Statistics 210B, Spring 1998

Class Notes

P.B. Stark

stark@stat.berkeley.edu

www.stat.berkeley.edu/∼stark/index.html

April 15, 1998

Tenth Set of Notes

1 More Multiplicity: Shrinkage Estimators

For references, see Stein, C., 1956. Inadmissibility of the usual estimator of the mean of a multivariate normal distribution, *Proc. Third Berkeley Symp. Math. Stat. Probab., 1*, 197-206, Univ. Calif. Press; James, W., and Stein, C., 1961. Estimation with quadratic loss, *Proc. Fourth Berkeley Symp. Math. Stat. Probab., 1*, 361-380, Univ. Calif. Press; Stein, C., 1981. Estimation of the mean of a multivariate normal distribution, *Ann. Stat., 9*, 1135-1151; Brandwein, A.C., and Strawderman, W.E., 1990. Stein estimation: the spherically symmetric case, *Statistical Sci., 5*, 356-369; Evans, S.N. and Stark, P.B., 1996. Shrinkage estimators, Skorokhod's problem, and stochastic integration by parts, *Ann. Stat., 24*, 809-815.

We have looked briefly at testing multiple hypotheses and at finding confidence intervals for multiple parameters. Here we consider estimating multiple parameters with squared error loss.

Let's recall a few results from earlier in the course: A minimax estimator is Bayes for the least favorable prior and Bayes estimators are admissible (if they are unique).

1.1 Minimaxity of the sample mean for estimating a normal mean

Let $X \sim N(\theta, \sigma^2)$, and let $\theta \in \Theta = \mathbf{R}$. Then X is admissible for $\theta \in \Theta$ under squared-error loss.

Before the proof, a few definitions and results.

The *Fisher Information* at θ is

$$\begin{aligned} I(\theta) &= E\left[\frac{\partial}{\partial\theta}\log p_\theta(X)\right]^2 \\ &= \int\left(\frac{p'_\theta}{p_\theta}\right)^2 p_\theta d\mu. \end{aligned} \tag{1}$$

The larger $I(\theta_0)$, the more easily θ_0 can be distinguished from neighboring values. The information for n iid measurements is n times the information for a single measurement (the factor p_θ occurs n times).

For $N(\theta, \sigma^2)$, $I(\theta) = \sigma^{-2}$.

Theorem 1 *The information inequality (see Lehmann, TPE, Theorem 6.4). Let θ be a real-valued parameter; let Θ be an open interval; suppose the distributions $\{P_\theta\}_{\theta\in\Theta}$ have common support; assume that $p'_\theta(x) = \partial/\partial\theta p_\theta(x)$ exists and is finite for all x in the support of P_θ and all $\theta \in \Theta$; and assume that*

$$E_\theta\left[\frac{\partial}{\partial\theta}\log p_\theta(X)\right] = 0. \tag{2}$$

Let δ be any statistic with $E_\theta\delta^2 < \infty$ for which the derivative w.r.t. θ of

$$E_\theta(\delta) = \int \delta p_\theta d\mu \tag{3}$$

exists and can be obtained by differentiating under the integral. Then

$$\mathbf{Var}_\theta(\delta) \geq \frac{\left[\frac{\partial}{\partial\theta}E_\theta(\delta)\right]^2}{I(\theta)}. \tag{4}$$

Theorem 2 *Let $\{X_j\}_{j=1}^n$ be iid $N(\theta, 1)$. The sample mean $\bar{X} = \frac{1}{n}\sum_{j=1}^n X_j$ is admissible and minimax under squared-error loss for θ.*

Proof. (See Lehmann, TPE, pp265-267.) For any estimator δ,

$$R(\theta, \delta) = E(\delta - \theta)^2 = \mathbf{Var}_\theta(\delta) + b^2(\theta) \tag{5}$$

where $b(\theta)$ is the bias of δ at θ ($b(\theta) = E\delta(X) - \theta$). By the information inequality,

$$R(\theta, \delta) \geq b^2(\theta) + \frac{[1 + b'(\theta)]^2}{nI(\theta)}. \tag{6}$$

Here $I(\theta) = 1$. For the sample mean, $R(\theta, \bar{X}) = 1/n$. If the risk of δ is at most $1/n$ for all θ, then

$$b^2(\theta) + \frac{[1 + b'(\theta)]^2}{n} \leq \frac{1}{n}, \ \forall \theta. \tag{7}$$

Now because $|b(\theta)| \leq n^{-1/2}$, b is bounded. For $n = 1$, we have

$$1 + 2b'(\theta) + b'^2(\theta) \leq 1, \tag{8}$$

so $b'(\theta) \leq 0$: b is nonincreasing. The boundedness of b and the fact that $b' \leq 0$ imply that $b' \to 0$ as $\theta \to \pm\infty$. This, together with the bound 7 implies that $b(\theta) \to 0$ as $\theta \to \pm\infty$; hence b is constant, and equal to zero. Thus for any δ,

$$R(\theta, \delta) \geq \frac{1}{n} \ \forall \theta. \tag{9}$$

Thus $R(\theta, \delta) = 1/n$, so $\bar{X}$ is admissible, and minimax.

1.2 Dimensions 3 and above: Stein's theorem

Charles Stein (1956) showed the following surprising result. Let $X \sim N(\theta, I)$, $\theta \in \Theta = \mathbf{R}^p$, $p \geq 3$. Not only is X not minimax for θ under squared-error loss, X is inadmissible for θ. Heuristically, in dimensions three and above, the "noise" that the normal adds to the mean is more likely than not to result in a "data vector" further from the origin (larger in norm) than the parameter vector is. "Shrinking" the data towards the origin by a small amount tends to give a more accurate estimator than simply reporting the data.

A proof can be based on Stein's integration-by-parts Lemma, which we saw earlier (fifth set of notes).

Let's take $X \sim N(\theta, I)$, $\theta \in \Theta = \mathbf{R}^p$, $p \geq 3$. We seek to estimate θ under squared-error loss. Define $\delta_a(x) \equiv (1 - a\|x\|^{-2})X$ (this is the James-Stein estimator; Stein's original result was for $\delta_{a,b}(x) = (1 - a/(b + \|x\|^2))X$). Note what $\delta_a(X)$ does: it estimates θ by a vector with the same direction as X, but that is shorter than X by a fraction a (the estimator "shrinks" X towards the origin). The estimate can have sign opposite that of X when $\|X\|$ is small; "positive part" estimators, of the form $(1 - a\|x\|^{-2})_+ X$, prevent the coordinates from crossing zero. Recall that in the case of estimating a bounded normal mean, we found that the optimal affine estimator "shrinks" towards the center of the interval θ is known to lie in. The surprise here is that even without any knowledge of where in $\mathbf{R}^p$ θ might lie, it helps to shrink towards the origin.

Theorem 3 *(James and Stein, 1961. See Brandwein and Strawderman, Thm. 4.1.) For $a \in (0, 2(p-2))$, $\delta_a(X)$ dominates X in squared-error loss. The estimator δ_{p-2} has uniformly smallest risk in the class.*

Proof. Let $Z \sim N(0, 1)$. The risk of X is

$$E\|X - \theta\|^2 = pEZ^2 = p. \tag{10}$$

Note that for $x \in \mathbf{R}^p$,

$$\frac{d}{dx_j}\|x\|^2 = 2x_j, \tag{11}$$

so

$$\begin{aligned} \frac{d}{dx_j}\|x\|^{-2} &= \frac{d}{dx_j}(\|x\|^2)^{-1} \\ &= -(\|x\|^2)^{-2} 2x_j \\ &= -2x_j\|x\|^{-4} \end{aligned} \tag{12}$$

Thus

$$R(\theta, \delta_a) = E\left\|(1 - a\|X\|^{-2})X - \theta\right\|^2$$

$$
\begin{aligned}
&= E\|X-\theta\|^2 + a^2 E\|X\|^{-2} - 2aE\left(X'(X-\theta)\|X\|^{-2}\right) \\
&= p + a^2 E\|X\|^{-2} - 2a\sum_{j=1}^{p} E\left(X_j(X_j-\theta_j)\|X\|^{-2}\right) \\
&= p + a^2 E\|X\|^{-2} - 2a\sum_{j=1}^{p} E\left(\frac{d}{dx_j}(X_j\|X\|^{-2})\right) \text{ (by Stein's lemma, componentwise)} \\
&= p + a^2 E\|X\|^{-2} - 2a\sum_{j=1}^{p} E\left(\|X\|^{-2} - 2X_j^2\|X\|^{-4}\right) \\
&= p + a^2 E\|X\|^{-2} - 2apE\|X\|^{-2} - 4aE\left(\|X\|^2\|X\|^{-4}\right) \\
&= p + a^2 E\|X\|^{-2} - 2apE\|X\|^{-2} - 4aE\|X\|^{-2} \\
&= p + (a^2 - 2a(p-2))E\|X\|^{-2}. \qquad (13)
\end{aligned}
$$

Note that $E\|X\|^{-2} > 0$. For $a \in (0, 2(p-2))$, $a^2 - 2a(p-2) < 0$, so for $a \in (0, p-2]$, δ_a dominates X for all $\theta \in \mathbf{R}^p$.

$$
\arg \min_{a \in (0,2(p-2))} \{a^2 - 2a(p-2)\} = p-2, \qquad (14)
$$

so δ_{p-2} has the uniformly smallest risk in this class. It can be shown further that $R(0, \delta_{p-2}) = 2$, $\forall p \geq 3$.

Note that there is nothing special about shrinking towards the origin: shrinking towards any other $\theta' \in \mathbf{R}^p$ by $a \in (0, p-2]$ also dominates X (take $\delta(x) = \theta' + (1 - (p-2)\|x - \theta'\|^{-2})(x - \theta')$).

Now suppose we observe instead $\{X_j\}_{j=1}^n$ iid $N(\theta, \sigma^2)$. We still seek to estimate $\theta \in \mathbf{R}$ under squared-error loss. The sample mean $\bar{X}$ is sufficient for θ, so this is risk-equivalent to estimating θ from $\bar{X} \sim N(\theta, \sigma^2/n)$; $\bar{X}$ is thus admissible and minimax.

1.3 Other Distributions

Stein's result that the sample mean is inadmissible in higher dimensions has been generalized in a variety of ways. The most general result so far is due to S.N. Evans and P.B. Stark (1996, Shrinkage estimators, Skorokhod's problem, and stochastic integration by parts, *Ann. Stat., 24*, 809-815.), who showed that in dimensions three and higher, the sample mean is inadmissible for the mean for all distributions that can be characterized as a stopping time

of Brownian motion. That condition is equivalent to saying that we observe $X = \theta + Z$ with Z satisfying $EZ = 0$, $EZ^2 < \infty$, and

$$E\|Z + \theta\|^{2-d} \leq \|\theta\|^{2-d}. \tag{15}$$

This includes all symmetric distributions, as well as many others (including some supported on fractal sets, for example). The condition 15, intuitively speaking, says that the "noise" Z on average makes $\|X\|^2 > \|\theta\|^2$ (not exactly—the exponent is not 2), which is when one would expect shrinking towards the origin to help. The proof is essentially a generalization of Stein's "integration by parts" lemma to stochastic integrals of Brownian motion.

Th amount of shrinkage that is optimal for other distributions depends on the distribution. The results in Evans and Stark are for the estimator

$$\delta(x) = (1 - a(1 + \|x\|^2)^{-1})x, \tag{16}$$

which is like Stein's original estimator with $b = 1$, but similar techniques work (with additional assumptions) for δ_a. They establish the existence of $a \in \mathbf{R}+$ sufficiently small that the shrinkage estimator dominates X; in the special cases that the support of Z is within a ball, or does not intersect a ball, they construct values of a that work.

Exercise. In the normal linear regression model, we observe $X \sim N(A \cdot \theta, \sigma^2 I)$, where A is a known $n \times p$ matrix with $A' \cdot A$ nonsingular, $\theta \in \mathbf{R}^p$ is unknown, $X \in \mathbf{R}^n$, and $\sigma^2 > 0$. The least-squares estimate of θ is

$$\hat{\theta}_{\mathrm{LS}} = (A' \cdot A)^{-1} A' X. \tag{17}$$

Consider estimating θ subject to squared-error "prediction" loss

$$L(\hat{\theta}, \theta) = \|A \cdot \hat{\theta} - A \cdot \theta\|^2. \tag{18}$$

Show that $\hat{\theta}_{\mathrm{LS}}$ is inadmissible for θ.