Combining Biased and Unbiased Estimators in High Dimensions

Bill Strawderman

Rutgers University

(joint work with Ed Green, Rutgers University)
OUTLINE:

I. Introduction

II. Some remarks on Shrinkage Estimators

III. Combining Biased and Unbiased Estimators

IV. Some Questions/Comments

V. Example
I. Introduction:

Problem: Estimate a vector valued parameter $\theta$ ($\text{dim } \theta = p$).

We have multiple (at least 2) estimators of $\theta$, at least one of which is unbiased.

How do we combine the estimators?

Suppose, e.g., $X \sim \mathcal{N}_p(\theta, \sigma^2)$, and $Y \sim \mathcal{N}_p(\theta + \eta, \tau^2)$,

i.e., $X$ is unbiased and $Y$ is biased.

(and $X$ and $Y$ are independent)

Can we effectively combine the information in $X$ and $Y$ to estimate $\theta$?
Example from Forestry: Estimating basal area per acre of tree stands

(Total cross sectional area at a height of 4.5 feet)

Why is it important? It helps quantify the degree of above ground competition in a particular stand of trees.

Two sources of data to estimate basal area

a. X, sample based estimates. (Unbiased)

b. Y, Estimates based on regression model predictions (Possibly biased)

Regression model based estimators are often biased for parameters of interest since they are often based on non-linearly transformed responses which become biased on transforming back to the original scale.

In our example it is the log(basal area) which is modeled (linearly).

Dimension of X and Y in our application range from 9 to 47 (Number of stands)
The Usual Combined Estimator:

The “usual” combined estimators assume both estimators are unbiased and average with weights which are inversely proportional to the variances, i.e.,

\[ \delta^*(X, Y) = w_1 X + (1 - w_1) Y = \frac{\tau^2 X + \sigma^2}{\sigma^2 + \tau^2} \]

Can we do something sensible when Y is suspected of being biased?

Homework Problem: a. Find a biopharmaceutical application
II. Some remarks on Shrinkage Estimation:

Why Shrink: Some Intuition

Let $X = (X_1, \ldots, X_p)$ be a random vector in $\mathbb{R}^p$, with

$E[X_i] = \theta_i$, and $\text{Var}[X_i] = \sigma^2$, and $\text{Cov}(X_i, X_j) = 0$ ($i \neq j$)

**PROBLEM:** Estimate

$(\theta_1, \ldots, \theta_p)$ with loss $\sum_{i=1}^{p} (\delta_i - \theta_i)^2$ and Risk $= \sum_{i=1}^{p} E[\delta_i(X) - \theta_i]^2$.

Consider linear estimates of the form

$\delta_i(X_1, \ldots, X_p) = (1-a)X_i$, $i = 1, \ldots, p$, or $\delta(X) = (1-a)X$.

*Note:* $a = 0$ corresponds to the “usual” unbiased estimator $X$, but is it the best choice?
WHAT IS THE BEST $a$? Risk of $(1-a)X =$

$$\sum_{i=1}^{p} E[\delta_i(X) - \theta_i]^2 = \sum_{i=1}^{p} E[(1-a)X_i - \theta_i]^2 = \sum_{i=1}^{p} \{ \text{Var}[(1-a)X_i] + [E(1-a)X_i - \theta_i]^2 \}$$

$$p \text{Var}(1-a)X_i + \sum_{i=1}^{p} (a\theta_i)^2 = p(1-a)^2 \sigma^2 + 2(1-a)\sum_{i=1}^{p} \theta_i^2 = Q(a)$$

The best $a$ corresponds to

$$Q'(a) = 0 = -2\{(1-a)p\sigma^2 + a\theta \| \theta \|^2 \leftrightarrow a = p\sigma^2/(p\sigma^2 + \| \theta \|^2)$$

Hence the best linear “estimator” is

$$(1 - p\sigma^2/(p\sigma^2 + \| \theta \|^2))X,$$

which depends on $\theta$. BUT

$$E[\|X\|^2] = p\sigma^2 + \| \theta \|^2$$ so we can estimate the best $a$ as $p\sigma^2/\|X\|^2$,

and the resulting approximate “best linear estimator” is

$$(1 - p\sigma^2/\|X\|^2)X$$
The James-Stein Estimator is \((1 - (p - 2)\sigma^2 / \|X\|^2)X\), which is close to the above.

Note that the argument doesn’t depend on normality of X.

In the normal case, theory shows that the James Stein estimator has lower risk than the usual unbiased (UMVUE, MLE, MRE) estimator, X, provided that \(p \geq 3\).

In fact if the true \(\theta\) is 0 then the risk of the James-Stein Estimator is \(2\sigma^2\) which is much less than the risk of X (which is identically \(p\sigma^2\))
A slight extension: Shrinkage toward a fixed point, \( \theta_0 \):

\[
\theta_0 + \left(1 - \frac{(p-2)\sigma^2}{\|X - \theta_0\|^2}\right)(X - \theta_0) = X - \frac{(p-2)\sigma^2(X - \theta_0)}{\|X - \theta_0\|^2}
\]

also dominates \( X \) and has risk \( 2\sigma^2 \) if \( \theta = \theta_0 \)

\[
Risk = p\sigma^2 - (p-2)^2\sigma^4 E[1/\|X - \theta_0\|^2] < p\sigma^2
\]
III. Combining Biased and Unbiased Estimates:

Suppose we wish to estimate a vector $\theta$, and we have 2 independent estimators, $X$ and $Y$. Suppose, e.g., $X \sim \mathcal{N}_p(\theta, \sigma^2 I)$, and $Y \sim \mathcal{N}_p(\theta + \eta, \tau^2 I)$, i.e., $X$ is unbiased and $Y$ is biased.

Can we effectively combine the information in $X$ and $Y$ to estimate $\theta$?

ONE ANSWER: YES, Shrink the unbiased estimator, $X$, toward the biased estimator $Y$.

A James-Stein type combined estimator:

$$\delta(X,Y) = Y + (1 - (p - 2)\sigma^2 / \|X - Y\|^2)(X - Y) = X - \frac{(p-2)\sigma^2}{\|X - Y\|^2}(X - Y)$$
IV. Some Questions/Comments:

1. Risk of $\delta(X,Y) = \frac{1}{2} \left( \left\| X - Y \right\|^2 | Y \right) - \frac{1}{2} \left( \left\| X - Y \right\|^2 | Y \right)

$$E[R(\theta, \delta(X,Y))|Y] = p\sigma^2 - (p - 2)^2 \sigma^4 E[1/\|X - Y\|^2 | Y]$$

$$= p\sigma^2 - (p - 2)^2 \sigma^4 E[1/\|X - Y\|^2]$$

$$\leq p\sigma^2 - (p - 2)^2 \sigma^4 / E[\|X - Y\|^2]$$

$$\leq R(\theta, X) = p\sigma^2$$

Hence the combined estimator beats $X$ no matter how badly biased $Y$ is.
2. Why not shrink Y towards X instead of shrinking X towards Y?

Answer: Note that if X and Y are not close together, $\delta(X, Y)$ is close to X and not Y. This is desirable since Y is biased.

If we shrunk Y toward X, the combined estimator would be close to Y when X and Y are far apart.

This is not desirable, again, since Y is biased.
3. How does $\delta(X, Y)$ compare to the “usual method” of combining unbiased estimators, i.e., weighing inversely proportionally to the variances, $\delta^*(X, Y) = (\tau^2 X + \sigma^2 Y)/(\sigma^2 + \tau^2)$?

ANSWER: The risk of the JS combined estimator is slightly greater than the risk of the optimal linear combination when Y is also unbiased.

(JS is, in fact, an approximation to the best linear combination.)

Hence if Y is unbiased ($\eta=0$), the JS estimator does a bit worse than the “usual” linear combined estimator.

The loss in efficiency (if $\eta=0$) is particularly small when the ratio $\text{var}(X)/\text{Var}(Y)$ is not large and $p$ is large.

But it does much better if the bias of Y is significant.
Risk (MSE) Comparison of Estimators: $X$, Usual linear Combination ($\delta_{\text{combined}}$), and JS Combination ($\delta_{p-2}$);
Dimension $p = 25$; $\sigma = 1$, $\tau = 1$, equal bias for all coordinates
4. Is there a Bayes/Empirical Bayes Connection

Answer: Yes. The combined estimator can be interpreted as an Empirical Bayes Estimator under the prior structure

$$\eta \sim N(0, \gamma^2 I), \quad \theta \sim N(0, k^2 I), \quad \text{with } k \to \infty \quad (\text{i.e. locally uniform})$$
5. How does the combined JS estimator compare with the usual JS estimator (i.e., shrinking toward a fixed point)?

Answer: The risk functions cross. Neither is uniformly better.

Roughly, the combined JS estimator is better than the usual JS estimator if $||\eta||^2 + \rho \tau^2$ is small compared to $||\theta||^2$. 
6. Is there a similar method if we have several different estimators $X$ and $Y_i$, $i=1, \ldots, k$.

Answer: Yes. Multiple Shrinkage; suppose $Y_i \sim N(\theta + \eta_i, \tau_i^2 I)$

A Multiple Shrinkage Estimator of the form

$$X - (p-2)\sigma^2\left[ \sum_{i=1}^{k} \frac{(X-Y_i)}{\|X-Y_i\|^2} / \sum_{i=1}^{k} \frac{1}{\|X-Y_i\|^2} \right]$$

will work and have somewhat similar properties.

In particular, it will improve on the unbiased estimator $X$. 
7. How do you handle the unknown scale ($\sigma^2$) case?

Answer: Replace $\sigma^2$ in the JS combined (and uncombined) estimator by $\text{SSE}/(\text{df}+2)$.

Note that, interestingly, the scale of $Y$ ($\tau^2$) is not needed to calculate JS combined but that it is needed for the usual linear combined estimator.
8. Is normality of X (and/or Y) essential?

Answer: Whether $\sigma^2$ is known or not, normality of Y is not needed for the combined JS estimator to dominate X. (Independence of X and Y is needed)

Additionally, if $\sigma^2$ is not known, then Normality of X is not needed as well!

That is, (in the unknown scale case) the Combined JS estimator dominates the usual unbiased estimator, X, simultaneously for all spherically symmetric sampling distributions.

Also, the shrinkage constant $(p-2)\text{SSE}/(df+2)$ is (simultaneously) uniformly best.
V. EXAMPLE: Basal Area per Acre by Stand (Loblolly Pine)

Data:

<table>
<thead>
<tr>
<th>Company</th>
<th>Number of Stands (p=dim)</th>
<th>Number of Plots</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47</td>
<td>653</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>143</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>330</td>
</tr>
</tbody>
</table>

Number of Plots/Stand: Average = 17, Range = 5 – 50

“True” $\theta_{ij}$ and $\sigma_i^2$ calculated on basis the of all data in all plots (i= company, j=stand)
Simulation:

Compare Three Estimators: X, dcomb, and dJS+ for several different sample sizes,

\[ m = 5, 10, 30, 100 \] (plots/stand)

\[ X = \text{mean of } m \text{ measured average basal areas} \]

\[ Y = \text{estimated mean based on a linear regression of } \log(\text{basal area}) \]

\[ (\text{on } \log(\text{height}), \log(\text{number of trees}), (\text{age})^{-1}) \]

Generally X came in last each time
MSE $\delta_{JS+}$ /MSE $\delta_{comb}$ for Company 1 (solid), Company 2 (dashed) and Company 3 (dotted) for various sample sizes (m)
References:

James and Stein (61), 4th Berkeley Symposium

Green and Strawderman (91) JASA

Green and Strawderman (90), Forest Science

George (86) Annals of Statistics

Fourdrinier, Strawderman, and Wells (98) Annals of Statistics

Fourdrinier, Strawderman, and Wells (03) J. Mult. Analysis

Maruyama and Strawderman (05) Annals of Statistics

Fourdrinier and Cellier (95) J. Multivariate Analysis
Some Risk Approximations (Upper Bounds):

The Unbiased Estimator, $X$

$R(\theta, X) = p\sigma^2 - 0$

Usual James Stein

$R(\theta, \delta_{JS}(X)) \sim p\sigma^2 - (p-2)^2\sigma^4 / [\|\theta\|^2 + p\sigma^2]$

James Stein Combination

$R(\theta, \delta_{JS}(X,Y)) \sim p\sigma^2 - (p-2)^2\sigma^4 / [\|\eta\|^2 + p\sigma^2 + p\tau^2]$

Usual Linear Combination

$R(\theta, \delta_{comb}(X,Y)) = p\sigma^2 - p\sigma^4 / [\sigma^2 + \tau^2] + \{\sigma^2 / [\sigma^2 + \tau^2]\}^2\|\eta\|^2$