1 Outline

1. Ideal Risk & Oracles
2. Risk Inflation Criterion & Oracle Inequalities
3. Risk Inflation for Thresholding Rules
4. Risk Inflation for model selection via $C_p$

2 Ideal risk & oracles

As usual we have the model

$$y = X\beta + z$$

and choose among submodels $S \subset \{1, \ldots, p\}$.

Define $\hat{\beta}[S]$ to be the least-squares regression coefficients on covariates in $S$, and similarly $\hat{\mu}[S] = X\hat{\beta}[S]$. Then we have risk:

$$R(\mu, \hat{\mu}[S]) = \mathbb{E}||\hat{\mu}[S] - \mu||^2 = ||P_S\mu - \mu||^2 + |S|\sigma^2$$

where $P_S$ is the projection onto the linear space spanned by those covariates in $S$.

The ideal risk is the one coming from the best model

$$R^I(\mu) = \min_S R(\mu, \hat{\mu}[S])$$

In practice this is unachievable since we cannot compute $R(\mu, \hat{\mu}[S])$, but it gives us a benchmark which we may hope to mimic.

Note that if $\beta$ is $k$-sparse, i.e. $||\beta||_0 \leq k$, then we can do at least as well as taking as our model the support of $\beta$.

$$R^I(\mu) \leq k\sigma^2$$

For very sparse $\beta$, then, we could potentially do much better than $p\sigma^2$ if we had access to an “oracle” telling us the best model.
Example: Now suppose $X = I$, so that $\beta = \mu$ and $y \sim N(\mu, \sigma^2 I)$.

Then

$$R(\mu, \hat{\mu}[S]) = \sum_{i \notin S} \mu_i^2 + |S|\sigma^2$$

so for a fixed model size $|S|$, the best subset consists of the largest components of $\mu$, i.e.

$$\min_{|S|=s} R(\mu, \hat{\mu}[S]) = \sum_{i>s} |\mu(i)|^2 + s\sigma^2$$

So the best model size is

$$\min \sum_{i>s} |\mu(i)|^2 + s\sigma^2$$

Minimizing the last quantity over $s$ yields the ideal estimator $\hat{\mu}_i^I$ with risk

$$R^I(\mu) = \sum \min(\mu_i^2, \sigma^2)$$

which has a very simple form:

$$\hat{\mu}_i^I = \begin{cases} y_i & |\mu_i| > \sigma \\ 0 & \text{o.w.} \end{cases}$$

That is, if $\mu_i$ is “worth estimating,” we should take $\hat{\mu}_i = y_i$; otherwise, $\hat{\mu}_i = 0$. This way, we pay the minimum between the squared bias and the variance.

### 3 Risk inflation criterion

We have seen that if we had an oracle, we could do much better than the MLE.

Foster and George [1994] asked a fundamental question: can we get a real estimator whose risk is close to the ideal risk, whatever $\mu$ is?

The answer is yes: we can get within a log factor of the ideal risk.

Suppose that we minimize

$$\min \|y - X\beta\|^2 + \lambda_p^2 \sigma^2 \|\hat{\beta}\|_0$$

Here, $\|\beta\|_0$ is $\ell_0$ counting norm equal to $\sum_i 1(\beta_i \neq 0)$.

If $\lambda_p^2$ is on the order of $2 \log p$, then $\forall \mu \in \mathbb{R}^n$, we have

$$R(\mu, \hat{\mu}) \leq C_0 (2 \log p) [\sigma^2 + R^I(\mu)],$$

where $C_0$ can be computed explicitly. This result says that we can mimic the risk achieved by an oracle up to a logarithmic factor, which is at most logarithmic in the dimension.

Problem (??) is combinatorially hard and there is no known algorithm running in polynomial time that would solve it in the general case.
4 Risk inflation for thresholding rules

When $X$ is orthogonal, the solution to (??) is given by hard thresholding. There are two important kinds of thresholding rules, hard and soft.

The hard thresholding (“keep or kill”) rule is

$$\eta_H(y) = \begin{cases} 
0 & |y| \leq \lambda \\
y & \text{o.w.}
\end{cases}$$

The minimization problem with $X = I$ we have defined above uses hard thresholding at $\lambda = \lambda_p \sigma$.

The soft thresholding rule, on the other hand, is

$$\eta_S(y) = \begin{cases} 
0 & |y| \leq \lambda \\
y - \lambda & y > \lambda \\
y + \lambda & y < -\lambda
\end{cases}$$

The solution to the minimization problem with $X = I$ and the $\ell_1$ norm instead of the $\ell_0$ norm is given by soft-thresholding.

We plot the two thresholding rules below:
To see why the hard-thresholding rule is the solution to (??) in the case where $X^TX = I$, note that the objective functional is the same as

$$\|\hat{\beta} - X^T y\|^2 + \lambda^2 \sigma^2 \|\hat{\beta}\|_0.$$ 

Set $\nu = X^T y$, then

$$\min \sum_i (\nu_i - \hat{\beta}_i)^2 + 2\lambda^2 \sigma^2 \|\hat{\beta}\|_0 \iff \min \sum_i (\nu_i - \hat{\beta}_i)^2 + 2\lambda^2 \sigma^2 1(\hat{\beta}_i \neq 0).$$

We can consider every coordinate individually to see:

$$\hat{\beta}_i = \begin{cases} 
0 & |\nu_i| \leq \sqrt{2}\lambda\sigma \\
\nu_i & |\nu_i| > \sqrt{2}\lambda\sigma. 
\end{cases}$$

**Theorem 1 (Foster & George / Donoho & Johnstone).** Let $\hat{\mu}$ be either a soft or hard thresholding estimator with $\lambda = \sigma\sqrt{2\log p}$. Then

$$\mathbb{E}|\mu - \hat{\mu}|^2 \leq [2\log p + \delta] \left[ \sigma^2 + \sum_i \min(\mu_i^2, \sigma^2) \right]$$

with $\delta = 1$ for soft thresholding, 1.2 for hard. We recognize $\sum_i \min(\mu_i^2, \sigma^2)$ as the ideal risk $R^I(\mu)$.

Note that this inequality is non-asymptotic and holds for any $\mu$.

This type of inequality is called an oracle inequality, and it states that the risk of the thresholding rule is at most $2\log p$ larger than the ideal risk.

**Interpretation:** If we take $\lambda = \sqrt{2\log p}$, we are making sure that $\mathbb{P}(\hat{\mu} = 0 | \mu = 0) \to 1$ as $p \to \infty$. This threshold is similar to the Bonferroni threshold for global testing.

Note that if $\mu = 0$, the RHS reduces to $[2\log p + \delta] \sigma^2$. Then if we take $\lambda$ smaller than $\sigma\sqrt{2\log p}$, we will let in too many covariates and the variance term will kill our hopes of achieving this risk.

If we take $\lambda$ larger, on the other hand, we will simply get a worse constant.

## 5 Risk inflation for model selection via $C_p$

One natural question is whether a model fitted via the $C_p$ criterion can also mimic the oracle. Finding the best $C_p$ model is equivalent to finding that estimate solving

$$\arg\min_{\hat{\beta}} \|y - X\hat{\beta}\|^2 + 2\sigma^2 \|\hat{\beta}\|_0 \quad (3)$$

Formally, if we find the solution $\beta^*$ to (??), then $S^* = \{i : \hat{\beta}_i^* \neq 0\}$. Conversely, if we know the optimal $S^*$, then the solution to (??) is the LS estimate computed from variables in $S^*$.

We can analyze the performance of the $C_p$ approach in the case where $X^TX$ is diagonal; i.e. the predictor variables are orthogonal to each other. Here we shall assume that the columns of the
design matrix are normalized so that $X^TX = I$. This is no loss of generality and everything is similar in the case where the columns are un-normalized. Note that the two models

$$y \sim \mathcal{N}(X\beta, \sigma^2 I) \quad \text{and} \quad X^Ty \sim \mathcal{N}(\beta, \sigma^2 I)$$

are statistically equivalent so we will work with the latter.

We know that the solution is of the form

$$\hat{\beta}_i = \begin{cases} 0 & |y_i| \leq \sqrt{2}\sigma \\ y_i & |y_i| > \sqrt{2}\sigma. \end{cases}$$

In other words, for each $i$ we choose to incur one of two costs:

- If $y_i$ is small ($|y_i| \leq \sqrt{2}\sigma$) then roughly, we opt to incur a cost of $\beta_i^2$ by estimating $\mu_i$ as 0, and thus avoid the variance cost.
- If $y_i$ is large ($|y_i| > \sqrt{2}\sigma$) then roughly, it becomes worthwhile to estimate $\beta_i$ by $y_i$, then incurring a variance cost, but avoiding the bias cost.

There are two consequences when using a threshold of $\sqrt{2}\sigma$:

**Consequence 1:** Suppose we have $p_0$ null variables with $\beta_j = 0$. Then $C_p$ will nevertheless select a large number of them:

$$\mathbb{E} \left[ \frac{\# \{ \hat{\beta}_i \neq 0 : \beta_i = 0 \}}{p_0} \right] = \mathbb{P}(|Z| > \sqrt{2}) \approx 0.16$$

About 16% of them enter the model!

**Consequence 2:** We may not care about model selection per se, i.e. in identifying those variables in the model and those that are not. The problem, however, is that the prediction error is also quite large:

$$\sum_{i=1}^{p} \mathbb{E}(\hat{\beta}_i - \beta_i)^2 = \sigma^2 \sum_{i} r_H(\lambda = \sqrt{2}, \beta_i)$$

where $r_H(\lambda, \mu)$ is the risk of a hard-thresholding rule for estimating a scalar mean $\mu$ from $y \sim \mathcal{N}(\mu, 1)$. Figure ? plots the risk of the hard-thresholding rule. We can see that

$$0.57 \leq r_H(\mu, \sqrt{2}) \leq 1.66$$

where the minimum is achieved (not surprisingly) at $\mu = 0$.

This says that the risk of the $C_p$ model is at least

$$\text{Risk } C_p \geq 0.57p\sigma^2.$$ 

**Hence the risk inflation may be very large (proportional to $p$)** and we can see that the $C_p$ criterion cannot mimic an oracle the way a hard-thresholding procedure at the Bonferroni level does. What’s particularly discomforting is that the risk of the selected model is at least 60% that of the full model, whatever the sparsity level or the size of the right model.
Figure 1: Risk of hard-thresholding rule at $\lambda = \sqrt{2}$ for estimating $\mu$ from $y \sim \mathcal{N}(\mu, 1)$. 