STATS 300C: Theory of Statistics **Spring 2022**

Lecture 17 — May 24, 2022

Lecturer: Prof. Emmanuel Candès *Editor: Parth Nobel, Scribe: William Hartog*

Warning: *These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès's course from 2018 and 2022, and 2021 scribe notes written by Albert Chiu, Disha Ghandwani, and Zhaomeng Chen.*

Reading: Include Reading

Agenda: James-Stein Continued and Full Conformal

1. Empirical Bayes interpretation of James-Stein (JS) estimator
2. Extension: shrinkage to arbitrary point
3. Baseball Example
4. Full Conformal

17.1 An Empirical Bayes Interpretation of the James-Stein Estimator (1)

17.1.1 Setup

Consider the Bayesian setup,

$$\mu_i \sim \mathcal{N}(0, \tau^2) \tag{17.1}$$

$$X \mid \mu \sim \mathcal{N}(\mu, \sigma^2 I). \tag{17.2}$$

17.1.2 Empirical Bayes estimator

Given the data X, the posterior of μ is distributed normal with mean linear in X and reduced variance. In particular,

$$\mu|X \sim \mathcal{N}(\frac{\nu}{\sigma^2}X, \nu I), \tag{17.3}$$

where $\frac{1}{\nu} = \frac{1}{\tau^2} + \frac{1}{\sigma^2} = \frac{\tau^2+\sigma^2}{\tau^2\sigma^2}$. We can show this to be true by either directly deriving it, or by analogy appealing to linear regression.[1]

[1] (μ, X) is bivariate normal, so $\mu|X$ is normal. $E[\mu \mid X]$ is simply a linear regression problem, and a result from univariate regression tells us the coefficient of X is $cov(\mu, X)/var(X) = \nu/\sigma^2$.

The Bayes estimator, which minimizes the Bayes risk, is simply the mean of the posterior,

$$\hat{\mu}_B = \nu/\sigma^2 X = \big(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\big)X. \tag{17.4}$$

Note the following:

- This is a shrinkage estimator. $\frac{\sigma^2}{\sigma^2+\tau^2} > 0$, so we are always shrinking the MLE $\hat{\mu}_{MLE} = X$ toward zero.
- Unlike the factor by which we multiple the MLE to get the James-Stein estimator $\hat{\mu}_{JS} = \big(1 - \frac{(p-2)\sigma^2}{\|X\|^2}\big)X$, the weight for (17.4) case cannot be negative. That is, if $X_i > 0$, then $\hat{\mu}_{B,i}$ will also be positive, and vice versa if $X_i < 0$.
- If $\sigma = \tau$, then we shrink the MLE halfway toward zero. In signal processing terms, we can think of τ as a parameter indicating the strength of the signal (with larger values indicating stronger signal) and σ as the amount of the noise. If the two are equal, the signal to noise ratio (SNR) is 1.

17.1.3 Bayes Risk

Proposition 1. Assuming 17.1 and 17.2, the Bayes risk is

$$R_{\hat{\mu}_B} := E\|\hat{\mu}_B - \mu\|^2 = R_{MLE}\frac{\tau^2}{\tau^2 + \sigma^2}, \tag{17.5}$$

where R_{MLE} denotes the risk of the MLE.

Proof. Let us first rewrite the difference between the estimator and the parameter as

$$\hat{\mu}_B - \mu = (1 - \rho)(X - \mu) - \rho\mu, \tag{17.6}$$

where $\rho = 1 - \nu = \frac{\sigma^2}{\sigma^2+\tau^2}$. We have,

$$E[\|\hat{\mu}_B - \mu\|^2 \mid \mu] = (1 - \rho^2)E[(X - \mu)^2 \mid \mu] - \rho^2\|\mu\|^2 \tag{17.7}$$

$$= (1 - \rho)^2 p\sigma^2 - \rho^2\|\mu\|^2. \tag{17.8}$$

The second equality follows from the facts that $X - \mu \mid \mu \sim \mathcal{N}(0, \sigma^2 I)$ and $Tr(\sigma^2 I) = p\sigma^2$. Taking an outer expectation and integrating over μ, we get the desired result:

$$E[\|\hat{\mu}_B - \mu\|^2] = E[E[\|\hat{\mu}_B - \mu\|^2 \mid \mu]] \tag{17.9}$$

$$= (1 - \rho)^2 p\sigma^2 + \rho^2 p\tau^2 \tag{17.10}$$

$$= p\sigma^2\Big[\frac{\tau^2}{\tau^2 + \sigma^2}\Big]. \tag{17.11}$$

□

Note the following:

- Clearly, $R_B < R_{MLE}$ always.
- If $\tau^2 \ll \sigma^2$ (that is, the SNR is low), then the risk improves significantly, which makes sense since we want to shrink more when the noise is higher. If the noise is low then the MLE should intuitively perform well.
- If $\sigma = \tau$ (SNR=1), then the risk is halved.

17.1.4 Connection to James-Stein

Let us assume that the Bayesian model is correct and σ is known but τ is not. We cannot directly compute the shrinkage factor $\frac{\tau^2}{\tau^2+\sigma^2}$, but perhaps we can estimate it using the data.

Since $X_i = \mu_i + z_i$, where $z_i \sim \mathcal{N}(0, \sigma^2)$, is a sum of independent normals, we have that $X_i \sim \mathcal{N}(0, \tau^2 + \sigma^2)$. This then implies that

$$\|X\|^2 \sim (\tau^2 + \sigma^2)\chi^2_p. \tag{17.12}$$

Combining this result with the fact that $E\left[\frac{p-2}{\chi^2_p}\right] = 1$, we arrive at an unbiased estimator for the shrinkage factor:

$$\frac{(p-2)\sigma^2}{\|X\|^2}. \tag{17.13}$$

If we substitute 17.13 for the unknown shrinkage factor in 17.4, we recover the James-Stein estimator $\hat{\mu}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\|X\|^2}\right)X$. Moreover, we can derive the risk for $\hat{\mu}_{JS}$ in this setting.

Theorem 1. Assuming 17.1 and 17.2, the risk of $\hat{\mu}_{JS}$ is

$$E\|\hat{\mu}_{JS} - \mu\|^2 = p\sigma^2\left[\frac{\tau^2}{\tau^2+\sigma^2}\right] + \frac{2\sigma^2}{\tau^2+\sigma^2} \tag{17.14}$$

$$= R_{\hat{\mu}_B} + \frac{2\sigma^2}{\tau^2+\sigma^2} \tag{17.15}$$

$$= \left(1 + \frac{2\sigma^2}{p\tau^2}\right)R_{\hat{\mu}_B}. \tag{17.16}$$

That is, the risk of JS is bigger than the Bayes risk by a factor of $\left(1 + \frac{2\sigma^2}{p\tau^2}\right)$. This can be quite small if p is large (in fact, it goes to zero as $p \to \infty$). As an example, if $\tau = \sigma$ and $p = 20$, the factor is 1.1.

17.2 Extension: Shrinking Toward an Arbitrary Point

Thus far, we have considered estimators that shrink toward zero, but we need not do so. As it turns out, we can shrink toward an arbitrary point μ_0.

Definition 1. For any point μ_0, we can define an estimator that shrinks toward it

$$\hat{\mu}^{\mu_0}_{JS} = \mu_0 + \left(1 - \frac{(p-2)\sigma^2}{\|X - \mu_0\|^2}\right)(X - \mu_0). \tag{17.17}$$

Proposition 2. $\hat{\mu}^{\mu_0}_{JS}$ also dominates the MLE everywhere.

Proof. Define $Y := X - \mu_0$. Then $Y \sim \mathcal{N}(\mu - \mu_0, \sigma^2)$, and $\hat{\mu}_{JS}$ dominates the MLE for estimating $\mu - \mu_0$. This is equivalent to saying that $\hat{\mu}^{\mu_0}_{JS}$ dominates the MLE for estimating μ. □

17.2.1 Shrinking Toward the Group Mean

In practice, instead of arbitrarily picking some point, it might instead make sense to chose $\mu_0 = \bar{X}$ so as to adapt to the true center of μ_i.

Assume the following Bayesian setup:

$$\mu_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \tau^2) \tag{17.18}$$

$$X \mid \mu \sim \mathcal{N}(\mu, \sigma^2 I), \tag{17.19}$$

with σ known and μ_0, τ unknown.

The marginal distribution of our data is

$$X_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \tau^2 + \sigma^2), \tag{17.20}$$

and the posterior of our means is

$$\mu_i \mid X_i \sim \mathcal{N}\big(\mu_0 + (1-\rho)(X_i - \mu_0), 1 - \rho\big), \tag{17.21}$$

where again $\rho = \frac{\sigma^2}{\sigma^2+\tau^2}$.

$\mu + (1-\rho)(X_i - \mu_0)$ is the Bayes estimator, but μ_0 is unknown. Taking an empirical Bayes approach, we can use the unbiased estimator $\bar{X}$ in its place. Similarly, we can use the sample variance $S = \sum_i (X_i - \bar{X})^2 \sim (\tau^2 + \sigma^2)\chi^2_{p-1}$ to estimate τ^2. This then gives us the estimator

$$\hat{\mu}^{\bar{X}}_{JS} = \bar{X} + \left(1 - \frac{(p-3)\sigma^2}{S}\right)(X_i - \bar{X}).^{[2]} \tag{17.22}$$

If $p > 3$, this estimator dominates the MLE everywhere.[3]

17.3 Baseball Example

Efron and Morris in (1) examine the batting averages of 18 baseball players with exactly 45 at-bats as of April 26, 1970. We assume a player has some "true" batting average which is well approximated by their batting average in the remainder of the season and wish to estimate this quantity using their first 45 at-bats. Since batting averages are binomial, we can use the normal approximation

$$X_i \sim \mathcal{N}(\theta_i, \frac{1}{45}\theta_i(1-\theta_i)). \tag{17.23}$$

Problematically, the variance depends on the mean. One solution is to make a variance stabilizing transformation,

$$Y_i = \sqrt{45}\arcsin(2X_i - 1) \overset{apprx}{\sim} \mathcal{N}(\mu_i, 1), \tag{17.24}$$

[2]Note that we have fewer degrees of freedom in χ^2 distribution since we use the data to estimate $\bar{X}$, so this expression contains $p-3$ instead of $p-2$.

[3]We require $p > 3$ instead of $p > 2$ because $\hat{\mu}^{\bar{X}}_{JS} = \hat{\mu}_{MLE}$ if $p = 3$.

where $\mu_i = \sqrt{45}\arcsin(2\theta_i - 1)$.

Then the JS estimator is

$$\hat{\mu}_{JS} = \bar{Y} + \Big(1 - \frac{15}{\|Y - \bar{Y}\|^2}\Big)(Y - \bar{Y}). \tag{17.25}$$

Comparing JS to the MLE, we see a dramatic improvement in MSE. In particular,

$$\|\hat{\mu}_{JS} - \mu\|^2 = 5.01 \tag{17.26}$$
$$\|\hat{\mu}_{MLE} - \mu\|^2 = 17.56 \tag{17.27}$$
$$. \tag{17.28}$$

Transforming back to the original parameter,

$$\|\hat{\theta}_{JS} - \theta\|^2 = .022 \tag{17.29}$$
$$\|\hat{\theta}_{MLE} - \theta\|^2 = .077 \tag{17.30}$$
$$. \tag{17.31}$$

Note however that the improvement is not uniform: for three players, the JS estimate is worse. The improvement in overall MSE results from the improvement in the remaining 15 players.

This should not be surprising, since the JS phenomenon only applies to overall risk. If we are interested in one or two specific players, the MLE is admissible. Only if we are interested in minimizing overall loss for more than two players does JS dominate the MLE.

17.4 Full conformal

17.4.1 Prediction intervals

In Lecture 8, we talked about the cases in which we use sophisticated machine learning algorithms to make predictions in situations that have tremendous consequences. When the cost of making a wrong prediction is very high, for example being denied college admission, it's important to understand the reliability of the algorithms we use. A very good way of understanding the reliability or the uncertainty of future prediction is to be able to return prediction intervals.

Suppose we have training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and a test point $(X_{n+1}, ?)$ (we want to predict the label corresponding to X_{n+1}). The data are assumed to be exchangeable, e.g. i.i.d. from some distribution P_{XY}. Our goal is to construct marginal distribution free prediction interval $\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$ for any (unknown) distribution P_{XY} and any sample size n. For example, we want to be able to say things like "Based on the candidate's high school identifier and GPA, SAT scores, and other attributes, the college GPA is predicted in the [3.4,3.8] range".

We first state the following quantile lemma, which will be useful for our later development.

Definition 2. Define the quantile function to be

$$\text{Quantile}(\beta; F) = \inf\{z : \mathbb{P}\{Z \leq z\} \geq \beta\},$$

where F is the cumulative distribution function corresponding to $\mathbb{P}$.

For a multiset $v_{1:n} = \{v_1, \cdots, v_n\}$, define

$$\text{Quantile}(\beta; v_{1:n}) = \text{Quantile}(\beta; \frac{1}{n}\sum_{i=1}^{n} \delta_{v_i}).$$

Proposition 3 (Quantile Lemma)**.** If $V_1, \cdots, V_{n+1}$ are exchangeable, then for any $\beta \in (0, 1)$,

$$\mathbb{P}(V_{n+1} \leq \text{Quantile}(\beta; V_{1:n} \cup \{\infty\})) \geq \beta.$$

If ties between $V_1, \cdots, V_{n+1}$ occur with probability 0, then above probability is at most $\beta + 1/(n+1)$.

The key to above lemma is the observation that the rank of V_{n+1} is uniform over $\{1, \cdots, n+1\}$ (due to the exchangeability of $V_1, \cdots, V_{n+1}$). See the following proof:

Proof. We note that setting Q to be the quantile in the above lemma, the following are equivalent:

$$V_{n+1} \leq Q \Leftrightarrow \frac{\#\{j : V_j \leq V_{n+1}\}}{n+1} \leq \beta \Leftrightarrow \text{rank}(V_{n+1}) \leq 1 + \beta(n+1).$$

Since $\text{rank}(V_{n+1}) \sim \text{DUnif}\{1, ..., n+1\}$ we have that

$$\text{BP}(\text{rank}(V_{n+1}) \leq 1 + \beta(n+1)) \leq \frac{1 + \beta(n+1)}{n+1} = \beta + \frac{1}{n+1},$$

using exchangeability and

$$\text{BP}(\text{rank}(V_{n+1}) \leq 1 + \beta(n+1)) = \frac{\lfloor 1 + \beta(n+1) \rfloor}{n+1} \geq \beta$$

by definition of the discrete uniform. The case with ties doesn't change anything for the lower bound argument.

□

17.4.2 Conformal prediction

Suppose we have a training set $Z_i = (X_i, Y_i)$, $i = 1, \cdots, n$, and a (non-conformity) score function $\mathcal{S}$ with two arguments, a point (x, y) and a multiset Z. A low score of $\mathcal{S}((x, y), Z)$ indicates that (x, y) "conforms" to Z, whereas a high value indicates that (x, y) is atypical relative to the points in Z.

A common example of the score function is given by $\mathcal{S}((x, y), Z) = |y - \hat{\mu}(x)|$, where $\hat{\mu} : \mathbb{R}^d \to \mathbb{R}$ is a regression function fitted by running an algorithm $\mathcal{A}$ on (x, y) and Z. We will also assume $\mathcal{A}$ treats its arguments symmetrically.

Now suppose we are given a test point x and we want to calculate the prediction interval $\hat{C}_n(x)$ for a potential observation whose covariate has value x.

Define $V_i^{(x,y)} = \mathcal{S}(Z_i, Z_{-i} \cup (x,y))$, $i = 1, \cdots, n$, and $V_{n+1}^{(x,y)} = \mathcal{S}((x,y), Z_{1:n})$. Now for a value y, we include y in $\hat{C}_n(x)$ if $V_{n+1}^{(x,y)} \leq \text{Quantile}(1-\alpha; V_{1:n}^{(x,y)} \cup \{\infty\}))$, where $V_{1:n}^{(x,y)} = \{V_1^{(x,y)}, \cdots, V_n^{(x,y)}\}$.

The main result is the following:

Theorem 2. (Vovk et al. 2005, Lei et al. 2018) Assume that $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, i = 1, \cdots, n+1$ are exchangeable. For any score function $\mathcal{S}$, and any $\alpha \in (0,1)$, $\hat{C}_n$ satisfies

$$\mathbb{P}(Y_{n+1} \in \hat{C}_n(X_{n+1})) \geq 1 - \alpha.$$

If ties between V_i's occur with probability 0, then above probability is at most $1 - \alpha + 1/(n+1)$ (we can also make it equal to $1-\alpha$ by randomization).

This theorem is a simple consequence of the quantile lemma. The only thing to notice is that since Z_i's are exchangeable, so are the $V_i^{(X_{n+1}, Y_{n+1})}$'s.

Bibliography

[1] Bradley Efron and Carl Morris Stein's estimation rule and its competitors – an empirical Bayes approach Journal of the American Statistical Association, 1973.