Statistics 210B, Spring 1998

Class Notes

P.B. Stark

stark@stat.berkeley.edu

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

March 19, 1998

Fifth Set of Notes

1 More on the Bounded Normal Mean

Lemma 1 *Stein's Lemma. (See Evans and Stark, 1996.* Ann. Stat., 24*, 809-815, for a substantial generalization.) Suppose* $X \sim N(\theta, 1)$*, and that* $\delta(\cdot)$ *is differentiable, with* $E_\theta|\delta'(X)| < \infty$, $\lim_{x\to\pm\infty} \delta(x)\exp\{-(x-\theta)^2/2\} = 0$*, and that* $E_\theta[\delta(X)(X-\theta)]$ *is finite. Then*

$$E_\theta[\delta(X)(X-\theta)] = E_\theta\delta'(X). \tag{1}$$

Proof. Let $\phi(x)$ be the standard normal density. Integrate by parts:

$$\begin{aligned} E_\theta[\delta(X)(X-\theta)] &= \int_{-\infty}^{\infty} \delta(x)(x-\theta)\phi(x-\theta)dx \\ &= \int_{-\infty}^{\infty} \delta(x)(x-\theta)\phi(x)dx \\ &= -\delta(x)\phi(x)|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \delta'(x)\phi(x)dx \\ &= \int_{-\infty}^{\infty} \delta'(x)\phi(x)dx \\ &= E_\theta\delta'(X). \end{aligned} \tag{2}$$

Lemma 2 *Consider estimating the mean θ of a normal distribution with unit variance using an estimator δ that satisfies the conditions of Stein's Lemma. For squared-error loss,*

$$R(\theta, \delta) = 1 - E_\theta(2\psi'(X) - \psi^2(X)), \tag{3}$$

where $\psi(x) = x - \delta(x)$.

Proof.

$$\begin{aligned} R(\theta, \delta) &= E_\theta(\theta - \delta(X))^2 \\ &= E_\theta\left(\theta - X + X - \delta(X)\right)^2 \\ &= E_\theta\left(-(X-\theta) + \psi(X)\right)^2 \\ &= E_\theta\left((X-\theta)^2 - 2(X-\theta)\psi(X) + \psi(X)^2\right) \\ &= 1 - E_\theta\left(2\psi'(X) - \psi^2(X)\right), \end{aligned} \tag{4}$$

using the lemma in the last step.

Bickel (1981, *Ann. Stat.*, *9*, 1301-1309) studies the minimax problem as the bound goes to infinity and finds the asymptotic form of the least-favorable prior. We have a single observation $X \sim N(\theta, 1)$, with $\theta \in \Theta = [-m, m]$. The loss function $\ell(\theta, a) = |\theta - a|^2$. The action space is $\mathcal{A} = \mathbf{R}$ (which we might as well limit to $[-m, m]$ if we can). The "natural" estimator is $\delta^0(x) = x$, and the maximum likelihood estimator is the truncation estimator:

$$\delta_{MLE}(x) = \begin{cases} m, & x \geq m \\ x, & |x| < m, \\ -m, x \leq -m \end{cases} \tag{5}$$

Clearly, the risk of δ^0 is unity for all θ, the maximum risk of the MLE is for $\theta = 0$, and the minimum risk is for $\theta = \pm m$.

Lemma 3 *(Bickel, 1981; special case of Brown, 1971) Suppose $X \sim N(\theta, 1)$, $\theta \in \Theta$, $\theta \sim \pi$. Let f_π be the density of the marginal distribution of X:*

$$f_\pi(x) = \phi \star \pi(x) = \int_{-\infty}^{\infty} \phi(x - \theta)\pi(d\theta). \tag{6}$$

The Bayes risk for π for squared-error loss is

$$r(\pi) = 1 - \int_{-\infty}^{\infty} \frac{(f'_\pi(x))^2}{f_\pi(x)} dx. \tag{7}$$

Proof. The Bayes estimator for prior π is the posterior mean of θ given x. The derivative of $f_\pi(x)$ with respect to x is

$$\begin{aligned} \frac{d}{dx} f_\pi(x) &= \int_{-\infty}^{\infty} \frac{d}{dx}\phi(x-\theta)\pi(d\theta) \\ &= \int_{-\infty}^{\infty} (-(x-\theta))\phi(x-\theta)\pi(d\theta) \\ &= -x f_\pi(x) + \int_{-\infty}^{\infty} \theta\phi(x-\theta)\pi(d\theta) \\ &= (-x + E(\theta|x)) f_\pi(x), \end{aligned} \tag{8}$$

so the posterior mean of θ given x is

$$\delta(x) = \delta_\pi(x) = x + f_\pi'(x)/f_\pi(x). \tag{9}$$

Applying the previous lemma, with $\psi(x) = x - \delta(x) = -f_\pi'(x)/f_\pi(x)$ gives the risk at θ of this estimator for squared-error loss to be

$$\begin{aligned} 1 - E_\theta\left(2\psi'(X) - \psi^2(X)\right) &= 1 - E_\theta\left(-2f_\pi''(X)/f_\pi(X) + 2(f_\pi'(X))^2/f_\pi^2(X) - (f_\pi'(X))^2/f_\pi^2(X)\right) \\ &= 1 - E_\theta\left(f_\pi'(X))^2/f_\pi^2(X) - 2f_\pi''(X)/f_\pi(X)\right). \end{aligned} \tag{10}$$

Taking the expectation with respect to π to find the Bayes risk, and using Fubini's Theorem yields

$$\begin{aligned} r(\pi) = E_\pi R(\theta, \delta_\pi) &= 1 - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{(f_\pi'(x))^2 - 2f_\pi''(x)f_\pi(x)}{f_\pi^2(x)}\phi(x-\theta)\pi(d\theta)dx \\ &= 1 - \int_{-\infty}^{\infty} \frac{(f_\pi'(x))^2 - 2f_\pi''(x)f_\pi(x)}{f_\pi^2(x)} f_\pi(x)dx \\ &= 1 - \int_{-\infty}^{\infty} \frac{(f_\pi'(x))^2}{f_\pi(x)}dx + 2f_\pi'(x)\big|_{x=-\infty}^{\infty}. \end{aligned} \tag{11}$$

Because f_π is the result of convolution with a Gaussian density, its derivatives of all orders vanish as $x \to \pm\infty$, so we have

$$r(\pi) = 1 - \int_{-\infty}^{\infty} \frac{(f_\pi'(x))^2}{f_\pi(x)}dx. \tag{12}$$

Note that for a distribution F with absolutely continuous density f, the Fisher information is

$$I(F) = \int_{-\infty}^{\infty} \frac{(f'(x))^2}{f(x)}dx, \tag{13}$$

so the equality just established is $r(\pi) = 1 - I(\Phi \star \pi)$, where Φ is the standard normal distribution. Thus we have a relation between the Bayes risk for a given prior on θ and the Fisher information of the marginal distribution of the observation X for that prior. The least favorable prior is that for which the Fisher information of $\Phi \star \pi$ is minimal.

Let $\rho(m)$ be the minimax risk for squared-error loss with $\Theta = [-m, m]$, and let $r(\pi)$ be the Bayes risk for squared-error loss using prior π on θ. Bickel (1981) uses the relation between the Bayes risk and the Fisher information, and properties of the Fisher information, to show that if π_1 is the distribution on $[-1, 1]$ with density

$$g_1(s) = \begin{cases} \cos^2(s\pi/2), & |s| \le 1 \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

and π_m is the distribution on $[-m, m]$ with density

$$g_m(s) = m^{-1} g_1(s/m), \tag{15}$$

then the priors $\{\pi_m\}$ are approximately least favorable.

Theorem 1 *(Bickel, 1981, Theorem 2.1). As $m \to \infty$,*

$$\rho(m) = r(\pi_m) + o(m^{-2}), \tag{16}$$

and

$$r(\pi_m) = 1 - \frac{\pi^2}{m^2} + o(m^{-2}). \tag{17}$$

Let π_m^0 be the least favorable prior when $\Theta = [-m, m]$, and let $\pi_1^{(m)}$ be the distribution obtained by scaling π_m^0 to $[-1, 1]$: $\pi_1^{(m)}(s) = \pi_m^0(ms)$. Then $\pi_1^{(m)}$ converges weakly to π_1.

Perhaps surprisingly, the Bayes estimators $\delta_{\pi_m}(x)$ are not asymptotically minimax (indeed, $\limsup_m R(m, \delta_{\pi_m}) > 1$, so it does not dominate the naive estimator $\delta(x) = x$). Bickel (1981) also shows how to modify the estimator to be asymptotically minimax to order m^{-2}.

Casella and Strawderman (1981, *Ann. Stat.*, *9*, 870-878) and Gatsonis, MacGibbon and Strawderman (1987, *Stat. Prob. Lett.*, *6*, 21-30) address the estimation of a bounded normal mean, using squared-error loss. The former paper looks at Bayes estimators for 2-point and

3-point priors, and shows that they are minimax when the bound on the mean is small; the latter shows that a uniform prior performs surprisingly well in a minimax sense.

Let π_m^0 put mass 1/2 on $\pm m$. Because the loss is squared-error, the Bayes estimator δ_m^0 against that prior is the posterior mean of θ given x, which we can calculate. The marginal density of X is

$$f_X(x) = \frac{1}{2\sqrt{2\pi}} \left[e^{-(x-m)^2/2} + e^{-(x+m)^2/2} \right]. \tag{18}$$

Proceeding blithely without concern for rigor, the posterior "density" of θ given $X = x$ is

$$f_{\theta|X=x}(\theta) = \frac{\delta_{\theta-m} e^{-(x-m)^2/2} + \delta_{\theta+m} e^{-(x-m)^2/2}}{e^{-(x-m)^2/2} + e^{-(x+m)^2/2}}, \tag{19}$$

where δ_x is the Dirac delta measure (a point-mass at $x = 0$). The posterior mean is

$$\begin{aligned} \delta_m^0(x) &= E(\theta|X = x) \\ &= \left[\frac{me^{-(x-m)^2/2} - me^{-(x-m)^2/2}}{e^{-(x-m)^2/2} + e^{-(x+m)^2/2}} \right] \\ &= m\frac{e^{mx} - e^{-mx}}{e^{mx} + e^{-mx}} \\ &= m \tanh(mx). \end{aligned} \tag{20}$$

Similarly, let π_m^α put mass α at zero, and mass $(1-\alpha)/2$ at $\pm m$. The Bayes estimator for that prior is

$$\delta_m^\alpha(x) = \frac{(1-\alpha) m \tanh(mx)}{1 - \alpha + \alpha \exp(m^2/2) \mathrm{sech}(mx)}. \tag{21}$$

Finally, let π_m be the uniform prior on $[-m, m]$. Let's find the corresponding Bayes estimator. The conditional distribution of X given θ is as before, and the marginal density of X is

$$\frac{1}{2m} \int_{\theta=-m}^{m} \phi(x-\theta) d\theta = \frac{1}{2m} (\Phi(x-m) - \Phi(x+m)), \tag{22}$$

where $\phi(\cdot)$ is the standard normal density and $\Phi(\cdot)$ is the standard normal cdf. The posterior density of θ is

$$f(\theta|x) = 1_{|\theta| \le m} \frac{\phi(x-\theta)}{\Phi(x-m) - \Phi(x+m)}. \tag{23}$$

The Bayes estimator $\delta_m(x)$ is the posterior mean, namely

$$\frac{\int_{-m}^{m} \theta \phi(\theta - x) d\theta}{\Phi(x-m) - \Phi(x+m)}. \tag{24}$$

Let's work on the numerator:

$$\int_{-m}^{m} \theta\phi(\theta - x)d\theta = \int_{-m}^{m} (\theta - x)\phi(\theta - x)d\theta + \int_{-m}^{m} x\phi(\theta - x)d\theta$$
$$= e^{-(x-m)^2/2} - e^{-(x+m)^2/2} + x(\Phi(x-m) - \Phi(x+m)). \quad (25)$$

Thus 24 is

$$\delta_m(x) = x + \frac{e^{-(x-m)^2/2} - e^{-(x+m)^2/2}}{\Phi(x-m) - \Phi(x+m)}. \quad (26)$$

One use the results following Stein's lemma to calculate the risks differently.

Lemma 4 *(Casella and Strawderman, 1981, Lemma 3.1) The Bayes estimator δ_m^0 has maximum risk*

$$\max_{\theta \in [-m,m]} R(\theta, \delta_m^0) = \max\left(R(0, \delta_m^0), R(m, \delta_m^0)\right). \quad (27)$$

Proof. Let $\delta(x) = \delta_m^0(x)$.

$$\delta'(x) = (d/dx)\delta(x) = m(d/dx)\tanh(mx)$$
$$= m^2 - m^2\tanh^2(mx)$$
$$= m^2 - \delta^2(x), \quad (28)$$

and

$$\delta''(x) = (d^2/dx^2)\delta(x) = -2\delta(x)\delta'(x). \quad (29)$$

Note that in general, if $\delta(x)$ is differentiable (any other conditions needed?) and θ is a location parameter,

$$\lim_{a\to 0} \frac{E_{\theta+a}\delta(X) - E_\theta\delta(X)}{a} = \lim_{a\to 0} \frac{E_\theta\delta(X+a) - E_\theta\delta(X)}{a}$$
$$= \lim_{a\to 0} \frac{E_\theta(\delta(X+a) - \delta(X))}{a}$$
$$= E_\theta \lim_{a\to 0} \frac{\delta(X+a) - \delta(X)}{a}$$
$$= E_\theta\delta'(X). \quad (30)$$

The risk of δ at θ is

$$E_\theta|\theta - \delta(X)|^2 = \theta^2 - 2\theta E_\theta\delta(X) + E_\theta\delta^2(X). \quad (31)$$

The derivative of the risk w.r.t. θ is (applying 30)

$$\begin{aligned} d/d\theta R(\theta, \delta_m^0) &= 2\theta - 2E_\theta\delta(X) - 2\theta E_\theta\delta'(X) + 2E_\theta\delta(X)\delta'(X) \\ &= 2E_\theta\left[(\theta - \delta(X))(1 - \delta'(X))\right] \\ &= 2E_\theta\left[((\theta - X) + (X - \delta(X)))(1 - \delta'(X))\right] \\ &= E_\theta\left[(X - \delta(X)) - \delta'(X)(X + \delta(X))\right]. \end{aligned} \tag{32}$$

Casella and Strawderman use Karlin's change of sign lemma to show that this can have at most three sign changes; recall that E_θ is a variation-diminishing transformation for $N(\theta, 1)$, so the result follows if the argument of the expectation has at most three sign changes. The argument is

$$(x - \delta(x)) - \delta'(x)(x + \delta(x)), \tag{33}$$

which vanishes at $x = 0$. Its other zeros solve

$$x - \delta(x) = \delta'(x)(x + \delta(x)), \tag{34}$$

or equivalently (for $x \neq 0$)

$$1 - \delta(x)/x = \delta'(x)(1 + \delta(x)/x). \tag{35}$$

For $x > 0$, $\delta(x)/x$ and $\delta'(x)$ are positive and decreasing. Thus for $x > 0$, $(1 - \delta(x)/x)$ is increasing, and $\delta'(x)(1+\delta(x)/x)$ is decreasing, so 35 has at most one solution for $x > 0$. Note that $\delta'(x)$ is an even function, and $\delta(x)$ and x are odd functions, so $(x-\delta(x))-\delta'(x)(x+\delta(x))$ is odd, and it has at most one zero for $x < 0$, and thus at most three zeros counting the one at $x = 0$. Both $\delta(x)$ and $\delta'(x)$ are bounded, so the argument is negative for $x \to -\infty$ and positive for $x \to +\infty$; hence, the sign sequence of the argument is $-+-+$. By the change of sign lemma, the expectation also has at most three sign changes, in the same order. Again because the argument is odd,

$$E_0\left[(X - \delta(X)) - \delta'(X)(X + \delta(X))\right] = 0. \tag{36}$$

Thus the risk is stationary at $\theta = 0$. The derivative of the risk at $\theta > 0$ is the negative of the derivative of the risk at $-\theta$ (as you can see from symmetry). Thus a local extremum of the risk for some $\theta > 0$ must be a minimum, and hence the maximum risk is either at $\theta = 0$ or at $\theta = \pm m$. This proves the lemma.

If we can establish that the Bayes risk of the estimator is equal to its maximum risk, the estimator is minimax.

Lemma 5 *Casella and Strawderman, 1981, Lemma 3.2. The function*

$$f(m) = R(0, \delta_m^0) - R(m, \delta_m^0) \tag{37}$$

has only one sign change as m goes from 0 to ∞. The sign change is from negative to positive, so there is a unique $m_0 \in \mathbf{R}$ s.t. $f(m) \leq 0$, $\forall m \leq m_0$.

Proof. The difference in risks is

$$\begin{aligned} f(m) &= E_0(\delta(X) - 0)^2 - E_m(\delta(X) - m)^2 \\ &= E_0(m \tanh(mX))^2 - E_m(m \tanh(mx) - m)^2 \\ &= m^2 \left[E_0 \tanh^2(mX) - E_m(1 - \tanh(mx))^2 \right] \\ &\equiv m^2 g(m). \end{aligned} \tag{38}$$

Differentiating $g(m)$ gives

$$\frac{d}{dm} g(m) = 2E_0(X \tanh(mX)\mathrm{sech}^2(mX)) + 2E_m((X+m)(1-\tanh(mX))\mathrm{sech}^2(mX)). \tag{39}$$

The second expectation is of an argument with but one sign change, from negative to positive, so if its expectation is positive at $\theta = 0$, it is positive for $\theta = m \geq 0$ (the distribution of the argument will be stochastically larger). Define $Q = X \tanh(mX)\mathrm{sech}^2(mX)$. We have

$$\begin{aligned} \frac{d}{dm} g(m) &\geq 2E_0(X \tanh(mX)\mathrm{sech}^2(mX)) + 2E_0((X + m)(1 - \tanh(mX))\mathrm{sech}^2(mX)) \\ &= 2E_0 Q + 2E_0 \left((X + m)\mathrm{sech}^2(mX) - Q \right) \\ &= 2E_0 \left((X + m)\mathrm{sech}^2(mX) \right) \\ &= 2mE_0 \mathrm{sech}^2(mX) + 2E_0 X \mathrm{sech}^2(mX) \\ &= 2mE_0 \mathrm{sech}^2(mX) \\ &\geq 0 \text{ for } m \geq 0. \end{aligned} \tag{40}$$

(The penultimate step uses the fact that $X\text{sech}^2(mX)$ is an odd function.) This proves the lemma, and takes us to one of the main theorems of Casella and Strawderman:

Theorem 2 *Casella and Strawderman, 1981, Theorem 3.1. If* $X \sim N(\theta, 1)$, $\theta \in \Theta = [-m, m]$, $0 \leq m \leq m_0$, *then* $\delta_m^0(x) = m \tanh(mx)$ *is minimax for squared-error loss, and* τ_m^0 *is a least-favorable prior.*

Proof. The two lemmas show that for $m \leq m_0$, $\max_{\theta \in \Theta} R(\theta, \delta_m^0) = R(m, \delta_m^0)$. But for the prior π_m^0 that assigns mass 1/2 to $\pm m$, the Bayes risk is

$$r(\pi_m^0, \delta_m^0) = \frac{1}{2} R(-m, \delta_m^0) + \frac{1}{2} R(m, \delta_m^0) = R(m, \delta_m^0). \tag{41}$$

We have already seen that if the Bayes risk of the Bayes estimator for a given prior equals the maximum risk of the Bayes estimator over the parameter set, the Bayes estimator is minimax. We are done.

Casella and Strawderman go on to show that π_m^0 is not least favorable for $m > m_0 \approx 1.05$. As $m \uparrow$, the least favorable prior concentrates at a larger and larger number of discrete points in $[-m, m]$. The three-point priors π_m^α are minimax for some range of values of m, including at least about $m \in [1.4, 1.6]$.