Shrinkage estimation

Regularized estimation, James-Stein type estimation and Empirical Bayes methods

Christoph Knappik

09.06.2006

Complex models - large p, small n

Applying statistical methods to analyze biological systems can be a tricky task. We often have only limited data to fit complex models and commonly used estimators (like ML) are not efficient enough in this context.
Generally there are three ways to deal with this problem. We can use:

- Bayes inference
- penalized maximum likelihood estimation or
- shrinkage estimators

Complex models - large p, small n

In the coming 45 minutes you will learn more about

- Stein's estimator
- Stein's paradoxon
- How to apply Stein's estimator to 'real' data
- The concept of shrinkage for regularized estimation

Multivariate normal distribution and MLE for the mean

Suppose that for given θ_i

$$X_i \mid \theta_i \overset{ind}{\sim} N(\theta_i, 1), \quad i = 1, ..., k \geq 3.$$

The unknown vector of means $\boldsymbol{\theta} \equiv (\theta_1, ..., \theta_k)$ is to be estimated with loss being the sum of squared component errors

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \equiv \sum_{i=1}^{k} (\hat{\theta}_i - \theta_i)^2$$

where $\hat{\boldsymbol{\theta}} \equiv (\hat{\theta}_1, ..., \hat{\theta}_k)$ is the estimate of $\boldsymbol{\theta}$.

The MLE's risk

The MLE which is also the sample mean,
$\boldsymbol{\delta}^{\mathbf{0}}(\boldsymbol{X}) \equiv \boldsymbol{X} \equiv (X_1, ..., X_k)$ has constant risk k (=MSE).

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathbf{0}}) \equiv E_\theta \sum_{i=1}^{k} (X_i - \theta_i)^2 = k$$

Stein's estimator

James and Stein introduced the estimator

$$\boldsymbol{\delta}^{\mathbf{1}}(\boldsymbol{X}) = (\delta_1^1(\boldsymbol{X}), ..., \delta_k^1(\boldsymbol{X})) \quad \text{for} \quad k \geq 3$$

$$\delta_i^1(\boldsymbol{X}) \equiv \mu_i + (1 - (k-2)/S)(X_i - \mu_i), \quad i = 1, ..., k$$

with

- $\boldsymbol{\mu} \equiv (\mu_1, ..., \mu_k)'$ any initial guess at $\boldsymbol{\theta}$
- $S \equiv \sum (X_j - \mu_j)^2$

Stein's estimator dominates the ML estimator

A simple calculation shows that $\delta_i(\boldsymbol{X})$ is a weighted sum of X_i and μ_i:

$$\delta_i^1(\boldsymbol{X}) = \lambda \mu_i + (1 - \lambda) X_i, \quad (\lambda = \frac{k-2}{S}).$$

$\boldsymbol{\delta}^{\mathbf{1}}(\boldsymbol{X})$ has risk

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathbf{1}}) \equiv E_\theta \sum_{i=0}^{k} (\delta_i^1(\boldsymbol{X}) - \theta_i)^2 \leq k - \frac{(k-2)^2}{k - 2 + \sum (\theta_i - \mu_i)^2} \leq k$$

Stein's estimator - Remember

Remember

Using the MLE to estimate the mean of a multivariate normal distribution is a not an optimal choice! For $k \geq 3$ the ML estimator is inadmissible.
As you will see later, empirical Bayes estimators like Stein's reduce the total risk by a large margin compared to the sample mean's risk.

Stein's estimator in an empirical Bayes context

Rev. T. Bayes

The empirical Bayes context - a priori and a posteriori distribution of θ_i

$\delta_i^1(\boldsymbol{X}) \equiv \mu_i + (1-(k-2)/S)(X_i - \mu_i), \quad i = 1, ..., k$ arises quite naturally in an empirical Bayes context. If the $\{\theta_i\}$ themselves are a sample from a prior distribution,

$$\theta_i \overset{ind}{\sim} N(\mu_i, \tau^2), \quad i = 1, ..., k$$

then the Bayes estimate of θ_i is the a posteriori mean of θ_i given the data

$$\delta_i^*(X_i) = E\theta_i \mid X_i = \mu_i + (1 - \underbrace{(1+\tau^2)^{-1}}_{\lambda})(X_i - \mu_i)$$

The empirical Bayes context - estimation of τ^2

In the empirical Bayes situation τ^2 is unknown but it can be estimated because marginally the $\{X_i\}$ are independently normal with means $\{\mu_i\}$ and

$$S = \sum (X_j - \mu_j)^2 \sim (1+\tau^2)\chi_k^2$$

Since $k \geq 3$, the unbiased estimate

$$E(k-2)/S = 1/(1+\tau^2)$$

is available.

The empirical Bayes context - derivation of Stein's estimator

Substitution of $(k-2)/S$ for the unknown $1/(1+\tau^2)$ in

$$\delta_i^*(X_i) = E\theta_i \mid X_i = \mu_i + (1-(1+\tau^2)^{-1})(X_i - \mu_i)$$

results in

$$\mu_i + (1-(k-2)/S)(X_i - \mu_i) \equiv \delta_i^1(\boldsymbol{X})$$

$\delta_i^1(\boldsymbol{X})$ has risk

$$E_\tau E_\theta(\delta_i^1(\boldsymbol{X}) - \theta_i)^2 = 1 - (k-2)/k(1+\tau^2)$$

The empirical Bayes context - Stein's estimator's risk

$$E_\tau E_\theta(\delta_i^1(\boldsymbol{X}) - \theta_i)^2 = 1 - (k-2)/k(1+\tau^2)$$

is to be compared to the corresponding risks of

- 1 for the MLE and
- $1 - 1/(1+\tau^2)$ for the Bayes estimator

Thus if k is moderate or large δ_i^1 is nearly as good as the Bayes estimator, but it avoids the possible gross errors of the Bayes estimator if τ^2 is misspecified.

The empirical Bayes context - positive part Stein

A simple way to improve δ_i^1 is to use $min\{1, (k-2)/S\}$ as an estimate of $1/(1+\tau^2)$ instead of $E(k-2)/S$.
This results in

$$\delta_i^{1+}(\boldsymbol{X}) = \mu_i + (1-(k-2)/S)^+(X_i - \mu_i)$$

with $a^+ \equiv max(0, a)$.

It can be proofed that $R(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathbf{1+}}) < R(\boldsymbol{\theta}, \boldsymbol{\delta}^{\mathbf{1}}) \quad \forall \boldsymbol{\theta}$.

The empirical Bayes context - Remember

Remember

- Stein's estimator dominates the MLE for $k \geq 3$
- Stein's estimator can be interpreted as an empirical Bayes estimator

Using Stein's estimator to predict batting averages

Using Stein's estimator to predict batting averages

The data

The batting averages of 18 major league players through their first 45 official at bats of the 1970 season (The samplesize of $n = 45$ was chosen to assure a satisfactory approximation of the binomial by the normal distribution).

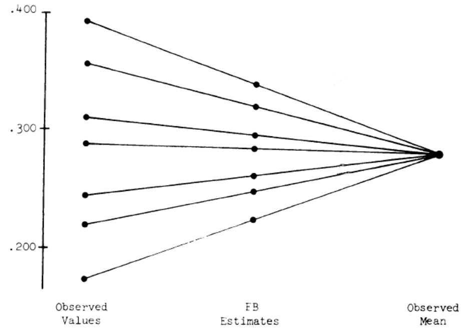
The challenge

Predict each player's batting average over the reminder of the season (70 - almost 600 at bats) using only the data of the first 45 at bats.

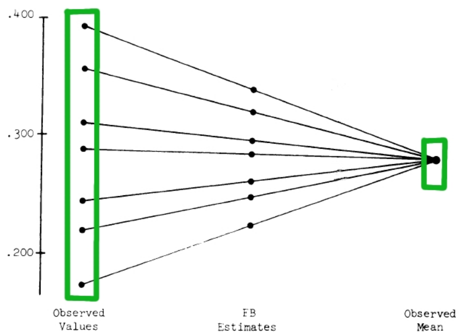
The solution

Using Stein's estimator as a shrinkage estimator.

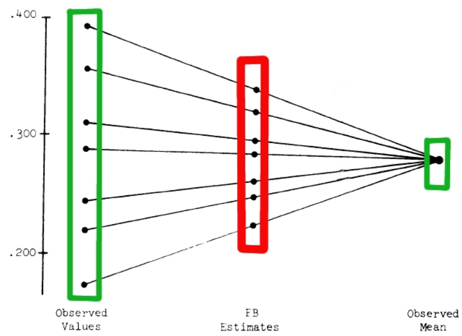
The concept of shrinking



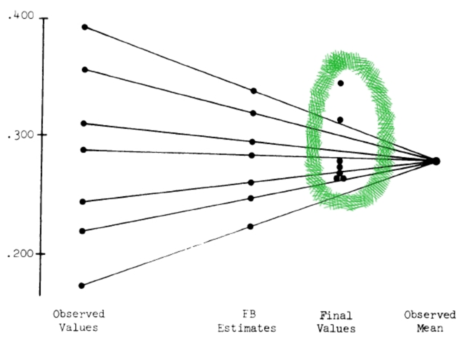
The concept of shrinking



The concept of shrinking



The concept of shrinking - Regression towards the mean



The estimation in detail - the data

i	Player	Y_i = batting average for first 45 at bats (1)	p_i = batting average for remainder of season (2)	At bats for remainder of season (3)
1	Clemente (Pitts, NL)	.400	.346	367
2	F. Robinson (Balt, AL)	.378	.298	426
3	F. Howard (Wash, AL)	.356	.276	521
4	Johnstone (Cal, AL)	.333	.222	275
5	Berry (Chi, AL)	.311	.273	418
6	Spencer (Cal, AL)	.311	.270	466
7	Kessinger (Chi, NL)	.289	.263	586
8	L. Alvarado (Bos, AL)	.267	.210	138
9	Santo (Chi, NL)	.244	.269	510
10	Swoboda (NY, NL)	.244	.230	200
11	Unser (Wash, AL)	.222	.264	277
12	Williams (Chi, AL)	.222	.256	270
13	Scott (Bos, AL)	.222	.303	435
14	Petrocelli (Bos, AL)	.222	.264	538
15	E. Rodriguez (KC, AL)	.222	.226	186
16	Campaneris (Oak, AL)	.200	.285	558
17	Munson (NY, AL)	.178	.316	408
18	Alvis (Mil, NL)	.156	.200	70

Transformation to adjust the variance

Let Y_i be the batting average of Player $i, i = 1, ..., 18 (k = 18)$ after $n = 45$ at bats. Further asume that

$$nY_i \overset{ind}{\sim} Bin(n, p_i), \quad i = 1, ..., 18$$

with p_i the true season batting average, so $EY_i = p_i$. To stabilize the variance of Y_i at nearly unit variance the arc-sin transformation is used: $X_i \equiv f_{45}(Y_i)$ with

$$f_n(y) \equiv (n)^{\frac{1}{2}} arcsin(2y - 1).$$

Estimation of μ from the data

From the central limit theorem for the binomial distribution and the continuity of f_n we have approximately

$$X_i \mid \theta_i \overset{ind}{\sim} N(\theta_i, 1), \quad i = 1, ..., k$$

with mean θ_i of X_i given approximately by $\theta_i = f_n(p_i)$.

We can now use Stein's estimator, but we also want to estimate the common unknown value $\mu = \sum \mu_i / k$ by $\overline{X} = \sum X_i / k$, shrinking all X_i toward $\overline{X}$.

Estimation of $\boldsymbol{\theta}$ using the Bayes rule

Using the Bayes rule shown earlier the resulting estimate of the i-th component θ_i of $\boldsymbol{\theta}$ is therefore

$$\tilde{\delta}_i^1 = \overline{X} + (1 - (k-3)/V)(X_i - \overline{X})$$

with $V \equiv \sum (X_i - \overline{X})^2$ and with $k - 3 = (k-1) - 2$ as the appropriate constant since one parameter is estimated.

And with risk

$$R(\boldsymbol{\theta}, \tilde{\boldsymbol{\delta}}^{\mathbf{1}}) \leq k - \frac{(k-3)^2}{k - 3 + \sum(\theta_i - \overline{\theta})^2}, \quad \overline{\theta} \equiv \sum \theta_i / k$$

Results

For our data the estimate of $1/(1+\tau^2)$ is $(k-3)/V = .791$, $\hat{\tau} = .514$ and $\overline{X} = -3.275$ so

$$\tilde{\delta}_i^1(\boldsymbol{X}) = \hat{\theta}_i = .791\overline{X} + .209X_i = .209X_i - 2.59.$$

i	Batting average for season remainder p_i	Maximum likelihood estimate Y_i	Retransform of Stein's estimator $\hat{p}_i^1$
1	.346	.400	.290
2	.298	.378	.286
3	.276	.356	.281
4	.222	.333	.277
5	.273	.311	.273
6	.270	.311	.273
7	.263	.289	.268
8	.210	.267	.264
9	.269	.244	.259
10	.230	.244	.259
11	.264	.222	.254
12	.256	.222	.254
13	.303	.222	.254
14	.264	.222	.254
15	.226	.222	.254
16	.285	.200	.249
17	.316	.178	.244
18	.200	.156	.239

Results

i	Batting average for season remainder p_i	Maximum likelihood estimate Y_i	Retransform of Stein's estimator $\hat{p}_i^1$
1	.346	.400	.290
2	.298	.378	.286
3	.276	.356	.281
4	.222	.333	.277
5	.273	.311	.273
6	.270	.311	.273
7	.263	.289	.268
8	.210	.267	.264
9	.269	.244	.259
10	.230	.244	.259
11	.264	.222	.254
12	.256	.222	.254
13	.303	.222	.254
14	.264	.222	.254
15	.226	.222	.254
16	.285	.200	.249
17	.316	.178	.244
18	.200	.156	.239

Results

The results are striking:

- $\boldsymbol{X}$ has total squared prediction error of **17.56** but
- $\tilde{\boldsymbol{\delta}}^{\mathbf{1}}(\boldsymbol{X})$ has total squared prediction error of only **5.01**
- $\tilde{\delta}^1(X)$ is closer than X_i to θ_i for 15 batters
- The use of „limited translation estimators“ (which we do not cover here) can further improve the results

Some things to take home

While the technical details might not be most important for this seminar, there are quite a few things you should remember about Stein's estimator:

- Stein's Estimator provides a simple way of doing regularized inference
- The MLE is inadmissible for estimating the mean of a multivariate normal distribution (Stein's paradoxon)
- Stein's estimator is available as empirical Bayes estimator
- Stein's estimator can be used as shrinkage estimator