta} = 0$
2. compute $W(\hat{\beta})$ and $z(\hat{\beta})$ based on the current value of $\hat{beta}$
3. solve
$$\min_\beta \frac{1}{2}(z - X\beta)^T W(z - X\beta) + \lambda||\beta||_1$$
4. repeat step 2 and 3 until $\hat{\beta}$ does not change more than some pre-specified threshold.

For example, logistic regression and cox's proportional hazards for censored survival data.
In order to carry out post-selection inference in this setting, we treat the final iterate as a weighted least square regression.
Suppose we reach a fixed point in the above iterations. The active block has the form :

$$X^T_M W(z - X_M\hat{\beta}_M) = \lambda s_M$$

If $\bar{\beta}_M$ solves

$$X^T_M W(z - X_M\bar{\beta}_M) = 0$$

then we have

$$\bar{\beta}_M = \hat{\beta}_M + \lambda(X^T_M W X_M)^{-1} s_M$$

We see that $\bar{\beta}_M$ is defined by one Newton-Raphson step in the selected model from $\hat{\beta}_M$
If we had not used the data to select variables M and signs s_M then

$$\bar{\beta}_M \approx N(\beta^*_M, (X^T_M W X_M)^{-1})$$

. However selection with LASSO has imposed the active constraint $\{y : \text{sign}(\bar{\beta}_M(y) - \lambda(X^T_M W X_M)^{-1} s_M) = s_M\}$. We will discuss it in the general form.

2.3 A more general form and an asymptotic justification

We assume p is fixed (avoid high-dimensional regime). Similar to generalized regression model, we have one-step estimator:

$$\bar{\beta}_M = \hat{\beta}_M + \lambda I_M(\hat{\beta}_M)^{-1} s_M$$

Where $I_M(\bar{\beta}_M)$ is the $|M| \times |M|$ is observed Fisher information matrix of the submodel M evaluated at $\bar{\beta}_M$ If we had not used the data to select variables M and signs s_M then

$$\bar{\beta}_M \approx N(\beta^*_M, I_M(\hat{\beta}_M)^{-1})$$

where $I_M(\bar{\beta}_M)$ is the "plug-in" estimate of the asymptotic covariance of $\bar{\beta}_M$, with the population value being $E_F[I_M(\bar{\beta}_M)]^{-1}$. However selection with LASSO has imposed the active constraint

$$\{\text{diag}(s_M)[\bar{\beta}_M - I_M(\hat{\beta}_M)^{-1}\lambda s_M] \geq 0\}$$

Similar to the Gaussian case, polyhedral lemma yield asymptotically exact selective inference for the selected event $\{(\hat{M}, s_{\hat{M}}) = (M, s_M)\}$ by construction of a pivotal quantity

$$g(\bar{\theta}_M; \lim_{n\to\infty} nE_F[I_M(n^{\frac{1}{2}\theta^*_M})]^{-1}; A, b)$$

where A and b can be derived from the active constraint, and $\lim_{n\to\infty} nE_F[I_M(n^{\frac{1}{2}\theta^*_M})]^{-1}$ can be approximated by using a plug-in estimate of variance.
The article provides the result that the pivot is asymptotically $U(0,1)$ as $n \to \infty$ conditioned on having selected model M with signs s_M. Then we can easily derive the post-selection p-values and confidence intervals for each predictor.

3 Simulation

3.1 Logistic regression

To assess performance in the $l1$-penalized logistic model we generated Gaussian features with pairwise correlation 0.2 in two scenarios: $n = 30, P = 10$ and $n = 40, P = 60$. Then y was generated as $P(Y = 1|x) = \frac{1}{1+exp(-x^T\beta)}$. There are two signal settings: null ($\beta = 0$) and non-null. Finally in each case we tried two methods for choosing the regularization parameter λ: a fixed value yielding a moderately sparse model and cross-validation. The figures show the cumulative distribution function of the resulting P-values over 1000 simulations. Thus a function above the 45 degree line indicates an anti-conservative test in the null setting and a test with some power in the non-null case. We see the adjusted P-values are close to uniform under the null in every case and show power in the non-null setting. Even with a cross-validation-based choice for λ the type I error seems to be controlled, although we have no theoretical support for this finding. In Figure 1 we also plot the naive P-values from GLM theory: as expected they are very anti-conservative.

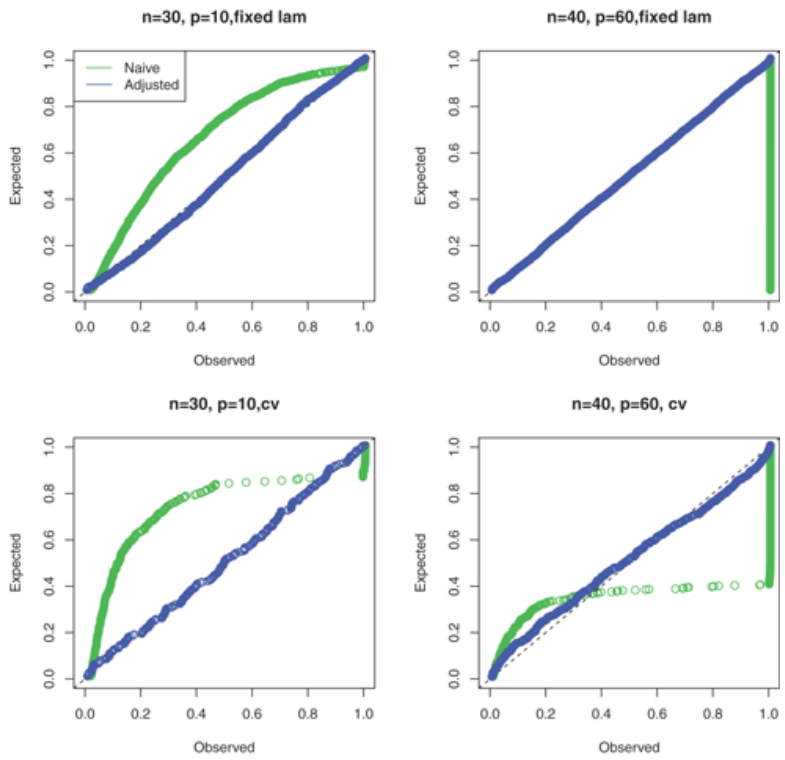

Figure 1: P-values for the logistic regression model, null setting. The top panels use a fixed λ, whereas the bottom ones use cross-validation to choose λ

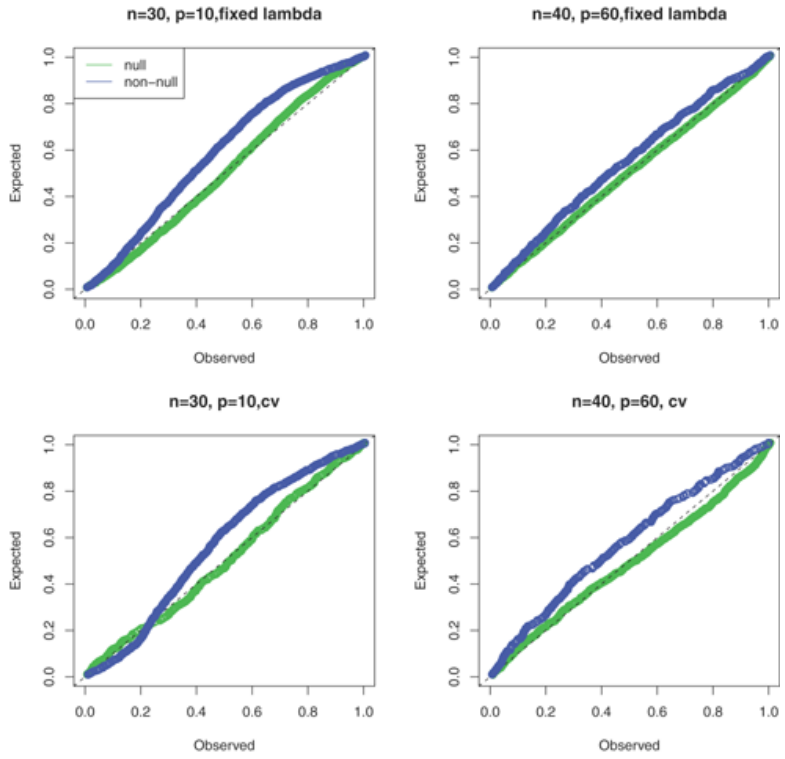

Figure 2: P-values for the logistic regression model, non-null setting. The top panels use a fixed λ, whereas the bottom ones use cross-validation to choose λ

3.2 Cox regression

Liver disease: 276 observations of patients, 17 predictors (clinic indicators)
We applied Cox's proportional hazards model. Figures 3 and 4 show the results. As expected the

adjusted P-values are larger than the naive ones and the corresponding selection (confidence) intervals tend to be wider.

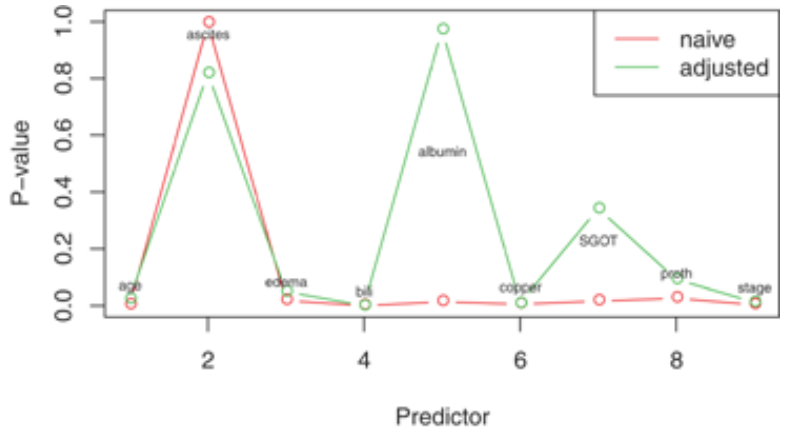

Figure 3: P-values for Cox model applied to the liver data

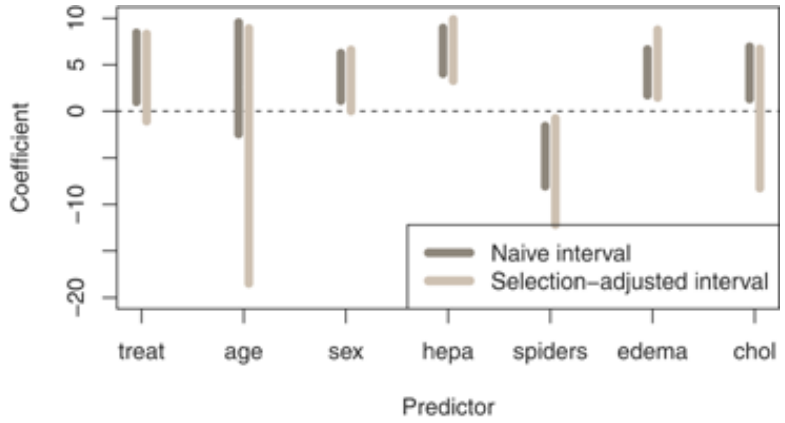

Figure 4: Selection intervals for Cox model applied to the liver data.

4 Conclusion

Model selection and inference have long been regarded as conflicting goals in regression. This paper has proposed a general framework for post-selection inference that conditions on which model was selected , that is, the event$\{\hat{M} = M\}$. We characterize this event for the LASSO and derive optimal and exact p-value and confidence intervals for each predictor conditional on the event$\{\hat{M} = M\}$. With this general framework, we can form post-selection intervals for regression coefficients, equipping practitioners with a way to obtain valid intervals even after model selection.

References

[1] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927, 06 2016.

[2] J. Taylor and R. Tibshirani. Post-selection inference for $l1$-penalized likelihood models. *Canadian Journal of Statistics*, 2017.