RIC: Risk Inflation Criterion

October 24, 2012

1 Admistrivia

- Sorry about being slow on grading homework
- I’ve been moving to NYC—so behind on lots of stuff

2 Game plan for homework 4

Problem of domain adaption. Examples:

- train on NYT to test on Washington post
- train on microphones, test on telephone
• train on men, test on women

• train on CS jargon, test on wine jargon

Commonalities in these problems:

• All high dimension

• All high signal

• all need dimension reduction

**Variable selection**

\[ \text{estimate } \beta_i \text{ by } = \begin{cases} \\
0 & \text{some times} \\
\hat{\beta}_i & \text{other times} \\
\end{cases} \]

• most of the time zero is a better answer

• Sometimes estimation is better answer

But when to do each?

**Variable selection**

Heuristic: If \( t_{X_i} \) is large, then estimate \( \beta_{X_i} \) otherwise set it equal to zero. (Tukey called this a testimator.):
estimate $\beta_i$ by $\{ \begin{array}{ll}
0 & |t_{X_i}| < c \\
\hat{\beta}_i & |t_{X_i}| \geq c \end{array}$

(Fun with latex. I just installed the arrow package, so it isn’t completely under my control yet.)

How do tell the two estimators apart (a notation issue):

$$\hat{\beta}_i = \{ \begin{array}{ll}
0 & |t_{X_i}| < \hat{\beta}_{i \text{mle}} \\
\hat{\beta}_{i \text{mle}} & |t_{X_i}| \geq \hat{\beta}_{i \text{mle}} \end{array}$$

for the classical estimator $\hat{\beta}_{i \text{mle}}$ or $\hat{\beta}_{i \text{least squares}}$, or even $\hat{\beta}_{i \text{JMP}}$ would all work, but the “maximum likelihood estimator” name sounds the coolest and so is most often used.

**Possible values of cut off**

Complete table of possible values of the cutoff:
<table>
<thead>
<tr>
<th>cut off</th>
<th>Name</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>c = 0</td>
<td>MLE</td>
<td>i.e. overall least squares</td>
</tr>
<tr>
<td>c = 1</td>
<td>min s²</td>
<td>or Maximized adjusted R-squared</td>
</tr>
<tr>
<td>c = 2</td>
<td>AIC</td>
<td>default in R</td>
</tr>
<tr>
<td>c = 2</td>
<td>CP</td>
<td></td>
</tr>
<tr>
<td>c = log n</td>
<td>BIC</td>
<td>where n is the number of samples</td>
</tr>
<tr>
<td>c = 2 log p</td>
<td>RIC</td>
<td>where p is the number of variables tried</td>
</tr>
<tr>
<td>c = 2 log(p/q)</td>
<td>FDR</td>
<td>where q is the correct number of variables</td>
</tr>
<tr>
<td>c = ∞</td>
<td>“null”</td>
<td></td>
</tr>
</tbody>
</table>

3 So which to use?

Permutation idea

You have Y, X₁, X₂, …, Xₚ in your JMP table. You don’t want to include any X’s which are noise. So add a bunch of “noise” X’s to your jmp table by permuting the existing X’s. So,
Now regress $Y$ on all the $X$’s, new and old. If you get a new $X$, you know it has no connection to $Y$, so it is noise. So if you don’t want any extranious variables, stop as soon as you get a noisy $X$.

- RIC = stop on first noisy variable

- MDR = stop when more than .05 of the variables are noisy.

**Theoretical idea**

Back to our original goal, minimize risk:

$$Risk \equiv \min E(Y_{\text{future}} - \hat{Y}_{\text{future}})^2$$

Can we prove which of these estimators makes this error the smallest? No.
• Under the null model, using $c = \infty$ is best.

• Under a “rich” model, using $c = 0$ is best.

• Under models somewhere in between using $c$ somewhere in be-
tween is best.

• Cross validation might be a good trick here.

• {technical note: Empirical bayes is an aproach which tries to
guess this value of $c$ to use, but it still can’t prove it is right.}

4 Stepwise regression reminders

4.1 Overfitting

• In sample always gets better the longer you run stepwise regres-
sion

• But out-of-sample might not

• Classic “over fitting curve.”

4.2 Where is the minimum?

• Three stylized models:
- Null is basically right. (Best is few variables)
- All variables are useful. (Best is LS)
- All other settings

- Which domain are we in? We don’t know!

4.3 **Wasting 1/2 your data**

- We could generate these out of sample plots by wasting 1/2 of our data.
- Called cross validation
- The problem is that this is inefficient, and will miss patterns that should be found
- Goal: do what cross validation does, without paying for it

**Risk inflation idea**

Idea: Don’t mess up the easy problem.

- If someone can get a real small risk, a good estimator should do well also
- For hard problems, no one will fault you if you screw up
• Easy problems are ones with few parameters, hence they are simple, useful and important models to understand carefully

• Hard problems have lots of parameters, so we don’t think about them extensively

Idea: Ockham’s razor:

• Do well on small models

• Only use big models when forced to

Both recommend having low error for easy problems. So let’s make the “risk” on easy problem have higher importance. I.e. inflate it and make it larger.

4.4 Risk inflation definition

The risk inflation of an estimator is its risk compared to the best regression model.

Story version: Supposed 20 years later, science knows which variables should have been used when you solved your regression problem. Looking back on it, you say, “Gee I wish I’d only used, X17 and X23, that would have been really smart.” If you had done that, your error would have been 50, but instead you used a AIC estimator, so your error was 400. This gives you a risk inflation of 8.
Definition:

\[ RI \equiv \frac{\text{Risk}}{\text{best Risk}} \]

More symbols, which estimator are we talking about? (\( \hat{\beta} \)) Which data set? (\( \beta \))

\[ RI(\hat{\beta}, \beta) \equiv \frac{\text{Risk}(\hat{\beta}, \beta)}{\text{best Risk}} \]

What does best mean?

\[ RI(\hat{\beta}, \beta) \equiv \frac{\text{Risk}(\hat{\beta}, \beta)}{\min_{\text{all } 2^p \text{ models}} \text{Risk}(\hat{\beta}^{mle}, \beta)} \]

\{\text{note: We can actually precompute the best:}\}

\[ RI(\hat{\beta}, \beta) \equiv \frac{\text{Risk}(\hat{\beta}, \beta)}{q(\beta)\sigma^2} \]

\}

9
<table>
<thead>
<tr>
<th>cut off</th>
<th>Name</th>
<th>Risk Inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>MLE</td>
<td>( p )</td>
</tr>
<tr>
<td>1</td>
<td>( \min s^2 )</td>
<td>.8( p )</td>
</tr>
<tr>
<td>2</td>
<td>AIC/Cp</td>
<td>.57( p )</td>
</tr>
<tr>
<td>( \log n )</td>
<td>BIC</td>
<td>( \infty )</td>
</tr>
<tr>
<td>2( \log p )</td>
<td>RIC</td>
<td>2( \log p )</td>
</tr>
<tr>
<td>4( \log p )</td>
<td>2*RIC</td>
<td>4( \log p )</td>
</tr>
<tr>
<td>6( \log p )</td>
<td>3*RIC</td>
<td>6( \log p )</td>
</tr>
<tr>
<td>1( \log p )</td>
<td>.5*RIC</td>
<td>( \sqrt{p} )</td>
</tr>
<tr>
<td>2( \log(p/q) )</td>
<td>FDR</td>
<td>best!</td>
</tr>
<tr>
<td>( \infty )</td>
<td>“null”</td>
<td>( \infty )</td>
</tr>
</tbody>
</table>

Clearly the “\( \infty \)”’s are bad!

**Driving force**

Two issues need to be balanced:

- Over fitting (putting in to many variables)
- missing signal

Hence a trade off:

- As you increase your penalty, you start missing signal. So keep penalty small.
• As you decrease your penalty, you start adding zeros. So keep penalty large.

We can graph each of these separately. (see slides)

Take home messages from Risk Inflation:

• Stepwise regression can work well

• Use $\sqrt{2\log p}$

• In JMP, use Prob-to-enter = $1/p$ is a good approximation

• Puts in more variables than Bonferroni (.05/$p$).

• Lots and lots of variables are fine

5 Too many variables for JMP/R

What if you have too many variables for JMP to handle? What if there are too many for R to handle? SAS?

Now we are talking big data.