2.7 Stein's phenomenon and James-Stein estimators. Let $|y| := (y_1^2 + \cdots y_d^2)^{1/2}$ for $y \in \mathbb{R}^d$. Consider the normal location family $N(\mu, I)$, $\mu \in \mathbb{R}^d$, on $\mathbb{R}^d$, having density $(2\pi)^{-d/2}\exp(-|x-\mu|^2/2)$ with respect to Lebesgue measure dx, where I is the $d \times d$ identity matrix. The problem is to estimate the unknown μ from an observation x. Here for simplicity n is taken equal to 1. If we had n i.i.d. observations $X_1, \ldots, X_n$, then $\overline{X} := (X_1 + \cdots X_n)/n$ is a minimal, Lehmann-Scheffé sufficient statistic having distribution $N(\mu, I/n)$, and $\sqrt{n}\overline{X}$ is such a statistic with distribution $N(\sqrt{n}\mu, I)$, so the situation would not be essentially different.

The observation x is an unbiased estimator of μ, in other words $Ex_i = \mu_i$ for $i = 1, \ldots, d$. The information inequality holds in this case and gives for each i, $E_{\mu_i}(T_i - \mu_i)^2 \geq 1$ for each unbiased estimator T_i of μ_i. Thus $E_\mu(|T-\mu|^2) \geq d$ for each unbiased estimator T of μ. This lower bound is attained by $T = x$.

It turns out, however, that for $d \geq 3$, $T = x$ is an inadmissible estimator of μ for squared-error loss. This fact is called "Stein's phenomenon," after Charles Stein, who discovered it. Let

$$J(x) := \left(1 - \frac{d-2}{|x|^2}\right)x,$$

called a *James-Stein* estimator of μ. Then for $d \geq 3$, $r(\mu, J) < r(\mu, x)$ for all μ, as will be proved for $d = 3$. It's very surprising that although the coordinates $x_1, \ldots, x_d$ are independent for any μ, the x_j for $j \neq i$ are useful in estimating μ_i. For $d = 2$, simply $J(x) \equiv x$. For $d = 1$, J would be a bad estimator with infinite risk for squared-error loss because of a singularity at $x = 0$. Note however that for $d \geq 3$, due to the factor $|x|^{d-1}$ in the volume element in spherical coordinates, $|x/|x|^2| = |x|^{-1}$ and $|x/|x|^2|^2 = |x|^{-2}$ are integrable for any normal law despite being unbounded near 0.

The estimator J is not admissible either; $[\max(0, 1-(d-2)|x|^{-2})]x$ is a better estimator, but it is still not admissible (see the notes). Here, it will just be shown that Stein's phenomenon occurs for $d = 3$ with the James-Stein estimator J.

2.7.1 Proposition. For $d = 3$ and the estimators $J(x) := (1 - |x|^{-2})x$ and x for the mean μ in the normal location family $\{N(\mu, I) : \mu \in \mathbb{R}^3\}$, we have $r(\mu, J) < r(\mu, x)$ for all μ. Thus x is an inadmissible estimator of μ.

Proof. Clearly $r(\mu, x) = 3$ for all μ. It will be enough to show that for all μ,

$$f(\mu) := r(\mu, J) = g(\mu) := 3 - E_\mu(|x|^{-2}). \tag{2.7.2}$$

It seems to be difficult to prove this directly, so it will be done by an indirect method as follows. First, $f(\mu) = f_1(|\mu|^2)$ for some function f_1, since for any orthogonal transformation (3×3 orthogonal matrix) U from $\mathbb{R}^3$ to $\mathbb{R}^3$, $J(Ux) \equiv UJ(x)$ and $N(\mu, I) \circ U^{-1} = N(U\mu, I)$, so by the image measure theorem (RAP, 4.1.11),

$$r(U\mu, J) = E_{U\mu}|J - U\mu|^2 = \int |J(x) - U\mu|^2 dN(U\mu, I)(x)$$

$$= \int |J(x) - U\mu|^2 d[N(\mu, I) \circ U^{-1}](x) = \int |J(Uy) - U\mu|^2 dN(\mu, I)(y)$$

$$= \int |U(J(y)-\mu)|^2 dN(\mu,I)(y) \;=\; \int |J(y)-\mu|^2 dN(\mu,I)(y) \;=\; r(\mu,J).$$

For any μ and μ' in $\mathbb{R}^3$ with $|\mu| = |\mu'|$, there is an orthogonal U with $U\mu = \mu'$, so indeed $f(\mu)$ is a function, say f_1, of $|\mu|^2$. It is also easily seen that f and f_1 are continuous, where integrals can be bounded using spherical coordinates as mentioned above.

Next, $g(\mu) \;\equiv\; g_1(|\mu|^2)$ for some function g_1, by a similar but shorter sequence of equations, where g and g_1 are also continuous. Specifically,

$$E_{U\mu}(|x|^{-2}) \;=\; (2\pi)^{-d/2} \int |x|^{-2} \exp(-|x-U\mu|^2/2)dx$$

$$= \; (2\pi)^{-d/2} \int |y|^{-2} \exp(-|y-\mu|^2/2)dy \;=\; E_\mu(|x|^{-2}).$$

For $A > 0$ let E_A denote the expectation of functions with respect to the $N(0, AI)$ distribution. It will be shown that for all $A > 0$,

$$E_A f \;=\; 3 - \frac{1}{A+1} \;=\; E_A g. \tag{2.7.3}$$

For μ with law $N(0, AI)$ and, given μ, x having law $N(\mu, I)$, the pair (x, μ) have a jointly normal distribution on $\mathbb{R}^6$, where the three 2-vectors (x_i, μ_i) for $i = 1, 2, 3$ are i.i.d. Let $E_{(A)}$ be expectation for this joint distribution. We have $E_{(A)} x_i = 0$ and $E(x_i^2|\mu_i) \;=\; \mu_i^2 + 1$, so $E_{(A)}(x_i^2) \;=\; A + 1$, while by Lemma 2.1.1

$$E_{(A)}(x_i\mu_i) \;=\; E_A E(x_i\mu_i|\mu_i) \;=\; E_A[\mu_i E(x_i|\mu_i)] \;=\; E_A\mu_i^2 \;=\; A.$$

Thus each (x_i, μ_i) has the bivariate normal law

$$N\left(0, \begin{pmatrix} A+1 & A \\ A & A \end{pmatrix}\right),$$

where 0 is the two-dimensional 0 vector. Now, we have

$$E_A g \;=\; 3 - E_A E_\mu |x|^{-2} \;=\; 3 - E_{(A)}|x|^{-2}. \tag{2.7.4}$$

If y has law $N(0, I)$ on $\mathbb{R}^3$, then by spherical coordinates,

$$E|y|^{-2} \;=\; (2\pi)^{-3/2} 4\pi \int_0^\infty \exp(-r^2/2)dr \;=\; 1.$$

For x with law $N(0, cI)$, $x/c^{1/2}$ has law $N(0, I)$, so

$$E(|x|^{-2}) \;=\; E\left(\frac{1}{c}\left|\frac{x}{c^{1/2}}\right|^{-2}\right) \;=\; \frac{1}{c}.$$

Thus in (2.7.4), $E_{(A)}(|x|^{-2}) \;=\; 1/(A+1)$, and (2.7.3) holds for g.

We have for each i, $E_{(A)}(\mu_i|x_i) = bx_i$ where, using Lemma 2.1.1 again,

$$A = E_{(A)}(\mu_i x_i) = E_{(A)}[E_{(A)}(\mu_i x_i|x_i)] = E_{(A)}(x_i b x_i) = b(A+1)$$

gives $b = b_A := A/(A+1)$. Thus given x_i, μ_i has law $N(b_A x_i, \tau_A^2)$ where $E_A\mu_i^2 = A = \tau_A^2 + b_A^2(A+1)$ implies $\tau_A^2 = A/(A+1)$. In other words given x, $\mu = b_A x + \zeta$ where ζ is independent of x and has law $N(0, AI/(A+1))$. Thus for conditional expectations given x we have

$$E_{(A)}\{|J(x) - \mu|^2|x\} = E_{(A)}\{|(1-|x|^{-2})x - \frac{A}{A+1}x - \zeta|^2|x\}$$

$$= \frac{3A}{A+1} + \left[1 - |x|^{-2} - \frac{A}{A+1}\right]^2 |x|^2,$$

so for the unconditional expectation,

$$E_{(A)}(|J(x) - \mu|^2) = \frac{3A}{A+1} + (A+1)^{-2}E_{(A)}|x|^2 - \frac{2}{A+1} + E_{(A)}|x|^{-2}$$

$$= \frac{3A}{A+1} + \frac{3}{A+1} - \frac{2}{A+1} + \frac{1}{A+1} = 3 - \frac{1}{A+1},$$

proving (2.7.3) for f, and so finishing its proof.

Now, for the family of laws $N(0, AI)$, $A > 0$, for μ in $\mathbb{R}^3$, $|x|^2$ is a Lehmann-Scheffé sufficient statistic from the exponential form of the density (Theorem 2.5.10). By the Lehmann-Scheffé property it follows that $f_1 = g_1$ almost everywhere and (2.7.2) follows, completing the proof. □

If we let μ have a prior distribution $N(0, AI)$, it follows from the above proof that bx is a Bayes estimator of μ for $b = A/(A+1)$, since for squared-error loss the Bayes estimator is the mean of the posterior distribution (by Proposition 2.6.1). So, up to equality almost surely, bx is the unique Bayes estimator. Thus for $0 < b < 1$, bx is admissible by Theorem 1.2.5. Such an estimator, however, has a large bias and large risk when $|\mu|$ is large. Letting $b \uparrow 1$ we see that the inadmissible estimator x is a limit of admissible estimators bx. It is much harder to give admissible estimators of μ which are *better than* x for $d \geq 3$ (see the Notes).

The estimator x, or $\overline{X}$ for any n, for μ is admissible for $d = 1$ (Lehmann, 1991, pp. 265-267 gives two proofs), for squared-error loss and many other loss functions. Stein (1956) showed that $\overline{X}$ is admissible for $d = 2$.

Recalling the notion of minimax decision rule, as defined in Sec. 1.2, the estimator x, or $\overline{X}$ for any n, is a minimax estimator of μ for any dimension d. To see this one can use again the fact that bx for $0 < b < 1$ is admissible. We have $r(\mu, bx) = db^2 + (1-b)^2|\mu|^2$, which is minimized with respect to μ when $\mu = 0$ (or $b = 1$), with $r(0, bx) = db^2$. Letting $b \uparrow 1$ ($A \to +\infty$) we see that the minimax risk is d, which is the risk of x for all μ. The supremum over μ of the risk for a James-Stein estimator, or any other estimator better than x, is also d. For the James-Stein estimator one can see that the risk $r(\mu, J)$ approaches d as $|\mu| \to \infty$. On the other hand the estimators bx for $0 < b < 1$ are not minimax, in fact $\sup_\mu r(\mu, bx) = +\infty$. For $d \geq 3$, there is a large class of minimax estimators (Baranchik, 1970; Lehmann, 1991, Theorem 4.6.3).

PROBLEMS

1. Show that for $N(\mu, I)$ on $\mathbb{R}^d$, the estimator bx for μ is inadmissible if $b < 0$ or $b > 1$.
2. Show that for a fixed vector $v \neq 0$ in $\mathbb{R}^d$, to estimate μ in $N(\mu, I)$, an estimator $bx + v$ is
   (a) never admissible for $b = 1$,
   (b) always admissible for $0 < b < 1$. *Hint:* $bx + v = b(x - w) + w$ where $w = v/(1 - b)$. Consider a prior $N(w, AI)$ for suitable A.
3. For normal distributions $N(\mu, I)$ on $\mathbb{R}^d$, $\mu \in \mathbb{R}^d$, if μ has a prior distribution $N(0, I)$ and X_1, X_2, and X_3 are observed, assumed to be i.i.d. $N(\mu, I)$, find the posterior distribution of μ.

NOTES

This section is based on the exposition in Lehmann (1991, Secs. 4.5 and 4.6). Stein (1956) discovered his phenomenon and James and Stein (1961) gave their estimator. Strawderman (1971) gave a rather complicated estimator better than x, i.e. minimax estimator, which is admissible for $d \geq 6$, also mentioned by Lehmann (1991, p. 304).

REFERENCES

Baranchik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Statist.* **41**, 642-645.

James, W., and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1**, 311-319.

Lehmann, Erich L. (1991). *Theory of Point Estimation*, 2d ed. Wadsworth, Pacific Grove, CA.

Stein, Charles (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1**, 197-206.

Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42**, 385-388.