Selective inference after cross-validation

Joshua R. Loftus

RSS test error estimation

- For K-fold cv, data partitioned (randomly) into $D_1, \ldots, D_K$. For each $k = 1, \ldots, K$, hold out D_k as a test set while training a model on the other $K - 1$ folds. Form estimate RSS_k of out-of-sample prediction error. Average these estimates over test folds.
- Use to choose model complexity: evaluate $RSS_{k,s}$ for various sparsity choices s. Pick s minimizing the cv-RSS estimate.

Examples

For each training set...

- Train LASSO models on a grid of λ values. Or fit sequentially with GLMNET. Choose λ^* minimizing cv-RSS. Finally, fit a LASSO model at λ^* on the whole data.
- Run forward stepwise with maxsteps S. For $s = 1, \ldots, S$ evaluate the test error $RSS_{k,s}$. Average to get RSS_s. Pick s^* minimizing this. Run forward stepwise on the whole data for s^* steps.

Can we do selective inference for the final models chosen this way?

Quadratic constraints example: forward stepwise

Key observation

The inequalities $\text{RSS}(i_s) \leq \text{RSS}(j)$ characterizing the variable added at step s can all be written in the form

$$y^T(Q_{i_s} - Q_{j,s})y \geq 0 \quad \forall j \neq i_s$$

Define $E_{s,j} := \{y : y^T(Q_{i_s} - Q_{j,s})y \geq 0\}$.

Characterization of the selection event

The event E that forward stepwise chooses model m can be written

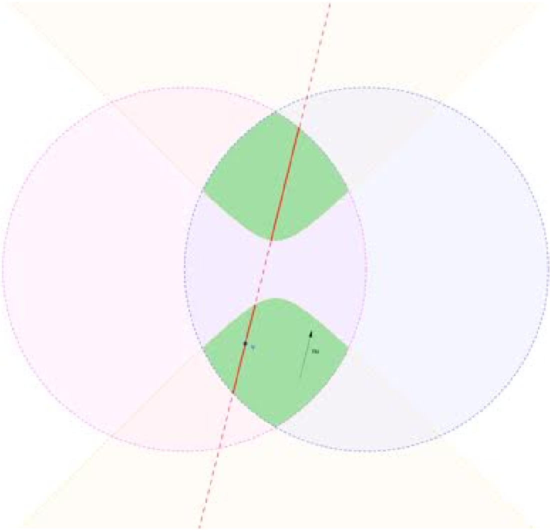
$$E = \bigcap_{s=1}^{S} \bigcap_{j \neq i_s, \ldots, i_1} E_{s,j}$$

Reduction to one dimension

- The geometry of E (intersection of quadratics) is complicated. For example, verifying if an ellipsoid contains an intersection of ellipsoids is NP-complete (Boyd et. al. LMI book, Section 3.7.2). Our quadratics may not even be $\succeq 0$.
- To test for variable i_s, reduce to a one-dimensional problem. Define $T\chi_{i_s} = \|Py\|_2$ for some projection P. Let $U = Py/S$ and $Z = y - Py$. Support of $T\chi_{i_s}$

$$M_{i_s} = \{t \geq 0 : Ut + Z \in E\}$$

Cartoon of selection event and one-dimensional slice



Cross-validation

- Let f, g index CV test folds.
- On fold f, model m_f at step s, and $-f$ denoting the training set for test fold f (complement of f).
- Define $P_{f,s} := X^f_{m_f,s}(X^{-f}_{m_f,s})^\dagger$ (not a projection)
- $s = \arg\min_s \sum_{f=1}^K \|y^f - P_{f,s}y^{-f}\|_2^2$
- Sums of squares... maybe it's a quadratic form?

The key result

Blockwise quadratic form of cv-RSS

Define $Q^s_{ff} := \sum_{g \neq f} (P_{g,s})_f^T (P_{g,s})_f$ and

$$Q^s_{fg} := -(P_{f,s})_g - (P_{g,s})_f^T + \sum_{\substack{h=1 \\ h \notin \{f,g\}}}^{K} (P_{h,s})_f^T (P_{h,s})_g^T$$

Then with y_K denoting the observations ordered by CV-folds,

$$\text{cv-RSS}(s) = y_K^T Q^s y_K$$

Proof.

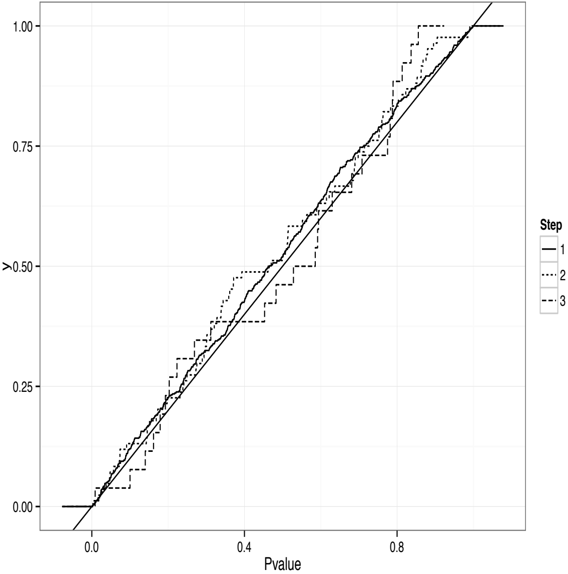
Algebra $\square$

Does it work?

Well yes, it's a theorem.

End of talk. Thank you for listening!

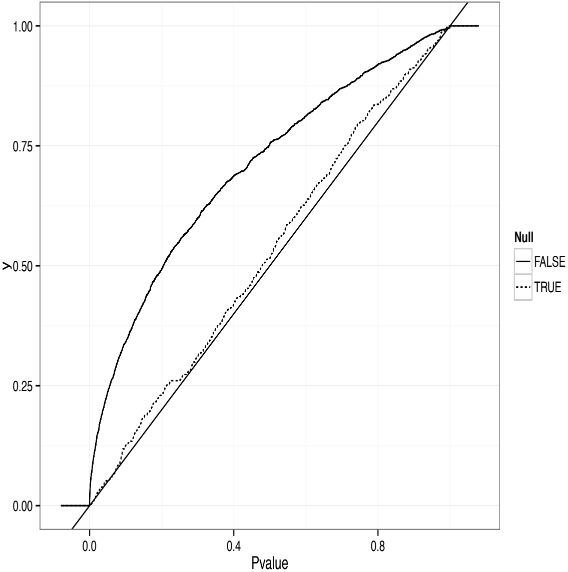
Global null, K = 5, n = 50, p = 100, steps = 10



About 16% chose sparsity > 1

Less than 0.5% chose sparsity > 3

5-sparse, K = 5, SNR = 7, n = 50, p = 100, steps = 10



Screened 90%

Another 4% 4/5

Limitations / ongoing work

- Forward stepwise used linearity $\hat{y}^f = Py^{-f}$. Ridge has same form. Lasso has additional constant terms, but works the same otherwise.
- Computationally expensive. Might be able to optimize a little more.
- Tuning parameters: K and maxsteps (or λ grid).
- Unknown σ. Estimate by CV / use selective F test.