Exercise 2.23 for another striking example).

Example 2.4.6 For reasons explained in Note 2.8.2, we consider the following estimator

$$\delta_2(x) = \begin{cases} \left(1 - \dfrac{2p-1}{||x||^2}\right) x & \text{if } ||x||^2 \geq 2p-1, \\ 0 & \text{otherwise,} \end{cases}$$

to estimate θ when $x \sim \mathcal{N}_p(\theta, I_p)$. This estimator, called the *positive-part James–Stein estimator*, is evaluated under *quadratic loss*,

$$\mathrm{L}(\theta, d) = ||\theta - d||^2.$$

2.4.1

Figure 2.4.2 gives a comparison of the respective risks of δ_2 and $\delta_1(x) = x$, maximum likelihood estimator, for $p = 10$. This figure shows that δ_2 cannot be minimax, since the maximum risk of δ_2 is above the (constant) risk of δ_1, that is, $R(\theta, \delta_2) = \mathbb{E}_\theta[||\theta - \delta_2(x)||^2] = p$. (We show in Section 2.4.3 that δ_1 is actually minimax in this case.) But the estimator δ_2 is definitely superior on the most interesting part of the parameter space, the additional loss being in perspective quite negligible. ||

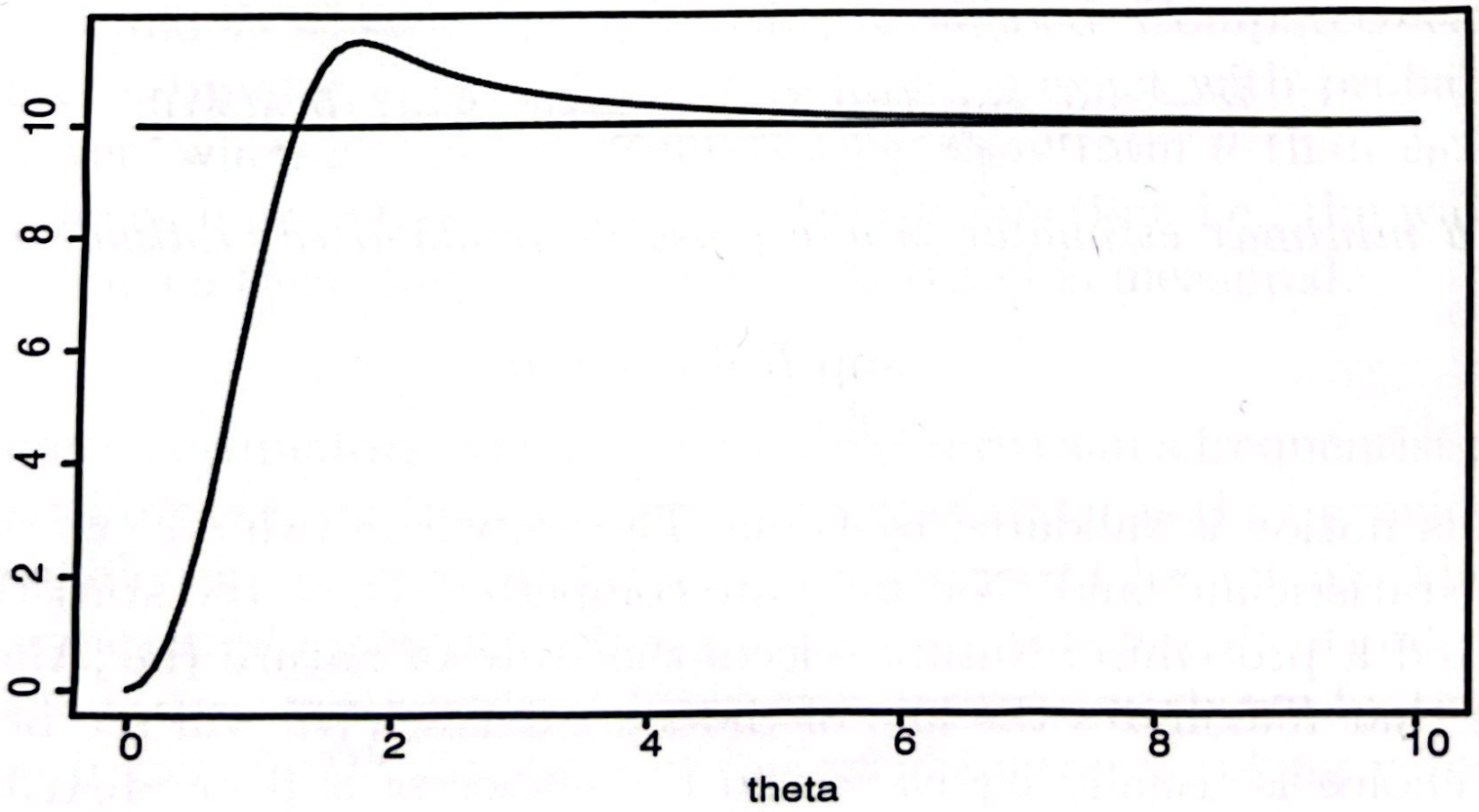

Figure 2.4.1. *Comparison of the risks of the estimators* δ_1 *and* δ_2.

2.8.2 *The Stein effect*

If there is a unique minimax estimator, this estimator is admissible, according to Proposition 2.4.20. Conversely, if a minimax estimator δ_0 is inadmissible, there are other minimax estimators that improve upon δ_0 (under some minor regularity conditions, see Brown (1976)). In particular, if the constant risk minimax estimator is inadmissible, this is the worst minimax estimator in the sense that every other minimax estimator has a uniformly smaller risk. Until 1955, it was assumed that the least-squares estimator, $\delta_0(x) = x$, when $x \sim \mathcal{N}_p(\theta, I_p)$, was admissible and, since its risk is constant, that it was the unique minimax estimator. Stein (1955a) showed that this result only holds for $p = 1, 2$ and hence discovered *"the Stein effect"* phenomenon, that is, the exhibition of apparently paradoxical domination results for usual estimators.

Formally, the Stein paradox is as follows. If a standard estimator $\delta^*(x) = (\delta_0(x_1), \ldots, \delta_0(x_p))$ is evaluated under weighted quadratic loss

$$\sum_{i=1}^{p} \omega_i (\delta_i - \theta_i)^2, \tag{2.8.2}$$

with $\omega_i > 0$ $(i = 1, \ldots, p)$, there exists p_0 such that δ^* is not admissible for $p \geq p_0$, *although the components* $\delta_0(x_i)$ *are separately admissible to estimate*

the θ_i*'s*. The Stein effect can be explained through the use of the joint loss (2.8.2), that allows the dominating estimator to borrow strength from the other components, even when they are independent and deal with totally different estimation problems. The literature on the Stein effect and related phenomena is now too extensive for us to give here a comprehensive covering of the results in this field. We refer the reader to Judge and Bock (1978), Lehmann (1983), and Berger (1985a) for a more detailed bibliography and we develop in Chapter 10 a Bayesian analysis of the Stein effect. This note briefly presents the main results about the Stein effect from a frequentist point of view.

First, while Stein's (1955a) proof of inadmissibility was nonconstructive, James and Stein (1961) exhibited an estimator that uniformly dominates $\delta_0(x) = x$ under quadratic loss for $p \geq 3$ in the normal case, i.e., such that, for every θ,

$$p = \mathbb{E}_\theta[||\delta_0(x) - \theta||^2] > \mathbb{E}_\theta[||\delta^{JS}(x) - \theta||^2].$$

This estimator,

$$\delta^{JS}(x) = \left(1 - \frac{p-2}{||x||^2}\right) x, \tag{2.8.3}$$

is now called the *James–Stein* estimator. Note the strange behavior of δ^{JS} when x gets near 0. The factor $1 - \frac{p-2}{||x||^2}$ becomes negative and even goes to $-\infty$ as $||x||$ goes to 0. However, δ^{JS} still dominates δ_0 for all θ's. (This is a consequence of Theorem 2.8.1 below.) Baranchick (1970) corrected this paradoxical behavior by showing that the *truncated* estimators ($p - 2 \leq c \leq 2(p-2)$)

$$\begin{aligned} \delta_c^+(x) &= \left(1 - \frac{c}{||x||^2}\right)^+ x \\ &= \begin{cases} (1 - \frac{c}{||x||^2})x & \text{if } ||x||^2 > c, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{2.8.4}$$

uniformly dominate their nontruncated counterparts and, in particular, that δ_{p-2}^+ was improving on δ^{JS}. They are, moreover, noncomparable (as c varies). This class of estimators is important because, although made of inadmissible estimators (see Chapter 8), estimators that dominate the truncated James–Stein estimator are quite difficult to derive and do not bring significant improvement in terms of risk (see Shao and Strawderman (1996)). On the contrary, the truncated (or *positive-part*) James–Stein estimators improve quite significantly on the least-squares estimator, as illustrated in Figure 2.4.2 for $p = 10$ and $c = 2p - 1$.

Following James and Stein (1961), more general classes of estimators dominating δ_0 have been proposed by Alam (1973), Berger and Bock (1976), Judge and Bock (1978), Stein (1981), George (1986a,b), and Brandwein et al. (1992). These estimators are called *shrinkage estimators* because, as (2.8.3) and (2.8.4), they shrink x toward 0. Stein effects have also been exhibited for distributions other than the normal distribution and losses other than the quadratic loss by Berger (1975), Brandwein and Strawderman (1980), Hwang (1982a), Ghosh et al. (1983), Bock (1985), Haff and Johnston (1987), Srivastava and Bilodeau (1988), Brandwein and Strawderman (1990). Some restrictions on the classes of shrinkage estimators have been proposed, in order to integrate the admissibility requirement (Brown (1971), Alam (1973),

Strawderman (1974), Brown (1975), Berger and Srinivasan (1978), Brown and Hwang (1982), Das Gupta and Sinha (1986), Brown (1988), and Fraisse et al. (1998)). Bondar (1987) has shown that the improvement (in terms of risk) brought by the shrinkage estimators is only significant on a limited part of the parameter space, but George (1986a,b) introduced the concept of *multiple shrinkage* estimators to extend the region where the improvement occurs (see Exercise 10.38).

The Stein effect is also *robust* in the sense that it depends mainly on the loss function, rather than on the exact distribution of the observations, as shown by Brown (1975), Shinozaki (1980, 1984), Berger (1980a,b), Das Gupta (1984), Bilodeau (1988), Cellier, Fourdrinier and Robert (1989), Brandwein and Strawderman (1990) or Kubokawa, Robert and Saleh (1991, 1992, 1993). It is not limited to point estimation, but also occurs for confidence regions (Stein (1962a), Hwang and Casella (1982, 1984), Casella and Hwang (1983, 1987), Robert and Casella (1990), Hwang and Ullah (1994)) and in accuracy (or loss) estimation (Johnstone (1988), Rukhin (1988a,b), Lu and Berger (1989a,b), Robert and Casella (1993), Fourdrinier and Wells (1994)), George and Casella (1994). However, Gutmann (1982) established that the Stein effect cannot occur in finite parameter spaces. Brown (1971) (see also Srinivasan (1981), Johnstone (1984), and Eaton (1992)) showed that admissibility is related to the recurrence of a stochastic process associated with the estimator and Brown (1980) shows the surprising result (called *Berger's phenomenon*, from Berger (1980b)) that there always exists a loss function such that the *boundary* between admissibility and inadmissibility for the usual estimator is an *arbitrary dimension* p_0.

This overview does not do justice to the richness of the work on the Stein effect. The advances realized in this field in the last thirty years have greatly benefitted Decision Theory in general, and Bayesian Decision Theory in particular. In fact, one of the major impacts of the Stein paradox has been to signify the end of a Golden Age for classical Statistics, since it shows that the quest for *the* best estimator, that is, the unique minimax admissible estimator, is hopeless, unless one restricts the class of estimators to be considered, or incorporates some prior information. The works on the Stein effect have thus led to the progressive abandonment of *unbiasedness*, to a deeper understanding of minimaxity and admissibility, and to the improvement of frequentist techniques of risk computation (with Stein's (1973) idea of *unbiased estimator of the risk*). However, its main consequence has been to reinforce the Bayesian-frequentist interface,[5] by inducing frequentists to call for Bayesian techniques (see, for instance, Bock's (1988) idea of *pseudo-Bayes estimators*) and Bayesians to robustify their estimators in terms of frequentist performances and prior uncertainty (Berger (1980b, 1982a, 1984a), George (1986a,b), Lu and Berger (1989a,b), Berger and Robert (1990)). We refer the reader to the books mentioned above as well as to Brandwein and Strawderman (1990) and Lehmann and Casella (1998) for additional references.

We conclude this note with the proof of the inadmissibility of $\delta_0(x) = x$ in the estimation of θ for *spherically symmetric* distributions, i.e., distributions with densities $f(||x - \theta||)$ in $\mathbb{R}^p$ $(p \geq 3)$. References on these distributions which

[5] A typical example is provided by the development of the *empirical Bayes* techniques (see Chapter 10).

generalize the normal distribution in linear regression models are given in Kelker (1970), Eaton (1986) and Fang and Anderson (1990) (see also Exercise 1.1). This result was first established in Cellier et al. (1989).

Theorem 2.8.1 *Consider* $z = (x^t, y^t)^t \in \mathbb{R}^p$, *with distribution*

$$z \sim f(||x-\theta||^2 + ||y||^2), \tag{2.8.5}$$

and $x \in \mathbb{R}^q$, $y \in \mathbb{R}^{p-q}$. *An estimator*

$$\delta_h(z) = (1 - h(||x||^2, ||y||^2))x$$

dominates δ_0 *under quadratic loss if there exist* α, $\beta > 0$ *such that:*

(1) $t^\alpha h(t,u)$ *is a nondecreasing function of* t *for every* u;

(2) $u^{-\beta} h(t,u)$ *is a nonincreasing function of* u *for every* t; *and*

(3) $0 \le (t/u)h(t,u) \le \dfrac{2(q-2)\alpha}{p-q-2+4\beta}$.

The above conditions on h are thus independent of f in (2.8.5), which does not need to be known, and, moreover, they are identical to those obtained in the normal case (see Brown (1975)). The occurrence of the Stein effect is then robust in the class of spherically symmetric distributions with finite quadratic risk.

Proof. Conditions (1) and (2) imply

$$\begin{cases} t\frac{\partial}{\partial t}h(t,u) \ge -\alpha h(t,u), \\ u\frac{\partial}{\partial u}h(t,u) \le \beta h(t,u). \end{cases}$$

The risk of δ_h can be developed as follows:

$$\begin{aligned} R(\theta,\delta_h) &= \mathbb{E}_\theta\left[\sum_{i=1}^q \left\{x_i - \theta_i - h(||x||^2,||y||^2)x_i\right\}^2\right] \\ &= \mathbb{E}_\theta\left[\sum_{i=1}^q (x_i-\theta_i)^2\right] - 2\mathbb{E}_\theta\left[\sum_{i=1}^q h(||x||^2,||y||^2)x_i(x_i-\theta_i)\right] \\ &\quad + \mathbb{E}_\theta\left[h^2(||x||^2,||y||^2)||x||^2\right]. \end{aligned}$$

An integration by parts shows that

$$\begin{aligned} &\int_{-\infty}^{+\infty} h(||x||^2,||y||^2)x_i(x_i-\theta_i)f(||x-\theta||^2+||y||^2)\,dx_i \\ &\quad = \int_{-\infty}^{+\infty} \frac{\partial}{\partial x_i}\left[h(||x||^2,||y||^2)x_i\right]\bar{F}(||x-\theta||^2+||y||^2)\,dx_i, \end{aligned}$$

with

$$\bar{F}(t) = \int_t^{+\infty} f(u)du.$$

Therefore,

$$\begin{aligned} &\mathbb{E}_\theta\left[\sum_{i=1}^q h(||x||^2,||y||^2)x_i(x_i-\theta_i)\right] \\ &\quad = \int_{\mathbb{R}^p}\left[qh(||x||^2,||y||^2) + 2h_1'(||x||^2,||y||^2)||x||^2\right]\bar{F}(||x-\theta||^2+||y||^2)\,dz, \end{aligned}$$

where $h_1'(t,u) = \frac{\partial}{\partial t}h(t,u)$. Similarly,

$$\begin{aligned}
\mathbb{E}_\theta[h^2(||x||^2,||y||^2)||x||^2] &= \mathbb{E}_\theta\left[\frac{||x||^2}{||y||^2}h^2(||x||^2,||y||^2)||y||^2\right] \\
&= \int_{\mathbb{R}^p} ||x||^2 \sum_{j=1}^{p-q} \frac{\partial}{\partial y_j}\left(h^2(||x||^2,||y||^2)\frac{y_j}{||y||^2}\right)\bar{F}(||x-\theta||^2+||y||^2)\,dz \\
&= \int_{\mathbb{R}^p} ||x||^2\Big[4h(||x||^2,||y||^2)h_2'(||x||^2,||y||^2)||x||^2 \\
&\quad + (p-q-2)h^2(||x||^2,||y||^2)\frac{1}{||y||^2}\Big]\bar{F}(||x-\theta||^2+||y||^2)\,dz,
\end{aligned}$$

where $h_2'(t,u) = \frac{\partial}{\partial u}h(t,u)$. The difference of the risks is then

$$\begin{aligned}
&R(\theta,\delta_0) - R(\theta,\delta_h) = \\
&\int_{\mathbb{R}^p}\Big\{2\Big[qh(||x||^2,||y||^2) + 2h_1'(||x||^2,||y||^2)||x||^2\Big]||x||^2h(||x||^2,||y||^2) \\
&\quad\Big[4h_2'(||x||^2,||y||^2) - (p-q-2)h(||x||^2,||y||^2)\frac{1}{||y||^2}\Big]\Big\} \\
&\qquad \times\bar{F}(||x-\theta||^2+||y||^2)\,dz \\
&\geq \int_{\mathbb{R}^p} h(||x||^2,||y||^2)\Big[-h(||x||^2,||y||^2)\frac{||x||^2}{||y||^2}(p-q-2+4\beta) \\
&\quad +2(q-2\alpha)\Big]\bar{F}(||x-\theta||^2+||y||^2)\,dz > 0,
\end{aligned}$$

which concludes the proof. □□

Notice that this domination result includes as a particular case the estimation of a normal mean vector when the variance is known up to a multiplicative factor, i.e., the problem originally considered in James and Stein (1961). When $h(t,u) = au/t$, the bound on a is $2(q-2)/(p-q+2)$, as obtained in James and Stein (1961).