Statistical Learning and Selective Inference

Ganchao Wei

September 30, 2021

Overview

1. Introduction: Selectie Inference
2. Forward Stepwise Regression
3. The LASSO
4. PCA

Introduction: Selectie Inference

Example 1: "Strong" correlation

- Two measurements A and B, with correlation 0.9. Awesome!
- But... If it is chosen from the best of 1000 measurements?
- Not impressive, even if all 1000 measurements were uncorrelated

Example 2: Clinical trial, two treatments

- If test statistic $z = (\bar{y}_2 - \bar{y}_1)/s = 2.5$? p-value = 0.01, Significant!
- But...If I tried many treatments and report only ones for which $|z| > 2$?
- $P(|z| > 2.5||z| > 2) \approx 27\%$
- Corrected p-value = 0.27, not significant.

Introduction: Selectie Inference

Selective inference: the assessment of significance and effect sizes from a dataset after mining the same data to find these associations.

–

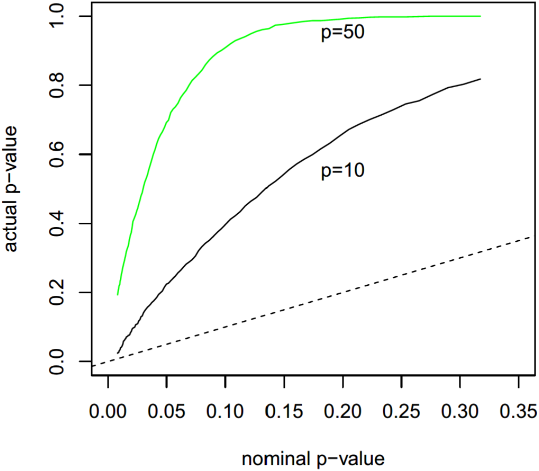
Exaggeration!

Forward Stepwise Regression

- Linear regression: N observations, (x_i, y_i)
- Large number of predictors ⇒ predictor selection ⇒ traditional way: (forward) stepwise regression (LASSO later)
- $RSS = \sum (y_i - \hat{y})^2$
  compare $R_k = \frac{1}{\sigma^2}(RSS_{k-1} - RSS_k)$ to χ_1^2 distribution
- **Problem**: assume models were prespeicifed before seeing the data ⇒ R_k will be larger than χ_1^2

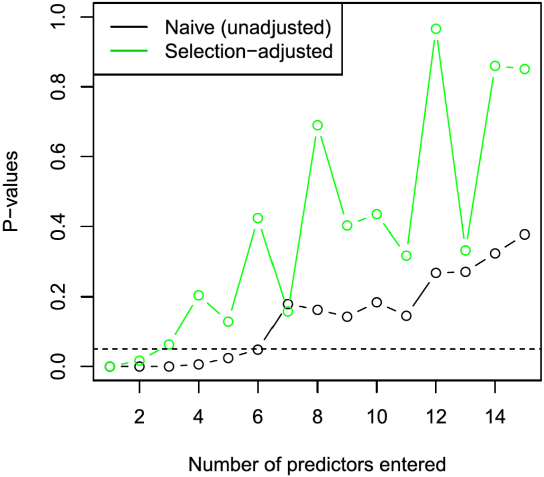
Forward Stepwise Regression

Simulation: N = 100, first step of forward stepwise regression.
correct p-value: record the max value of R_k achieved each time.



Forward Stepwise Regression

Example: HIV data, $n = 1073$ samples and $p = 240$ mutation sites. Randomly select 100 samples and 30 sites. By forward stepwise regression, predictors enter in the order (5,9,25,8,16,21,...).



Forward Stepwise Regression

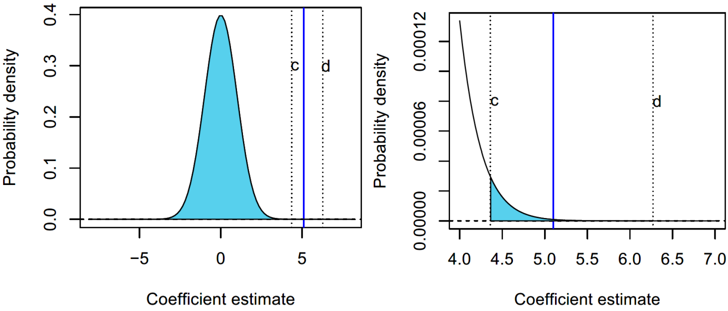
- Naive p-values $\Rightarrow$ 6 strong predictors
- After adjustment, only 2 or 3 are significant
- How to adjust? Do Monte Carlo directly? $\Rightarrow$ cumbersome
- Luckily, we can do things in closed form.

Forward Stepwise Regression

- Assume have taken 2 steps, entering x_5 and x_9.
- standard: $\hat{\beta} \sim N(\beta, \tau^2)$. Assume we had only these 2 predictors available.
- This is not the case: select the strongest 2 from 30 predictors available.
- Write things in polyhedral form $Ay \leq b$. A and b depend on the data and selected variables.
- Each stage represents a competition among all p variables. A and b reconstruct the competition and check whether new outcomes y^* yields the same result.
- results of polyhedral selection: truncated normal $\hat{\beta} \sim TN^{c,d}(\beta, \tau^2)$

Forward Stepwise Regression

Truncated normal $\hat{\beta} \sim TN^{c,d}(\beta, \tau^2)$. $\hat{\beta} = 5.1$, $c = 4.3$ and $d = 6.3$.



(a) (b)

Ignoring selection effects ⇒ significant.
After adjustment ⇒ moderate evidence

FDR and Sequential Stopping Rule

Question: when should we stop adding variables?
Control $FDR = E(V/R)$: $\hat{k} = max\{k : -1/k \sum_{i=1}^{k} \log 1 - pv_i \leq \alpha\}$,
where $\{pv_i\}$ are successive p-values.

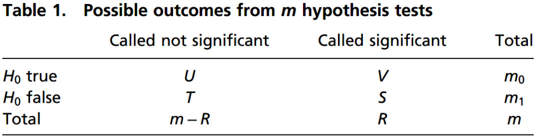


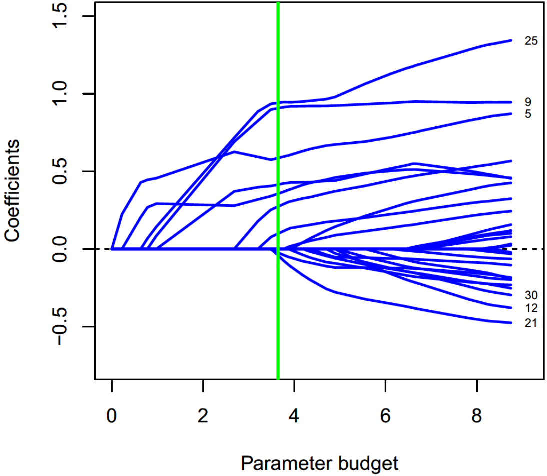
Table 1. Possible outcomes from *m* hypothesis tests

	Called not significant	Called significant	Total
H_0 true	U	V	m_0
H_0 false	T	S	m_1
Total	$m - R$	R	m

The LASSO

Modern approach for model selection: the LASSO

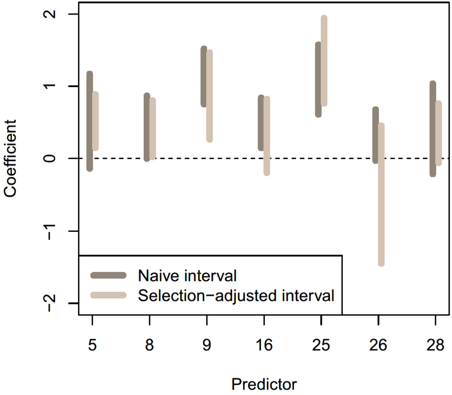
- Still use the HIV data as an example.
- tune the parameter by cross-validation: 9 predictors



The LASSO

Can still use the polyhetral region of the form $Ay \leq b$

- For fixed predictors and λ, the vector of response values y^* that would yield the same active set can be written as $Ay^* \leq b$
- A and b depend on active set and λ, but not y.
- selection-adjusted intervals



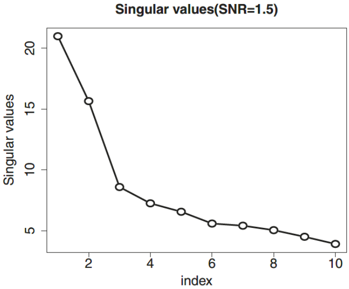
PCA

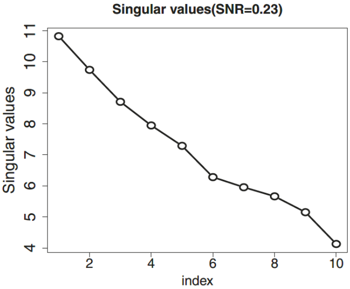
How to select the number of components?

- Traditional way: scree plot, based on the "elbow" point of eigenvalue
- But... It fails when there are too many noises.
- We can choose the leading eigenvectors as previous, i.e. calculate the adjusted p-values
- calculate p-values is more informative: use more information in the correlation matrix, rather than just the eigenvalues.

PCA

The adjusted p-values in the right: (0.030, 0.064, 0.222, 0.286, 0.197, 0.831, 0.510, 0.185, 0.126)


Singular values(SNR=1.5)


Singular values(SNR=0.23)