Dominique Fourdrinier,

William E. Strawderman

and Martin T. Wells

Shrinkage estimation

June 25, 2014

Springer

Contents

Chapter 1

Spherically symmetric distributions

1.1 The multivariate normal distribution

For theoretical and practical reasons, the normal distribution plays a central role in Statistics. The central limit theorem is one reason for its importance; given $X_1,\ldots,X_n$ i.i.d. random variables with mean μ and variances $\sigma^2 < \infty$ it states that $(\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ converges in distribution to the standard normal distribution. Hence whatever the distribution of the X_i's, the distribution of the sample mean $\bar{X}_n$ can be approximate by a normal distribution with mean equal to μ and variance equal to σ^2/n. Essentially the same theorem has been used to provide theoretical justification for the empirical fact that many observed quantities tend to be approximately normally distributed.

In this section, we recall basic properties of the univariate and multivariate normal distributions. By definition, the univariate normal distribution $\mathscr{N}(\theta,\sigma^2)$ with mean $\theta \in \mathbb{R}$ and variance $\sigma^2 > 0$ has density $x \mapsto 1/(\sqrt{2\pi}\sigma)\exp\{-(x-\theta)^2/(2\sigma^2)\}$ with respect to Lebesgue measure in $\mathbb{R}^1$. For technical reasons, we also include the case where $\sigma^2 = 0$ which corresponds to the point mass at θ. In this case, the distribution is singular and has no density with respect to Lebesgue measure. As the distribution of any random vector $X \in \mathbb{R}^n$ is characterized by the distribution of all linear functions of the form $a^t X$ for $a \in \mathbb{R}^n$, the following multivariate extension is natural (see Johnson and Kotz [1972]).

Definition 1.1. A random vector $X \in \mathbb{R}^n$ has a normal distribution if, for all $a \in \mathbb{R}^n$, $a^t X$ is distributed as an univariate normal distribution.

Note that the means and variances of the individual components exist by definition and so does the covariance matrix. Also the characteristic function of a univariate standard normal $\mathscr{N}(0,1)$ random variable X is given by $\varphi_{X_i}(t) = E[\exp\{itX_i\}] = \exp\{-t^2/2\}$. Hence, if $X = (X_1, \ldots, X_n)$ where the X_i are i.i.d. standard normal, the characteristic function of X is equal to $\varphi_X(u) = E[\exp\{iu^t X\}] = \exp\{-u^t u/2\}$. Furthermore, if $Y = AX + \theta$ where A is a $p \times n$ matrix and θ a $p \times 1$ vector, $\varphi_Y(v) = E[\exp\{iv^t(\theta + AX)\}] = \exp\{iv^t\theta\}\exp\{-v^t\Sigma v/2\}$ with $\Sigma = AA^t$ the covariance matrix of Y and θ is the mean vector of Y.

This shows that the distribution of Y is determined by its mean vector θ and its covariance matrix Σ. Hence Definition 1.1 is not vacuous and the multivariate normal distribution exists for any mean vector θ and any positive semi-definite covariance matrix Σ (take $A = \Sigma^{1/2}$). We denote this distribution by $\mathscr{N}_p(\theta, \Sigma)$.

It follows from the form of the above characteristic function that, if X is distributed as $\mathscr{N}_n(\theta, \Sigma)$ and B is a $q \times n$ matrix and ν is a $q \times 1$ vector, then $Z = BX + \nu$ is distributed as $\mathscr{N}_q(\theta_Z, \Sigma_Z)$ where $\theta_Z = B\theta + \nu$ and $\Sigma_Z = B\Sigma B^t$. Hence, in particular, all marginal distributions are normal. Specifically decomposing

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathscr{N}_n\left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

with $\dim X_i = \dim \theta_i = n_i$ and where $\Sigma_{i,j}$ is $n_i \times n_j \quad (1 \leq i, j \leq 2)$, we have $X_i \sim \mathscr{N}_{n_i}(\theta_i, \Sigma_{ii})$ and X_1 is independent of X_2 if and only if $\Sigma_{12} = 0$.

We can find the conditional distribution of X_1 given X_2 as follows. Suppose there exists an $n_1 \times n_2$ matrix A such that $X_1 - AX_2$ is independent of X_2. Then the distribution of $X_1 - AX_2$ is normal with mean $\theta_1 - A\theta_2$ and covariance matrix $\Sigma_{11} - A\Sigma_{21}$. Hence the conditional distribution of X_1 given X_2 is normal with mean $\theta_1 + A(X_2 - \theta_2)$ and covariance matrix $\Sigma_{11} - A\Sigma_{21}$. However such an A is easy to find since $\mathrm{cov}(X_1 - AX_2, X_2) = \Sigma_{12} - A\Sigma_{22}$. If Σ_{22} is non-singular, $A = \Sigma_{12}\Sigma_{22}^{-1}$. If Σ_{22} is singular then $A = \Sigma_{12}\Sigma_{22}^{-}$, where Σ_{22}^{-} is a generalized inverse, works since the range of Σ_{12} is the range of Σ_{22} (see Muirhead [1982]).

We now consider the existence of a density with respect to the Lebesgue measure on $\mathbb{R}^n$ for a random vector X distributed as $\mathscr{N}_n(\theta, \Sigma)$. Note that, when Σ is singular, there exists $a \in \mathbb{R}^n$ such that $a \neq 0$ and $\Sigma a = 0$ and hence, for any such a, $V(a^t X) = a^t \Sigma a = 0$. It follows that $a^t X = a^t \theta$ almost surely and hence that $X - \theta$ is almost surely in a proper subspace of $\mathbb{R}^n$; thus the distribution of X is singular and X has no density in $\mathbb{R}^n$.

If however Σ is non-singular, then a density exists. To see this let $\Sigma = AA^t$ for some non-singular $n \times n$ matrix and let $X = AZ + \theta$ where Z is a vector of i.i.d. standard normal random variables in R^1, as in the comment after Definition 1.1. The standard change of variables formula gives the density of X as

$$\frac{1}{(2\pi)^n |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\theta)^t \Sigma^{-1}(x-\theta) \right\}. \tag{1.1}$$

It is important to note that, when $\Sigma = \sigma^2 I_n$ and $\theta = 0$, the density (1.1) is a function of $\|x\|^2$. Consequently, for any orthogonal transformation where $\|x\|$ is the standard Euclidean norm H (i.e. $H^t H = I_n$), the distribution of $Y = HX$ is the same as that of X. Many properties of the normal $\mathcal{N}_n(0, \sigma^2 I_n)$ follow from this invariance property and hold for other distributions similarly invariant. We formalize this in the following definition.

Definition 1.2. A random vector $X \in \mathbb{R}^n$ (equivalently the distribution of X) is orthogonally invariant if, for any orthogonal transformation H, the distribution of $Y = HX$ is the same as that of X.

As a simple example of the utility of this notion, note that, if X is orthogonally invariant and $P[X = 0] = 0$, then the unit vector, which lies on the unit sphere, $X/\|X\|$ is orthogonally invariant as well. We will see in Section 1.2 that there exists only one distribution orthogonally invariant on the unit sphere. It follows that, for any function φ from R^n into R^k, the distribution of $\varphi(X/\|X\|)$ does not depend on the distribution of X as long as X is orthogonally invariant. One of the best known and most useful of such statistics is the Fisher statistic

$$\frac{\|\pi_1(X)\|^2/k_1}{\|\pi_2(X)\|^2/k_2}$$

where π_1 and π_2 are orthogonal projections from R^n onto orthogonal subspaces of dimension k_1 and k_2, respectively.

1.2 The uniform distribution on a sphere

We already noticed the existence of an orthogonally invariant distribution on the unit sphere S of $\mathbb{R}^n$. A closely related alternative approach is through the uniform measure σ_R on the sphere S_R of radius R centered at 0 which can be defined for any Borel set Ω of S_R, as

$$\sigma_R(\Omega) = \frac{n}{R}\lambda(\{ru \in \mathbb{R}^n \mid 0 < r < R,\ u \in \Omega\}) \tag{1.2}$$

where λ is Lebesgue measure on $\mathbb{R}^n$. Thus the measure of Ω is proportional to the Lebesgue measure of the cone spanned by Ω. The constant of proportionality n/R is standard and is chosen so that the total surface area of the sphere S_R agrees with the usual formulas relating $\sigma_R(S_R)$ to the volume of the ball B_R of radius R : $\sigma_R(S_R) = n/R\lambda(B_R)$ (for example, for $n = 2, \sigma_R(S_R) = 2/R\lambda(B_R) = 2\pi R$ and, for $n = 3, \sigma_R(S_R) = 3/R\lambda(B_R) = 4\pi R^2$). As a consequence $\sigma_R(S_R) = \sigma_1(S_1)R^{n-1}$.

The uniform distribution on S_R is naturally defined through σ_R.

Definition 1.3. The uniform distribution $\mathscr{U}_R$ on S_R is defined, for any Borel subset Ω of S_R, by

$$\mathscr{U}_R(\Omega) = \frac{\sigma_R(\Omega)}{\sigma_R(S_R)} = \frac{\sigma_R(\Omega)}{\sigma_1(S_1)R^{n-1}}. \tag{1.3}$$

The orthogonal invariance of $\mathscr{U}_R$ and σ_R follows immediately from the orthogonal invariance of Lebesgue measure λ. The following lemma establishes a uniqueness property for $\mathscr{U}_R$.

Lemma 1.1. *$\mathscr{U}_R$ is the unique orthogonally invariant distribution on S_R.*

Proof. We follow the lines of Cellier and Fourdrinier [1990] which are adapted from the proof given by Philoche [1977] and rely on the uniqueness of Haar measure on the group $\mathscr{O}$ of the orthogonal matrices (as it is developed, for instance, by Nachbin [1965]). More precisely, we use the fact that there exists a unique probability measure ν on $\mathscr{O}$ which is invariant under left and right translations, that is, which satisfies

$$\int_{\mathscr{O}} \phi(h^{-1}g)\, d\nu(g) = \int_{\mathscr{O}} \phi(g)\, d\nu(g)\,,$$

and

$$\int_{\mathscr{O}} \phi(gh^{-1})\, d\nu(g) = \int_{\mathscr{O}} \phi(g)\, d\nu(g)\,,$$

for any function ϕ defined on $\mathscr{O}$ and for any $h \in \mathscr{O}$. This is the so-called Haar measure on $\mathscr{O}$.

Clearly, it suffices to consider the case where $R = 1$. Let $\mathscr{C}(S_1)$ be the set of real values continuous functions on S_1. For any $f \in \mathscr{C}(S_1)$, for any $g \in \mathscr{O}$ and for any $x \in S_1$, define the functions $f_x(g)$ and $f_g(x)$ by

$$f_x(g) = f_g(x) = f(g^{-1}(x))\,.$$

As the group $\mathscr{O}$ operates transitively on S_1, for any $f \in \mathscr{C}(S_1)$, the integral

$$\int_{\mathscr{O}} f_x(g)\, d\nu(g)$$

does not depend on $x \in S_1$. Hence it is possible to define on S_1 a probability measure Q setting, for any $f \in \mathscr{C}(S_1)$,

$$\int_{S_1} f(x)\, dQ(x) = \int_{\mathscr{O}} f_x(g)\, d\nu(g)\,.$$

Now let P be an orthogonally invariant distribution on S_1 and fix $f \in \mathscr{C}(S_1)$. Then

$$\begin{aligned}
\int_{S_1} f(x)\,dP(x) &= \int_{\mathscr{O}} \left(\int_{S_1} f(x)\,dP(x) \right) d\nu(g) \\
&= \int_{\mathscr{O}} \left(\int_{S_1} f_g(x)\,dP(x) \right) d\nu(g) \\
&= \int_{S_1} \left(\int_{\mathscr{O}} f_x(g)\,d\nu(g) \right) dP(x) \\
&= \int_{\mathscr{O}} f_x(g)\,d\nu(g) \\
&= \int_{S_1} f(x)\,dQ(x).
\end{aligned}$$

Therefore $P = Q$, which establishes the unicity. □

The following result mentioned in Section 1.1 is then immediate.

Lemma 1.2. *If $X \in \mathbb{R}^n$ is an orthogonally invariant random vector such that $P[X = 0] = 0$ then $X/\|X\|$ is distributed as $\mathscr{U}_1$.*

It is worth noting that σ_R (and hence $\mathscr{U}_R$) can be expressed through the usual parametrization in terms of polar coordinates. Indeed let $V = (0,\pi)^{n-2} \times (0,2\pi)$ and for $(\theta_1,\dots,\theta_{n-1}) \in V$, $\varphi_R(\theta_1,\dots,\theta_{n-1}) = (x_1,\dots,x_n)$ with

$$\begin{aligned}
x_1 &= R \sin\theta_1 \, \sin\theta_2 \dots \sin\theta_{n-2} \, \sin\theta_{n-1} \\
x_2 &= R \sin\theta_1 \, \sin\theta_2 \dots \sin\theta_{n-2} \, \cos\theta_{n-1} \\
x_3 &= R \sin\theta_1 \, \sin\theta_2 \dots \cos\theta_{n-2} \\
&\vdots \\
x_{n-1} &= R \sin\theta_1 \, \cos\theta_2 \\
x_n &= R \cos\theta_1 .
\end{aligned} \tag{1.4}$$

Note that φ_R maps V onto S_R (except for the set A of σ_R-measure 0, $A = \{x = (x_1,\dots,x_n) \in \mathbb{R}^n \mid x_1 = 0, x_2 \leq 0 \text{ and} \|x\| = R\}$.

Lemma 1.3. *For any Borel subset Ω of S,*

$$\sigma(\Omega) = R^{n-1} \int_{\varphi_R^{-1}(\Omega)} \sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \dots \sin\theta_{n-2}\, d\theta_1\, d\theta_2 \dots d\theta_{n-1}. \tag{1.5}$$

Proof. The usual polar coordinates express x as $r\,\varphi_1(\theta_1,\dots,\theta_{n-1})$ and the set on the right hand side of (1.2) can be written as $]0,R] \times \varphi_R^{-1}(\Omega)$. Hence

$$\begin{aligned}
&\sigma_R(\Omega)\\
&= \frac{n}{R}\lambda\left((0,R) \times \varphi_R^{-1}(\Omega)\right)\\
&= \frac{n}{R}\int_0^R r^{n-1} \int_{\varphi_R^{-1}(\Omega)} \sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \dots \sin\theta_{n-2}\, d\theta_1 d\ \theta_2 \dots d\theta_{n-1}\, dr\\
&= R^{n-1} \int_{\varphi_R^{-1}(\Omega)} \sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \dots \sin\theta_{n-2}\, d\theta_1\, d\theta_2 \dots d\theta_{n-1}.
\end{aligned}$$

□

An immediate consequence is that, if X is distributed as $\mathscr{U}_R$, then the angles θ_i are independent with density proportional to $\sin^{n-i-1}\theta_i$ on $(0,\pi)$ for $1 \leq i \leq n-2$ and θ_{n-1} is uniform on $(0,2\pi)$. Note that $(\mathscr{U}_R)_{R>0}$ is a scale family of distributions in the sense that $\mathscr{U}_R(\Omega) = \mathscr{U}_1(\Omega/R)$ since, in (1.4) we have, $\varphi_R^{-1}(\Omega) = \varphi_1^{-1}(\Omega/R)$.

We will have occasion to use the following lemma which is just a re-expression in terms of σ_R of the usual formula for integration in polar coordinates.

Lemma 1.4. *For any Lebesgue integrable function h, we have*

$$\int_{\mathbb{R}^n} h(x)\, dx = \int_0^\infty \int_{S_R} h(x)\, d\sigma_R(x)\, dR.$$

Proof. Lemma 1.3 implies that

$$\int_{S_R} h(x)\, d\sigma_R(x)$$

equals

$$\int_V h(\varphi_R(\theta_1,\ldots,\theta_{n-1}))\, R^{n-1} \sin^{n-2}\theta_1 \ldots \sin\theta_{n-2}\, d\theta_1,\ldots,d\theta_{n-1}$$

and the result follows. □

Corollary 1.1. *The area measure of the unit sphere is given by*

$$\sigma_1(S_1) = \frac{2\,\pi^{n/2}}{\Gamma(n/2)}.$$

Proof. We apply Lemma 1.4 with $h(x) = 1/(2\pi)^{n/2} \exp\left\{-\|x\|^2/2\right\}$. Then

$$1 = \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{\|x\|^2}{2}\right\} dx = \int_0^\infty \frac{1}{(2\pi)^{n/2}} \exp\{-r^2/2\}\, \sigma_1(S_1)\, r^{n-1}\, dr$$

where we used the fact that $h(x)$ is a function of $\|x\|$ and that $\sigma_r(S_r) = \sigma_1(S_1)\, r^{n-1}$. Letting $t = r^2/2$ reduces the integral to a multiple of a gamma function. More precisely, we have

$$1 = \frac{\sigma_1(S_1)}{2\,\pi^{n/2}}\, \Gamma(n/2)$$

which is the desired result. □

It is worth noting that the normalizing constant $1/(2\pi)^{n/2}$ of the normal density is usually obtained through Lemma 1.4 in dimension 2.

It is convenient to extend the notions of uniform measure and distribution on S_R to any sphere in $\mathbb{R}^n$.

Definition 1.4. For any $R > 0$ and for any $\theta \in \mathbb{R}^n$, let $S_{R,\theta} = \{x \in \mathbb{R}^n \mid \|x - \theta\| = R\}$ the sphere of radius R and center θ. The uniform distribution $\mathscr{U}_{R,\theta}$ (respectively the uniform measure $\sigma_{R,\theta}$) on $S_{R,\theta}$ is the uniform distribution $\mathscr{U}_R$ (respectively the uniform measure σ_R) translated by θ i.e.

$$\mathscr{U}_{R,\theta}(\Omega) = \mathscr{U}_1\left(\frac{\Omega - \theta}{R}\right),$$

for any Borel set Ω of $S_{R,\theta}$. For completeness, we denote the point mass at θ as $\mathscr{U}_{0,\theta}$.

Note that the definition of $\mathscr{U}_{R,\theta}$ (and $\sigma_{R,\theta}$) can be extended to be a distribution (measure) on $\mathbb{R}^n$ by $\mathscr{U}_{R,\theta}(A) = \mathscr{U}_{R,\theta}(A \cap S_{R,\theta})$ for any Borel set A of $\mathbb{R}^n$.

Formula (1.5) is an example of what is sometimes called superficial (or natural) measure on a $n-1$ dimensional submanifold of $\mathbb{R}^n$. Briefly, let O be an open set in $\mathbb{R}^{n-1}$ and φ be a differentiable function mapping O into $\mathbb{R}^n$ with rank $n-1$. Let $g = \sqrt{\det(J^t J)}$ where J is the $n \times (n-1)$ Jacobian matrix of φ. Then the superficial measure σ on $\varphi(O)$ is defined by

$$\sigma(\Omega) = \int_{\varphi^{-1}(\Omega)} g(t_1, \ldots, t_{n-1})\, dt_1 \ldots dt_{n-1}$$

for any Borel set Ω in $\varphi(O)$.

It is easy to check that, for the transformation given by (1.4), the function g is the integrand in the right hand side of (1.5).

The superficial measure is connected in an essential way to Stokes' theorem which we will use extensively. Some details of this connection are given in Chapter 3.

1.3 Spherically symmetric distributions

We now turn our interest to general orthogonally invariant distributions in $\mathbb{R}^n$ and a slightly more general notion of spherically symmetric distributions.

Definition 1.5. A random vector $X \in \mathbb{R}^n$ (equivalently the distribution of X) is spherically symmetric about $\theta \in \mathbb{R}^n$ if $X - \theta$ is orthogonally invariant. We denote this by $X \sim ss(\theta)$.

Note that Definition 1.5 states that $X \sim ss(\theta)$ if and only if $X = Z + \theta$ where $Z \sim ss(0)$.

Furthermore, if P is a spherically symmetric distribution about θ, then

$$P(HC + \theta) = P(C + \theta),$$

for any Borel set C of $\mathbb{R}^n$ and any orthogonal transformation H.

The following proposition is immediate from the definition.

Proposition 1.1. *If a random vector $X \in \mathbb{R}^n$ is spherically symmetric about $\theta \in \mathbb{R}^n$ then, for any orthogonal transformation H, HX is spherically symmetric about $H\theta$ ($X - \theta$ has the same distribution as $HX - H\theta$).*

The connection between spherical symmetry and uniform distributions on spheres is indicated in the following theorem.

Theorem 1.1. *A distribution P in $\mathbb{R}^n$ is spherically symmetric about $\theta \in \mathbb{R}^n$ if and only if there exists a distribution ρ in $\mathbb{R}_+$ such that $P(A) = \int_{\mathbb{R}_+} \mathscr{U}_{r,\theta}(A)\, d\rho(r)$ for any Borel set A of $\mathbb{R}^n$. Furthermore, if a random vector X has such a distribution P, then the radius $\|X - \theta\|$ has distribution ρ (called the radial distribution).*

Proof. Sufficiency is immediate since the distribution $\mathscr{U}_{r,\theta}$ is spherically symmetric about θ for any $r \geq 0$.

It is clear that for the necessity it suffices to consider $\theta = 0$. Let X be distributed as P where P is $ss(0)$, let $\nu(x) = \|x\|$ and let ρ be the distribution of ν. Now, for any Borel sets A in $\mathbb{R}^n$ and B in $\mathbb{R}_+$ and for any orthogonal transformation H, we have using basic properties of conditional distributions

$$\begin{aligned}
\int_B P(H^{-1}(A) \mid \nu = r)\, d\rho(r) &= P(H^{-1}(A) \cap \nu^{-1}(B)) \\
&= P(H^{-1}(A \cap H(\nu^{-1}(B)))) \\
&= P(A \cap H(\nu^{-1}(B))) \\
&= P(A \cap \nu^{-1}(B)) \\
&= \int_B P(A \mid \nu = r)\, d\rho(r)
\end{aligned}$$

where we used orthogonal invariance of the measure P and the function ν. Since the above equality holds for any B, then, almost everywhere with respect to ρ, we have

$$P(H^{-1}(A) \mid \nu = r) = P(A \mid \nu = r)$$

or equivalently the conditional distribution given ν is orthogonally invariant on S_r. By unicity (see Lemma 1.1) it is the uniform distribution on S_r and the theorem follows. □

Corollary 1.2. *A random vector $X \in \mathbb{R}^n$ has a spherically symmetric distribution about $\theta \in \mathbb{R}^n$ if and only if X has the stochastic representation $X = RU$ where R ($R = \|X - \theta\|$) and U are independent, $R \geq 0$ and $U \sim \mathscr{U}$.*

Proof. In the proof of Theorem 1.1, we essentially show that the distribution of $(X-\theta)/\|X-\theta\|$ is $\mathscr{U}$ independently of $\|X-\theta\|$. This is the necessity part of the corollary. The sufficiency part is direct. □

The class of spherically symmetric distributions with a density with respect to Lebesgue measure is of particular interest. The form of this density and its connection with the radial distribution are the subject of the following theorem.

Theorem 1.2. *Let $X \in \mathbb{R}^n$ have a spherically symmetric distribution about $\theta \in \mathbb{R}^n$. Then the following two statements are equivalent.*

(1) *X has a density f with respect to Lebesgue measure in $\mathbb{R}^n$.*

(2) *$\|X-\theta\|$ has a density h with respect to Lebesgue measure in $\mathbb{R}_+$.*

Further, if (1) *or* (2) *holds, there exists a function g from $\mathbb{R}_+$ into $\mathbb{R}_+$ such that*

$$f(x) = g(\|x-\theta\|^2)\, a.e.$$

and

$$h(r) = \frac{2\pi^{n/2}}{\Gamma(n/2)} r^{n-1} g(r^2)\, a.e.$$

The function g is called the generating function and h the radial density.

Proof. The fact that (1) implies (2) follows directly from the representation of X in polar coordinates. We can also argue that (2) implies (1) in a similar fashion using the independence of $\|X-\theta\|$ and the angles and the fact that the angles have a density. The following argument shows this directly and, furthermore, gives the relationship between f, g and h.

It is clear that it suffices to assume that $\theta = 0$. Suppose then that $R = \|X\|$ has a density h. According to Theorem 1.1, for any Borel set A of $\mathbb{R}^n$, we have

$$\begin{aligned}
P(X \in A) &= \int_0^\infty \int_{S_r} \mathbb{1}_A\,(y)\, d\mathscr{U}_r(y)\, h(r)\, dr \\
&= \int_0^\infty \int_{S_r} \mathbb{1}_A(y)\, \frac{d\sigma_r\,(y)}{\sigma_1(S_1) r^{n-1}}\, h(r)\, dr \quad (\text{by } (1.3)) \\
&= \int_0^\infty \int_{S_r} \mathbb{1}_A(y)\, \frac{h(\|y\|)}{\sigma_1(S_1)\|y\|^{n-1}}\, d\sigma_r(y)\, dr \\
&= \int_{\mathbb{R}^n} \mathbb{1}_A(y) \frac{h(\|y\|)}{\sigma_1(S_1)\|y\|^{n-1}}\, dy \quad (\text{by Lemma}\,1.4) \\
&= \int_A \frac{h(\|y\|)}{\sigma_1(S_1)\|y\|^{n-1}}\, dy
\end{aligned}$$

This expresses that the random vector X has a density

$$f(x) = \frac{h(\|x\|)}{\sigma_1(S_1)\|x\|^{n-1}} = g(\|x\|^2)$$

with $h(r) = \sigma_1(S_1)\, r^{n-1}\, g(r^2)$, which is the announced formula for $h(r)$ since $\sigma_1(S_1) = 2\pi^{n/2}/\Gamma(n/2)$ by Corollary 1.1. □

We now turn our attention to the mean and the covariance matrix of a spherically symmetric distribution when they exist.

Theorem 1.3. *Let $X \in \mathbb{R}^n$ be a random vector with a spherically symmetric distribution about $\theta \in \mathbb{R}^n$.*

Then the mean of X exists if and only if the mean of $R = \|X - \theta\|$ exists, in which case $E[X] = \theta$.

The covariance matrix of X exists if and only if $E[R^2]$ is finite, in which case

$$\text{cov}(X) = 1/n\, E[R^2] I_n.$$

Proof. Note that $X = Z + \theta$ where $Z \sim ss(0)$ and it suffices to consider the case $\theta = 0$. By the stochastic representation $X = RU$ in Corollary 1.2 with $R = \|X\|$ independent of U and $U \sim \mathscr{U}$, the expectation $E[X]$ exists if and only if the expectations $E[R]$ and $E[U]$ exist. However $E[U]$ exists, since U is bounded, and is equal to 0 since $E[U] = E[-U]$ by orthogonal invariance.

Similarly $E[\|X\|^2] = E[R^2]E[\|U\|^2] = E[R^2]$, and hence the covariance matrix of X exists if and only if $E[R^2] < \infty$. Now

$$\mathrm{cov}(RU) = E[R^2]E[UU^t] = \frac{E[R^2]}{n} I_n.$$

Indeed $E[U_i^2] = E[U_j^2] = 1/n$ since U_i and U_j have the same distribution by orthogonal invariance and since $\sum_{i=1}^n U_i^2 = 1$. Furthermore $E[U_iU_j] = 0$, for $i \neq j$, since U_iU_j has the same distribution as $-U_iU_j$ by orthogonal invariance. □

An interesting and useful subclass of spherically symmetric distributions consists of the spherically symmetric unimodal distributions. We only consider absolutely continuous distributions.

Definition 1.6. A random vector $X \in \mathbb{R}^n$ with density f is unimodal if, for any $a \geq 0$, the set $\{x \in \mathbb{R}^2 \mid f(x) \geq a\}$ is convex.

Lemma 1.5. *Let $X \in \mathbb{R}^n$ be a spherically symmetric random vector about θ with generating function g. Then the distribution of X is unimodal if and only if g is nonincreasing.*

Proof. Suppose first that the generating function g is nonincreasing. Take the left continuous version of g. For any $a \geq 0$, defining $g^{-1}(a) = \sup\{y \geq 0 \mid g(y) = a\}$ we have

$$\{x \in \mathbb{R}^n \mid g(\|x\|^2) \geq a\} = \{x \in \mathbb{R}^n \mid \|x\|^2 \leq g^{-1}(a)\}$$

which is a ball of radius $\sqrt{g^{-1}(a)}$ and hence convex. Conversely suppose that the set $\{x \in \mathbb{R}^n \mid g(\|x\|^2) \geq a\}$ is convex for any $a \geq 0$ and let $\|x\| \leq \|y\|$. Then, for $x^t = y/\|y\|\|x\|$ we have $\|x^t\| = \|x\|$ and $x^t \in [-y, y]$ and hence, by unimodality assumption, $g(\|x\|^2) = g(\|x^t\|^2) \geq g(\|y\|^2)$. □

Theorem 1.1 showed that a spherically symmetric distribution is a mixture of uniform distributions on spheres. It is worth noting that, when the distribution is also unimodal, it is a mixture of uniform distributions on balls.

Theorem 1.4. *Let $X \in \mathbb{R}^n$ be a spherically symmetric random vector about $\theta \in \mathbb{R}^n$ with generating function g. Then the distribution of X is unimodal if and only if there exists a distribution ν in $\mathbb{R}_+$ with no point mass at* 0 *such that*

$$P[X \in A] = \int_{R_+} \mathscr{V}_{r,\theta}(A)\, d\nu(r) \tag{1.6}$$

for any Borel set A of $\mathbb{R}^n$, where $\mathscr{V}_{r,\theta}$ is the uniform distribution on the ball $B_{r,\theta} = \{x \in \mathbb{R}^n \mid \|x - \theta\| \leq r\}$.

Proof. It is clear that it suffices to consider the case where $\theta = 0$. Suppose first that formula (1.6) is satisfied. Then expressing

$$\mathscr{V}_{r,0}(A) = \frac{1}{\lambda(B_r)} \int_{B_r} \mathbb{1}_A(x)\, dx$$

gives

$$\begin{aligned}
P[X\in A] &= \int_{\mathbb{R}_+} \frac{1}{\lambda_{(}B_r)} \int_{B_r} \mathbb{1}_A(x)\, dx\, d\nu\ (r) \\
&= \int_{\mathbb{R}_+} \frac{1}{\lambda(B_r)} \int_0^r \int_{S_u} \mathbb{1}_A(x)\, d\sigma_u(x)\, du\, d\nu(r) \\
&= \int_{\mathbb{R}_+} \int_{S_u} \mathbb{1}_A(x) \int_u^\infty \frac{1}{\lambda(B_r)}\, d\nu(r)\, d\sigma_u(x)\, du
\end{aligned}$$

after applying Lemma 1.4 and Fubini's theorem. Then

$$\begin{aligned}
P[X\in A] &= \int_u^\infty \int_{S_u} \mathbb{1}_A(x)\, g(\|x\|^2)\, d\sigma_u(x)\, du \\
&= \int_A g(\|x\|^2)\, dx
\end{aligned}$$

again by Lemma 1.4 with the nonincreasing function

$$g(u^2) = \int_u^\infty \frac{1}{\lambda(B_r)}\, d\nu\ (r). \tag{1.7}$$

Hence according to Lemma 1.5, the distribution of X is unimodal.

Conversely, suppose that the distribution of X is unimodal. According to the above, this distribution will be a mixture of uniform distributions on balls if there exists a distribution ν on $\mathbb{R}_+$ with no point mass at 0 such that (1.7) holds. If such a distribution exists, (1.7) implies that ν can be expressed through a Stieltjes integral as

$$\nu(u) = \int_0^u \lambda(B_r)(-dg(r^2)).$$

It suffices therefore to show that ν is a distribution function on $\mathbb{R}_+$ with no point mass at 0. Note that, as g is nonincreasing, ν is the Stieltjes integral of a positive function with respect to a nondecreasing function and hence ν is nondecreasing. Since $\lambda(B_r) = \lambda(B_1)\, r^n = n\, \sigma_1(S_1)\, r^n$, an integration by parts gives

$$\nu(u) = \sigma_1(S_1) \int_0^u r^{n-1} g(r^2)\, dr - \lambda(B_1)^n g(u^2). \tag{1.8}$$

Note that the first term of the right hand side (1.8) is the distribution function of the radial distribution (see Theorem 1.2) and hence approaches 0 (respectively 1) when u approaches 0 (respectively ∞). Therefore to complete the proof it suffices to show that

$$\lim_{u \to 0} u^n g(u^2) = \lim_{u \to \infty} u^n g(u^2) = 0\ .$$

As

$$\int_0^\infty r^{n-1} g(r^2)\, dr < \infty$$

we have

$$\lim_{r \to \infty} \int_{r/2}^r r^{n-1} g(u^2)\, du = 0$$

and, by monotonicity of g, we have

$$\int_{r/2}^r u^{n-1} g(u^2)\ du \geq (r^2) \int_{r/2}^r u^{n-1}\, du = g(r^2)\, r^n \, \frac{1}{n} \left(1 - \frac{1}{2^n}\right).$$

Hence $\lim_{r \to \infty} r^n g(r^2) = 0$. The limit as r approaches 0 can be treated similarly and the result follows. □

It is possible to admit the possibility of a point mass at 0 for a spherically symmetric unimodal distribution, but we choose to restrict the class to absolutely continuous distributions. For a more general version of unimodality see Section 2.1 of Liese and Miescke [2008].

1.4 Elliptically symmetric distributions

By Definition 1.2, a random vector $X \in \mathbb{R}^n$ is orthogonally invariant if, for any orthogonal transformation H, HX has the same distribution as X. The notion of orthogonal transformation is relative to the classical scalar product $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$. It is natural to investigate orthogonal invariance with respect to orthogonal transformations relative to a general scalar product $\langle x, y\rangle_\Gamma = x^t \Gamma y = \sum_{1\leq i,j\leq n} x_i \Gamma_{ij} y_j$ where Γ is a symmetric positive definite $n \times n$ matrix. We define a transformation H to be $\Gamma-$orthogonal if it preserves the scalar product, in the sense that, for any $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, $\langle Hx, Hy\rangle_\Gamma = \langle x, y\rangle_\Gamma$ or, equivalently, if it preserves the associated norm $\|x\|_\Gamma = \sqrt{\langle x, x\rangle_\Gamma}$, that is, if $\|Hx\|_\Gamma = \|x\|_\Gamma$. Note that H is necessarily invertible since

$$\ker H = \{x \in \mathbb{R}^n / Hx = 0\} = \{x \in \mathbb{R}^n / \|Hx\|_\Gamma = 0\} = \{x \in \mathbb{R}^n / \|x\|_\Gamma = 0\} = \{0\}\,.$$

Then it can be seen that H is $\Gamma-$orthogonal if and only if $\langle Hx, y\rangle_\Gamma = \langle x, H^{-1}y\rangle_\Gamma$, for any $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ or, equivalently, if $H^t \Gamma H = \Gamma$.

In this context the Γ-sphere of radius $r \geq 0$ is defined as

$$S_r^\Gamma = \{x \in \mathbb{R}^n \mid x^t \Gamma x = r^2\}\,.$$

Definition 1.7. A random vector $X \in \mathbb{R}^n$ (equivalently the distribution of X) is Γ-orthogonally invariant if, for any Γ-orthogonal transformation H, the distribution of $Y = HX$ is the same as that of X.

We can define a uniform measure on the ellipse S_r^Γ in a manner analogous to (1.2) and the resulting measure is indeed Γ-orthogonally invariant. It is not however the superficial measure mentioned at the end of Section 1.2 but is, in fact, a constant multiple of this measure where the constant of proportionality depends on Γ and reflects the shape of the ellipse. Whatever the constant of proportionality is, it allows the construction of a unique uniform distribution on S_r^Γ as in (1.3). The uniqueness follows from the fact that the Γ-orthogonal transformations form a compact group. We can then adapt the material from Sections 1.2 and 1.3 to the case of a general positive definite matrix Γ. However we present an alternative development.

The following discussion indicates a direct connection between the usual orthogonal invariance and Γ-orthogonal invariance. Suppose, for the moment, that $X \in \mathbb{R}^n$ has a spherically symmetric density given by $g(\|x\|^2)$. Let Σ be a positive definite matrix and A be a non-singular matrix such that $AA^t = \Gamma$. Standard change of variables gives the density of $Y = AX$ as $|\Sigma|^{-1/2} g(y^t \Sigma^{-1} y)$. Let H be any Σ^{-1} orthogonal transformation and let $Z = HY$. The density of Z is $|\Sigma|^{-1/2} g(z^t \Sigma^{-1} z)$ since H^{-1} is also Σ^{-1} -orthogonal and hence $(H^{-1})^t \Sigma^{-1} H^{-1} = \Sigma^{-1}$. This suggests that, in general, $Y = \Sigma^{\frac{1}{2}} X$ is Σ^{-1}-orthogonally invariant if and only if X is orthogonally invariant. The following result establishes this general fact.

Theorem 1.5. *Let Σ be a positive definite $n \times n$ matrix. A random vector $Y \in \mathbb{R}^n$ is Σ^{-1}-orthogonally invariant if and only if $Y = \Sigma^{1/2} X$ with X orthogonally invariant.*

Proof. First note that, for any Σ^{-1}-orthogonal matrix H, $\Sigma^{-1/2} H \Sigma^{-1/2}$ is an (I_n)-orthogonal matrix since

$$
\begin{aligned}
(\Sigma^{-1/2}H\Sigma^{1/2})^t(\Sigma^{-1/2}H\Sigma^{1/2}) &= \Sigma^{1/2}H^t\Sigma^{-1}H\Sigma^{1/2} \\
&= \Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2} \\
&= I_n.
\end{aligned}
$$

Then, if X is orthogonally invariant, for any Borel set C, of $\mathbb{R}^n$ we have

$$
\begin{aligned}
P[H\Sigma^{1/2}X \in C] &= P[\Sigma^{-1/2}H\Sigma^{1/2}X \in \Sigma^{-1/2}C] \\
&= P[X \in \Sigma^{-1/2}C] \\
&= P[\Sigma^{1/2}X \in C].
\end{aligned}
$$

Hence $Y = \Sigma^{1/2}X$ is Σ^{-1}-orthogonally invariant.

Similarly, for any orthogonal matrix G, $\Sigma^{1/2}G\Sigma^{-1/2}$ is a Σ^{-1}-orthogonal matrix. So, if $Y = \Sigma^{1/2}X$ is Σ^{-1}-orthogonally invariant, then X is orthogonally invariant. □

Note that, if X is orthogonally invariant and its covariance matrix exists, it is of the form $\sigma^2 I_n$ by Theorem 1.3. Therefore, if $Y = \Sigma^{1/2}X$, the covariance matrix of Y is $\sigma^2\Sigma$, while, by Theorem 1.5, Y is Σ^{-1}-orthogonal invariant. In statistics, it is more natural to parametrize through a covariance matrix Σ than through its inverse and this motivates the following definition of elliptically symmetric distributions.

Definition 1.8. Let Σ be a positive definite $n \times n$ matrix. A random vector X (equivalently the distribution of X) is elliptically symmetric about $\theta \in \mathbb{R}^n$ if $X - \theta$ is Σ^{-1}-orthogonally invariant. We denote this by $X \sim es(\theta, \Sigma)$.

Note that, if $X \sim ss(\theta)$, then $X \sim es(\theta, I_n)$. If $Y \sim es(\theta, \Sigma)$, then $\Sigma^{-1/2}Y \sim ss(\Sigma^{-1/2}\theta)$.

In the following, we briefly present some results for elliptically symmetric distributions which follow from Theorem 1.5 and are the analogues of those in Sections 1.2 and 1.3. The proofs are left to the reader.

For the rest of this section, let Σ be a fixed positive definite $n \times n$ matrix and denote by $S_R^{\Sigma^{-1}} = \{x \in \mathbb{R}^n \mid x^t \Sigma^{-1} x = R^2\}$ the $(\Sigma^{-1}-)$ ellipse of radius R and by $\mathscr{U}_R^{\Sigma}$ the uniform distributions on $S_R^{\Sigma^{-1}}$.

Lemma 1.6. (1) *The uniform distribution $\mathscr{U}_R^{\Sigma}$ on $S_R^{\Sigma^{-1}}$ is the image under the transformation $Y = \Sigma^{\frac{1}{2}} X$ of the uniform distribution $\mathscr{U}_R$ on the sphere S_R, that is,*

$$\mathscr{U}_R^{\Sigma}(\Omega) = \mathscr{U}_R(\Sigma^{-\frac{1}{2}} \Omega)$$

for any Borel set Ω of $S_R^{\Sigma^{-1}}$.

(2) *If X is distributed as $\mathscr{U}_R^{\Sigma}$ then*

a) $\Sigma^{-1/2} X / (X^t \Sigma^{-1} X)^{1/2}$ *is distributed as $\mathscr{U}$,*

b) $X / (X^t \Sigma^{-1} X)^{1/2}$ *is distributed as $\mathscr{U}_1^{\Sigma}$.*

Theorem 1.6. *A random vector $X \in \mathbb{R}^n$ is distributed as $es(\theta, \Sigma)$ if and only if there exists a distribution $\rho \in \mathbb{R}_+$ such that*

$$P[X \in A] = \int_{\mathbb{R}_+} \mathscr{U}_{r,\theta}^{\Sigma}(A)\, d\rho\,(r)$$

for any Borel set A for $\mathbb{R}^n$, where $\mathscr{U}_{r,\theta}^{\Sigma}$ is the uniform distribution $\mathscr{U}_r^{\Sigma}$ translated by θ. Equivalently X has the stochastic representation $X = RU$ where $R\,(R = \|X - \theta\|_{\Sigma^{-1}} = ((x-\theta)^t \Sigma^{-1} (x-\theta))^{1/2})$ and U are independent, $R \geq 0$ and $U \sim \mathscr{U}_1^{\Sigma}$. For such X, the radius R has distribution ρ (called the radial distribution).

Theorem 1.7. *Let $X \in \mathbb{R}^n$ be distributed as $es(\theta, \Sigma)$. Then the following two statements are equivalent;*

(1) *X has a density f with respect to Lebesgue measure in $\mathbb{R}^n$*

(2) *$\|X - \theta\|_{\Sigma^{-1}}$ has a density h with respect to Lebesgue measure in $\mathbb{R}_+$.*

Further, if (1) *or* (2) *holds, there exists a function g from $\mathbb{R}_+$ into $\mathbb{R}_+$ such that*

$$f(x) = g(\|x - \theta\|^2_{\Sigma^{-1}})$$

and

$$h(r) = \frac{2\pi^{n/2}}{\Gamma(n/2)} |\Sigma|^{-1/2} r^{n-1} g(r^2)$$

(g is called the generating function).

Theorem 1.8. *Let $X \in \mathbb{R}^n$ be distributed as $es(\theta, \Sigma)$. Then the mean of X exists if and only if the mean of $R = \|X - \theta\|_{\Sigma^{-1}}$ exists, in which case $E[X] = \theta$. The covariance matrix exists if and only if $E[R^2]$ is finite, in which case $\operatorname{cov}(X) = 1/nE[R^2]\,\Sigma$.*

Theorem 1.9. *Let $X \in \mathbb{R}^n$ be distributed as $es(\theta, \Sigma)$ with generating function g. Then the distribution of X is unimodal if and only if g is nonincreasing. Equivalently there exists a distribution $\nu \in \mathbb{R}_+$ with no point mass at* 0 *such that*

$$P[X \in A] = \int_{\mathbb{R}_+} \mathscr{V}^{\Sigma}_{r,\theta}(A)\, d\nu(r)$$

for any Borel set A of $\mathbb{R}^n$, where $\mathscr{V}^{\Sigma}_{r,\theta}$ is the uniform distribution on the ball (solid ellipse)

$$B^{\Sigma}_{r,\theta} = \{x \in \mathbb{R}^n \mid \|x - \theta\|_{\Sigma^{-1}} \leq r\}.$$

1.5 Marginal and conditional distributions for *s.s.d.*

In this section, we study marginal and conditional distributions of spherically symmetric distributions. We first consider the marginal distributions for a uniform distribution on S_R.

Theorem 1.10. *Let $X = (x_1^t, x_2^t)^t \sim \mathscr{U}_R$ in $\mathbb{R}^n$ where dim $X_1 = p$ and dim $X_2 = n - p$. Then, for $1 \leq p < n, X_1$ has an absolutely continuous spherically symmetric distribution with generating function g_R given by*

$$g_R(\|x_1\|^2) = \frac{\Gamma(\frac{n}{2})R^{2-n}}{\Gamma((n-p)/2)\pi^{p/2}} (R^2 - \|x_1\|^2)^{(n-p)/2-1} \mathbb{1}_{B_R}(x_1). \tag{1.9}$$

Proof. The proof is based on the fact that $RY/\|Y\| \sim \mathscr{U}_R$, for any random variable Y with a spherically symmetric distribution (see Lemma 1.2), in particular $\mathscr{N}_n(0, I_n)$, and on the fact that X_1 has an orthogonally invariant distribution in $\mathbb{R}^p$. To see this invariance, note that, for any $p \times p$ orthogonal matrix H_1 and any $(n-p) \times (n-p)$ orthogonal matrix H_2, the matrix

$$H = \begin{pmatrix} H_1 & 0 \\ 0 & H_2 \end{pmatrix},$$

is a block diagonal $n \times n$ orthogonal matrix. Hence

$$\left[H \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} H_1 X_1 \\ H_2 X_2 \end{pmatrix} \right] \tag{1.10}$$

is distributed as $(x_1^t, x_2^t)^t$ and it follows that $H_1 X_1 \sim X_1$ and so X_1 is orthogonally invariant.

Therefore, if $Y = (Y_1^t, Y_2^t)^t \sim \mathscr{N}_n(0, I_n)$, then $\|Y_1\|^2$ is independent of $\|Y_2\|^2$ and, according to standard results, $Z = \|Y_1\|^2/\|Y\|^2$ has a beta distribution, that is $Be(p/2,(n-p)/2)$. It follows that $Z' = \|X_1\|^2/\|X\|^2 = \|X_1\|^2/R^2$ has the same distribution since both $X/\|X\|$ and $Y/\|Y\|$ have distribution $\mathscr{U}_R$.

Thus $\|X_1\|^2 = R^2 Z'$ has a $Be(p/2,(n-p)/2)$ density scaled by R^2. By a change of variable, the density of $\|X_1\|$ is equal to

$$h_R(r) = \frac{2}{B(p/2,(n-p)/2)} \, \frac{r^{p-1}(R^2-r^2)^{(n-p)/2-1}}{R^{n-2}} \, \mathbb{1}_{(0,R)}(r).$$

Hence, by Theorem 1.2, X_1 has a density given by (1.9). □

Corollary 1.3. *If $X \sim \mathscr{U}_R$ in $\mathbb{R}^n$ and Π is an orthogonal projection onto any space V of dimension $1 \leq p < n$, the distribution of $Y = \Pi X$ is an absolutely continuous orthogonally invariant distribution on V with generating function g_R given by (1.9).*

Corollary 1.4. *If $X \sim ss(\theta)$ and Π is an orthogonal projection onto any space V of dimension $1 \leq p < n$, the distribution of $Y = \Pi X$ is an absolutely continuous spherically symmetric distribution $ss(\Pi\theta)$ on V with generating function given by $\int g_R(\|y - \pi\theta\|^2)\, d\nu(R)$ where ν is the radial distribution of X and g_R is given by (1.9).*

Note that Corollary 1.4 implies that all projections of X onto a space of dimension p have identical distributions. Unimodality properties of the densities of projections are given in the following result.

Corollary 1.5. *The density of a projection of a $ss(\theta)$ distribution onto a subspace of dimension p is unimodal whenever $n - p \geq 2$. Further the density of the projection* *of $\mathscr{U}_{R,\theta}$ onto a subspace of dimension $n-2$ is the uniform distribution on $B_{R,\theta}$ in $\mathbb{R}^{n-2}$.*

In this book, we will have relatively much more need for the marginal distributions than the conditional distributions of spherical symmetric distributions. For results on conditional distributions, we refer the reader to Fang and Zhang [1990] and to Fang, Kotz and Ng [1990]. We will however have use for the following result.

Theorem 1.11. *Let $X \sim \mathscr{U}_{R,\theta}$ in $\mathbb{R}^n$ and let π be an orthogonal projection onto a space V of dimension p and $\pi^{\perp}$ be the orthogonal projection onto $V^{\perp}$ (the $n-p$ orthogonal complement of V). Then the conditional distribution of πX given $\pi^{\perp} X$ is the uniform distribution on the sphere in V of radius $(R^2 - \|\pi^{\perp} X - \pi^{\perp}\theta\|^2)^{1/2}$ centered at $\pi\theta$.*

Proof. First, it is clear that the support of the conditional distribution of πX given $\pi^{\perp} X$ is the sphere in V of radius $(R^2 - \|\pi^{\perp} X - \pi^{\perp}\theta\|^2)^{1/2}$ centered at $\pi\theta$. It suffices to show that the translated distribution centered at 0 is orthogonally invariant. By an orthogonal transformation of X assume that $X = (x_1^t, x_2^t)^t$ where $\pi X = X_1$ and $\pi^{\perp} X = X_2$. For any orthogonal transformation H on $\mathbb{R}^p$, the block diagonal transformation with blocks H and I_{n-p} is orthogonal in $\mathbb{R}^n$ and hence $((HX_1)^t, x_2^t)^t \sim (x_1^t, x_2^t)^t$. Therefore the distribution of X_1 given X_2 is orthogonally invariant and the lemma follows. □

1.6 Characterizations of the normal distribution

There is a large literature on characterizations of the normal distribution. A classical reference is Kagan, Linnik and Rao [1973]. We give only a small sample of these characterizations. The first result gives a characterization in terms of normality of linear transformation.

Theorem 1.12. *Let $X \sim ss(\theta)$ in $\mathbb{R}^n$. If A is any fixed linear transformation of positive rank such that AX has a normal distribution then X has a normal distribution.*

Proof. First note that it suffices to consider the case $\theta = 0$. Furthermore it suffices to prove the result for $X \sim ss(0)$ since an elliptically symmetric distribution is the image of a spherically symmetric distribution by a non-singular transformation. Note also that, if $X \sim ss(0)$, its characteristic function $\varphi_X(t) = \Psi(t^t t)$ since, for any orthogonal transformation H, the characteristic function φ_{HX} of HX satisfies

$$\varphi_{HX}(t) = \varphi_X(H^t t) = \varphi_X(t).$$

Now the characteristic function φ_{AX} of AX equals

$$\varphi_{AX}(t) = E[\exp\{it^t AX\}] = E[\exp\{i(A^t t)^t X\}] = \Psi(t^t AA^t t) \tag{1.11}$$

Also, by Theorem 1.3, $\mathrm{Cov}(X) = E[R^2]/nI_n$. Hence $\mathrm{Cov}(AX) = E[R^2]/nAA^t$ and the fact that AX is normal implies that $E[R^2] < \infty$ and that $\mathrm{Cov}(AX) = \alpha A, A^t$ for $\alpha \geq 0$. This implies that $\varphi_{AX}(t) = \exp\{-\alpha t^t AA^t t/2\}$. Therefore, by (1.11), $\Psi(z) = \exp\{-\alpha z/2\}$ and hence $\varphi_X(t) = \exp\{-\alpha t^t t/2\}$. □

Corollary 1.6. *Let $X \sim es(\theta)$ in $\mathbb{R}^n$. If any orthogonal projection Π has normal distribution (and, in particular, any marginal) then X has a normal distribution.*

The next theorem gives a characterization in terms of independence of linear projections.

Theorem 1.13. *Let $X \sim es(\theta)$ in $\mathbb{R}^n$. If A and B are any two fixed linear transformations of positive rank such that AX and BX are independent then X has a normal distribution.*

Proof. As in the proof of Theorem 1.12, we can assume that $X \sim es(0)$. Then the characteristic function φ_X of X is $\varphi_X(t) = \Psi(t^t t)$. Hence the characteristic function φ_{AX} and φ_{BX} of AX and BX are $\varphi_{AX}(t_1) = \Psi(t_1^t AA^t t_1)$ and $\varphi_{BX}(t_2) = \Psi(t_2^t BB^t t_2)$. By independence of AX and BX, we have

$$\Psi(t_1^t AA^t t_1 + t_2^t BB^t t_2) = \Psi(t_1^t AA^t t_1)\Psi(t_2^t BB^t t_2).$$

Since A and B are of positive rank this implies that, for any $u \geq 0$ and $v \geq 0$,

$$\Psi(u+v) = \Psi(u)\Psi(v).$$

This equation is known as Hamel's equation and its only continuous solution is $\Psi(u) = e^{\alpha u}$ for some $\alpha \in \mathbb{R}$ (see *e.g*. Feller page 305, 1971). Hence $\varphi_X(t) = e^{\alpha t^t t}$ for some $\alpha \leq 0$ since φ_X is a characteristic function. It follows that X has a normal distribution. □

Corollary 1.7. *Let $X \sim es(\theta)$ in $\mathbb{R}^n$. If any two projections (in particular, any two marginals) are independent then X has a normal distribution.*

Chapter 2

Decision Theory Preliminaries

2.1 Introduction

In this chapter, we introduce statistical and decision theoretic terminology and results that will be used throughout the book. We assume that the reader is familiar with the basic statistical notions of parametric families of distributions, likelihood functions, maximum likelihood estimation, sufficiency, completeness and unbiaseness at the level of, for example, Casella and Berger [2001], Shao [2003], or Bickel and Doksum [2006]. In the following, we will discuss, largely without proof, some results in Bayesian decision theory, minimaxity, admissibility, invariance, and general linear models that will be used later in the book.

2.2 Bayesian decision theory

In this section, we introduce loss functions, risk functions, and some results in Bayesian decision theory.

Suppose $X \sim f_\theta(x)$ where $f_\theta(x)$ is a density with respect to a σ -finite measure μ on $\mathscr{X}$ a measurable subset of $\mathbb{R}^n$ ($\mathscr{X}$ is the sample space) and θ is in Ω a measurable subset of $\mathbb{R}^p$ (Ω is the parameter space). We require that $f_\theta(x)$ be jointly measurable on $\mathscr{X} \times \Omega$.

In the problem of estimating a measurable function $g(\theta)$ from $\mathbb{R}^p$ into $g(\Omega) \subset \mathbb{R}^k$, an estimator is a measurable function $\delta(X)$ from $\mathbb{R}^n$ into $\mathscr{D} \subset \mathbb{R}^k$ ($\mathscr{D}$ is the decision space). Typically we would require $\mathscr{D} \subset \Omega$ but, occasionally, it is more convenient to allow $\mathscr{D}$ to contain $g(\Omega)$.

The measure of closeness of an action $d \in \mathscr{D}$ to the "true value " of $g(\theta)$ is given by a (jointly measurable) loss function $L(\theta, d)$, where $L(\theta, g(\theta))$ and $L(\theta, d) \geq 0$. Hence there is no loss if the "correct decision" $d = g(\theta)$ is made and a nonnegative loss whatever decision is made. A larger value of the loss corresponds to a worse decision.

A simple example for the case of $g(\Omega) \subset \mathbb{R}^1$ and $\mathscr{D} \subset \mathbb{R}^1$ is $L(\theta, d) = (d - g(\theta))^2$, so called squared error loss. Another common choice is $L(\theta, d) = |d - g(\theta)|$ or, more generally, $L(\theta, d) = \rho(g(\theta), d)$ where $\rho(g(\theta), g(\theta)) = 0$ and $\rho(g(\theta), d)$ is monotone nondecreasing in d when $d \geq g(\theta)$, and monotone nonincreasing in d when $d \leq g(\theta)$, a so called bowl-shaped loss.

In higher dimensions, when $\mathscr{D} \subset \mathbb{R}^k$ and $\Omega \subset \mathbb{R}^k$, similar examples would be

$$L(\theta, d) = ||d - g(\theta)||^2 = \sum_{i=1}^{k} (d_i - g_i(\theta))^2$$

(the sum of squared errors loss or quadratic loss),

$$L(\theta, d) = \sum_{i=1}^{k} |d_i - g_i(\theta)|$$

(the sum of absolute errors loss) and

$$L(\theta, d) = (d - g(\theta))^t Q(d - g(\theta)),$$

where Q is a positive semidefinite matrix (the weighted quadratic loss).

To help in the assessment of estimators (or, more generally, decision procedures), it is useful to introduce the risk function $\mathscr{R}(\theta, \delta) = E_\theta[L(\theta, \delta(X))]$. The risk function only depends on the estimator $\delta(\cdot)$ (and not just on its value, $\delta(x)$, at a particular observation, $X = x$) and, of course, on θ.

Frequentist decision theory is mainly concerned with the choice of estimators which, in some sense, make $\mathscr{R}(\theta, \delta)$ small. Bayesian decision theory, in particular, is largely focused on minimizing the average of $\mathscr{R}(\theta, \delta)$ with respect to some (positive) weight function (measure) π, referred to as the prior measure or prior distribution.

It suffices for our purpose to suppose that the prior measure π is a finite measure on Ω and, without loss of generality, to assume it is a probability measure (i.e. $\pi(\Omega) = 1$).

Definition 2.1. [Bayes procedures] For any (measurable) function δ from $\mathscr{X}$ into $\mathscr{D}$ the Bayes risk of δ is

$$\begin{aligned} r(\pi, \delta) &= \int_\Omega \mathscr{R}(\theta, \delta)\, d\pi(\theta) \\ &= \int_\Omega \left[\int_{\mathscr{X}} L(\theta, \delta(x))\, f_\theta(x)\, d\mu(x) \right] d\pi(\theta). \end{aligned} \tag{2.1}$$

A (proper) Bayes procedure, $\delta_\pi(X)$, with respect to the (proper) prior π, is any estimator δ_π such that

$$r(\pi) = r(\pi, \delta_\pi) = \inf_\delta r(\pi, \delta)\,. \tag{2.2}$$

The quantity $r(\pi)$ is referred to as the Bayes risk of π or simply the Bayes risk.

In certain settings, it is not necessary to require that π be a finite measure but only to require that there exists a $\delta(X)$ such that (2.1) is finite. Note also that the joint measurability of $f_\theta(X)$, and also of $L(\theta, \delta(X)$, implies that the double integral in (2.1) makes sense.

It is helpful to define joint and marginal distributions as follows.

Definition 2.2. (1) The joint distribution of (X, θ) is

$$P[X \in A, \theta \in B] = \int_B \left[\int_A f_\theta(x)\, d\mu(x) \right] d\pi(\theta)\,. \tag{2.3}$$

(2) The marginal distribution of θ is the prior distribution $\pi(\cdot)$ since

$$P[\theta \in B] = \int_B \left[\int_{\mathscr{X}} f_\theta(x)\, d\mu(x) \right] d\pi(\theta) = \int_B d\pi(\theta) = \pi(B)\,. \tag{2.4}$$

(3) The marginal distribution of X is

$$\begin{aligned} M(A) &= P[X \in A] \\ &= \int_\Omega \left[\int_A f_\theta(x)\, d\mu(x) \right] d\pi(\theta) \\ &= \int_A \left[\int_\Omega f_\theta(x)\, d\pi(\theta) \right] d\mu(x) \quad \text{by Fubini's theorem} \\ &= \int_A m(x)\, d\mu(x) \end{aligned} \tag{2.5}$$

where

$$m(x) = \int_{\Omega} f_{\theta}(x)\, d\pi(\theta)\,.$$

Hence it follows that the marginal distribution of X is defined and is absolutely continuous with respect to μ, and has density m.

Definition 2.3. The posterior distribution of θ given x is defined such that (for $m(x) \neq 0$)

$$d\pi(\theta|x) = \frac{f_{\theta}(x)}{m(x)}\, d\pi(\theta)\,. \tag{2.6}$$

Note that the posterior distribution as defined in (2.6) is absolutely continuous with respect to the measure π, and hence, has density

$$\frac{f_{\theta}(x)}{m(x)}$$

with respect to π. It is well defined for all x such that $m(x) > 0$, and hence M-almost everywhere.

The above observations and (again) Fubini's theorem allow an immediate convenient re-expression of (2.1).

Lemma 2.1. *The Bayes risk in (2.1) may be expressed as*

$$\begin{aligned} r(\pi,\delta) &= \int_{\mathscr{X}} \left[\int_{\Omega} L(\theta,\delta(x))\, d\pi(\theta|x) \right] dM(x) \\ &= \int_{\mathscr{X}} \left[\int_{\Omega} L(\theta,\delta(x))\, d\pi(\theta|x) \right] m(x)\, d\mu(x). \end{aligned} \tag{2.7}$$

It follows that a Bayes estimate $\delta_{\pi}(x)$ may be calculated, for μ-almost every x, by minimizing the so-called posterior loss function or posterior expected loss of δ.

Lemma 2.2. *Suppose that there exists an estimator with finite Bayes risk and that, for M-almost every x, there exists a value $\delta_\pi(x)$ minimizing*

$$E[L(\theta,\delta(X))|x] = \int_\Omega L(\theta,\delta(x))\,\frac{f_\theta(x)}{m(x)}\,d\pi(\theta). \tag{2.8}$$

Then $\delta_\pi(X)$ is a Bayes estimator and $E[L(\theta,\delta(X))|x]$ is said the posterior risk.

Corollary 2.1. *Under the assumptions of Lemma 2.2,*

(1) *if $L(\theta,d)) = (d-g(\theta))^t Q(d-g(\theta))$ where Q is positive (semi) definite, the Bayes estimator is given by*

$$\delta_\pi(X) = E[g(\theta)|X]\,;$$

(2) *if $L(\theta,d)) = (d-g(\theta))^t Q(\theta)(d-g(\theta))$ where $Q(\theta)$ is positive definite, the Bayes estimator is given by*

$$\delta_\pi(X) = (E[Q(\theta)|X])^{-1}E[Q(\theta)\,g(\theta)|X]\,.$$

Uniqueness of the Bayes estimator follows under the assumption of strict convexity $L(\theta,d)$ in d, finiteness of the integrated risk of $\delta_\pi(X)$ and absolute continuity of μ with respect to the marginal distribution M of X (i.e. μ and M are mutually absolutely continuous).

It is often convenient to deal with prior measures π that are not finite. In such case, typically there is no procedure $\delta(\cdot)$ for which (2.1) is finite, but it is often the case that the posterior distribution given formally by (2.6) exists and is a finite measure which can be normalized to be a probability distribution. In such a case, an estimator $\delta_\pi(X)$ minimizing (2.8) is called a generalized Bayes (or formal Bayes) estimator.

Example 2.1. [Normal location families] Suppose $X \sim \mathscr{N}_p(\theta, \sigma^2 I)$ with σ^2 known and the prior measure π (not necessarily finite) satisfies

$$m(x) = \int_{\mathbb{R}^p} \left(\frac{1}{\sqrt{2\pi}\,\sigma} \right)^p \exp\left(-\frac{1}{2\sigma^2} ||x-\theta||^2 \right) d\pi(\theta) < \infty$$

for all $x \in \mathbb{R}^p$. If the loss is of the form $L(\theta, d) = (d - g(\theta))^t Q(d - g(\theta))$, the Bayes (or generalized Bayes) estimator is given by

$$\begin{aligned}
\delta_\pi(X) &= E[\theta|X] \\
&= X + \frac{\int_{\mathbb{R}^p} (\theta - X) \exp\left(-\frac{1}{2\sigma^2}||X-\theta||^2\right) d\pi(\theta)}{\int_{\mathbb{R}^p} \exp\left(-\frac{1}{2\sigma^2}||X-\theta||^2\right) d\pi(\theta)} \\
&= X + \sigma^2 \frac{\nabla m(X)}{m(X)} \qquad (2.9)
\end{aligned}$$

where interchange of integration and differentiation is justified by standard results for exponential families. (See Brown [1986] and also Lemma A.1 in the Appendix. Expression (2.9) is due to Brown [1971] and is also useful in analyzing risk properties of Bayes estimators. Similar expressions for spherically symmetric location families will be developed in Chapters 5 and 6.

Consider now the posterior risk $E[\|\theta - \delta_\pi(X))\|^2 \big| x]$. According to (2.9), we have

$$\begin{aligned}
E[\|\theta - \delta_\pi(X))\|^2 \big| x] &= E\left[\left\| \theta - X - \sigma^2 \frac{\nabla m(X)}{m(X)} \right\|^2 \Big| x \right] \\
&= E\left[\|\theta - X\|^2 + \sigma^4 \left\| \frac{\nabla m(X)}{m(X)} \right\|^2 - 2\sigma^2 \frac{\nabla m(X)}{m(X)} \cdot (\theta - X) \Big| x \right] \\
&= E\left[\|\theta - X\|^2 - \sigma^4 \left\| \frac{\nabla m(X)}{m(X)} \right\|^2 \Big| x \right]
\end{aligned}$$

Now

$$E\left[\|\theta - X\|^2 \middle| x\right] = p\,\sigma^2 + \sigma^4\,\frac{\Delta m(x)}{m(x)}$$

since

$$\begin{aligned}
\frac{\Delta m(x)}{m(x)} &= \frac{\Delta \int_{\mathbb{R}^p} \exp\left(-\frac{1}{2\sigma^2}||x-\theta||^2\right)}{\exp\left(-\frac{1}{2\sigma^2}||x-\theta||^2\right)} \\
&= \frac{\Delta \int_{\mathbb{R}^p} \exp\left(-\frac{1}{2\sigma^2}||x-\theta||^2\right)}{\exp\left(-\frac{1}{2\sigma^2}||x-\theta||^2\right)} \\
&= \frac{\int_{\mathbb{R}^p} (\frac{||x-\theta||^2}{\sigma^4} - \frac{p}{\sigma^2}) \exp\left(-\frac{1}{2\sigma^2}||x-\theta||^2\right)}{\exp\left(-\frac{1}{2\sigma^2}||x-\theta||^2\right)} \\
&= E\left[\frac{\|\theta - X\|^2}{\sigma^4} - \frac{p}{\sigma^2} \middle| x\right].
\end{aligned}$$

Therefore the postrior risk equals

$$E[\|\theta - \delta_\pi(X))\|^2 | x] = p\,\sigma^2 + \sigma^4 \left\{ \frac{\Delta m(x)}{m(x)} - \left\| \frac{\nabla m(X)}{m(X)} \right\|^2 \right\}. \tag{2.10}$$

Now suppose $\theta \sim \mathcal{N}_p(\nu, \tau^2 I_p)$ (i.e. π is a normal distribution with mean vector ν and covariance matrix $\tau^2 I_p$). Then $m(x)$ equals

$$\begin{aligned}
&\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^p \left(\frac{1}{\sqrt{2\pi}\,\tau}\right)^p \\
&\int_{\mathbb{R}^p} \exp\left(-\frac{1}{2\sigma^2}||x-\theta||^2\right) \exp\left(-\frac{1}{2\tau^2}||\theta-\nu||^2\right) d\theta \\
&= \left(\frac{1}{\sqrt{2\pi}\sqrt{\sigma^2+\tau^2}}\right)^p \exp\left(-\frac{1}{2(\sigma^2+\tau^2)}||x||^2\right)
\end{aligned}$$

(the convolution of $\mathcal{N}_p(0, \sigma^2 I_p)$ and $\mathcal{N}_p(\nu, \tau^2 I_p)$ is $\mathcal{N}_p(\nu, (\sigma^2+\tau^2) I_p)$. Hence the Bayes estimator is

$$
\begin{aligned}
\delta_\pi(X) &= X + \frac{\sigma^2\,(-(X-\nu))}{\sigma^2+\tau^2} \\
&= \frac{\tau^2}{\sigma^2+\tau^2}\,X + \frac{\sigma^2}{\sigma^2+\tau^2}\,\nu \\
&= \nu + \frac{\tau^2}{\sigma^2+\tau^2}\,(X-\nu) \qquad (2.11) \\
&= \nu + \left(1 - \frac{\sigma^2}{\sigma^2+\tau^2}\right)(X-\nu)\,.
\end{aligned}
$$

If the general prior distribution π is Lebesgue measure ($d\pi(\theta) = d\theta$), then $m(X) \equiv 1$ and the generalized Bayes estimator is given by

$$\delta_\pi(X) = X + \sigma^2 \frac{\nabla 1}{1} = X\,.$$

It is often convenient, both theoretically and for computational reasons, to express (proper and generalized) prior distributions hierarchically, typically in two or three stages. The first stage of the hierarchy is often a conjugate prior, i.e. one such that the posterior distribution is in the same class as the prior distribution. In Example 2.1, the class of $\theta \sim \mathscr{N}_p(\nu, \tau^2 I_p)$ priors is a conjugate family since the posterior is given by

$$\theta | x \sim \mathscr{N}_p\left(\frac{\tau^2 x + \sigma^2 \nu}{\sigma^2+\tau^2}, \frac{\sigma^2\,\tau^2}{\sigma^2+\tau^2}\,I_p\right).$$

At the second stage, one could put a prior (or generalized prior) distribution on the first stage prior variance τ^2. A convenient way to do this in certain settings (see, for example, Chapter 4 where this device is used to produce improved shrinkage estimators for the normal model) is as follows.

Suppose the first stage prior variance τ^2 is expressed as $\tau^2 = \sigma^2\,(1-\lambda)/\lambda$ for $0 < \lambda < 1$. Then $\sigma^2 + \tau^2 = \sigma^2/\lambda$ and

$$\frac{\tau^2 x + \sigma^2 \nu}{\sigma^2 + \tau^2} = (1-\lambda)\, x + \lambda\, \nu = \nu + (1-\lambda)\,(\lambda - \nu)\,.$$

Hence a second stage prior $H(\lambda)$ with prior density $h(\lambda)$ for $0 < \lambda < 1$ (hierarchical generalized or proper) leads to the marginal density

$$m(x) = \int_0^1 \left(\frac{\lambda}{2\,\pi\,\sigma^2}\right)^{p/2} \exp\left(-\frac{\lambda}{2\,\sigma^2}\,||x-\nu||^2\right) h(\lambda)\, d\lambda$$

and the Bayes estimator

$$\begin{aligned}\delta_\pi(X) &= X + \sigma^2\,\frac{\nabla m(X)}{m(X)} \\ &= X - \frac{\int_0^1 \lambda^{p/2+1} \exp\left(-\frac{\lambda}{2\sigma^2}||X-\nu||^2\right) h(\lambda)\, d\lambda}{\int_0^1 \lambda^{p/2} \exp\left(-\frac{\lambda}{2\sigma^2}||X-\nu||^2\right)}(X-\nu) \\ &= \nu + E[(1-\lambda)|X]\,(X-\nu)\,.\end{aligned}$$

Empirical Bayes estimators are closely related to hierarchical Bayes estimators. If the first stage prior $\pi(\theta|\tau)$ is viewed as specifying a class of priors indexed by a parameter τ, then the first stage marginal

$$m(x|\tau) = \int f_\theta(x)\, d\pi(\theta|\tau)$$

may be viewed as a likelihood depending on the data x and the parameter τ. One may choose to estimate the parameter τ in a classical frequentist way such as a maximum likelihood estimator (MLE) or perhaps a UMVU estimator, and then calculate a Bayes estimator by the first stage Bayesian model substituting the estimated λ. Such estimators are called Empirical Bayes estimators.

For example, in the above normal model, the first stage marginal distribution (parametrized by τ^2 with ν fixed and known) is

$$X|\tau^2 \sim \mathscr{N}_p(\nu, (\sigma^2 + \tau^2)I)\,.$$

Since ν is fixed and known, $||X-\nu||^2$ is a complete sufficient statistic and the MLE of τ^2 is $\hat{\tau}^2 = \max(0, \sigma^2 - ||X-\nu||^2/p)$ giving an empirical Bayes estimate of θ (based on (2.11))

$$\begin{aligned}\delta^{EB}(X) &= \nu + \frac{\hat{\tau}^2}{\sigma^2 + \hat{\tau}^2}(X-\nu) \\ &= \nu + \left(1 - \frac{p\,\sigma^2}{||X-\nu||^2}\right)_+ (X-\nu)\,.\end{aligned}$$

Alternatively, the UMVU estimator of $1/(\sigma^2+\tau^2)$ is $1/(\widehat{\sigma^2+\tau^2}) = (p-2)/||X-\nu||^2$, so a different empirical Bayes estimator based on (2.11) would be

$$\nu + \left(1 - \frac{(p-2)\,\sigma^2}{||X-\nu||^2}\right)(X-\nu).$$

The first of these is a version of the James-Stein positive part estimator while the second is the classical James-Stein estimator. The risk properties of these estimators are examined in Chapter 3 and 4.

2.3 Minimaxity

In the development of Bayes estimators, the risk function was integrated with respect to a prior. Minimax estimation takes another approach and does not depend on a prior.

Definition 2.4. An estimator $\delta_0(X)$ is minimax if

$$\sup_{\theta\in\Omega} \mathscr{R}(\theta, \delta_0) = \inf_{\delta\in\mathscr{D}} \sup_{\theta\in\Omega} \mathscr{R}(\theta, \delta)\,,$$

where $\mathscr{D}$ is the class of all estimators.

It is occasionally useful to take $\mathscr{D}$ to be a subset of the class of all estimators (for example, all linear estimators) in which case δ_0 would be said to be minimax in $\mathscr{D}$.

We give two results which have proved useful for finding minimax estimators (see Lehmann and Casella [1998] for proofs).

Lemma 2.3. *If a proper prior π has an associated Bayes estimator $\delta_\pi(X)$ and if $\sup_{\theta\in\Omega}\mathscr{R}(\theta,\delta_\pi) = r(\pi,\delta_\pi)(= r(\pi))$, then $\delta_\pi(X)$ is minimax. The prior π is also least favorable in the sense that $r(\pi',\delta_\pi) \leq r(\pi,\delta_\pi)$ for all prior distributions π'.*

One easy and useful corollary of this result is that a Bayes estimator with constant (finite) risk is minimax. The second result is more useful in the case where the parameter space is noncompact.

Lemma 2.4. *If $\delta_0(X)$ is an estimator such that $\sup_{\theta\in\Omega}\mathscr{R}(\theta,\delta_0) = r$ and if there exists a sequence of priors (π_n) such that $\lim_{n\to\infty} r(\pi_n,\delta_{\pi_n}) = r$ then $\delta_0(X)$ is minimax. The sequence of priors (π_n) is what is known as a least favorable sequence in the sense that, for any prior π, we have $r(\pi) \leq r$.*

This second result is useful for establishing minimaxity of the usual estimator X in the normal location problem.

Example 2.2. [Minimaxity of X for $X \sim \mathscr{N}_p(\theta,\sigma^2 I)$, σ^2 known]

Let $X \sim \mathscr{N}_p(\theta,\sigma^2 I)$ with σ^2 known and loss equal to $L(\theta,d) = ||d-\theta||^2$. Suppose the sequence of priors, (π_n), on θ is $\mathscr{N}_p(0,nI)$. Then the posterior distribution is $\mathscr{N}_p(n/(n+\sigma^2)X, n\sigma^2/(n+\sigma^2)I)$, the posterior risk is $n\sigma^2/(n+\sigma^2)\,p$ which is also the Bayes risk. Since $r(\pi_n) = n\sigma^2/(n+\sigma^2)\,p \to p\sigma^2 \equiv \mathscr{R}(\theta, X)$, it follows that X is minimax.

Example 2.3. [Minimaxity of X for $X \sim f(||X-\theta||^2)$]

Similarly, if $X \sim f(||X-\theta||^2)$ where $E[||X-\theta||^2)] = p\sigma^2 < \infty$, then the sequence of priors $\pi_n(\theta) = f^{*n}(\theta)$, where f^{*n} is the n-fold convolution of f with itself, leads to a proof that X is minimax. To see this, note that, if $U_1, \ldots, U_n$ are i.i.d. copies of $U_0 \sim f(||u||^2)$, then $\theta = \sum_{i=1}^{n} U_i \sim f^{*n}(||\theta||^2)$. Also $U_0 = X - \theta \sim f(||u_0||^2)$ and is independent of $\theta = \sum_{i=1}^{n} U_i$. Also $X = (X-\theta) + \theta = \sum_{i=0}^{n} U_i$. It follows that the Bayes estimator corresponding to π_n may be represented as

$$\begin{aligned}
\delta_{\pi_n}(X) &= E[\theta|X] \\
&= E\left[\sum_{i=1}^{n} U_i \middle| \sum_{i=0}^{n} U_i\right] = nE\left[U_1 \middle| \sum_{i=0}^{n} U_i\right] \\
&= \frac{n}{n+1} E\left[\sum_{i=0}^{n} U_i \middle| \sum_{i=0}^{n} U_i\right] \\
&= \frac{n}{n+1} X\,.
\end{aligned}$$

The corresponding Bayes risk is

$$\begin{aligned}
E^{\theta}[E^{X|\theta}[||\delta_{\pi_n}(X)-\theta||^2]] &= E^{\theta}\left[E^{X|\theta}\left[\left|\left|\frac{n}{n+1}X-\theta\right|\right|^2\right]\right] \\
&= E^{\theta}\left[pVar\left[\frac{n}{n+1}X_1\right]+\sum_{i=1}^{p}\left(\frac{n}{n+1}X_i-\theta_i\right)^2\right] \\
&= E^{\theta}\left[p\left(\frac{n}{n+1}\right)^2\sigma^2+\sum_{i=1}^{p}\left(\frac{1}{n+1}\theta_i\right)^2\right] \\
&= E^{\theta}\left[p\left(\frac{n}{n+1}\right)^2\sigma^2+\left(\frac{1}{n+1}\right)^2||\theta||^2\right] \\
&= p\left(\frac{n}{n+1}\right)^2\sigma^2+\left(\frac{1}{n+1}\right)^2 pn\sigma^2 \\
&= p\sigma^2\frac{n}{n+1}\longrightarrow p\sigma^2 = E[||X-\theta||^2].
\end{aligned}$$

Hence X is minimax and (π_n) is a least favorable sequence.

Example 2.4. [Minimaxity of X in the unknown σ^2 case]

In this example, we assume $(X,U)\sim\mathscr{N}_{p+k}((\theta,0)^t,\sigma^2 I)$ when $\dim X=\dim\theta=p$ and $\dim U=\dim 0=k$. Suppose the loss is $||\delta-\theta||^2/\sigma^2$. We need the following easy result (see Lehmann and Casella [1998]).

Lemma 2.5. *Suppose $\delta(X)$ is minimax in a problem for $X\sim f$ with $f\in\mathscr{F}_0$. Suppose $\mathscr{F}_0\subset\mathscr{F}_1$ and $\sup_{f\in\mathscr{F}_0}\mathscr{R}(f,\delta)=\sup_{f\in\mathscr{F}_1}\mathscr{R}(f,\delta)$. Then $\delta(X)$ is minimax for $f\in\mathscr{F}_1$.*

The argument of Example 2.2 suffices to show that X is minimax for any fixed σ^2. Since the risk of X is constant and equal to p for the entire family when σ^2 is unknown, Lemma 2.5 proves that X is minimax.

2.4 Admissibility

An admissible estimator is one which cannot be uniformly improved upon in terms of risk. An inadmissible estimator is one for which an improved estimator can be found. More formally we have the definition.

Definition 2.5. (1) $\delta(X)$ is inadmissible if there exists an estimator $\delta'(X)$ for which $\mathscr{R}(\theta,\delta') \leq \mathscr{R}(\theta,\delta)$ for all $\theta \in \Omega$, with strict inequality for some θ.

(2) $\delta(X)$ is admissible if it is not inadmissible.

The most direct method to prove that an estimator is inadmissible is to find a better one. Much of this book is concerned with exactly this process of finding and developing improved estimators, typically by combining information from all coordinates. Hence, in a certain sense, we are more concerned with inadmissibility issues than with admissibility.

Proving admissibility can often be difficult but there are a few basic techniques that can sometimes be applied with reasonable ease. The most basic is the following.

Lemma 2.6. *A unique (proper) Bayes estimator is admissible. (Here uniqueness is meant in the sense of probability 1 for all f_θ, $\theta \in \Omega$).*

Typically, a minimax estimator in a location parameter problem is not proper Bayes so Lemma 2.6 will be of little help in this setting. Some form of a technique due to Blyth [1951] is typically used to demonstrate admissibility in such problem. The techniques relies on the fact that, if $\delta_0(X)$ is inadmissible and $\delta'(X)$ is a better procedure, then for any (proper) prior distribution π, and corresponding Bayes estimator $\delta_\pi(X)$, the ratio

$$0 \leq \frac{r(\pi,\delta_0 - r(\pi,\delta')}{r(\pi,\delta_0) - r(\pi,\delta_\pi)} \leq 1 \tag{2.12}$$

since both numerator and denominator are nonnegative (the numerator since $\delta'(X)$ dominates $\delta_0(X)$ and the denominator since $\delta_\pi(X)$ is Bayes with respect to π) and the denominator is at least as large as the numerator (since $\delta_\pi(X)$ is Bayes).

To prove admissibility it suffices to find a sequence of priors for which the ratio in (2.12) approaches ∞. This is accomplished by finding a sequence of priors for which the denominator approaches zero at a faster rate than is possible for the numerator. If all risk functions are continuous (e.g. this is the case for exponential families and squared error loss), then $\mathscr{R}(\theta,\delta_0) - \mathscr{R}(\theta,\delta') > \varepsilon$, for $\varepsilon > 0$, in some η neighborhood of at least one point θ_0. We demonstrate the technique in the one dimensional normal case.

Theorem 2.1. *If $X \sim \mathcal{N}(\theta,1)$ and loss is $L(\theta,d) = (d-\theta)^2$, then $\delta(X) = X$ is admissible.*

Proof. Let (π_n) be a sequence of $\mathcal{N}(0,n)$ priors distributions. Then the risk of X is $\mathscr{R}(\theta,X) \equiv 1$ and, if $\delta'(X)$ is an estimator which dominates X by at least ε in a η neighborhood of θ_0, then the numerator of Blyth's ratio (2.12) is such that

$$\begin{aligned}
r(\pi_n,X) - r(\pi_n,\delta^t) &= \int_{-\infty}^{\infty} \frac{\mathscr{R}(\theta,X) - \mathscr{R}(\theta,\delta^t)}{\sqrt{2\pi n}} \exp\left(-\frac{\theta^2}{2n}\right) d\theta \\
&\leq \frac{\varepsilon}{\sqrt{2\pi n}} \int_{\theta_0-\eta/2}^{\theta_0+\eta/2} \exp\left(-\frac{\theta^2}{2n}\right) d\theta \\
&\leq \frac{\varepsilon\eta}{\sqrt{2\pi n}} (1-\gamma),
\end{aligned}$$

for some $n > n_0$ and some $0 < \gamma < 1$. On the other hand, the Bayes risk $r(\pi_n)$ is given by $n/(n+1)$ and hence the denominator of (2.12) is

$$1 - \frac{n}{n+1} = \frac{1}{n+1}.$$

Therefore, for the ratio in (2.12), we have

$$\frac{r(\pi_n, X) - r(\pi_n, \delta^t)}{r(\pi_n, X) - r(\pi_n, \delta_{\pi_n})} \geq \left(\frac{\varepsilon\, \eta\, (1-\gamma)}{\sqrt{2\,\pi\, n}} \right) \bigg/ \frac{1}{n+1} \longrightarrow \infty.$$

This contradiction establishes the admissibility of X. □

The above argument, for a sequence of normal priors with covariance equal to multiples of the identity, breaks down in 2-dimensions and completely fails in 3 and higher dimensions. An alternative sequence for 2-dimensions was found by James and Stein [1961] to demonstrate admissibility, while Stein [1956] showed inadmissibility of X for all $p \geq 3$. Brown [1966] showed the dimension cutoff of $p =$ 3 for inadmissibility of the best equivariant estimator ($\delta(X) = X$ in the spherically symmetric case) was quite general.

Brown [1971] and Brown and Hwang [1982] give conditions on the generalized prior which give admissibility under quadratic loss for normal and exponential families respectively. Brown's result [1971] resolves most of admissibility issues in the multivariate normal case (with σ^2 known). Here is a version of Brown's result.

Theorem 2.2. *Let $X \sim \mathscr{N}_p(\theta, \sigma^2 I)$. Suppose π is a generalized prior distribution and loss is $L(\theta, d) = ||d - \theta||^2$. Define, for $||x|| = r$,*

$$\bar{m}(r) = \int m(x)\, d\mathscr{U}_r(x)$$

and

$$\bar{m}(r) = \int (1/m(x))\, d\mathscr{U}_r(x)$$

where $\mathscr{U}_r$ *is the uniform distribution on the sphere of radius* r *and* $m(x)$ *is the marginal distribution.*

(1) *(Admissibility) If* $||\delta_\pi(x) - x||$ *is uniformly bounded and*

$$\int_c^\infty (r^{p-1}\,\bar{m}(r))^{-1}\, dr = \infty$$

for some $c > 0$*, then* $\delta_\pi(X)$ *is admissible.*

(2) *(Inadmissibility) If*

$$\int_c^\infty r^{1-p}\,\underline{m}(r)\, dr < \infty$$

for some $c > 0$*, then* $\delta_\pi(X)$ *is inadmissible.*

ADD BROWN AND HWUANG PROOF

2.5 Invariance

Large classes of problems are invariant under a variety of groups of transformations on the sample space $\mathscr{X}$ and associated groups of transformations acting on the parameter space Ω and the action space $\mathscr{D}$. In such cases, it seems natural to search for (optimal) procedures which behave in a manner consistent with the group structure. There is also a (generalized) Bayes connection, in that optimal procedures, when they exist, can be viewed as generalized Bayes estimators with respect to right invariant Haar measure which may be considered a natural “objective” prior. Almost all the problems considered in this book are invariant under the location or location-scale group (when σ^2 is unknown).

We give a brief discussion of some of the general theory. Suppose $X \in \mathscr{X} \sim P_\theta$ with $\theta \in \Omega$ and G is a group of $1-1$ transformations on $\mathscr{X}$. Let $\theta' = \bar{g}\theta$ and suppose that $\bar{g}$ is a $1-1$ transformations on Ω. This statistical problem is said invariant if $X \sim P\theta$ implies that $gX \sim P_{\bar{g}\theta}$, for all $g \in G$ and $\theta \in \Omega$. Note that $\bar{G} = \{\bar{g} \mid g \in G\}$ also forms a group of $1-1$ transformations on Ω.

As an example, suppose the distributions of X form a location parameter in $\mathbb{R}^p$ with density $f(x-\theta)$. The location group G in $\mathbb{R}^p$ consists of transformations of the form $g_a : x \mapsto x + a$ where $a \in \mathbb{R}^p$. If $X \sim f(x-\theta)$ then $X' = g_a(X) \sim f(x' - (\theta + a))$ so that $\bar{g}_a : \theta \mapsto \theta + a$. In this case, $\bar{G}$ and G essentially coincide although they act on different (but equivalent) spaces.

If the statistical problem is to estimate $h(\theta)$, under the loss function $L(\theta, d)$, the problem is said to be invariant if there is a g^* acting on the action space $\mathscr{D}$ corresponding to each $g \in G$ such that $L(\theta, d) = L(\bar{g}\theta, g^* d)$ for every $\theta \in \Omega$, $d \in \mathscr{D}$ and $g \in G$. In the above location problem, if $h(\theta) = \theta$, $L(\theta, d) = \rho(||d - \theta||^2)$, the transformation g^*_a corresponding to g_a is $g^*_a : d \mapsto d + a$, so that essentially $G = \bar{G} = G^*$.

An estimator, δ, is said to be equivariant if $\delta(gX) = g^*\delta(X)$ for all $g \in G$ and $X \in \mathscr{X}$. In the above location problem, this implies that $\delta(X + a) = \delta(X) + a$. Choosing $a = -X$, this implies $\delta(X) = X + \delta(0)$.

A key property of equivariant estimators in an invariant problem is the following.

Lemma 2.7. *If the problem is invariant and δ is equivariant, then the risk of δ is constant on orbits of $\bar{G}$, i.e. $\mathscr{R}(\bar{g}\theta,\delta) = \mathscr{R}(\theta,\delta)$ for all $\theta \in \Omega$ and $\bar{g} \in \bar{G}$. If the group $\bar{G}$ acting on Ω is transitive (i.e. for all $\theta_1, \theta_2 \in \Omega$, there exists $\bar{g} \in \bar{G}$ such that $\theta_2 = \bar{g}\theta_1$) then it follows that the risk of an equivariant estimator is constant on Ω.*

Proof. The lemma immediately follows from the equalities

$$\begin{aligned}
\mathscr{R}(\bar{g}\theta,\delta) &= E_{\bar{g}\theta}[L(\bar{g}\theta,\delta(X))] \\
&= E_{\bar{g}\theta}[L(\bar{g}\theta,\delta(gX))] \qquad \text{since } gX \sim P_{\bar{g}\theta} \\
&= E_{\bar{g}\theta}[L(\bar{g}\theta,g^*\delta(X))] \qquad \text{since } \delta \text{ is equivariant} \\
&= \mathscr{R}(\theta,\delta)\,.
\end{aligned}$$

□

The group $\bar{G}$ is transitive in the location problem since any $\theta_1, \theta_2 \in \mathbb{R}^p$, $\theta_2 = \theta_1 + (\theta_2 - \theta_1)$ so that $\bar{g}_{\theta_2-\theta_1}\theta_1 = \theta_2$.

The risk constancy of equivariant estimators gives hope of finding a best one, or minimum risk equivariant (MRE) estimator, since all that is required is the existence of an estimator that attains the infimum among the set of constant risks. The following lemma settles the issue for the location problem with quadratic loss; the proofs of these results are given in Lehmann and Casella [1998].

Lemma 2.8. (1) *For the multivariate location problem with loss $L(\theta,d) = ||d - \theta||^2$, the MRE $\delta_0(X)$ exists and is unique provided $E_0[||X||^2] < \infty$.*

(2) $\delta_0(X) = X - E_0[X]$

(3) $\delta_0(X) = \int_{\mathbb{R}^p} \theta\, f(X-\theta)\, d\theta \Big/ \int_{\mathbb{R}^p} f(X-\theta)\, d\theta$, *i.e.* δ_0 *is the generalized Bayes estimator with respect to Lebesgue measure on* $\mathbb{R}^p$ *(this is known as the Pitman estimator).*

(4) *The MRE coincides with the UMVUE of* θ *provided the UMVUE is equivariant.*

(5) *If the distribution of* X *is spherically symmetric about* θ *(*$X \sim f(||x-\theta||^2)$*), then* $\delta_0(X) = X$.

Things are somewhat simpler in the spherically symmetric case.

Lemma 2.9. *If* $X \sim f(||x-\theta||^2)$ *and* $L(\theta, d) = ||d-\theta||^2$ *then*

(1) $\delta_0(X) = X$ *is MRE;*

(2) *the MRE is also UMVUE provided the family of distributions is complete.*

For the location-scale family

$$(X,U) \sim \frac{1}{\sigma^{p+k}} f\left(\frac{||x-\theta||^2 + ||u||^2}{\sigma^2}\right),$$

the results for estimation of the parameter θ under loss $L((\theta,\sigma^2),d) = ||d-\theta||^2/\sigma^2$ is quite similar. In particular, the family is invariant under the group of transformations $g_{a,b,P}(x,u) = (a+bx, bPu)$ where $a \in \mathbb{R}^p$, $b > 0$, P orthogonal, is such that $\bar{g}_{a,b,P}(\theta,\sigma^2) = (a+b\theta, b^2\sigma^2)$, and thus $\overline{G}$ is transitive. Similarly, $g^*_{a,b,P}\, d = a + b d$ and $\delta(X,U)$ is equivariant if $\delta(a+bX, bPU) = a + b\,\delta(X,U)$.

Choosing P such that $Pu = (||u||, 0, \ldots, 0)^t$, $b = 1/||u||$, and $a = -x/||u||$, implies

$$\begin{aligned}\delta(X,U) &= \left[\frac{X}{||U||} + \delta(0,(1,0,\ldots,0)^t)\right] \Big/ \left[\frac{1}{||U||}\right] \\ &= X + c\,||U||\end{aligned}$$

where $c = \delta(0,(1,0,\ldots,0)^t) \in \mathbb{R}^p$ is arbitrary.

Lemma 2.10. *Suppose (X,U) has the density function*

$$\frac{1}{\sigma^{p+k}} f\left(\frac{||x-\theta||^2 + ||u||^2}{\sigma^2}\right)$$

and the loss is

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

Then

(1) $\delta_0(X) = X$ *is MRE and unbiased.*

(2) *The MRE is also the UMVUE provided the family of distributions is complete.*

(3) $\delta_0(X)$ *is generalized Bayes with respect to the right invariant prior on* $(\theta,\sigma^2) \in \mathbb{R}^p \times \mathbb{R}^+$*, that is,*

$$\delta_0(X) = \frac{\int\int \frac{\theta}{\sigma^2} \frac{1}{\sigma^{p+k}} f\left(\frac{||x-\theta||^2+||u||^2}{\sigma^2}\right) \frac{1}{\sigma^2}\, d\theta\, d\sigma^2}{\int\int \frac{1}{\sigma^2} \frac{1}{\sigma^{p+k}} f\left(\frac{||x-\theta||^2+||u||^2}{\sigma^2}\right) \frac{1}{\sigma^2}\, d\theta\, d\sigma^2}.$$

Minimaxity of the MRE of the location parameter in the location and location-scale families follows also from the so-called Hunt-Stein theorem since the location and location-scale groups are amenable. See Kiefer [1957], Robert [1994], Lehmann and Casella [1998], Bondar and Milnes [1981] and Eaton [1989], for details.

2.6 The general linear model

This section is devoted to the general linear model, its canonical form and the issues of estimation, sufficiency and completeness.

2.6.1 The canonical form of the general linear model

Much of this book will be devoted to some form of the following problem. Let $(X^t, U^t)^t$ be a partitioned random vector in $\mathbb{R}^n$ with a spherically symmetric distribution around a vector partitioned as $(\theta^t, 0^t)^t$, where $\dim X = \dim \theta = p$ and $\dim U = \dim 0 = k$ with $p + k = n$. Such a distribution arises from a fixed orthogonally invariant random vector $(X_0^t, U_0^t)^t$ and a fixed scale parameter σ through the transformation

$$(X^t, U^t)^t = \sigma\,(X_0^t, U_0^t)^t + (\theta^t, 0^t)^t\,, \tag{2.13}$$

so that the distribution of $((X-\theta)^t, U^t)^t$ is orthogonally invariant. We also refer to θ as a location parameter.

We will assume that the covariance matrix of $(X^t, U^t)^t$ exists, which is equivalent to the finiteness of the expectation $E[R^2]$ where $R = (\|X-\theta\|^2 + \|U\|^2)^{1/2}$ is its radius (in this case, we have $\mathrm{cov}(X^t, U^t)^t = E[R^2]\, I_n / n$ where I_n is the identity matrix). Then it will be convenient to assume that the radius $R_0 = (\|X_0\|^2 + \|U_0\|^2)^{1/2}$ of $(X_0^t, U_0^t)^t$ satisfies $E[R_0^2] = n$ since we have

$$\mathrm{cov}(X^t, U^t)^t = \sigma^2\,\mathrm{cov}(X_0^t, U_0^t)^t = \sigma^2\, I_n\,.$$

Note that, when it is assumed that the distribution in (2.13) is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^n$, the corresponding density may be represented as

$$\frac{1}{\sigma^n}\, g\!\left(\frac{\|z-\theta\|^2 + \|u\|^2}{\sigma^2}\right), \tag{2.14}$$

where g is the generating function.

This model also arises as the canonical form of the following seemingly more general model, the general linear model. For an $n \times p$ matrix V (often referred to as the design matrix and assumed here of full rank p), suppose that an $n \times 1$ vector Y is observed such that

$$Y = V\beta + \varepsilon, \tag{2.15}$$

where β is a $p \times 1$ vector of (unknown) regression coefficients and ε is an $n \times 1$ vector with a spherically symmetric distribution about 0. A common alternative representation of this model is $Y = \eta + \varepsilon$ where ε is as above and η is in the column space of V.

Using partitioned matrices, let $G = (G_1^t \; G_2^t)^t$ be an $n \times n$ orthogonal matrix partitioned such that the first p rows of G (*i.e.* the rows of G_1 considered as column vectors) span the column space of V. Now let

$$\begin{pmatrix} X \\ U \end{pmatrix} = GY = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} V\beta + G\varepsilon = \begin{pmatrix} \theta \\ 0 \end{pmatrix} + G\varepsilon \tag{2.16}$$

with $\theta = G_1 V\beta$ and $G_2 V\beta = 0$ since the rows of G_2 are orthogonal to the columns of V. It follows from the definition that $(X^t, U^t)^t$ has a spherically symmetric distribution about $(\theta^t, 0^t)^t$. In this sense, the model given in the first paragraph is the canonical form of the above general linear model.

This model has been considered by various authors such as Cellier, Fourdrinier and Robert [1989], Cellier and Fourdrinier [1995], Maruyama [2003], Maruyama and Strawderman [2005], Fourdrinier and Strawderman [2010]. Also, Kubokawa and Srivastava in [2001] adressed the multivariate case where θ is a mean matrix (in this case X and U are matrices as well).

2.6.2 Least squares, unbiased and shrinkage estimation

Consider the model in (2.16). Since the columns of G_1^t (the rows of G_1) and the columns of V span the same space, there exists a nonsingular $p \times p$ matrix A such that

$$V = G_1^t A, \text{ which implies } A = G_1 V, \tag{2.17}$$

since $G_1 G_1^t = I_p$. So

$$\theta = A\beta, \text{ that is, } \beta = A^{-1}\theta\,. \tag{2.18}$$

Noting that $V^t V = A^t\, G_1\, G_1^t A = A^t A$, it follows that estimation of θ by $\hat{\theta}(X,U)$ under loss

$$L(\theta,\hat{\theta}) = (\hat{\theta}-\theta)^t(\hat{\theta}-\theta) = \|\hat{\theta}-\theta\|^2 \tag{2.19}$$

is equivalent to estimation of β by

$$\hat{\beta}(Y) = A^{-1}\,\hat{\theta}(G_1Y,G_2Y) = (G_1V)^{-1}\,\hat{\theta}(G_1Y,G_2Y) \tag{2.20}$$

under loss

$$L^*(\beta,\hat{\beta}) = (\hat{\beta}-\beta)^t A^t A(\hat{\beta}-\beta) = (\hat{\beta}-\beta)^t V^t V(\hat{\beta}-\beta) \tag{2.21}$$

in the sense that the resulting risk functions are equal, that is,

$$R^*(\beta,\hat{\beta}) = E[L^*(\beta,\hat{\beta}(Y))] = E[L(\theta,\hat{\theta})] = R(\theta,\hat{\theta})\,.$$

Actually, the corresponding loss functions are equal. To see this, note that

$$\begin{aligned} L^*(\beta,\hat{\beta}(Y)) &= (\hat{\beta}(Y)-\beta)^t A^t A(\hat{\beta}(Y)-\beta) \\ &= (A(\hat{\beta}(Y)-\beta))^t (A(\hat{\beta}(Y)-\beta)) \\ &= (\hat{\theta}(X,U)-\theta)^t (\hat{\theta}(X,U)-\theta) \\ &= L(\theta,\hat{\theta}(X,U)), \end{aligned}$$

where (2.20) and (2.18) were used for the third equality.

Note that, clearly, the above equivalence between estimation of θ, the mean vector of X, and estimation the regression coefficients β holds for the respective invariant losses

$$L(\theta,\hat{\theta},\sigma^2) = \frac{1}{\sigma^2}(\hat{\theta}-\theta)^t(\hat{\theta}-\theta) = \frac{1}{\sigma^2}\|\hat{\theta}-\theta\|^2 \qquad (2.22)$$

and

$$L^*(\beta,\hat{\beta},\sigma^2) = \frac{1}{\sigma^2}(\hat{\beta}-\beta)^t A^t A(\hat{\beta}-\beta) = \frac{1}{\sigma^2}(\hat{\beta}-\beta)^t V^t V(\hat{\beta}-\beta)\,. \qquad (2.23)$$

Note also that the correspondence (2.20) can be reversed as

$$\hat{\theta}(X,U) = A\,\hat{\beta}(G_1^t X + G_2^t U) = G_1 X\,\hat{\beta}(G_1^t X + G_2^t U) \qquad (2.24)$$

since, according to (2.16),

$$Y = G^t \begin{pmatrix} X \\ U \end{pmatrix} = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}^t \begin{pmatrix} X \\ U \end{pmatrix} = (G_1^t\, G_2^t) \begin{pmatrix} X \\ U \end{pmatrix} = G_1^t X + G_2^t U\,. \qquad (2.25)$$

There is also a correspondence between estimation of θ and estimation of η in the following alternative representation of the general linear model. Here

$$\eta = G^t \begin{pmatrix} \theta \\ 0 \end{pmatrix} = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}^t \begin{pmatrix} \theta \\ 0 \end{pmatrix} = (G_1^t \, G_2^t) \begin{pmatrix} \theta \\ 0 \end{pmatrix} = G_1^t \, \theta + G_2^t \, 0 = G_1^t \, \theta$$

and also

$$G_1 \eta = G_1 \, G_1^t \, \theta = \theta \, .$$

It follows that estimation of $\theta \in \mathbb{R}^p$ by $\hat{\theta}(X,U)$ under loss $\|\hat{\theta} - \theta\|^2$ (that is, loss (2.19)) is equivalent to estimation of η in the column space of V under loss $\|\hat{\eta} - \eta\|^2$ by

$$\hat{\eta}(Y) = G_1^t \, \hat{\theta}(G_1 Y, G_2 Y) \tag{2.26}$$

in the sense that the risks functions are equal. The easy demonstration is left to the reader.

Consider the first correspondence expressed in (2.20) and (2.24) between estimators in Models (2.15) and (2.16). We will see that it can be made completely explicit for a wide class of estimators. First, note that the matrix G_1 can be easily obtained by the Gram-Schmidt orthonormalization process or by the QR decomposition of the design matrix X, where Q is an orthogonal matrix such that $Q^t V = R$ and R is a $n \times p$ upper triangular matrix (so that $G_1 = Q_1^t$ and $G_2 = Q_2^t$. Secondly, a particular choice of A can be made which gives rise to a close form of G_1.

To see this, let

$$A = (V^t V)^{1/2} \tag{2.27}$$

(one of the square root of $V^t V$, which is invertible since V has full rank) and set

$$G_1 = A \, (V^t V)^{-1} V^t = (V^t V)^{-1/2} V^t \, . \tag{2.28}$$

Then we have

$$G_1 V = A \quad \text{and} \quad V = G_1^t A, \tag{2.29}$$

in addition,

$$G_1 G_1^t = (V^t V)^{-1/2} V^t V (V^t V)^{-1/2} = I_p \,. \tag{2.30}$$

Hence, as in (2.17), Equality (2.29) expresses that the columns of G_1^t (the rows of G_1) span the same space as the columns of V, noticing that (2.30) means that these vectors are orthogonal. Therefore, completed G_1^t through the Gram-Schmidt orthonormalization process, we obtain an orthogonal matrix $G = (G_1^t G_2^t)^t$, with G_1 in (2.28), such that

$$G V = \begin{pmatrix} (V^t V)^{1/2} \\ 0 \end{pmatrix} . \tag{2.31}$$

The relationship linking A and G_1 in (2.28) is an alternative to (2.17) and is true in general, that is,

$$G_1 = A \, (V^t V)^{-1} \, V^t \text{ or equivalently } A = (V^t G_1^t)^{-1} V^t V \,.$$

Indeed, as noticed above, we have $V^t V = A^t A$ so that $(V^t V)^{-1} \, (A^t A) = I_p$ and hence $(V^t V)^{-1} A^t = A^{-1}$, which implies $V \, (V^t V)^{-1} A^t = V A^{-1} = G_1^t A A^{-1} = G_1^t$, according to (2.17).

As a consequence, if $\hat{\beta}_{ls}$ is the least squares estimator of β, we have

$$\hat{\beta}_{ls}(Y) = (V^t V)^{-1} \, V^t \, Y \tag{2.32}$$

so that the corresponding estimator $\hat{\theta}_0$ of θ is the projection $\hat{\theta}_0(X, U) = X$ since

$$\hat{\theta}_0(X,U) = A\,\hat{\beta}_{ls}(Y) = A\,(V^t V)^{-1} V^t\,Y = G_1\,Y = V\,. \tag{2.33}$$

From this correspondence, the estimator $\hat{\theta}_0(X,U) = X$ of θ is often viewed as the standard estimator. Note that, with the choice of A in (2.27), we have the close form

$$\hat{\beta}_{ls}(Y) = (V^t V)^{-1/2} X\,. \tag{2.34}$$

Furthermore, the correspondence between $\hat{\theta}(X,U)$ and $\hat{\beta}_{ls}(Y)$ can be specified when $\hat{\theta}(X,U)$ depends on U only through $\|U\|^2$, in which case, with a slight abuse of notation, we write $\hat{\theta}(X,U) = \hat{\theta}(X,\|U\|^2)$. Indeed, first note that

$$\begin{aligned}
\|X\|^2 &= V^t V \\
&= (A\,\hat{\beta}_{ls}(Y))^t (A\,\hat{\beta}_{ls}(Y)) \\
&= (\hat{\beta}_{ls}(Y))^t\,A^t A\,(\hat{\beta}_{ls}(Y)) \\
&= (\hat{\beta}_{ls}(Y))^t\,V^t V\,(\hat{\beta}_{ls}(Y)) \\
&= (V\,\hat{\beta}_{ls}(Y))^t V\,(\hat{\beta}_{ls}(Y)) \\
&= \|V\,\hat{\beta}_{ls}(Y)\|^2\,.
\end{aligned} \tag{2.35}$$

On the other hand, according to (2.16), we have $\|X\|^2 + \|U\|^2 = \|GY\|^2 = \|Y\|^2$ and hence

$$\begin{aligned}
\|U\|^2 &= \|Y\|^2 - \|X\|^2 \\
&= \|Y\|^2 - \|V\,\hat{\beta}_{ls}(Y)\|^2 \\
&= \|Y - V\,\hat{\beta}_{ls}(Y)\|^2
\end{aligned} \tag{2.36}$$

since $Y - V\,\hat{\beta}_{ls}(Y)$ is orthogonal to $V\,\hat{\beta}_{ls}(Y)$. Consequently, according to (2.20) and (2.17), Equations (2.36) and (2.33) give that the estimator $\hat{\beta}(Y)$ of β corresponding to the estimator $\hat{\theta}(X, \|U\|^2)$ of θ is

$$\hat{\beta}(Y) = (G_1\,V)^{-1}\,\hat{\theta}(G_1\,V\,\hat{\beta}_{ls}(Y), \|Y - V\,\hat{\beta}_{ls}(Y)\|^2)\,. \tag{2.37}$$

Note that, when ones chooses G_1 as in (2.28), $\hat{\beta}(Y)$ in (2.37) has the closed form

$$\begin{aligned}\hat{\beta}(Y) &= (V^t\,V)^{-1/2}\,\hat{\theta}\Big((V^t\,V)^{1/2}\,\hat{\beta}_{ls}(Y), \|Y - V\,\hat{\beta}_{ls}(Y)\|^2\Big)\\ &= (V^t\,V)^{-1/2}\,\hat{\theta}\Big((V^t\,V)^{-1/2}V^tY, \|Y - V\,\hat{\beta}_{ls}(Y)\|^2\Big)\,.\end{aligned} \tag{2.38}$$

In particular, we can see, through (2.35), that the robust Stein-type estimators of θ, that is,

$$\hat{\theta}_r(X, \|U\|^2) = \left(1 - a\,\frac{\|U\|^2}{\|X\|^2}\right) X \tag{2.39}$$

(robust since, for appropriate values of the positive constant a, they dominate X whatever the spherically symmetric distribution is, as we will see in Chapter 5 (Cellier, Fourdrinier and Robert [1989] and Cellier and Fourdrinier [1995]) have as correspondence the robust estimators of β

$$\begin{aligned}\hat{\beta}_r(Y) &= (G_1\,V)^{-1}\left(1 - a\,\frac{\|Y - V\,\hat{\beta}_{ls}(Y)\|^2}{\|V\,\hat{\beta}_{ls}(Y)\|^2}\right) G_1\,V\,\hat{\beta}_{ls}(Y)\\ &= \left(1 - a\,\frac{\|Y - V\,\hat{\beta}_{ls}(Y)\|^2}{\|V\,\hat{\beta}_{ls}(Y)\|^2}\right) \hat{\beta}_{ls}(Y)\end{aligned} \tag{2.40}$$

(note that the two $G_1\,V$ terms simplify). According to the correspondence seen above between the risk functions of the estimators of θ and the estimators of β, using these estimators in (2.40) is then a good alternative to the least squares estimator:

they dominate the least squares estimator of β simultaneously for all spherically symmetric error distributions with a finite second moment (see also Fourdrinier and Strawderman [1996] for the use of these robust estimators when σ^2 is known).

2.6.3 Sufficiency in the general linear model

Suppose $(X^t, U^t)^t$ has a spherically symmetric distribution about $(\theta^t, 0^t)^t$ with $\dim X = \dim \theta = p > 0$ and $\dim U = \dim 0 = k > 0$. Furthermore suppose that the distribution is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^n$ $(n = p + k)$. The corresponding density may be represented as in (2.14) We refer to θ as a location vector and to σ as a scale parameter. Such a distribution arises from a fixed orthogonally invariant random vector $(X_0^t, U_0^t)^t$ with generating function g through the transformation

$$\begin{pmatrix} X \\ U \end{pmatrix} = \sigma \begin{pmatrix} X_0 \\ U_0 \end{pmatrix} + \begin{pmatrix} \theta \\ 0 \end{pmatrix}.$$

Each of θ, σ^2 and $g(\cdot)$ may be known or unknown. The most interesting cases from a statistical standpoint are the following.

Suppose θ and σ^2 are unknown and $g(\cdot)$ is known. It follows immediately from the factorization theorem that $(X, \|U\|^2)$ is sufficient. It is intuitively clear that this statistic is also minimal sufficient since $\dim(X, \|U\|^2) = \dim(\theta, \sigma^2)$. Here is a proof of that fact.

Theorem 2.3. *Suppose that $(X^t, U^t)^t$ is distributed as (2.14). Then the statistic $(X, \|U\|^2)$ is minimal sufficient for (θ, σ^2) when g is known.*

Proof. By Theorem 6.14 of Casella and Berger [2001], it suffices to show that if, for all $(\theta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+$,

$$\frac{g\left(\frac{\|x_1-\theta\|^2+\|u_1\|}{\sigma^2}\right)}{g\left(\frac{\|x_2-\theta\|^2+\|u_2\|^2}{\sigma^2}\right)} = c \tag{2.41}$$

where c is a constant then $x_1 = x_2$ and $\|u_1\|^2 = \|u_2\|^2$. Note that $0 < c < \infty$ since otherwise (2.14) cannot be a density.

Letting $\tau^2 = 1/\sigma^2$, (2.41) can be written, for all $\tau > 0$, as

$$g(\tau^2 v_1^2) = c g(\tau^2 v_2^2) \tag{2.42}$$

where $v_1^2 = \|x_1 - \theta\|^2 + \|u_1\|^2$ and $v_2^2 = \|x_2 - \theta\|^2 + \|u_2\|^2$ for each fixed $\theta \in \mathbb{R}^p$. First we will show that $v_1^2 = v_2^2$. Note that

$$\begin{aligned} 1 &= \int_{\mathbb{R}^p \times \mathbb{R}^k} g(\|x\|^2 + \|u\|^2)\, dx\, du \\ &= K \int_0^\infty r^{p+k-1} g(r^2)\, dr && (cf.\text{ Theorem 1.2}) \\ &= K v^{p+k} \int_0^\infty \tau^{p+k-1} g(v^2 \tau^2)\, d\tau && (2.43) \end{aligned}$$

for any $v > 0$. Then it follows from (2.42) and (2.43) that

$$\begin{aligned} 1 &= K v_1^{p+k} \int_0^\infty \tau^{p+k-1} g(v_1^2 \tau^2)\, d\tau \\ &= c K v_1^{p+k} \int_0^\infty \tau^{p+k-1} g(v_2^2 \tau^2)\, d\tau \\ &= c \frac{v_1^{p+k}}{v_2^{p+k}}. \end{aligned} \tag{2.44}$$

Let $F(b) = \int_0^b \tau^{p+k-1} g(\tau^2)\, d\tau$ and choose b such that F is strictly increasing at b. Suppose $v_1 > v_2$. Then, for any $v > 0$,

$$F(b) = v^{p+k} \int_0^{b/v} \tau^{p+k-1} g(v^2\tau^2)\, d\tau$$

and hence

$$\begin{aligned}\int_0^{b/v_1} \tau^{p+k-1} g(v_1^2\tau^2) d\tau &= \frac{F(b)}{v_1^n} \\ &= c\int_0^{b/v_1} \tau^{p+k-1} g(v_2^2\tau^2) d\tau \\ &< c\int_0^{b/v_2} \tau^{p+h-1} g(v_2^2\tau^2) d\tau \\ &= c\,\frac{F(b)}{v_2^n}.\end{aligned}$$

It follows that $c\,v_1^{p+k}/v_2^{p+h} > 1$ which contradicts (2.44). A similar argument would give $c\,v_1^{p+k}/v_2^{p+h} < 1$ for $v_1 < v_2$ and hence $v_1 = v_2$. Now setting $\theta = \frac{x_1+x_2}{2}$ in the expressions for v_1 and v_2 implies $\|u_1\|^2 = \|u_2^2\|$. It then follows that $\|x_1 - \theta\|^2 = \|x_2 - \theta\|^2$ for all $\theta \in \mathbb{R}^p$ which implies $x_1 = x_2$ by setting $\theta = x_2$ (or x_1). □

In the case where θ is unknown and σ^2 is known, and the distribution is multivariate normal, X is minimal sufficient (and complete). However, in the non-normal case $(X, \|U\|^2)$ is typically minimally sufficient.

2.6.4 Completeness for the general linear model

In the case where both θ and σ^2 are unknown and g is known, the minimal sufficient statistic $(X, \|U\|^2)$ can be either complete or incomplete depending on g.

If g corresponds to a normal distribution the statistic is complete by standard results for exponential families. However, when the generating function is of the form $K g(t) \mathbb{1}_{(r_1,r_2)}(t)$ with $0 < r_1 < r_2 < \infty$ where K is the normalizing constant, $(X, \|U\|^2)$ is not complete. In fact incompleteness of $(X, \|U\|^2)$ follows from the fact that the minimal sufficient statistic, when θ is known, σ^2 is unknown and g is known, is incomplete. We provide details below. To show this note the following result.

Lemma 2.11.(1) *If $X \sim f(x-\theta)$ with $\theta \in \mathbb{R}^p$ where f has compact support, then X is not complete for θ.*

(2) *If $X \sim 1/\sigma f(x/\sigma)$, where f has support contained in an interval $[a,b]$ with $0 < a < b < \infty$, then X is not complete for σ.*

Before giving the proof of Lemma 2.11, note that, if the generating function is of the form $K g(t) \mathbb{1}_{[r_1,r_2]}(t)$ for $0 < r_1 < r_2 < \infty$ and the value of θ is assumed to be known and equal to θ_0, then $T = \|X - \theta_0\|^2 + \|U\|^2$ is minimal sufficient and has density of the form $K/\sigma^{p+k} t^{(p+k)/2} g(t/\sigma^2) \mathbb{1}_{[r_1\sigma^2, r_2\sigma^2]}(T)$.

Therefore T is not a complete statistic for σ^2 by Lemma 2.11 (2). It follows that there exists a function $h(\cdot)$, not equal to zero a.e. such that $E_\sigma[h(T)] = 0$ for all $\sigma > 0$. Since $E_{\sigma^2}[h(\beta T) = E_{\beta\sigma^2}[h(T)]$, it follows that $E_{\sigma^2}[h(\beta T)] = 0$ for all $\sigma^2 > 0, \beta > 0$, and also $M(t) = \int_0^1 E_{\sigma^2}[h(\beta t)] m(\beta) d\beta = 0$ for any function $m(\cdot)$ for which the integral exists. In particular, this holds when $m(\cdot)$ is the density of Beta $(k/2, p/2)$ random variable (where finiteness of the integral is guaranteed since

$E_{\sigma^2}[h(\beta T)$ is continuous in β). Now, since $B = \|U|^2/T$ has a Beta $(k/2, p/2)$ distribution, $\|U\|^2 = BT$, and $M(\sigma^2) = E_{\sigma^2}[h(BT) = E_{\sigma^2}[h(\|U\|^2)] \equiv 0$.

Since the distribution of $\|U\|^2$ does not depend on θ, it follows that, when both θ and σ^2 are unknown, $E_{\theta,\sigma^2}[h(\|U\|^2)] \equiv 0$ and hence $(X, \|U\|^2)$, while minimal sufficient, is not complete for the case of a generating function of the form $g(t)\mathbb{1}_{[r_1,r_2]}(T)$ with $0 < r_1 < r_2 < \infty$.

Note that whenever θ is unknown and σ^2 is known and $(X, \|U\|^2)$ is minimal sufficient (so the distribution is not normal, since then X would be minimal sufficient) $\|U\|^2$ is ancillary and hence the minimal sufficient statistic is not complete.

Proof of Lemma 2.11. First, note that part (2) follows from part (1) by the standard technique of transforming a scale family to a location family by taking logs.

We will show the incompleteness of a location family when F has bounded support. Note that, if F is a cdf with bounded support contained inside $[a,b]$, the c.f. $\hat{f}$ is analytic in $\mathbb{C}$ (the entire complex plane) and is of order 1 (i.e., $|\hat{f}(\eta)|$ is $O\exp(r^{1+\varepsilon})$ for all $\varepsilon > 0$ and is not $O(\exp(r^{1-\varepsilon})$ for any $\varepsilon > 0)$.

Without loss of generality assume $0 < a < b < \infty$. Then

$$\begin{aligned}
|\hat{f}(\eta)| &\leq \int_a^b \exp(|\eta|X)\,dF(x)\\
&\leq \exp(b|\eta|)\int_a^b dF(x)\\
&= \exp(b|\eta|)\\
&= O(\exp(|\eta|^{1+\varepsilon})).
\end{aligned}$$

for all $\varepsilon > 0$. Also, if $\eta = -iv$ for $v > 0$, then

$$|\hat{f}(\eta)| = \int_a^b \exp(vx)\,dF(x) \geq \exp(av)\int_a^b dF(x) = \exp(av)\,.$$

But, $\exp(av)$ is not $O(\exp(v^{1-\varepsilon})$ for any $\varepsilon > 0$. Hence $\hat{f}(\eta)$ is of order 1.

In the step above, we used $0 < a < b < \infty$. Note that if either a and/or b is negative then the distribution of X is equal to the distribution of $\mathrm{z} + \theta_0$ where θ_0 is negative and where the distribution of z satisfies the assumptions of the theorem. Hence

$$E\exp(i\eta x) = E\exp(i\eta z)e^{i\eta\theta_0},\text{ so } |E\exp(i\eta x)| \leq \exp(|\eta|b)\exp(|i\eta||\theta_0|).$$

which is $O\exp(|\eta|^{1+\varepsilon})$ for all $\varepsilon > 0$.

Similarly, for $\eta = -iv$ (recall $\theta_0 < 0$),

$$\begin{aligned}|E\exp(i\eta x)| &= E\exp(tvz)\exp(-v\theta_0)\\ &\geq e^{v|\theta_0|}\exp(av)\\ &= \exp(v(a+|\theta_0|)\end{aligned}$$

and this is not $O(\exp^{v^{1-\varepsilon}})$ for any $\varepsilon > 0$.

Note that $\hat{f}(\eta)$ exists in all of $\mathbb{C}$ since F has bounded support and then is analytic by standard results in complex analysis (See e.g. Rudin).

Theorem 2.4. *If $X \sim F(x)$ where the cdf F has bounded support and F is not degenerate, then the characteristic function $\hat{f}(\eta)$ has at least one zero in $\mathbb{C}$.*

Proof. This follows almost directly from the Hadamard factorization theorem which implies that a function $\hat{f}(z)$ which is analytic in all of $\mathbb{C}$ and of order 1 is of the form $\hat{f}(z) = \exp(az+b)P(z)$, where $P(z)$ is the so called canonical product formed from the zeros of $\hat{f}(z)$, and where $P(0) = 1$ and when $P(z) = 0$ for each such root. (See e.g., Titchmarsh for an extended discussion of the form of $P(z)$). Therefore either $\hat{f}(z)$ has no zeros, in which case $\hat{f}(z) = \exp(az)$ (since $\hat{f}(0) = 1 = e^b \Rightarrow b = 0$) and $P(z) \equiv 1$, or $\hat{f}(z)$ has at least one zero. The case where $\hat{f}(z) = \exp(az)$ corresponds to the degenerate case where $\exp(az) = \hat{f}(z) = E\exp(izx)$ with $P[X = -ia] = 1$. Since F is assumed not to be degenerate $\hat{f}(z)$ must have at least 1 zero by the unicity of the Fourier transform. □

Theorem 2.5. *Suppose $X \sim f(x-\theta)$ where $f(\cdot)$ has bounded support. Then X is not a complete statistic for θ (alternatively, $f(x-\theta)$ is not complete.)*

Proof. By Theorem 2.4, there exists η_0 such that

$$\hat{f}(\eta_0) = \int_{-\infty}^{\infty} \exp(i\eta_0 x) f(x)\, dx = 0.$$

This implies that for any $\theta \in \mathbb{R}$,

$$\begin{aligned}
0 &= \left(\int_{-\infty}^{\infty} \exp(i\eta_0 x) f(x)\, dx\right) \exp^{i\eta_0\theta} \\
&= \int_{-\infty}^{\infty} \exp(ix(\eta_0+\theta)) f(x)\, dx \\
&= \int_{-\infty}^{\infty} \exp(i\eta_0 x) f(x-\theta)\, dx \\
&= E_\theta[\exp(i\eta_0 X) = E_\theta[\exp(i(a_0+b_0 i))X] \\
&= E_\theta \exp(i\eta_0 X)\exp(-b_0 X)] \\
&= E_\theta[\exp(-b_0 X)\{\cos a_0 x + i \sin a_0 x\}]
\end{aligned}$$

Hence, for any $\theta \in \mathbb{R}$, we have $E_\theta[\exp(-b_0 x)\cos(a_0 x)] \equiv 0$.

Additionally $E_\theta[|\exp(-b_0x)\cos(a_0x)|] < \infty$ for all θ since $f(\cdot)$ has bounded support. The theorem then follows, since $h(X) = e^{-b_0X}\cos a_0X$ is an unbiased estimator of 0 which is not equal to 0 almost surely for each θ. □

Chapter 3

Estimation of a normal mean vector I

3.1 Introduction

This chapter is concerned with estimating the p-dimensional mean vector of a multivariate normal distribution under quadratic loss. Most of the chapter will be concerned with the case of a known covariance matrix of the form $\Sigma = \sigma^2 I_p$ and "usual quadratic loss," $L(\theta, \delta) = \|\delta - \theta\|^2 = (\delta - \theta)^t(\delta - \theta)$. Generalizations to known general covariance matrix Σ, and to general quadratic loss, $L(\theta, \delta) = (\delta - \theta)^t Q(\delta - \theta)$, where Q is a $p \times p$ symmetric non-negative definite matrix will also be considered.

Let $X \sim N_p(\theta, \sigma^2 I_p)$ where σ^2 is assumed known and it is desired to estimate the unknown vector $\theta \in \mathbb{R}^p$. The "usual" estimator of θ is $\delta_0(X) = X$, in the sense that it is the Maximum Likelihood Estimator (MLE), the Uniformly Minimum Variance Unbiased Estimator (UMVUE), the Least Squares Estimator (LSE), and under a wide variety of loss functions it is the Minimum Risk Equivariant Estimator (MRE), and is minimax. The estimator $\delta_0(X)$ is also admissible under a wide class of loss functions if $p = 1$ or 2. However Stein [1956] showed that X is inadmissible if $p \geq 3$ for the loss $L(\theta, \delta) = \|\delta - \theta\|^2$. This result was surprising at the time and has led to a large number of developments in multi-parameter estimation. One important aspect of this "Stein phenomenon" (also known as the Stein paradox, see Efron and Morris [1977]) is that it illustrates the difference between estimating one component at a time and simultaneously estimating the whole mean vector. Indeed if we wish to estimate any particular component, θ_i, of the vector θ, then the estimator $\delta_{0i}(X) = X_i$ remains admissible whatever the value of p (see for example Lehmann and Casella [1998], Lemma 5.2.12). James and Stein [1961] showed that the estimator $\delta_a^{JS}(X) = (1 - a\,\sigma^2/\|X\|^2)$ dominates $\delta_0(X)$ for $p \geq 3$ provided $0 < a < 2\,(p-2)$. They also showed that the risk of $\delta_{p-2}(X) = \big(1 - (p-2)\,\sigma^2/\|X\|^2\big)\,X$ at $\theta = 0$ is equal to $2\,\sigma^2$ for all $p \geq 3$ indicating that substantial gain in risk over the usual estimator is possible for large p, since the risk of $\delta_0(X)$ is equal to the constant $p\,\sigma^2$.

In Section 3.2, we will give some intuition into why improvement over $\delta_0(X)$ should be possible in higher dimensions and how much improvement might be expected. Section 3.3 is devoted to Stein's unbiased estimation of risk technique which provides the technical basis of many results in the area of multi-parameter estimation. Section 3.4 is devoted to establishing improved procedures such as the James-Stein estimator.

In Section 3.5, we will provide a link between Stein's lemma and Stokes' theorem while, in Section 3.6, we will give an insight into the reason of Stein's phenomenon in terms of non linear partial differential operators.

3.2 Some intuition into Stein estimation

3.2.1 *Best linear estimators*

Suppose X is a p-dimensional random vector such that $E[X] = \theta$ and $Cov(X) = \sigma^2 I$ where θ is unknown and σ^2 is known. We do not require at this point that X have a normal distribution. Consider estimators of θ of the form $\delta_a(X) = (1-a)X$ under quadratic loss $L(\theta,\delta) = \|\delta - \theta\|^2 = \sum_{i=1}^p (\delta_i - \theta_i)^2$. The risk of $\delta_a(X)$ is given by

$$\begin{aligned}
R(\theta,\delta_a) &= E\left[\sum_{i=1}^p ((1-a)X_i - \theta_i)^2\right] \\
&= \sum_{i=1}^p Var((1-a)X_i) + \sum_{i=1}^p (E[(1-a)X_i - \theta_i])^2 \\
&= (1-a)^2 p\sigma^2 + a^2 \sum_{i=1}^p \theta_i^2 \\
&= (1-a)^2 p\sigma^2 + a^2 \|\theta\|^2 .
\end{aligned}$$

The optimal choice of a, a_{opt}, which minimizes $R(\theta,\delta_a)$ is obtained by differentiating $R(\theta,\delta_a)$ with respect to a and equating the result to 0, i.e.

$$\begin{aligned}
\frac{\partial}{\partial a} R(\theta,\delta_a) &= -2(1-a)p\sigma^2 + 2a\|\theta\|^2 \\
&= 0
\end{aligned}$$

or

$$a_{opt} = \frac{p\sigma^2}{p\sigma^2 + \|\theta\|^2} .$$

We see that a_{opt} depends on the unknown θ but, since

$$E\|X\|^2 = p\sigma^2 + \|\theta\|^2 ,$$

we may estimate a_{opt} as

$$\hat{a}_{opt} = \frac{p\sigma^2}{\|X\|^2},$$

and hence approximate the best linear "estimator"

$$\delta_{a_{opt}}(X) = \left(1 - \frac{p\sigma^2}{p\sigma^2 + \|\theta\|^2}\right) X$$

by

$$\hat{\delta}_{a_{opt}}(X) = \left(1 - \frac{p\sigma^2}{\|X\|^2}\right) X.$$

This is in fact a James-Stein type estimator

$$\hat{\delta}_{a_{opt}}(X) = \delta_p^{JS}(X)$$

which is close to the optimal James-Stein estimator (as we will see in Section 3.4 $\delta_{p-2}^{JS}(X)$ is optimal if X is normal). Hence the James-Stein estimator can be viewed as an approximation to the best linear "estimator" which adapts to the value of $\|\theta\|^2$.

It is worth noting that $a_{opt} = p\sigma^2/(p\sigma^2 + \|\theta\|^2)$ can typically be better estimated for large values of p since $E\|X\|^2/p = \sigma^2 + \|\theta\|^2/p$ and (if we assume X_i are symmetric about θ_i and that the $X_i - \theta_i$ are independent)

$$Var\left(\frac{\|X\|^2}{p}\right) = \frac{Var(X_1 - \theta_1)^2}{p} + \frac{4\|\theta\|^2\sigma^2}{p^2}$$

which tends to 0 uniformly as $p \to \infty$ provided $\|\theta\|^2/p$ is bounded. This helps to explain why there is a dimension effect and that it is easier to find dominating estimators for large p.

It is also interesting to note that normality plays no role in the above discussion indicating that we can expect James-Stein type estimators to improve on $\delta_0(X)$ in a fairly general location vector setting.

Note also, since the estimators are generally shrinking X toward 0, we expect the largest gains in risk to occur at $\theta = 0$. In particular the risk of $\delta_{a_{opt}}(X)$ at the true value of θ is given by

$$\begin{aligned} R(\theta, \delta_{a_{opt}}) &= \frac{p\sigma^2 \|\theta\|^2}{p\sigma^2 + \|\theta\|^2} \\ &= R(\theta, X) \frac{\|\theta\|^2}{p\sigma^2 + \|\theta\|^2}. \end{aligned}$$

Hence when $\|\theta\|^2$ is large, there is very little savings in risk, but when $\|\theta\|^2$ is close to 0, the savings is substantial.

We will see later (in Section 3.4) that this is also true for the James-Stein estimator in the sense that there is very little saving in risk for large $\|\theta\|^2$ but substantial savings for small $\|\theta\|^2$ and especially for large p.

3.2.2 Some geometrical insight

The argument here follows closely the discussion presented by Brandwein and Strawderman in [1991]. We again suppose $E[X] = \theta \in \mathbb{R}^p$ and $Cov(X) = \sigma^2 I_p$ with σ^2 known. Since $E[\|X\|^2] = \|\theta\|^2 + p\sigma^2$ it seems that X is too long as an estimator of θ and that perhaps the projection of θ onto X or something close to it would be a better estimator than X. Again, the projection of θ onto X will depend on θ and so will not be a valid estimator, but perhaps we can find a reasonable approximation.

Since the projection of θ on X has the form $(1-a)X$ we are trying to approximate the constant a. Note $E(\theta - X)^t\theta = 0$, and hence we expect θ and $X - \theta$ to be nearly orthogonal which implies that we expect $0 < a < 1$.

In what follows, we assume that θ and $X - \theta$ are exactly orthogonal. The situation is shown in Figure 1.

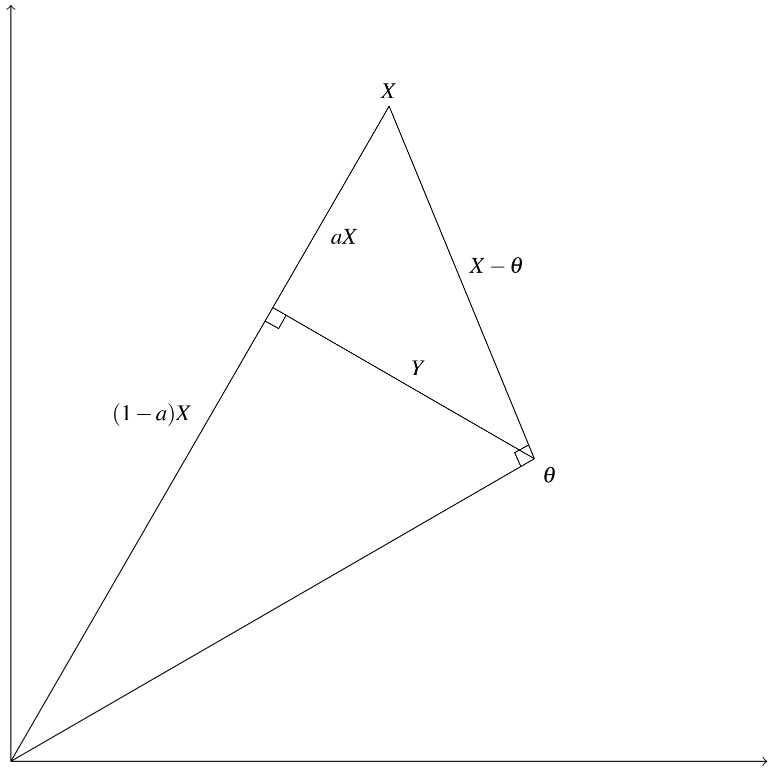

Fig. 3.1 Observation vector X in p dimensions with mean θ orthogonal to $X - \theta$

From the two right triangles in Figure 3.1 we note

$$\|(1-a)X\|^2 + \|Y\|^2 = \|\theta\|^2 \quad \text{and} \quad \|aX\|^2 + \|Y\|^2 = \|X-\theta\|^2.$$

Since

$$E\|X\|^2 = \|\theta\|^2 + p\sigma^2 \quad \text{and} \quad E\|X-\theta\|^2 = p\sigma^2\,,$$

reasonable approximations are

$$\|\theta\|^2 \cong \|X\|^2 - p\sigma^2 \quad \text{and} \quad \|X-\theta\|^2 \cong p\sigma^2\,.$$

Hence we have as approximations

$$\|(1-a)X\|^2 + \|Y\|^2 \cong \|X\|^2 - p\sigma^2 \quad \text{and} \quad \|aX\|^2 + \|Y\|^2 \cong p\sigma^2\,.$$

Subtracting to eliminate $\|Y\|^2$, that is,

$$\|(1-a)X\|^2 - \|aX\|^2 = (1-2a)\|X\|^2 \cong \|X\|^2 - 2p\sigma^2\,,$$

we obtain $a \cong p\sigma^2/\|X\|^2$. Hence we may approximate the projection of θ on X as

$$(1-a)X \cong \left(1 - \frac{p\sigma^2}{\|X\|^2}\right) X = \delta_p^{JS}(X)\,,$$

i.e. the same James-Stein estimator suggested in Subsection 3.2.1. Once again, note that normality plays no role in the discussion.

3.2.3 *The James-Stein estimator as an empricial Bayes estimator*

Assume in this subsection that $X \sim N_p(\theta, \sigma^2 I_p)$ (with σ^2 known) and that the prior distribution on θ is $N_p(0, \tau^2 I_p)$. As indicated in Section 2.2, the Bayes estimator of θ for quadratic loss is the posterior mean of θ given by $\delta(X) = E[\theta \mid X] = (1 - \sigma^2/(\tau^2 + \sigma^2))X$.

If we now assume that τ^2 is unknown we can derive an empirical Bayes estimator as follows; the marginal distribution of X is $N_p(0,(\sigma^2+\tau^2)I_p)$ and hence $\|X\|^2$, which is distributed as $(\sigma^2+\tau^2)$ times a chi-square with p degrees of freedom, is a complete sufficient statistics for $\sigma^2+\tau^2$. It follows that $(p-2)/\|X\|^2$ is the UMVUE of $1/(\sigma^2+\tau^2)$ and hence that $\delta^{JS}_{p-2}(X)=(1-(p-2)\,\sigma^2/\|X\|^2)\,X$ can be viewed as an empirical Bayes estimator of θ.

Here we have explicitly used the assumption of normality but a somewhat analogous argument can be given for a general multivariate location family.

3.3 Improved estimators via Stein's lemma

In this section, we restrict attention to the case where $X \sim N_p(\theta,\sigma^2 I_p)$ with σ^2 known and where the loss function is $L(\theta,\delta)=\|\delta-\theta\|^2$. We will be concerned with developing expressions for the risk function of a general estimator of the form $\delta(X)=X+\sigma^2\,g(X)$ for some function g from R^p into R^p. This development is due to Stein [1973] and [1981].

First note that

$$\begin{aligned} L(\theta,\delta) &= \|X+\sigma^2\,g(X)-\theta\|^2 \\ &= \|X-\theta\|^2+\sigma^4\,\|g(X)\|^2+2\,\sigma^2\,(X-\theta)^t g(X)\,. \end{aligned}$$

Since the risk of X is finite, $E[\|X-\theta\|^2]=p\,\sigma^2$, it follows from the Cauchy-Schwarz inequality that the risk of δ is finite if and only if $E_\theta[\|g(X)\|^2]<\infty$. Under this condition the risk function of δ is given by

$$R(\theta,\delta) = p\sigma^2 + \sigma^4 E_\theta[\|g(X)\|^2] + 2\sigma^2 E_\theta[(X-\theta)^t g(X)]\,.$$

Stein's lemma allows an alternative expression for the last expectation, i.e. $E_\theta[(X-\theta)^t g(X)] = \sigma^2 E_\theta[\mathrm{div} g(X)]$ where $\mathrm{div} g(X) = \sum_{i=1}^p \frac{\partial}{\partial X_i} g_i(X)$ under suitable conditions on g. The great advantage that Stein's lemma gives is that the risk function can be expressed as the expected value of a function of only X (and not θ), that is,

$$R(\theta,\delta) = E_\theta[p\sigma^2 + \sigma^4\,\|g(X)\|^2 + 2\sigma^4\,\mathrm{div} g(X)]$$

and hence the expression

$$p\sigma^2 + \sigma^4\left[\|g(X)\|^2 + 2\,\mathrm{div} g(X)\right]$$

can be interpreted as an unbiased estimate of the risk of δ. Actually, as X is a complete sufficient statistic, this unbiased estimator is the uniformly minimum variance unbiased estimator of the risk.

To see that $E_\theta[(X-\theta)^t g(X)] = \sigma^2 E_\theta[\mathrm{div} g(X)]$ is quite easy if $g(X)$ is sufficiently smooth. For example, suppose that $p=1$ and $g(X)$ is continuously differentiable and is such that $\lim_{x\to\pm\infty} g(x)\exp\{-(x-\theta)^2/2\sigma^2\} = 0$, which can be seen to occur if $E_\theta[|g'(X)|] < \infty$ (see e.g. Hoffmann [1992]).

Then, under that last condition, a simple integration by parts gives

$$\begin{aligned}
E_\theta[(X-\theta)g(X)] &= \frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}(x-\theta)g(x)\exp\{-(x-\theta)^2/2\sigma^2\}\,dx \\
&= \frac{1}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}\sigma^2 g(x)\left(\frac{-d}{dx}\exp\{-(x-\theta)^2/2\sigma^2\}\right)dx \\
&= \frac{\sigma^2}{(2\pi\sigma^2)^{1/2}}\int_{-\infty}^{\infty}g'(x)\exp\{-(x-\theta)^2/2\sigma^2\}\,dx \\
&= \sigma^2 E_\theta[g'(X)]\,.
\end{aligned}$$

The extension to p dimensions is straightforward if we assume that each coordinate function $g_i(X)$, $i=1,\dots p$ is continuously differentiable.

However we wish to include estimators such as the James-Stein estimators

$$\delta_a^{JS}(X) = \left(1-\frac{a\sigma^2}{\|X\|^2}\right)X \tag{3.1}$$

where the coordinate functions of $g(X) = (a\sigma^2/\|X\|^2)X$ are not everywhere continuously differentiable, since it explodes at 0. For this reason, Stein considered a weaker regularity condition for his identity to hold, that he called almost differentiability. He essentially required that $g(X) = (g_1(X), g_2(X), \dots, g_p(X))$ is such that, for each $i = 1, \dots, p$, the coordinate $g_i(X)$ is absolutely continuous in X_i for almost every $X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_p$. His precise statement is the following.

Definition 3.1. A function h from $\mathbb{R}^p$ into $\mathbb{R}$ is said to be almost differentiable if there exists a function $\nabla h = (\nabla_1 h, \dots, \nabla_p h)$ from $\mathbb{R}^p$ into $\mathbb{R}^p$ such that, for all $z \in \mathbb{R}^p$,

$$h(x+z) - h(x) = \int_0^1 z^t\,\nabla h(x+t\,z)\,dt\,,$$

for allmost all $x \in \mathbb{R}^p$.

A function $g = (g_1, \ldots, g_p)$ from $\mathbb{R}^p$ into $\mathbb{R}^p$ is said to be almost differentiable if all its coordinates g_i's are.

Almost differentiability is equivalent to the following notion of weak differentiability which is of more common use in analysis and which provides a more explicit criterion.

INTRODUCE LOCAL INTEGRABILITY, SOBOLEV SPACE ...

Definition 3.2. A function h from $\mathbb{R}^p$ into $\mathbb{R}$ is said to be weakly differentiable if there exist p locally integrable functions $\nabla_1 h, \ldots \nabla_p h$ such that, for any $i = 1, \ldots, p$,

$$\int_{\mathbb{R}^p} h(x)\, \frac{\partial \varphi}{\partial x_i}(x)\, dx = - \int_{\mathbb{R}^p} \nabla_i h(x)\, \varphi(x)\, dx \tag{3.2}$$

for any infinitely differentiable function φ with compact support from $\mathbb{R}^p$ into $\mathbb{R}$.

A function $g = (g_1, \ldots, g_p)$ from $\mathbb{R}^p$ into $\mathbb{R}^p$ is said to be weakly differentiable if all its coordinates g_i's are.

Note that weak differentiability is a global, not local, property. The functions $\nabla_i h$ in Definitions 3.1 and 3.2 coincide almost everywhere and are referred to as the weak partial derivatives of h . They are denoted, as the usual derivatives, by $\partial/\partial x_i$. The vector $\nabla h = (\nabla_1 h, \ldots, \nabla_p h) = (\partial/\partial x_1, \ldots, \partial/\partial x_p)$ denotes the weak gradient of h and the scalar div$g = \sum_{i=1}^p \nabla_i g_i$ denotes the weak divergence of g. We give, in the Appendix, a proof of the equivalence between these definitions and some additional details.

We show next that, as expected, the shrinkage function of the James-Stein estimator is weakly differentiable when $p \geq 3$. Note first that the function

$$h(x) = \frac{x}{\|x\|^2}$$

is locally integrable for $p \geq 2$. Indeed, for any ball B_R of radius R centered at 0 we have to determine the finiteness of

$$I = \int_{B_R} \frac{|x_j|}{\|x\|^2} \, dx$$

for any $j = 1, \ldots, p$. Now, it is easy to see, through polar coordinates, that

$$I \leq \int_{B_R} \frac{1}{\|x\|^2} \, dx < \infty \;\Leftrightarrow \int_0^R \frac{1}{r} r^{p-1} \, dr < \infty ,$$

that is,

$$\int_0^R r^{p-2} \, dr < \infty \Leftrightarrow p - 2 > -1 \Leftrightarrow p \geq 2 .$$

As for the derivatives, since they are defined almost everywhere, and since $x \mapsto x/\|x\|^2$ is infinitely derivable outside a neighbourhood of zero, the usual partial derivatives are the candidates to be the weak derivatives and it suffices to consider their local integrability. For fixed i and j between 1 and p, classical calculations give

$$\frac{\partial}{\partial x_i}\left(\frac{x_j}{\|x\|^2}\right) = \begin{cases} \frac{1}{\|x\|^2} + x_i \frac{\partial}{\partial x_i}\left(\frac{1}{\|x\|^2}\right) \text{ if } i = j \\ x_j \frac{\partial}{\partial x_i}\left(\frac{1}{\|x\|^2}\right) \text{ if } i \neq j \end{cases}$$

$$= \begin{cases} \frac{1}{\|x\|^2} - \frac{2x_i^2}{\|x\|^4} \text{ if } i = j \\ -\frac{2x_j x_i}{\|x\|^4} \text{ if } i \neq j \end{cases} .$$

Now we can see that

$$\int_{B_R} \left| \frac{-2x_j x_i}{\|x\|^4} \right| \, dx < \infty \Leftrightarrow \int_{B_R} \left| \frac{1}{\|x\|^2} - \frac{2x_i^2}{\|x\|^4} \right| \, dx < \infty$$

and these integrals are finite if and only if we have

$$\int_{B_R} \frac{1}{\|x\|^2}\, dx < \infty\,.$$

Clearly, this finiteness is guaranteed if and only if

$$\int_0^R \frac{1}{r^2}\, r^{p-1}\, dr < \infty\,,$$

that is,

$$\int_0^R r^{p-3}\, dr < \infty \Leftrightarrow p-3 > -1 \Leftrightarrow p \geq 3\,.$$

In that context, note that, for any $x \neq 0$,

$$\begin{aligned}
\operatorname{div}\left(\frac{X}{\|X\|^2}\right) &= \sum_{i=1}^{p} \frac{\partial}{\partial x_i}\left(\frac{x_i}{\|x\|^2}\right) \\
&= \sum_{i=1}^{p}\left(\frac{\|x\|^2 - 2x_i^2}{\|x\|^4}\right) \\
&= \frac{p-2}{\|x\|^2}\,.
\end{aligned} \tag{3.3}$$

We give now a precise statement of Stein's lemma for almost (or, equivalently, weakly) differentiable functions in the lines of Stein [1981]. Note that we will see, in Section 3.5, that it is closely related to Stokes' theorem, which will provide an alternative proof.

Theorem 3.1. *(Stein's lemma) Let $X \sim N_p(\theta, \sigma^2 I_p)$ and let $g(X)$ be an almost or weakly differentiable function from $\mathbb{R}^p$ into $\mathbb{R}^p$. Then*

$$E_\theta[(X-\theta)^t g(X)] = \sigma^2\, E_\theta[\operatorname{div} g(X)]\,, \tag{3.4}$$

provided either $E_\theta[\|(X-\theta)^t g(X)\|] < \infty$ or $E_\theta[|\operatorname{div} g(X)|] < \infty$.

Proof. We follow closely the Stein's approach proving the result for $p = 1$ and then for general p. Note that the proof will show that the finiteness conditions of the expectations in the statement of the theorem are equivalent.

First let $p = 1$. Let $Y \sim N(0,1)$ and let g be an absolutely continuous function from $\mathbb{R}$ into $\mathbb{R}$, i.e., there exists a function denoted by g', defined almost everywhere with respect to Lebesgue measure and satisfying

$$\forall\, y \in \mathbb{R},\ g(y) = \int_{-\infty}^{y} g'(z)\, dz\,.$$

Note that, setting, for any $y \in \mathbb{R}$,

$$\phi(y) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\}$$

we have, for any $y \in \mathbb{R}$,

$$\phi'(y) = -y\,\phi(y)\,,$$

so that $\phi(y)$ can be written as

$$\phi(y) = \int_{-\infty}^{y} -z\,\phi(z)\, dz = \int_{y}^{\infty} z\,\phi(z)\, dz\,.$$

Assuming $E\left[|g'(Y)|\right] < \infty$ guarantees that all of the following integrals exist. We have

$$\begin{aligned} E[g'(Y)] &= \int_{-\infty}^{\infty} g'(y)\ \phi(y)\ dy \\ &= \int_{-\infty}^{0} g(y) \int_{-\infty}^{y} -z\,\phi(z)\ dz\ dy + \int_{0}^{\infty} g'(y) \int_{y}^{\infty} z\,\phi(z)\ dz\ dy \\ &= \int_{-\infty}^{0} -z\,\phi(z) \int_{z}^{0} g'(y)\ dy\ dz + \int_{0}^{\infty} z\,(\phi(z) \int_{0}^{z} g'(y)\ dy\ dz \end{aligned}$$

by Fubini's theorem. Then, expressing the integrals of the derivatives,

$$
\begin{aligned}
E\left[g'(Y)\right] &= \int_{-\infty}^{\infty} z\,\phi(z)[g(z)-g(0)]\,dz \\
&= \int_{-\infty}^{\infty} z\,g(z)\,\phi(z)\,dz \\
&= E[Y\,g(Y)]. \qquad (3.5)
\end{aligned}
$$

Now, let $X = \sigma Y + \theta$ where $\sigma > 0$ and $\theta \in \mathbb{R}$, so that $X \sim N(\theta, \sigma^2)$. Then, if h is the function associated to g defined, for any $x \in \mathbb{R}$, by

$$
h(x) = g\left(\frac{x-\theta}{\sigma}\right)
$$

we have

$$
\begin{aligned}
E\left[h'(X)\right] &= E\left[\frac{1}{\sigma}\,g'\left(\frac{X-\theta}{\sigma}\right)\right] \\
&= \frac{1}{\sigma}\,E\left[g'(Y)\right] \\
&= \frac{1}{\sigma}\,E\left[Y\,g(Y)\right] \text{ by (3.5)} \\
&= \frac{1}{\sigma}\,E\left[\frac{X-\theta}{\sigma}\,g\left(\frac{X-\theta}{\sigma}\right)\right] \\
&= \frac{1}{\sigma^2}\,E\left[(X-\theta)\,h(X)\right].
\end{aligned}
$$

Hence

$$
E\left[(X-\theta)\,h(X)\right] = \sigma^2\,E[h'(X)]\,,
$$

which is Stein's lemma for $p = 1$.

For fixed $i = 1, \ldots, p$, consider g_i and set $X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)$. With a slight abuse of notation, letting $X = (X_i, X_{-i})$, we have, conditioning on X_{-i},

$$
E_\theta\left[(X_i-\theta_i)\,g_i(X_i,X_{-i})|X_{-i}\right] = \sigma^2\,E_\theta\left[\nabla_i g_i(X_i,X_{-i})|X_{-i}\right],
$$

using the independence between X_i and X_{-i} and the above result when $p = 1$. Hence, unconditioning, we obtain

$$E_\theta[(X_i - \theta_i)\, g_i(X)] = \sigma^2\, E_\theta[\nabla_i g_i(X)]\,.$$

Therefore, summing on i, gives

$$E_\theta\left[\sum_{i=1}^{p}(X_i - \theta_i)\, g_i(X)\right] = \sigma^2\, E_\theta\left[\sum_{i=1}^{p}\nabla_i g_i(X)\right]\,,$$

that is,

$$E_\theta[(X-\theta)^t\, g(X)] = \sigma^2\, E_\theta[\mathrm{div} g(X)]\,. \quad \square$$

The following corollary is immediate from Stein's lemma and the above discussion. (Recall: $L(\theta,d) = \|d-\theta\|^2$, $R(\theta,\delta) = E_\theta[L(\theta,\delta(X))] = E_\theta[\|\delta(X)-\theta\|^2]$).

Corollary 3.1. *Let $g(X)$ be a weakly differentiable function from $\mathbb{R}^p$ into $\mathbb{R}^p$ such that $E_\theta[\|g(X)\|^2] < \infty$. Then*

(1) $R(\theta, X + \sigma^2\, g(X)) = E_\theta[p\,\sigma^2 + \sigma^4\,(\|g(X)\|^2 + 2\,\mathrm{div} g(X))]$ *and* (2) $\delta(X) = X + \sigma^2\, g(X)$ *is minimax as soon as* $\|g(X)\|^2 + 2\,\mathrm{divg(X)} \leq 0$ *a.e. and dominates X provided additionally there is strict inequality on a set of positive measure.*

In the development above, it was tacitly assumed that the covariance matrix was known and equal to a multiple of the identity matrix $\sigma^2\, I_p$. Typically, this covariance is unknown and should be estimated. The next result extends Stein's integration by parts identity to the case where it is of the form $\sigma^2\, I_p$ with σ^2 unknown.

Lemma 3.1. *Let $X \sim \mathcal{N}(\theta, \sigma^2 I_p)$ and let S be a non negative random variable independent of X such that $S \sim \sigma^2 \chi^2_k$. Denoting by E_{θ,σ^2} the expectation with respect* *to the joint distribution of (X,S), we have, provided the corresponding expectations exist, the following two results:*

(1) *if $g(x,s)$ is a function from $\mathbb{R}^p \times \mathbb{R}_+$ into $\mathbb{R}^p$ such that, for any $s \in \mathbb{R}_+$, $g(\cdot,s)$ is weakly differentiable, then*

$$E_{\theta,\sigma^2}\left[\frac{1}{\sigma^2}(X-\theta)^t g(X,S)\right] = E_{\theta,\sigma^2}[\mathrm{div}_X g(X,S)]$$

where $\mathrm{div}_x g(x,s)$ *is the divergence of $g(x,s)$ with respect to x;*

(2) *if $h(x,s)$ is a function from $\mathbb{R}^p \times \mathbb{R}_+$ into $\mathbb{R}$ such that, for any $x \in \mathbb{R}^p$, $h(x,\|u\|^2)$ is weakly differentiable as a function of u, then*

$$E_{\theta,\sigma^2}\left[\frac{1}{\sigma^2}h(X,S)\right] = E_{\theta,\sigma^2}\left[2\frac{\partial}{\partial S}h(X,S) + (k-2)\,S^{-1}\,h(X,S)\right].$$

Proof. Part (1) is Stein's lemma 1981 (cf. [1981]). Part (2) can be seen as a particular case of Lemma 1 (2) (established for elliptically symmetric distributions) of Fourdrinier, Strawderman and Wells [2003], although we will present a direct proof.

The joint distribution of (X,S) can be viewed as resulting, in the setting of the canonical form of the general linear model, from the distribution of $(X,U) \sim \mathscr{N}((\theta,0),\sigma^2 I_{p+k})$ with $S = ||U||^2$. Then we can write

$$\begin{aligned} E_{\theta,\sigma^2}\left[\frac{1}{\sigma^2}h(X,S)\right] &= E_{\theta,\sigma^2}\left[\frac{1}{\sigma^2}U^t\,\frac{U}{||U||^2}\,h(X,||U||^2)\right] \\ &= E_{\theta,\sigma^2}\left[\mathrm{div}_U\left(\frac{U}{||U||^2}\,h(X,||U||^2)\right)\right] \end{aligned}$$

according to part (1). Hence, expanding the divergence term, we have

$$\begin{aligned} E_{\theta,\sigma^2}\left[\frac{1}{\sigma^2}\,h(X,S)\right] &= E_{\theta,\sigma^2}\left[\frac{k-2}{||U||^2}\,h(X,||U||^2)+\frac{U^t}{||U||^2}\,\nabla_U h(X,||U||^2)\right] \\ &= E_{\theta,\sigma^2}\left[\frac{k-2}{S}\,h(X,S)+2\,\frac{\partial}{\partial S}h(X,S)\right] \end{aligned}$$

since

$$\nabla_U h(X,||U||^2) = 2\,\frac{\partial}{\partial S}h(X,S)\Big|_{S=||U||^2}\,U\,.$$

□

In the next few sections we apply the above corollary to show domination of the James-Stein estimators and several others over the usual estimator in three and higher dimensions.

3.4 James-Stein estimators and other improved estimators

In this section, we apply the integration by parts results of Section 3.3 to obtain several classes of estimators that dominate the classical minimax estimator $\delta_0(X)$ in dimension 3 and higher. The estimators of James and Stein, Baranchik, and certain estimators shrinking toward subspaces are the main application of this section. Bayes (generalized, proper, and pseudo) are considered in the next section. Throughout this section, except for Theorem 3.4, let $X \sim N_p(\theta,\sigma^2 I_p)$ and loss be $L(\theta,\delta) = \|\delta - \theta\|^2$. According to Corollary 3.1 it suffices to find weakly differentiable functions g from R^p into R^p such that $E_\theta[\|g(X)\|^2] < \infty$ and $\|g(X)\|^2 + 2 \text{ div } g(X) \leq 0$ (with strict inequality on a set of positive measure) in order to show that $\delta(X) = X + \sigma^2 g(X)$ dominates X.

3.4.1 James-Stein estimators

The class of James-Stein estimators is given by

$$\delta_a^{JS}(X) = \left(1 - \frac{a\,\sigma^2}{\|X\|^2}\right) X. \tag{3.6}$$

The basic properties of $\delta_a^{JS}(X)$ are given in the following result.

Theorem 3.2. *Under the above model*

(1) *The risk of $\delta_a^{JS}(X)$ is given by*

$$R(\theta, \delta_a^{JS}) = p\,\sigma^2 + \sigma^4\,(a^2 - 2\,a\,(p-2))E_\theta\left[\frac{1}{\|X\|^2}\right] \tag{3.7}$$

for $p \geq 3$.

(2) *$\delta_a^{JS}(X)$ dominates $\delta_0(X) = X$ for $0 < a < 2\,(p-2)$ and is minimax for $0 \leq a \leq 2\,(p-2)$ for all $p \geq 3$.*

(3) *The uniformly optimal choice of a is $a = p-2$ for $p \geq 3$.*

(4) *The risk at $\theta = 0$ for the optimal James-Stein estimator $\delta_{p-2}^{JS}(X)$ is $2\,\sigma^2$ for all $p \geq 3$.*

Proof. Observe that $\delta_a^{JS}(X) = X + \sigma^2 g(X)$ where $g(X) = -a/\|X\|^2 X$. As noted in the Section 3.3, $g(X)$ is weakly differentiable if $p \geq 3$. Also $E_\theta[\|g(X)\|^2] = a^2 E_\theta[1/\|X\|^2]$ is finite if $p \geq 3$ since $\|X\|^2/\sigma^2$ has a non-central χ^2 distribution with p degrees of freedom and non-centrality parameter $\lambda = \|\theta\|^2/2\sigma^2$. Indeed by the usual Poisson representation of a non-central χ^2, we have $\|X\|^2/\sigma^2 \mid K \sim \chi^2_{p+2K}$ where $K \sim$ Poisson $(\lambda = \|\theta\|^2/2\sigma^2)$ and hence

$$E_\theta\left[\frac{\sigma^2}{\|X\|^2}\right] = E_\lambda\left[E\left[\frac{1}{\chi^2_{p+2K}}\middle|K\right]\right] = E_\lambda\left[\frac{1}{p+2K-2}\right] < \infty \tag{3.8}$$

if $p > 2$.

Also, according to (3.3), for any $x \neq 0$,

$$\operatorname{div}\left(\frac{x}{\|x\|^2}\right) = \frac{p-2}{\|x\|^2}. \tag{3.9}$$

Hence

$$\|g(x)\|^2 + 2\operatorname{div} g(x) = (a^2 - 2a(p-2))\frac{1}{\|x\|^2}$$

and by Corollary 3.1, for $p \geq 3$,

$$R(\theta, \delta_a^{JS}) = p\sigma^2 + \sigma^4(a^2 - 2a(p-2))E_\theta\left(\frac{1}{\|X\|^2}\right).$$

This proves (1).

Part (2) follows since $a^2 - 2a(p-2) < 0$ for $0 < a < 2(p-2)$ and hence for such $a > 0$,

$$R(\theta, \delta_a^{JS}) < p\sigma^2 = R(\theta, \delta_0). \tag{3.10}$$

The minimaxity claim for $0 \leq a \leq 2(p-2)$ follows by replacing $<$ by $\leq$ in (3.10). It is interesting to note that $R\big(\theta, \delta^{JS}_{2(p-2)}\big) \equiv R(\theta, \delta_0) \equiv p\sigma^2$.

Part (3) follows by noting that, for all θ, the risk of $R\big(\theta, \delta_a^{JS}\big)$ is minimized by choosing $a = p-2$ since this value minimizes the quadratic $a^2 - 2a(p-2)$.

To prove part (4) note that, when $\theta = 0$, $\|X\|^2/\sigma^2$ has a central chi-square distribution with p degrees of freedom. Hence $E_0\left[\sigma^2/\|X\|^2\right] = E\left[1/\chi^2_p\right] = (p-2)^{-1}$ and therefore, provided $p \geq 3$,

$$\begin{aligned} R(0,\delta_{p-2}^{JS}) &= p\,\sigma^2 + \left((p-2)^2 - 2\,(p-2)^2\right)\frac{\sigma^2}{p-2} \\ &= p\,\sigma^2 - (p-2)\,\sigma^2 \\ &= 2\,\sigma^2. \quad \square \end{aligned}$$

Hence we have that $\delta_{p-2}^{JS} = \left(1-(p-2)\,\sigma^2/\|X\|^2\right)X$ is the uniformly best estimator in the class of James-Stein estimators. This is the estimator that is typically referred to as the James-Stein estimator. Also note that at $\theta = 0$ the risk is $2\,\sigma^2$ regardless of p and so large savings in risk are possible in a neighborhood of $\theta = 0$ for large p.

We may use (3.8) to give upper and lower bounds for the risk of δ_a^{JS} based on the following lemma.

Lemma 3.2. *Let* $K \sim \text{Poisson}(\lambda)$. *Then, for* $b \geq 1$, *we have*

$$\frac{1}{b+\lambda} \leq E_\lambda\left[\frac{1}{b+K}\right] \leq \frac{\frac{1-e^{-\lambda}}{\lambda}}{(b-1)\frac{1-e^{-\lambda}}{\lambda}+1} \leq \frac{1}{b-1+\lambda}.$$

Proof. The first inequality follows directly from Jensen's inequality and the fact that $E_\lambda(K) = \lambda$. The second inequality follows since (also by Jensen's inequality)

$$\begin{aligned} E_\lambda\left[\frac{1}{b+K}\right] &= E_\lambda\left[\frac{\frac{1}{K+1}}{\frac{b-1}{K+1}+1}\right] \\ &\leq \frac{E_\lambda\left[\frac{1}{K+1}\right]}{(b-1)E_\lambda\left[\frac{1}{K+1}\right]+1} \\ &= \frac{\frac{1-e^{-\lambda}}{\lambda}}{(b-1)\frac{1-e^{-\lambda}}{\lambda}+1} \end{aligned}$$

since $E_\lambda\left[(K+1)^{-1}\right] = (1-\exp(-\lambda))/\lambda$.

Now, since $y/[(b-1)y+1]$ is increasing in y and $(1-\exp(-\lambda))/\lambda < \lambda^{-1}$, we have

$$\frac{\frac{1-e^{-\lambda}}{\lambda}}{(b-1)\,\frac{1-e^{-\lambda}}{\lambda}+1} \leq \frac{\frac{1}{\lambda}}{\frac{b-1}{\lambda}+1} = \frac{1}{b-1+\lambda}.$$

Hence the third inequality follows. □

The following bounds on the risk of δ_a^{JS} follow directly from (3.7), (3.8) and Lemma 3.2.

Corollary 3.2. *[Casella and Hwang (1982)] For $p \geq 4$ and $0 \leq a \leq 2\,(p-2)$, we have*

$$p\,\sigma^2 + \frac{(a^2 - 2\,a\,(p-2))\,\sigma^2}{p-2+\|\theta\|^2/\sigma^2} \leq R(\theta,\delta_a^{JS}) \leq p\,\sigma^2 + \frac{(a^2 - 2\,a\,(p-2))\,\sigma^2}{p-4+\|\theta\|^2/\sigma^2}.$$

We note in passing that the upper bound may be improved at the cost of added complexity by using the second inequality in Lemma 3.2. The improved upper bound has the advantage that it is exact at $\theta = 0$. The lower bound is also valid for $p = 3$ and is also exact at $\theta = 0$.

3.4.2 Positive-part and Baranchik-type estimators

James-Stein estimators are such that, when $\|X\|^2 < a\,\sigma^2$, the multiplier of X becomes negative and, furthermore, $\lim_{\|X\|\to 0} \|\delta_a^{JS}(X)\| = \infty$. It follows that, for any $K > 0$, there exits $\eta > 0$ such that $\|X\| < \eta$ implies $\|\delta_a^{JS}(X)\| > K$. Hence an observation that would lead to almost certain acceptance of $H_0 : \theta = 0$ gives rise to an estimate very far from 0. Furthermore the estimator is not monotone in the sense that a larger value of X for a particular coordinate may give a smaller estimate of the mean of that coordinate. For example, if $X = (X_0, 0, \dots, 0)$ and $-\sqrt{a\sigma^2} < X_0 < 0$, then $\left(1 - a\sigma^2/\|X\|^2\right) X_0 > 0$ while, if $0 < X_0 < \sqrt{a\sigma^2}$, then $\left(1 - a\sigma^2/\|X\|^2\right) X_0 > 0$.

This behavior is undesirable. One possible remedy is to modify the James-Stein estimator to its positive-part, namely

$$\delta_a^{JS+}(X) = \left(1 - \frac{a\sigma^2}{\|X\|^2}\right)_+ X \tag{3.11}$$

where $t_+ = \max(t, 0)$. The positive past estimate is a particular example of a Baranchik-type estimator of the form

$$\delta_{a,r}^B(X) = \left(1 - \frac{a\sigma^2\, r(\|X\|^2)}{\|X\|^2}\right) X \tag{3.12}$$

where, typically $r(\cdot)$ is continuous and nondecreasing. The $r(\cdot)$ function for δ_a^{JS+} is given by

$$r(\|X\|^2) = \begin{cases} \frac{\|X\|^2}{a\sigma^2} \text{ if } & 0 < \|X\|^2 < a\sigma^2 \\ 1 \quad\;\; \text{ if } & \|X\|^2 \geq a\sigma^2. \end{cases}$$

We show in this section that, under certain conditions, Baranchik-type estimators improve on X and that the positive-part James-Stein estimator improves on the James-Stein estimator as well.

We first give conditions under which Baranchik-type estimator improves on X.

Theorem 3.3. *The estimator given by (3.12) with $r(\cdot)$ absolutely continuous, is minimax for $p \geq 3$ provided*

(1) $0 < a \leq 2(p-2)$*;*

(2) $0 \leq r(\cdot) \leq 1$*; and*

(3) $r(\cdot)$ *is nondecreasing.*

Furthermore it dominates X provided that both inequalities are strict in (1) *or in* (2) *on a set of positive measure, or if $r'(\cdot)$ is strictly positive on a set of positive measure.*

Proof. Here $\delta^B_{a,r}(X) = X + \sigma^2\, g(X)$ where $g(X) = -a\, r(\|X\|)^2/\|X\|^2\, X$. As noted in the Appendix, $g(\cdot)$ is weakly differentiable and

$$\begin{aligned} \operatorname{div} g(X) &= -a\left\{r(\|X\|^2)\operatorname{div}\left(\tfrac{X}{\|X\|^2}\right) + \tfrac{X^t}{\|X\|^2}\nabla\, r(\|X\|^2)\right\} \\ &= -a\left\{r(\|X\|^2)\tfrac{(p-2)}{\|X\|^2} + 2\, r'(\|X\|^2)\right\}. \end{aligned} .$$

Hence

$$\begin{aligned} &\|g(X)\|^2 + 2\operatorname{div} g(X) \\ &= \frac{a^2\, r^2(\|X\|^2)}{\|X\|^2} - \frac{2\, a\,(p-2) r(\|X\|^2)}{\|X\|^2} - 4\, a\, r'(\|X\|^2) \\ &\le \frac{r(\|X\|^2)}{\|X\|^2}(a^2 - 2\, a\,(p-2)) - 4\, a\, r'(\|X\|^2) \\ &\le 0\,, \end{aligned} \tag{3.13}$$

the first inequality being satisfied by Conditions (2) while the last inequality uses all Conditions (1), (2) and (3). Hence minimaxity follows from Corollary 3.1. Under the additional conditions, it is easy to see that the above inequalities become strict on a set of positive measure so that domination over X is guaranteed. □

As an example of a dominating Baranchik-type estimator consider

$$\delta(X) = \left(1 - \frac{a\,\sigma^2}{b + \|X\|^2}\right) X$$

for $0 < a \le 2\,(p-2)$ and $b > 0$. Here $r(\|X\|^2) = \|X\|^2/(\|X\|^2 + b)$ and is strictly increasing.

The theorem also shows that the positive-part James-Stein estimator dominates X for $0 < a \leq 2\,(p-2)$. In fact, as previously noted, the positive-part James-Stein estimator even improves on the James-Stein estimator itself. This reflects the more general phenomenon that a positive-part estimator will typically dominate the non-positive-part version if the underlying density is symmetric and unimodal. Here is a general result along these lines.

Theorem 3.4. *Suppose X has a density $f(x-\theta)$ in $\mathbb{R}^p$ such that the function f is symmetric and unimodal in each coordinate separately for each fixed value of the other coordinates. Then, for any finite risk estimator of θ of the form*

$$\delta(X) = \left(1 - B\left(X_1^2, X_2^2, \ldots, X_p^2\right)\right) X\,,$$

the positive-part estimator

$$\delta_+(X) = \left(1 - B\left(X_1^2, X_2^2, \ldots, X_p^2\right)\right)_+ X$$

dominates $\delta(X)$ under any loss of the form $L(\theta,\delta) = \sum_{i=1}^p a_i(\delta_i - \theta_i)^2$ ($a_i > 0$ for all i) provided $P_\theta[B(X_1^2, X_2^R, \ldots, X_p^2) > 1] > 0$.

Proof. Note that the two estimators differ only on the set where $B(\cdot) > 1$. Hence the ith term in $R(\theta,\delta) - R(\theta,\delta_+)$ is

$$\begin{aligned} &a_i E_\theta\left[\left\{(1 - B(X_1^2, \ldots, X_p^2))^2 X_i^2 - 2\,\theta_i X_i\,(1 - B(X_1^2, \ldots, X_p^2)\right\} I_{B>1}(X)\right] \\ &\qquad > -2\,\theta_i\, a_i\, E_\theta\left\{X_i(1 - B(X_1^2, \ldots, X_p^2) I_{B>1}(X)\right\}. \end{aligned}$$

It suffices, therefore, to show that, for any non-negative function $H(X_1^2,\dots,X_p^2)$, $\theta_i E_\theta[X_i H(X_1^2,\dots,X_p^2)] \geq 0$. Hence it suffices to show by symmetry that, if $\theta_i \geq 0$, then $E_\theta[X_i \mid X_i^2 = t_i^2, X_j = t_j \; j \neq i] \geq 0$ for all i $(1 \leq i \leq p)$ and all $(t_1,\dots,t_p)$.

However this expression is proportional to

$$\begin{aligned} &\mid t_i \mid \left[f\left((t_i-\theta_1)^2,(t_2-\theta_2)^2,\dots,(\mid t_i \mid -\theta_i)^2,\dots,(t_p-\theta_p)^2\right)\right.\\ &\left. -f\left((t_1-\theta_1)^2,(t_2-\theta_2)^2,\dots,(-\mid t_i \mid -\theta_i)^2,\dots,(t_p-\theta_p)^2\right)\right] \geq 0 \end{aligned}$$

since, for $\theta_i \geq 0, (\mid t_i \mid -\theta_i)^2 \leq (-\mid t_i \mid -\theta_i)^2$ and since $f(X_1^2,X_2^2,\dots,X_p^2)$ is nonincreasing in each argument. Hence the theorem follows. □

For the remainder of the current section we return to the assumption that $X \sim N_p(\theta,\sigma^2 I_p)$.

The positive-part James-Stein estimators are inadmissible because of a lack of smoothness which precludes them from being generalized Bayes. The Baranchik class however contains "smooth" estimators which are generalized (and even proper) Bayes and admissible. They will play an important role in Chapter 4.

We close this subsection with a generalization of the Baranchik result in Theorem 3.3. It is apparent from the proof of the theorem that it is only necessary that the second expression in (3.13) be nonpositive (and negative on a set of positive measure) in order that $\delta(X)$ dominates X. In particular it is not necessary that $r(\cdot)$ be nondecreasing. The following result (see Efron and Morris [1976] and Fourdrinier and Ouassou[2000]) gives a necessary and sufficient condition for the unbiased estimator of risk difference, $R(\theta,\delta)-R(\theta,X)$, for $\delta(X) = \left(1 - a\,r(\|X\|^2)/\|X\|^2\right)X$ to be non-positive. The proof is by direct calculation.

Lemma 3.3. *Let $g(X) = -a\,r\left(\|X\|^2)/\|X\|^2\right)X$ where $r(y)$ is an absolutely continuous function from R^+ into R. Then on the set where $r(y) \neq 0$,*

$$\|g(x)\|^2 + 2\,\mathrm{div}g(x) = a\left\{\frac{a\,r^2(y)}{y} - \frac{2(p-2)r(y)}{y} - 4\,r'(y)\right\}$$
$$= -4a^2r^2(y)y^{\frac{p-2}{2}}\frac{d}{dy}\left[y^{-\frac{p-2}{2}}\left(\frac{1}{2(p-2)} - \frac{1}{a\,r(y)}\right)\right] \text{ a.e.},$$

where $y = \|x\|^2$.

The following corollary broadens the class of minimax estimators of Baranchik's form.

Corollary 3.3. *Suppose $\delta(X) = \left(1 - a\,r(\|X\|^2)/\|X\|^2\right)X$ with*

$$a\,r(y) = \left[\frac{1}{2\,(p-2)} + y^{(p-2)/2}H(y)\right]^{-1}$$

where $H(y)$ is absolutely continuous, nonnegative and nonincreasing. Then $\delta(X)$ is minimax provided $E_\theta\left[r^2(\|X\|^2)/\|X\|^2\right] < \infty$. If in addition $H(y)$ is strictly monotone on a set of positive measure where $r(y) \neq 0$, then $\delta(X)$ dominates X.

Proof. The result follows from Corollary 3.1 and 3.3 by noting that

$$H(y) = y^{-(p-2)/2}\left(\frac{1}{2(p-2)} - \frac{1}{a\,r(y)}\right).$$

□

An application of Corollary 3.3 gives a useful class of dominating estimators due to Alam [1973].

Corollary 3.4. *Let $\delta(X) = \left(1 - a\,f(\|X\|^2)/(\|X\|^2)^{\tau+1}\right)X$ where $f(y)$ is nondecreasing and absolutely continuous and where $0 \leq a\,f(y)/y^\tau < 2\,(p-2-2\,\tau)$ for* *some* $\tau \geq 0$*. Then* $\delta(X)$ *is minimax and dominates* X *if* $0 < a\,f(y)/y^{\tau}$ *on a set of positive measure.*

Proof. The proof follows from Corollary 3.3 by letting

$$a\,r(y) = \frac{a\,f(y)}{y^{\tau}} \quad \text{and} \quad H(y) = -y^{-(p-2)/2}\left(\frac{1}{2\,(p-2)} - \frac{y^{\tau}}{a\,f(y)}\right).$$

Clearly r is bounded so that $E_{\theta}\left[r^2(\|X\|^2)/\|X\|^2\right] < \infty$ and $H(y) \geq 0$. Also

$$\begin{aligned} H'(y) &= \frac{p-2}{2}y^{-p/2}\left(\frac{1}{2\,(p-2)} - \frac{y^{\tau}}{a\,f(y)}\right) - y^{-(p-2)/2}\left(\frac{-\tau\,y^{\tau-1}}{a\,f(y)} + \frac{y^{\tau}\,f^{\tau}(y)}{a\,f^2(y)}\right) \\ &\leq y^{-\frac{p}{2}}\left[\frac{1}{4} - y^2\frac{p-2-2\,\tau}{2\,a\,f(y)}\right] \\ &\leq 0 \end{aligned}$$

since $f'(y) \geq 0$ and $0 < a\,f(y)/y^{\tau} < 2\,(p-2-2\,\tau)$. □

A simple example of a minimax Baranchik-type estimator with a nonmonotone $r(\cdot)$ is given by $r(y) = y^{1-\tau}/(1+y)$ for $0 < \tau < 1$ and $0 < a < 2(p-2-2\,\tau)$. To see this, apply Corollary 3.4 with $f(y) = y/(1+y)$ and note that $f(y)$ is increasing and $0 \leq f(y)/y^{\tau} = r(y) \leq 1$. Note also that $r'(y) = y^{-\tau}[(1-\tau) - \tau\,y]/(1+y)^2$ and hence $r(y)$ is increasing for $0 < y < (1-\tau)/\tau^{-1}$ and decreasing for $y > (1-\tau)/\tau^{-1}$.

We will use the above corollaries in Chapter 4 to establish minimaxity of certain Bayes and generalized Bayes estimators.

3.4.3 Unknown variance

The following theorem provides an extension of the above results to the setting of an unknown variance.

Theorem 3.5. *Let $X \sim \mathcal{N}(\theta, \sigma^2 I_p)$ where θ and σ^2 are unknown and $p \geq 4$ and let S be a non negative random variable independent of X and such that $S \sim \sigma^2 \chi^2_k$. Consider an estimator of θ of the form $\varphi(X,S) = X + S\,g(X,S)$ with $E_{\theta,\sigma^2}[S^2\,||g(X,S)||^2] < \infty$, where E_{θ,σ^2} denotes the expectation with respect to the joint distribution of (X,S).*

Then an unbiased estimator of the risk $||\varphi(X,S) - \theta||^2/\sigma^2$ is

$$\delta_0(X,S) = p + S\left\{(k+2)\,||g(X,S)||^2 + 2\,\mathrm{div}_X g(X,S) + 2\,S\frac{\partial}{\partial S}||g(X,S)||^2\right\}. \tag{3.14}$$

Proof. According to the expression of $\varphi(X,S)$, its risk $R(\theta,\varphi)$ is the expectation of

$$\frac{1}{\sigma^2}\,||X-\theta||^2 + 2\,\frac{S}{\sigma^2}\,(X-\theta)^t g(X,S) + \frac{S^2}{\sigma^2}\,||g(X,S||^2. \tag{3.15}$$

Clearly

$$E_{\theta,\sigma^2}\left[\frac{1}{\sigma^2}\,||X-\theta||^2\right] = p$$

and Lemma 3.1 (1) and (2) express respectively that

$$E_{\theta,\sigma^2}\left[\frac{1}{\sigma^2}(X-\theta)^t g(X,S)\right] = E_{\theta,\sigma^2}[\mathrm{div}_{\mathrm{X}} g(X,S)]$$

and, with $h(x,s) = s^2\,||g(x,s)||^2$, that

$$E_{\theta,\sigma^2}\left[\frac{S^2}{\sigma^2}\,||g(X,S)||^2\right] = E_{\theta,\sigma^2}\left[S\left\{(k+2)\,||g(X,S)||^2 + 2\,S\frac{\partial}{\partial S}||g(X,S)||^2\right\}\right].$$

Therefore $R(\theta,\varphi) = E_{\theta,\sigma^2}[\delta_0(X,S)]$ with $\delta_0(X,S)$ given in (3.14), which means that $\delta_0(X,S)$ is an unbiased estimator of the risk $||\varphi(X,S)-\theta||^2/\sigma^2$. □

As an example, consider an extension of the Baranchik form in Corollary 3.3 where

$$\delta(X,S) = \left(1 - \frac{a\,S\,r(\|X\|^2/S)}{\|X\|^2}\right) X$$

with r nondecreasing and bounded between 0 and 1. Straightforward calculations show that the terms in curved brackets in (3.14) equals

$$a\,\frac{r(\|X\|^2/S)}{\|X\|^2}\big((k+2)\,ar(\|X\|^2/S) - 2\,(p-2)\big) \\ -4\,a\,\frac{r'(\|X\|^2/S)}{S}\,\big(1 + a\,r(\|X\|^2/S)\big)\,. \qquad (3.16)$$

Therefore, if $0 < a < 2\,(p-2)/(k+2)$, then $\delta(X,S)$ dominates X and hence is minimax. In the case $r \equiv 1$, the constant a is $(p-2)/(k+2)$. This is the estimator developed by Stein [1956] and James and Stein [1961] using direct methods. Note, when $r \equiv 1$, the derivative with respect to S in (3.16) equals 0. This is in contrast to the development above which gives a construction of a Baranchik class in the unknown variance setting.

3.4.4 Estimators that shrink toward a subspace

We saw in Subsection 3.4.1 that the James-Stein estimator shrinks toward $\theta = 0$ and that substantial risk savings are possible if θ is in a neighborhood of 0. If we feel that θ is close to some other value, say θ_0, a simple adaptation of the James-Stein estimator which shrinks toward θ_0 may be desirable. Such an estimator is given by

$$\delta^{JS}_{a,\theta_0}(X) = \theta_0 + \left(1 - \frac{a\,\sigma^2}{\|X-\theta_0\|^2}\right)(X-\theta_0). \tag{3.17}$$

It is immediate that $R(\theta, \delta^{JS}_{a,\theta_0}(X)) = R(\theta-\theta_0, \delta^{JS}_a)$ since

$$\begin{aligned} R(\theta, \delta^{JS}_{a,\theta_0}) &= E_\theta \|\theta_0 + \left(1 + \frac{a\,\sigma^2}{\|X-\theta_0\|^2}\right)(X-\theta_0) - \theta\|^2 \\ &= E_{\theta-\theta_0} \| \left(1 + \frac{a\,\sigma^2}{\|X\|^2}\right) X - (\theta-\theta_0)\|^2 \\ &= R(\theta-\theta_0, \delta^{JS}_a(X)). \end{aligned}$$

Hence, for $p \geq 3, \delta^{JS}_{a,\theta_0}$ dominates X and is minimax for $0 < a < 2\,(p-2)$, and $a = p-2$ is the optimal choice of a. Furthermore the risk of $\delta^{JS}_{a,\theta_0}(X)$ at $\theta = \theta_0$ is $2\,\sigma^2$ and so large gains in risk are possible in a neighborhood of θ_0. The same argument establishes the fact that, for any estimator, $\delta(X)$, we have $R(\theta, \theta_0 + \delta(X - \theta_0) = R(\theta-\theta_0, \delta(X))$. Hence any of the minimax estimators of Subsections 3.4.1 and 3.4.2 may be modified in this way and minimaxity will be preserved.

More generally, we may feel that θ is close to some subspace V of dimension $s < p$. In this case we may wish to shrink X toward the subspace V. One way to do this is to consider the class of estimators given by

$$P_V X + \left(1 - \frac{a\,\sigma^2 r(\|X - P_V X\|^2)}{\|X - P_V X\|^2}\right)(X - P_V X) \tag{3.18}$$

where $P_V X$ is the projection of X onto V.

A standard canonical representation is helpful. Suppose V is an s-dimensional subspace of $\mathbb{R}^p$ and $V^\perp$ is the p-s dimensional orthogonal complement of V. Let

$P = (P_1\ P_2)$ be an orthogonal matrix such that the s columns of the $p \times s$ matrix P_1 span V and the $p-s$ columns of the $p \times (p-s)$ matrix P_2 span $V^\perp$.

For any vector $z \in R^p$, let

$$W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix} = P^t z$$

where W_1 is $s \times 1$ and W_2 is $(p-s) \times 1$. Then $P_V z = P_1 W_1$ and $\|P_V z\|^2 = \|P_1 W_1\|^2 = \|W_1\|^2$. Also $P_{V^\perp} z = P_2 W_2$ and $\|P_{V^\perp} z\|^2 = \|P_2 W_2\|^2 = \|W_2\|^2$. Further, if $X \sim N_p(\theta, \sigma^2 I)$, then

$$P^t X = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_p\Bigg(\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \sigma^2 \begin{pmatrix} I_s & 0 \\ 0 & I_{p-s} \end{pmatrix} \Bigg)$$

where $\nu_1 = P_V \theta$ and $\nu_2 = P_{V^\perp} \theta$ so that

$$\|P_V X\|^2 = \|Y_1\|^2, \qquad \|P_{V^\perp} X\|^2 = \|Y_2\|^2 \|P_V (X-\theta)\|^2 = \|Y_1 - \nu_1\|^2$$

and

$$\|P_{V^\perp}(X-\theta)\|^2 = \|Y_2 - \nu_2\|^2.$$

The following result gives risk properties of the estimator (3.18).

Theorem 3.6. *Let V be a subspace of dimension $s \geq 0$. Then, for the estimator (3.18), we have*

$$R(\theta,\delta) = s\sigma^2 + E_{\nu_2}\left[\left\| \left(1 - \frac{a\sigma^2 r(\|Y_2\|^2}{\|Y_2\|^2} \right) Y_2 - \nu_2 \right\|^2 \right]$$

where Y_2 and ν_2 are as above. Further, if $p-s \geq 3$ and a and $r(y)$ satisfy the assumptions of Theorem 3.3 (or Corollary 3.3 or Corollary 3.4) with $p-s$ in place of p, then $\delta(X)$ is minimax and dominates X if the additional conditions are satisfied.

The proof involves showing that the risk decomposes into the sum of two components. The first component is essentially the risk of the usual estimator in a space of dimension s (i.e. of V) and the second represents the risk of a Baranchik-type estimator on a space of dimension $p-s$. The risk is

$$\begin{aligned}
R(\theta,\delta) &= E_\theta\left[\left|\left|P_V X + \left(1-\tfrac{a\sigma^2 r(\|X-P_V X\|^2)}{\|X-P_V X\|^2}\right)(X-P_V X) - \theta\right|\right|^2\right] \\
&= E_\theta\left[\left|\left|(P_V X - P_V\theta) + \left(1-\tfrac{a\sigma^2 r(\|X-P_V X\|^2)}{\|X-P_V X\|^2}\right)(X-P_V X) - (\theta - P_V\theta)\right|\right|^2\right] \\
&= E_\theta[\|P_V(X-\theta)\|^2] \\
&\qquad + E_\theta\left[\left|\left|\left(1-\tfrac{a\sigma^2 r(\|X-P_V X\|^2)}{\|X-P_V X\|^2}\right)(X-P_V X) - (\theta - P_V\theta)\right|\right|^2\right] \\
&= E_{\nu_1}[\|Y_1-\nu_1\|^2] + E_{\nu_2}[\|\left(1-\tfrac{a\sigma^2 r(\|Y_2\|^2}{\|Y_2\|^2}\right)Y_2-\nu_2\|^2] \\
&= s\sigma^2 + E_{\nu_2}[\|\left(1-\tfrac{a\sigma^2 r(\|Y_2\|^2}{\|Y_2\|^2}\right)Y_2-\nu_2\|^2]\,.
\end{aligned}$$

This gives the first part of the theorem. The second part follows since $Y_2 \sim N_{p-s}(\nu_2, \sigma^2 I_{p-s})$, with $p-s \geq 3$. □

For example, if we choose $r(y) \equiv 1$ the risk of the resulting James-Stein type estimator

$$P_V X + \left(1 - \frac{a\sigma^2}{\|X-P_V X\|^2}\right)(X-P_V X)$$

is

$$s\sigma^2 + (p-s)\sigma^2 + \sigma^4\,(a^2 - 2a\,(p-s-2))\,E_\theta\left[\frac{1}{\|X-P_V X\|^2}\right].$$

This estimator is minimax if $0 \le a \le 2\,(p-s-2)$ and dominates X if $0 < a < 2\,(p-s-2)$ provided $p-s \ge 3$. The uniformly best choice of a is $p-s-2$. If in fact $\theta \in V$, the risk of the corresponding optimal estimator is $(s+2)\,\sigma^2$, since in this case $\nu_2 = P_{V^\perp}\theta = 0$ and hence $E_\theta\left[\sigma^2/\|X - P_V X\|^{-2}\right] = E_0\left[\sigma^2/\|Y_2\|^{-2}\right] E\left[1/\chi^2_{p-s}\right] = (p-s-2)^{-1}$. If $\theta \not\in V$, then $\nu_2 \neq 0$ and $\|Y_2\|^2$ has a non-central chi-square distribution with $p-s$ degrees of freedom and non-centrality parameter $\|\nu_2\|^2/2\,\sigma^2$.

One of the first instances of an estimator shrinking toward a subspace is due to Lindley [1962]. He suggested that while we might not have a good idea as to the value of the vector θ, we may feel that the components are approximately equal. This suggests shrinking all the coordinates to the overall mean $\bar{X} = \frac{1}{p}\sum_{i=1}^{p} X_i$ which amounts to shrinking toward the subspace V of dimension one generated by the vector $\mathbf{1} = (1,\ldots,1)^t$. The resulting optimal James-Stein type estimator is

$$\delta(X) = \bar{X}\,\mathbf{1} + \left(1 - \frac{(p-3)\,\sigma^2}{\|X - \bar{X}\mathbf{1}\|^2}\right)(X - \bar{X}\,\mathbf{1}).$$

Here, the risk is equal to $3\,\sigma^2$ if in fact all coordinates of θ are equal. If the dimension of the subspace V is also at least 3 we could consider applying a shrinkage estimator to $P_V X$ as well.

It may sometimes pay to break up the whole space into a direct sum of several subspaces and apply shrinkage estimators separately to the different subspaces.

Occasionally it is helpful to shrink toward another estimator. For example, Green and Strawderman [1991] combined two estimators, one of which is unbiased, by shrinking the unbiased estimator toward the biased estimator to obtain a Stein-type improvement over the unbiased estimator.

3.5 A link between Stein's lemma and Stokes's theorem

That a relationship between Stein's lemma and Stokes' theorem (the divergence theorem) is not surprising. Indeed, on one hand, Stein's lemma expresses that, if X has a normal distribution with mean θ and covariance matrix proportional to the identity matrix, the expectation of the inner product of $X - \theta$ and a suitable function g is proportional to the expectation of the divergence of g. On the other hand, when the sets of integration are spheres $S_{r,\theta}$ and balls $B_{r,\theta}$ of radius $r \geq 0$ centered at θ, Stokes' theorem states that the integral of the inner product of the unit outward vector at $x \in S_{r,\theta}$, which is $(x-\theta)/\|x-\theta\|$, with respect to the uniform measure equals the integral of the divergence of g on $B_{r,\theta}$ with respect to the Lebesgue measure.

Most of the time in the literature, Stokes' theorem is considered for a more general open set Ω in $\mathbb{R}^p$ with boundary $\partial\Omega$ which could be less smooth than a sphere, but the function g is often smooth. For example Stroock [1990], considering bounded open set Ω in $\mathbb{R}^p$ for which there exists a function φ from $\mathbb{R}^p$ into $\mathbb{R}$ having continuous third order partial derivatives with the properties that $\Omega = \{x \in \mathbb{R}^p / \varphi(x) < 0\}$ and the gradient $\nabla\varphi$ of φ vanishes at no point where φ itself vanishes, requires that g has continuous first order partial derivatives in a neighborhood of the closure $\bar{\Omega}$ of Ω. For such an open set, its boundary is $\partial\Omega = \{x \in \mathbb{R}^p / \varphi(x) = 0\}$. Then Stroock states that

$$\int_{\partial\Omega} n^t(x)\, g(x) d\sigma(x) = \int_{\Omega} \text{div} g(x)\, dx \tag{3.19}$$

where $n(x)$ is the outer normal (the unit outward vector) to $\partial\Omega$ at $x \in \partial\Omega$ and σ is the surface measure (the uniform measure) on $\partial\Omega$. He provides an elegant proof of Stokes' theorem in (3.19) through a rigorous construction of the outer normal and the surface measure. It is beyond the scope of this book to reproduce Stroock's proof, especially as the link we wish to make with Stein's identity only needs to deal with Ω being a ball and so with $\partial\Omega$ being a sphere. Note that Stroock's conditions are satisfied for a ball of radius $r \geq 0$ centered at $\theta \in \mathbb{R}^p$ with the function $\varphi(x) = \|x-\theta\| - r$. In that context, Stokes' theorem expresses that

$$\int_{S_{r,\theta}} \left(\frac{x-\theta}{\|x-\theta\|}\right)^t g(x)\, d\sigma_{r,\theta}(x) = \int_{B_{r,\theta}} \operatorname{div} g(x)\, dx\,, \tag{3.20}$$

where $\sigma_{r,\theta}$ is the uniform measure on $S_{r,\theta}$.

However, as we have seen in the context of Stein's identity, it is often necessary to deal with functions which are not smooth and are, typically, weakly differentiable. The fact that (3.19) is still valid for such non smooth functions does not seem to have been discussed thoroughly in the literature. Nevertheless, it is stated in Kavian [1993]. Also it should be noticed that Lepelletier [2004], for more general open sets Ω than those considered by Stroock, gives an explicit proof (based on convergences in Sobolev spaces) of that extension in his dissertation. As a consequence, (3.20) is valid for weakly differentiable functions g.

In the following, we will show that Stein' identity can be derived in a straightforward way from this ball-sphere version of Stokes' theorem. Furthermore, and perhaps more interestingly, we will see that the converse is also true: Stein's identity (for which we have an independent proof in Section 3.3) implies directly the ball-sphere Stokes' theorem for weakly differentiable functions.

Let $X \sim N_p(\theta, \tau^2 I_p)$ and let $g(X)$ be a weakly differentiable function from $\mathbb{R}^p$ into $\mathbb{R}^p$ such that either $E_\theta[\|(X-\theta)^t g(X)\|] < \infty$ or $E_\theta[|\text{div} g(X)|] < \infty$. Integrating through uniform measures on spheres, we have

$$\begin{aligned} E_{\theta,\tau^2}[(X-\theta)^t g(X)] &= \int_{\mathbb{R}^p} (x-\theta)^t g(x) \, \frac{1}{(2\,\pi\,\tau^2)^{p/2}} \exp\left(-\frac{\|x-\theta\|^2}{2\,\tau^2}\right) dx \\ &= \int_0^\infty \int_{S_{r,\theta}} \left(\frac{x-\theta}{\|x-\theta\|}\right)^t g(x)\, d\sigma_{r,\theta}(x)\, \psi_{\tau^2}(r)\, dr \qquad (3.21) \end{aligned}$$

where

$$\psi_{\tau^2}(r) = \frac{1}{(2\,\pi\,\tau^2)^{p/2}}\, r \exp\left(-\frac{r^2}{2\,\tau^2}\right) \qquad (3.22)$$

and $\sigma_{r,\theta}$ is the uniform measure on $S_{r,\theta}$. Then applying Stokes' theorem in (3.20) to the inner most integral in (3.21) gives

$$E_{\theta,\tau^2}[(X-\theta)^t g(X)] = \int_0^\infty \int_{B_{r,\theta}} \text{div} g(x)\, dx\, \psi_{\tau^2}(r)\, dr\,. \qquad (3.23)$$

Now, applying Fubini's theorem to the right-hand side of (3.23), we have

$$\begin{aligned} \int_{B_{r,\theta}} \text{div} g(x)\, dx\, \psi_{\tau^2}(r)\, dr &= \int_{\mathbb{R}^p} \text{div} g(x) \int_{\|x-\theta\|}^\infty \psi_{\tau^2}(r)\, dr\, dx \\ &= \int_{\mathbb{R}^p} \text{div} g(x)\, \frac{1}{(2\,\pi\,\tau^2)^{p/2}} \left[-\tau^2 \exp\left(-\frac{r^2}{2\,\tau^2}\right)\right]_{\|x-\theta\|}^\infty dx \\ &= \tau^2 \int_{\mathbb{R}^p} \text{div} g(x)\, \frac{1}{(2\,\pi\,\tau^2)^{p/2}} \exp\left(-\frac{\|x-\theta\|^2}{2\,\tau^2}\right) dx \\ &= \tau^2\, E_{\theta,\tau^2}[\text{div} g(X)] \qquad (3.24) \end{aligned}$$

since, according to (3.22),

$$\frac{\partial}{\partial r}\left\{\frac{1}{(2\,\pi\,\tau^2)^{p/2}}\left[-\tau^2\exp\left(-\frac{r^2}{2\,\tau^2}\right)\right]\right\}=\psi_{\tau^2}(r)\,.$$

Therefore (3.23) and (3.24) express that

$$E_{\theta,\tau^2}[(X-\theta)^t g(X)]=\tau^2\,E_{\theta,\tau^2}[\mathrm{div}g(X)]\,, \tag{3.25}$$

which is Stein's identity.

Conversely, assume that Stein's identity in (3.25) is satisfied. As in (3.21), we have

$$E_{\theta,\tau^2}[(X-\theta)^t g(X)]=\int_0^\infty\int_{S_{r,\theta}}\left(\frac{x-\theta}{\|x-\theta\|}\right)^t g(x)\,d\sigma_{r,\theta}(x)\,\psi_{\tau^2}(r)\,dr \tag{3.26}$$

and also, as in (3.24), we have

$$\tau^2\,E_{\theta,\tau^2}[\mathrm{div}g(X)]=\int_0^\infty\int_{B_{r,\theta}}\mathrm{div}g(x)\,dx\,\psi_{\tau^2}(r)\,dr\,. \tag{3.27}$$

Hence it follows from (3.25), (3.26) and (3.27) that, for all $\tau^2>0$,

$$\int_0^\infty\int_{S_{r,\theta}}\left(\frac{x-\theta}{\|x-\theta\|}\right)^t g(x)\,d\sigma_{r,\theta}(x)\,\psi_{\tau^2}(r)\,dr=\int_0^\infty\int_{B_{r,\theta}}\mathrm{div}g(x)\,dx\,\psi_{\tau^2}(r)\,dr\,.$$

Therefore, since the family $(\psi_{\tau^2}(r))_{\tau^2>0}$ defined in (3.22)) is complete as an exponential family, we have

$$\int_{S_{r,\theta}}\left(\frac{x-\theta}{\|x-\theta\|}\right)^t g(x)\,d\sigma_{r,\theta}(x)=\int_{B_{r,\theta}}\mathrm{div}g(x)\,dx\,, \tag{3.28}$$

for almost every $r\geq 0$ (θ being fixed). Since these functions are continuous in r, (3.28) holds for all $r\geq 0$, which is Stokes' theorem in (3.20).

We have seen that, for balls and spheres, Stokes' theorem can be directly derived from Stein's identity, for weakly differentiable functions. This result will be particularly important for proving Stein type identities for spherically symmetric distributions in Chapters 5 and 6. Note that, in fact, we have obtained a stronger result. Indeed, it is actually shown that, any time Stein's identity is valid, then Stokes' theorem holds as well. This result is particularly interesting when the weak differentiability assumption is not met. For example, Fourdrinier, Strawderman and Wells [2006] noticed that this may be the case when dealing with a location parameter restricted to a cone; Stein's identity (3.4) holds but the weak differentiability of the functions at hand is not guaranteed.

3.6 Differential operators and dimension cut-off when estimating a mean

In the previous sections, we have seen that, in the normal case, when estimating the mean θ, the MLE X is admissible when $p \leq 2$, but inadmissible when $p \geq 3$. Although it is specific to the normal case, it can be extended to other distributional settings (such as exponential families) so that this dimension cut-off should reflect a more fundamental mathematical phenomenon. Below, we give an insight into such phenomena in terms of non linear partial differential operators.

Indeed, when estimating θ under quadratic loss, improvements on X through unbiased estimation techniques often involve a nonlinear partial differential operators for which the necessary nonpositivity will imply higher dimensions. This operator is of the form

$$\mathscr{R}g(x) = k\,\mathrm{div}g(x) + \|g(x)\|^2 \tag{3.29}$$

for a certain constant k, the sufficient improvement condition being typically

$$\mathscr{R}g(x) \leq 0 \tag{3.30}$$

for all $x \in \mathbb{R}^p$ (with strict inequality on a set of positive Lebesgue measure). We will see that Inequality (3.30) has no nontrivial solution g (i.e. g is not equal to 0 almost everywhere) when the space dimension p is less than or equal to 2, even if we look for solutions with smoothness conditions as weak as possible. Consequently, a necessary dimension condition for (3.30) to have solutions $g \not\equiv 0$ is $p \geq 3$.

Here follows the precise statement of this fact.

Theorem 3.7. *Let $k \in \mathbb{R}$ fixed. When $p \leq 2$, the only solution g in $L^2_{loc}(\mathbb{R}^p)$ of*

$$\mathscr{R}g(x) = k\,\mathrm{div}g(x) + \|g(x)\|^2 \leq 0\,, \tag{3.31}$$

for any $x \in \mathbb{R}$, is $g = 0$ (a.e.).

Note that, in Theorem 3.7, the search of solutions of Inequation (3.31) is addressed in a very general setting. Indeed the g's are first sought in the space of distributions $\mathscr{D}'(\mathbb{R}^p)$ (see Schwartz [1973] for a full account) so that the nature of the result is not due to whichever regularity conditions on the solutions we may choose. Nevertheless it is worth noticing that defining the non linear term $\|g\|^2$ in (3.31) as a distribution prompts to seek g in $L^2_{loc}(\mathbb{R}^p)$ (in which case the linear part of $\mathscr{R}g$ is well defined). This is still a wide space for possible solutions g.

The proof of Theorem 3.7 is based on the use of the following sequence of the so-called test functions. Let φ be a positive infinitely differentiable function on $\mathbb{R}_+$ bounded by 1, identically equal to 1 on $[0,1]$ and with support the interval $[0,2]$ ($\mathrm{supp}(\varphi) = [0,2]$). Associate to φ the sequence $(\varphi_n)_{n\geq 1}$ of infinitely differentiable functions from $\mathbb{R}^p$ into $[0,1]$ defined through

$$\forall n \geq 1 \quad \forall x \in \mathbb{R}^p \quad \varphi_n(x) = \varphi\left(\frac{||x||}{n}\right). \tag{3.32}$$

Clearly, for any $n \geq 1$, the function φ_n has compact support B_{2n}, the closed ball of radius $2n$ and centered at 0 in $\mathbb{R}^p$. Also an interesting property is that, for any $\beta \geq 1$ and for any $j = 1,\dots,p$,

$$\left|\frac{\partial \varphi_n^\beta}{\partial x_j}(x)\right| \leq \frac{K}{n}\,\varphi_n^{\beta-1}(x)\,. \tag{3.33}$$

Note that, as all the derivatives of φ vanish out of the compact [1, 2] and φ is bounded by 1, Inequality (3.33) can be refined in

$$\left|\frac{\partial \varphi_n^\beta}{\partial x_j}(x)\right| \leq \frac{K}{n}\;\mathbb{1}_{C_n}(x)\,. \tag{3.34}$$

where $\mathbb{1}_{C_n}$ is the indicator function of the annulus $C_n = \{x \in \mathbb{R}^p | n \leq ||x|| \leq 2n\}$.

Proof of Theorem 3.7. Let $g \in L^2_{loc}(\mathbb{R}^p)$ satisfying (3.31). Then, through the duality brackets between the space of distributions $\mathscr{D}'(\mathbb{R}^p)$ and the space $\mathrm{C}_0^\infty(\mathbb{R}^p)$ of infinitely differentiable functions on $\mathbb{R}^p$ with compact support, we have, for any $n \in \mathbb{N}^*$ and any $\beta > 0$,

$$\begin{aligned}\int_{\mathbb{R}^p} \|g(x)\|^2 \varphi_n^\beta(x)\,dx &\le -k\left\langle \operatorname{div} g, \varphi_n^\beta \right\rangle \\ &= -k\sum_{i=1}^p \left\langle \frac{\partial}{\partial x_i} g_i, \varphi_n^\beta \right\rangle \\ &= k\sum_{i=1}^p \left\langle g_i, \frac{\partial}{\partial x_i} \varphi_n^\beta \right\rangle . \qquad (3.35)\end{aligned}$$

Now, since the distribution g lies in $L^2_{loc}(\mathbb{R}^p)$, we can express (3.35) as

$$\begin{aligned}\int_{\mathbb{R}^p} \|g(x)\|^2 \varphi_n^\beta(x)\,dx &\le k\sum_{i=1}^p \int_{\mathbb{R}^p} g_i(x)\,\frac{\partial}{\partial x_i}\varphi_n^\beta(x)\,dx \\ &= k\int_{\mathbb{R}^p} g^t\,\nabla\varphi_n^\beta(x)\,dx \\ &\le k\int_{\mathbb{R}^p} \|g(x)\|\,\|\nabla\varphi_n^\beta(x)\|\,dx. \qquad (3.36)\end{aligned}$$

Then, using (3.33), it follows from (3.36) that there exists a constant $C > 0$ such that

$$\begin{aligned}\int_{\mathbb{R}^p} \|g(x)\|^2 \varphi_n^\beta(x)\,dx &\le \frac{C}{n}\int_{\mathbb{R}^p} \|g(x)\|\,\varphi_n^{\beta-1}(x)\,dx \\ &\le \frac{C}{n}\left(\int_{\mathbb{R}^p} \varphi_n^{\beta-2}(x)\,dx\right)^{1/2}\left(\int_{\mathbb{R}^p} \|g(x)\|^2\,\varphi_n^\beta(x)\,dx\right)^{1/2} (3.37)\end{aligned}$$

applying Schwarz's inequality with $\beta > 2$ and

$$\|g(x)\|\,\varphi_n^{\beta-1}(x) = \varphi_n^{\beta/2-1}(x)\;\|g(x)\|\,\varphi_n^{\beta/2}(x)\,.$$

Clearly (3.37) is equivalent to

$$\int_{\mathbb{R}^p} \|g(x)\|^2 \varphi_n^\beta(x)\,dx \le \frac{C^2}{n^2}\int_{\mathbb{R}^p} \varphi_n^{\beta-2}(x)\,dx. \qquad (3.38)$$

Thus, since $1\!\!1_{B_n} \le \varphi_n \le 1$ with $\operatorname{supp}\varphi_n \subset B_n$, restricting the first integral of (3.38) over B_n leads to

$$\int_{B_n} \|g(x)\|^2\,dx \le \frac{C^2}{n^4}\int_{B_n} dx = A\,n^{p-2} \qquad (3.39)$$

for some constant $A > 0$. Letting n go to infinity in (3.39) shows that, when $p < 2$, $g = 0$ almost everywhere, which proves the theorem in that case.

Consider now the case $p = 2$. Note that the above reasoning for $p = 2$ guarantees that $g \in L^2(\mathbb{R}^p)$. The result will follow in applying Inequality (3.34). Indeed it follows from (3.34) and the first inequality in (3.37) that, for some constant $C > 0$,

$$\begin{aligned}\int_{B_n} \|g(x)\|^2\, dx &\le \frac{C}{n} \int_{C_n} \|g(x)\|\, dx \\ &\le \frac{C}{n} \left(\int_{C_n} dx \right)^{1/2} \left(\int_{C_n} \|g(x)\|^2\, dx \right)^{1/2} \end{aligned} \tag{3.40}$$

by Schwarz's inequality. Now

$$\int_{C_n} dx \le \int_{B_{2n}} dx \propto n^2 \tag{3.41}$$

since $p = 2$. Hence (3.40) and (3.41) imply that, for some constant $A > 0$,

$$\int_{B_n} \|g(x)\|^2\, dx \le A \left(\int_{C_n} \|g(x)\|^2\, dx \right)^{1/2} . \tag{3.42}$$

As $g \in L^2(\mathbb{R}^p)$, we have

$$\lim_{n\to\infty} \int_{C_n} \|g(x)\|^2\, dx = 0$$

and hence (3.42) gives rise to

$$0 = \lim_{n\to\infty} \int_{C_n} \|g(x)\|^2\, dx = \int_{\mathbb{R}^p} \|g(x)\|^2\, dx ,$$

which implies that $g = 0$ almost everywhere and gives the desired result for $p = 2$. □

Such a dimension cut-off result which states that the usual Stein's inequality $2\,\mathrm{div} g(x) + ||g(x)||^2 \le 0$ (for any $x \in \mathbb{R}^p$) has no nontrivial solution g in $\left(L^2_{loc}(\mathbb{R}^p)\right)^p$

when $p \leq 2$, reinforces the fact that the MLE X is admissible in dimension $p \leq 2$ when estimating a normal mean. The above proof of Theorem 3.7 follows closely the approach of Blanchard and Fourdrinier [1999] (to which we refer to for a full account on that dimension cut-off phenomenon) where more general non linear partial differential inequalities are considered. It will be reused in Chapter 8 to prove that, for an inequality of the form $k\,\Delta\gamma(x) + \gamma^2(x) \leq 0$, the same dimension cut-off phenomenon occurs for $p \leq 4$ (there is no nontrivial solution γ in $\left(L^2_{loc}(\mathbb{R}^p)\right)^p$ when $p \leq 4$).

Chapter 4

Estimation of a normal mean vector II

4.1 Bayes minimax estimators

In this section, we derive a general sufficient condition for minimaxity of Bayes and generalized Bayes estimators when $X \sim N_p(\theta, \sigma^2 I_p)$, with known σ^2, and the loss function is $\|\delta - \theta\|^2$, due to Stein [1973, 1981]. The condition depends only on the marginal distribution and states that a generalized Bayes estimator is minimax if the square root of the marginal distribution is superharmonic. Alternative (stronger) sufficient conditions are that the prior distribution or the marginal distribution are superharmonic. We establish these results in Subsection 4.1.1 and apply them in Subsection 4.1.2 to establish classes of prior distributions which lead to minimax (generalized and proper) Bayes estimators. Subsection 4.1.3 will be devoted to minimax multiple shrinkage estimators.

Throughout this section, let $X \sim N_p(\theta, \sigma^2 I_p)$ and loss be $L(\theta, \delta) = \|\delta - \theta\|^2$. Let θ have the (generalized) prior distribution π and let the marginal density, $m(x)$, of X be

$$m(x) = K \int_{\mathbb{R}^p} e^{-\frac{\|x-\theta\|^2}{2\sigma^2}} \, d\pi(\theta). \tag{4.1}$$

Recall from Section 2.2 that the Bayes estimator corresponding to $\pi(\theta)$ is given by

$$\delta_\pi(X) = X + \sigma^2 \frac{\nabla m(X)}{m(X)}. \tag{4.2}$$

Since the constant K in (4.1) plays no role in (4.2) we will typically take it to be equal to 1 for simplicity. It may happen that an estimator will have the form (4.2) where $m(X)$ does not correspond to a true marginal distribution. In this case we will refer to such an estimator as a pseudo-Bayes estimator, provided $x \mapsto \nabla m(x)/m(x)$ is weakly differentiable. Recall that, if $\delta_\pi(X)$ is generalized Bayes, $x \mapsto m(x)$ is a positive analytic function and so $x \mapsto \nabla m(x)/m(x)$ is automatically weakly differentiable.

4.1.1 A sufficient condition for minimaxity of (proper, generalized, and pseudo) Bayes estimators

Stein [1973, 1981] gave the following sufficient condition for a generalized Bayes estimator to be minimax.

Theorem 4.1. *Under the model of this section, an estimator of the form (4.2) has finite risk if* $E_\theta\left[\|\nabla m(X)/m(X)\|^2\right] < \infty$ *and is minimax provided* $x \mapsto \sqrt{m(x)}$ *is superharmonic (i.e.,* $\Delta \sqrt{m(x)} \leq 0$*, for any* $x \in \mathbb{R}^p$*).*

Proof. Using Corollary 3.1 and the fact that $\delta_\pi(X) = X + \sigma^2 g(X)$ with $g(X) = \nabla m(X)/m(X)$, the estimator $\delta_\pi(X)$ has finite risk if $E_\theta\left[\|\nabla m(X)/m(X)\|^2\right] < \infty$.

Also, it is minimax provided, for almost any $x \in \mathbb{R}^p$,

$$\mathscr{D}(x) = \frac{\|\nabla m(x)\|^2}{m^2(x)} + 2\,\mathrm{div}\frac{\nabla m(x)}{m(x)} \le 0\,.$$

Now, for any $x \in \mathbb{R}^p$,

$$\mathscr{D}(x) = \frac{\|\nabla m(x)\|^2}{m^2(x)} + 2\,\frac{m(x)\,\Delta m(x) - \|\nabla m(x)\|^2}{m^2(x)}$$

where

$$\Delta m(x) = \sum_{i=1}^{p} \frac{\partial^2}{\partial x_i^2} m(x)$$

is the Laplacian of $m(x)$. Hence, by straightforward calculation,

$$\begin{aligned} \mathscr{D}(x) &= \frac{2\,m(x)\,\Delta m(x) - \|\nabla m(x)\|^2}{m^2(x)} \\ &= 4\,\frac{\Delta\sqrt{m(x)}}{\sqrt{m(x)}}\,. \end{aligned} \tag{4.3}$$

Therefore $\mathscr{D}(x) \le 0$ since $x \mapsto \sqrt{m(x)}$ is superharmonic. □

It is convenient to assemble the following results for the case of spherically symmetric marginals. The proof is straightforward and left to the reader.

Corollary 4.1. *Assume the prior density $\pi(\theta)$ is spherically symmetric around 0 (i.e., $\pi(\theta) = \pi(\|\theta\|^2)$). Then*

(1) *the marginal density m of X is spherically symmetric around 0 (i.e., $m(x) = m(\|x\|^2)$, for any $x \in \mathbb{R}^p$);*

(2) *the Bayes estimator equals*

$$\delta_\pi(X) = X + 2\,\sigma^2\,\frac{m'(\|X\|^2)}{m(\|X\|^2)}\,X$$

and has the form of a Baranchik estimator (3.12) with

$$a\,r(t) = -2\,\frac{m'(t)}{m(t)}\,t \qquad \forall t \geq 0\,;$$

(3) *The unbiased estimator of the risk difference between $\delta_\pi(X)$ and X is given by*

$$\mathscr{D}(X) = 4\,\sigma^4 \left\{ p\,\frac{m'(\|X\|^2)}{m(\|X\|^2)} + 2\,\|X\|^2\,\frac{m''(\|X\|^2)}{m(\|X\|^2)} - \|X\|^2 \left(\frac{m'(\|X\|^2)}{m(\|X\|^2)}\right)^2 \right\}.$$

While, in Theorem 4.1, minimaxity of $\delta_\pi(X)$ follows from the superharmonicity of $\sqrt{m(X)}$, it is worth noting that, in the setting of Corollary 4.1, it can be obtained from the concavity of $t \mapsto m^{1/2}(t^{2/(2-p)})$.

The following corollary is often useful. It shows that $\sqrt{m(X)}$ is superharmonic if $m(X)$ is superharmonic which in turn follows if the prior density $\pi(\theta)$ is superharmonic.

Corollary 4.2. (1) *A finite risk (generalized, proper or pseudo) Bayes estimator of the form (2.5.2) is minimax provided the marginal m is superharmonic (i.e. $\Delta m(x) \leq 0$, for any $x \in \mathbb{R}^p$).*

(2) *If the prior distribution has a density, π, which is superharmonic, then a finite risk generalized or proper Bayes estimator of the form (4.2) is minimax.*

Proof. Part (1) follows from the first equality in (4.3), which shows that superharmonicity of m implies superharmonicity of $\sqrt{m}$. Indeed more generally superharmonicity of m implies superharmonicity of any nondecreasing concave function of m.

Part (2) follows since, for any $x \in \mathbb{R}^p$,

$$\begin{aligned}
\Delta_x m(x) &= \Delta_x \int_{\mathbb{R}^p} \exp\left(-\frac{1}{2\,\sigma^2}\|x-\theta\|^2\right)\pi(\theta)\,d\theta \\
&= \int_{\mathbb{R}^p} \Delta_x \exp\left(-\frac{1}{2\,\sigma^2}\|x-\theta\|^2\right)\pi(\theta)\,d\theta \\
&= \int_{\mathbb{R}^p} \Delta_\theta \exp\left(-\frac{1}{2\,\sigma^2}\|x-\theta\|^2\right)\pi(\theta)\,d\theta \\
&= \int_{\mathbb{R}^p} \exp\left(-\frac{1}{2\,\sigma^2}\|x-\theta\|^2\right)\Delta_\theta \pi(\theta)\,d\theta
\end{aligned}$$

where the second equality follows from exponential family properties and the last equality is a Green's formula (see the Appendix). More generally any mixture of superharmonic functions is superharmonic. □

Note that the condition of finiteness of risk is superfluous for proper Bayes estimators (for which the convex hull of the support of π is $\mathbb{R}^p$) since the Bayes risk is bounded above by $p\,\sigma^2$, and Fubini's theorem assures that the risk function is finite a.e. (π). Continuity of the risk function implies finiteness for all θ in the convex hull of the support of π (see Berger [1985] and Lehmann and Casella [1998] for more discussion on finiteness and continuity of risk).

As an example of a pseudo-Bayes estimator, consider $m(X)$ of the form

$$m(X) = \frac{1}{(\|X\|^2)^b}\,.$$

The case $b = 0$ corresponds to $m(X) = 1$ which is the marginal corresponding to the "uniform" generalized prior distribution $\pi(\theta) \equiv 1$, which in turn corresponds to the generalized Bayes estimator $\delta_0(X) = X$. If $b > 0$, $m(X)$ is unbounded in a neighborhood of 0 and hence is not analytic. Thus $m(X)$ cannot be a true marginal (for any generalized prior). However

$$\nabla m(X) = \frac{-2b}{(\|X\|^2)^{b+1}} X \qquad \text{and} \qquad \frac{\nabla m(X)}{m(X)} = \frac{-2b}{\|X\|^2} X$$

which is weakly differentiable if $p \geq 3$ (see Section 3.3). Hence for $p \geq 3$, the James-Stein estimator

$$\delta_{2b}^{JS}(X) = \left(1 - \frac{2b\sigma^2}{\|X\|^2}\right) X$$

is a pseudo-Bayes estimator. Also a simple calculation gives

$$\Delta m(X) = \frac{(-2b)[p-2(b+1)]}{(\|X\|^2)^{b+1}}.$$

It follows that $m(X)$ is superharmonic for $0 \leq b \leq (p-2)/2$ and similarly that $\sqrt{m(X)}$ is superharmonic for $0 \leq b \leq p-2$. An application of Theorem 4.1 gives minimaxity for $0 \leq b \leq p-2$ which agrees with Theorem 3.2 (with $a = 2b$), while an application of Corollary 4.1 establishes minimaxity for only half of the interval, i.e. $0 \leq b \leq (p-2)/2$. Thus the corollary, while useful, is considerably weaker than the theorem.

Another interesting aspect of this example relates to the existence of proper Bayes minimax estimators for $p \geq 5$. Considering the behavior of $m(x)$ for $\|x\| \geq R$ for some positive R, note that

$$\int_{\|x\| \geq R} m(x)\, dx = \int_{\|x\| \geq R} \frac{1}{(\|X\|^2)^b}\, dX \propto \int_R^\infty \frac{r^{p-1}}{r^{2b}}\, dr = \int_R^\infty r^{p-2b-1}\, dr$$

and that this integral is finite if and only if $p - 2b < 0$. Thus integrability of $m(x)$ for $\|x\| \geq R$ and minimaxity of the (James-Stein) pseudo-Bayes estimator corresponding to $m(X)$ are possible if and only if $p/2 < b \leq p-2$ which implies $p \geq 5$.

It is also interesting to note that superharmonicity of $m(X)$ (i.e. $0 \leq b \leq (p-2)/2$) is incompatible with integrability of $m(x)$ on $\|x\| \geq R$ (i.e. $b > p/2$). This is illustrative of a general fact that a generalized Bayes minimax estimator corresponding to a superharmonic marginal cannot be proper Bayes (see Theorem 4.2).

4.1.2 Construction of (proper and generalized) minimax Bayes estimators

Corollary 4.1 provided a method of constructing pseudo-Bayes minimax estimators. In this section we concentrate on the construction of proper and generalized Bayes minimax estimators. The results in this section are primarily from Fourdrinier, Strawderman and Wells [1998]. Although Corollary 4.1 is helpful in constructing minimax estimators it cannot be used to develop proper Bayes minimax estimators as indicated in the example at the end of the previous section. The following result establishes that a superharmonic marginal (and hence a superharmonic prior density) cannot lead to a proper Bayes estimator.

Theorem 4.2. *Let m be a superharmonic marginal density corresponding to a prior π. Then π is not a probability measure.*

Proof. Assume π is a probability measure. Then it follows that m is an integrable, strictly positive and bounded function in C_∞ (the space of functions which have derivatives of all orders). Recall also from Example 2.1 of Section 2.2 that the posterior risk is given, for any $x \in \mathbb{R}^p$, by

$$p\,\sigma^2 + \sigma^4\,\frac{m(x)\,\Delta m(x) - \|\nabla m(x)\|^2}{m^2(x)}$$

and hence the Bayes risk is

$$r(\pi) = E^m\left[p\sigma^2 + \sigma^4 \frac{m(X)\Delta m(X) - \|\nabla m(X)\|^2}{m^2(X)}\right],$$

where E^m is the expectation with respect to the marginal density m. Also, denoting by E^π the expectation with respect to the prior π, we may use the unbiased estimate of risk to express $r(\pi)$ as

$$r(\pi) = E^\pi\left[E_\theta\left[p\,\sigma^2 + \sigma^4\,\tfrac{2\,m(X)\Delta m(X) - \|\nabla m(X)\|^2}{m^2(X)}\right]\right]$$

$$= E^m\left[p\,\sigma^2 + \sigma^4\,\tfrac{2m(X)\Delta m(X) - \|\nabla m(X)\|^2}{m^2(X)}\right],$$

since, by definition, the unbiased estimate of risk does not depend on θ. Hence, by taking the difference,

$$E^m\left[\frac{\Delta m(X)}{m(X)}\right] = 0\,.$$

Now, since the marginal m is superharmonic ($\Delta m(x) \leq 0$ for any $x \in \mathbb{R}^p$), strictly positive and in C_∞, it follows that $\Delta m \equiv 0$. Finally, the strict positivity and harmonicity of m implies that $m \equiv C$ where C is a positive constant (see Doob 1984), and hence, that $\int_{\mathbb{R}^p} m(X)\,dx = \infty$, which contradicts the integrability of m. □

We now turn to the construction of Bayes minimax estimators. Consider prior densities of the form

$$\pi(\theta) = k\int_0^\infty \exp\left(-\frac{\|\theta\|^2}{2\,\sigma^2\,v}\right) v^{-p/2}\,h(v)\,dv \qquad (4.4)$$

for some constant k and some nonnegative function of h on $\mathbb{R}^+$ such that the integral exists, i.e. $\pi(\theta)$ is a variance mixture of normal distributions. It follows from Fubini's theorem that, for any $x \in \mathbb{R}^p$,

$$m(x) = \int_0^\infty m_v(x)\, h(v)\, dv$$

where

$$m_v(x) = k \exp\left(-\frac{\|x\|^2}{2\,\sigma^2\,(1+v)}\right)(1+v)^{-p/2}\,.$$

Lebesgue's Dominated Convergence theorem ensures that we may differentiate under the integral sign and so

$$\nabla m(x) = \int_0^\infty \nabla m_v(x)\, h(v)\, dv \tag{4.5}$$

and

$$\Delta m(x) = \int_0^\infty \Delta m_v(x)\, h(v)\, dv \tag{4.6}$$

where

$$\nabla m_v(x) = -\frac{k}{\sigma^2} \exp\left(-\frac{\|x\|^2}{2\,\sigma^2\,(1+v)}\right)(1+v)^{-p/2-1}\, x$$

and

$$\Delta m_v(x) = -\frac{k}{\sigma^2}\left[p - \frac{\|x\|^2}{\sigma^2(1+v)}\right] \exp\left(-\frac{\|x\|^2}{2\,\sigma^2\,(1+v)}\right)(1+v)^{-p/2-1}$$

Then the following integral

$$I_k(Y) = \int_0^\infty e^{-\frac{y}{1+v}}\,(1+v)^{-j}\, h(v)\, dv$$

exists for $j \geq p/2$. Hence, with $y = \|x\|^2/2\sigma^2$, we have

$$\begin{aligned} m(x) &= kI_{p/2}(y) \\ \nabla m(x) &= -\frac{k}{\sigma^2}\, kI_{p/2+1}(y)x \\ \Delta m(x) &= -\frac{k}{\sigma^2}\left[pI_{p/2+1}(y) - 2yI_{p/2+2}(y)\right] \\ \|\nabla m(x)\|^2 &= 2\,\frac{k^2}{\sigma^2}\, yI^2_{\frac{p}{2}+1}(y). \end{aligned} \tag{4.7}$$

Note that

$$\frac{\|\nabla m(x)\|^2}{m^2(x)} = \frac{2}{\sigma^2}\,\frac{I^2_{p/2+1}(y)}{I^2_{\frac{p}{2}}(y)}\, y \le \frac{2y}{\sigma^2} = \frac{\|x\|^2}{\sigma^4}$$

since $I_{j+p}(y) \le I_j(y)$, and hence

$$E_0\left[\frac{\|\nabla m(x)\|^2}{m^2(x)}\right] \le E_0\left[\frac{\|x\|^2}{\sigma^4}\right] < \infty\,,$$

which, according to Theorem 4.1, guarantees the finiteness of the risk of the Bayes estimator $\delta_\pi(X)$ in (4.2). Furthermore the unbiased estimator of risk (4.3) can be expressed as

$$\begin{aligned} \mathscr{D}(X) &= -\tfrac{2}{\sigma^2}\left[pI_{p/2+1}(y) - 2yI_{p/2+2}(y)\right]/I_{p/2}(y) \\ &\quad -\tfrac{2}{\sigma^2}\left[yI^2_{p/2+1}(y)/I^2_{p/2}(y)\right] \\ &= \tfrac{2I_{p/2+1}(y)}{\sigma^2 I_{p/2}(y)}\left[\tfrac{2yI_{p/2+2}(y)}{I_{p/2+1}(y)} - p - \tfrac{yI_{p/2+1}(y)}{I_{p/2}(y)}\right]. \end{aligned} \tag{4.8}$$

Then the following intermediate result follows immediately from (4.2) and Theorem 4.1 since finiteness of risk has been guaranteed above.

Lemma 4.1. *The generalized Bayes estimator corresponding to the prior density (4.4) is minimax provided*

$$\frac{2I_{p/2+2}(y)}{I_{p/2+1}(y)} - \frac{I_{p/2+1}(y)}{I_{p/2}(y)} \le \frac{p}{y}\,. \tag{4.9}$$

The next theorem gives sufficient conditions on the mixing density $h(\cdot)$ so that the resulting generalized Bayes estimator is minimax.

Theorem 4.3. *Let h be a positive differentiable function such that the function $-(v+1)h'(v)/h(v) = l_1(v) + l_2(v)$ where $l_1(v) \leq A$ and is nondecreasing while $0 \leq l_2 \leq B$ with $A+2B \leq (p-2)/2$. Assume also that* $\lim_{v\to\infty} h(v)/(v+1)^{p/2-1} = 0$ *and that $\int_0^\infty \exp(-y/(1+v))\,(1+v)^{-p/2}\,h(v)\,dv < \infty$. Then the generalized Bayes estimator (4.2) for the prior density (4.4) corresponding to the mixing density h is minimax. Furthermore if h is integrable, the resulting estimator is also proper Bayes.*

Proof. Via integration by parts, we first find an alternative expression for

$$I_k(y) = \int_0^\infty \exp(-y/(1+v))\,(1+v)^{-k/2}\,h(v)\,dv$$

Letting $u = (1+v)^{-k+2}h(v)$ and $dw = (1+v)^{-2}\exp(-y/(1+v))\,dv$, so that $du = (-k+2)(1+v)^{-k+1}\,h(v) + (1+v)^{-k+2}h'(v)$ and $w = \exp(-y/(1+v))/y$, we have (for $k \geq p/2+1$)

$$\begin{aligned}
I_k(y) &= \frac{(1+v)^{-k+2}\exp(-y/(1+v))\,h(v)}{y}\Big|_0^\infty \\
&\qquad + \frac{k-2}{y}\int_0^\infty \exp\left(-\frac{y}{1+v}\right)(1+v)^{-k+1}\,h(v)\,dv \\
&\qquad\qquad - \frac{1}{y}\int_0^\infty \exp\left(-\frac{y}{1+v}\right)(1+v)^{-k+2}\,h'(v)\,dv \\
&= -\frac{e^{-y}\,h(0)}{y} + \frac{k-2}{y}I_{k-1}(y) \\
&\qquad - \frac{1}{y}\int_0^\infty \exp\left(-\frac{y}{1+v}\right)(1+v)^{-k+2}\,h'(v)\,dv\,. \qquad (4.10)
\end{aligned}$$

Applying (4.10) to both numerators in the left-hand side of (4.9) we have

$$
\begin{aligned}
&\frac{2}{I_{p/2+1}(y)}\left[\frac{-e^{-y}h(0)}{y}+\frac{p}{2y}I_{p/2+1}(y)-\frac{1}{y}\int_0^\infty \exp\left(-\frac{y}{1+v}\right)(1+v)^{-p/2}h'(v)\,dv\right]\\
&-\frac{1}{I_{p/2}(y)}\left[\frac{-e^{-y}h(0)}{y}+\frac{p-2}{2y}I_{p/2}(y)-\frac{1}{y}\int_0^\infty \exp\left(-\frac{y}{1+v}\right)(1+v)^{-p/2+1}h'(v)\,dv\right]\\
&\leq \frac{p+2}{2y}-\frac{2\int_0^\infty \exp\left(-\frac{y}{1+v}\right)(1+v)^{-p/2+2}h'(v)\,dv}{yI_{p/2+1}(y)}\\
&\qquad\qquad\qquad\qquad +\frac{\int_0^\infty \exp\left(-\frac{y}{1+v}\right)(1+v)^{-p/2+1}h'(v)\,dv}{yI_{p/2}(y)}
\end{aligned}
$$

since $I_{p/2+1}(y) < I_{p/2}(y)$.

Then it follows from Lemma 4.1 that $\delta_\pi(X)$ is minimax provided, for any $y \geq 0$,

$$J_p^y \leq p - \frac{p+2}{2} = \frac{p-2}{2},$$

where

$$J_p^y = -2E_{p/2+1}^y\left[(V+1)\frac{h'(V)}{h(V)}\right] + E_{p/2}^y\left[(V+1)\frac{h'(V)}{h(V)}\right]$$

and where $E_k^y[f(V)]$ is the expectation of $f(V)$ with respect to the random variable V with density $g_k^y(v) = \exp(-y/(1+v))\,(1+v)^{-k/2}h(v)/I_k(y)$. Setting $-(v+1)$ $h'(v)/h(v) = l_1(v) + l_2(v)$ and noting that $g_k^y(v)$ has monotone decreasing likelihood ratio in k, for fixed y, we have

$$
\begin{aligned}
J_p^y &= 2E_{p/2+1}^y[l_1(V)+l_2(V)] - E_{p/2}^y[l_1(V)+l_2(V)]\\
&\leq 2E_{p/2+1}^y[l_1(V)] - E_{p/2}^y[l_1(V)] + 2E_{p/2+1}^y[l_2(V)]
\end{aligned}
$$

since l_1 is nondecreasing. Then

$$J_p^y \leq E_{p/2}^y[l_1(V)] + 2E_{p/2+1}^y[l_2(V)] \leq A + 2B \leq \frac{p-2}{2}.$$

since $l_1 \leq A$ and $l_2 \leq B$ and by assumption on A and B. Hence the result follows. □

The above theorem gives sufficient conditions for minimaxity of a generalized Bayes estimator with respect to a prior $\pi(\theta)$ which is a variance mixture of normals with mixing density h. The following corollary allows the construction of mixing distributions so that the conditions of the theorem are met.

Corollary 4.3. *Let $\psi = \psi_1 + \psi_2$ be a continuous function such that $\psi_1 \leq C$ and is non decreasing, while $0 \leq \psi_2 \leq D$, and where $C \leq -2D$.*

Define, for $v > 0$, $h(v) = \exp\left[-\frac{1}{2}\int_{v_0}^{v} \frac{2\psi(u)+p-2}{v+1}\, du\right]$ where $v_0 \geq 0$. Assume also that $\lim_{v\to\infty} \frac{h(v)}{(1+v)^{p/2-1}} = 0$, and that $I_{\frac{p}{2}}(y) = \int_0^\infty e^{-\frac{y}{1+v}}(1+v)^{-\frac{p}{2}}\, h(v)\, dv < \infty$. Then the Bayes estimator corresponding to the mixing density h is minimax. Furthermore if h is integrable the estimator is proper Bayes.

Proof. A simple calculation shows that

$$-(v+1)h'(v)/h(v) = \psi_1(v) + \psi_2(v) + \frac{p-2}{2}.$$

Setting $l_1(v) = \psi_1(v) + \frac{p-2}{2}$ and $l_2(v) = \psi_2(v)$, the result follows from Theorem 4.1 with $A = \frac{p-2}{2} + C$ and $B = D$. □

Note that finiteness of $I_{\frac{p}{2}}(y)$ in Corollary 4.2 is assured if we strengthen the limit condition to $\lim_{V\to\infty} \frac{h(v)}{(1+v)^{p/2-1-\varepsilon}} = 0$ for some $\varepsilon > 0$, since this implies that, for some $M > 0$, $\frac{h(v)}{(1+v)^{p/2}} \leq \frac{M}{(1+v)^{1+\varepsilon}}$ for any $v > 0$ and thus

$$\begin{aligned}
I_{\frac{p}{2}}(y) &= \textstyle\int_0^\infty e^{-\frac{y}{1+v}}(1+v)^{-\frac{p}{2}}\,h(v)\,dv \\
&\leq \textstyle\int_0^\infty (1+v)^{-\frac{p}{2}}\,h(v)\,dv \\
&\leq \textstyle\int_0^\infty \frac{M}{(1+v)^{1+\varepsilon}}\,dv \\
&< \infty.
\end{aligned}$$

An interesting and useful class of examples results from the choice

$$\psi(v) = \alpha + \beta/v + \gamma/v^2 \tag{4.11}$$

for some (α, β, γ). A simple calculation shows

$$\begin{aligned}
h(v) &= \exp\left[-\int_{v_0}^{v} \frac{\alpha + \beta/u + \alpha/u^2 + \frac{p-2}{2}}{1+u}\,du\right] \\
&\propto C(v+1)^{\beta-\alpha-\gamma-\frac{p-2}{2}}\,v^{\gamma-\beta}\exp(\frac{\gamma}{v}).
\end{aligned} \tag{4.12}$$

Example 4.1. Strawderman [1971] Suppose $\alpha \leq 0$ and $B = \gamma = 0$ so that $h(v) \propto C(v+1)^{-\alpha-(\frac{p-2}{2})}$. Let $\psi_1(v) = \psi(v) \equiv \alpha$ and $\psi_2(v) \equiv 0$ so that $C = D = 0$. Then the minimaxity conditions of Corollary 4.1 require $\lim_{v\to\infty} \frac{h(v)}{(v+1)^{p/2-1}} = \lim_{v\to\infty}(v+1)^{-\alpha-(p-2)} = 0$ and this is satisfied if $\alpha > 2-p$. Also

$$\begin{aligned}
I_{\frac{p}{2}}(y) &= \textstyle\int_0^\infty e^{-\frac{y}{1+v}}(1+v)^{-\frac{p}{2}}\,h(v)\,dv \\
&\propto C\textstyle\int_0^\infty e^{-\frac{y}{1+v}}(1+v)^{-\alpha-p+1}\,dv \\
&\leq \textstyle\int_0^\infty (1+v)^{-\alpha-p+1}\,dv \\
&< \infty \text{ if } \alpha > 2-p \text{ as above.}
\end{aligned}$$

Hence in this case the corresponding generalized Bayes estimator is minimax if $\alpha - p < \alpha \leq 0$ (which requires $p \geq 3$).

Furthermore it is proper Bayes minimax if $\int_0^\infty (1+v)^{-\alpha-(\frac{p-2}{2})}\,dv < \infty$ which is equivalent to $2-\frac{p}{2} < \alpha \le 0$. This latter condition requires $p \ge 5$ and demonstrates the existence of proper Bayes minimax estimator for $p \ge 5$. We will see below that this is the class of priors studied in Strawderman [1971] under the alternative parametrization $\lambda = \frac{1}{1+v}$.

Example 4.2. Consider $\psi(v)$ given by (4.7) with $\alpha \le 0$, $\beta \le 0$ and $\gamma \le 0$. Here we take $\psi_1(v) = \psi(v), \psi_2(v) = 0$ and $C = D = 0$. The minimaxity conditions of Corollary 4.2 require $\lim_{v\to\infty} \frac{h(v)}{(v+1)^{p/2-1}} = \lim_{v\to\infty} (v+1)^{\beta-\alpha-\gamma-p+2} v^{\gamma-\beta} e^{(\frac{\gamma}{v})} = 0$. This implies $2-p < \alpha \le 0$. The finiteness condition on

$$\begin{aligned} I_{\frac{p}{2}}(y) &= \int_0^\infty e^{-\frac{y}{1+v}} (1+v)^{-\frac{p}{2}}\, h(v)\, dv \\ &\propto \int_0^\infty e^{-\frac{y}{1+v}} (v+1)^{\beta-\alpha-\gamma-p+1} v^{\gamma-\beta} e^{\frac{\gamma}{v}}\, dv \end{aligned}$$

also requires $2-p < \alpha \le 0$. Therefore minimaxity is ensured as soon as $2-p < \alpha \le 0$.

Furthermore the minimax estimator will be proper Bayes if

$$\int_0^\infty h(v)\, dv \propto \int_0^\infty (1+v)^{\beta-\alpha-\gamma-\frac{p-2}{2}} v^{\gamma-\beta} e^{\frac{\gamma}{v}}\, dv < \infty.$$

This holds if $2-\frac{p}{2} < \alpha \le 0$ as in Example 4.1.

Example 4.3. Suppose $\alpha \le 0$, $\beta > 0$, and $\gamma < 0$ and take $\psi_1(v) = \alpha + (\gamma/v)(1/+\beta/\gamma) I_{[0,-2\gamma/\beta]}(v)$, $\psi_2(v) = (\gamma/v)(1/v+\beta/\gamma) I_{[-2\gamma/\beta,\infty]}(v)$, $C = \alpha$ and $D = -\beta^2/4\gamma$.

Note first that $\psi_1(v)$ is monotone non decreasing and bounded above by α, and $0 \le \psi_2(v) \le -\beta^2/4\gamma$. Therefore we require $C = \alpha < -2D = \beta^2/2\gamma$. The conditions

$\lim_{v\to\infty} \frac{h(v)}{(v+1)^{\frac{p}{2}-1}} = 0$, and $\int_0^\infty e^{-\frac{y}{1+v}}(1+v)^{-p/2}h(v)\,dv < \infty$ are as in Example 4.2, $2-p<\alpha\le 0$.

Thus $\delta_\pi(X)$ is minimax for $2-p<\alpha\le\beta^2/2\gamma<0$. The condition for integrability of h is also as in Example 4.2, i.e. $2-\frac{p}{2}<\alpha\le\beta^2/2\gamma<0$.

In this example $\psi(v)$ is not monotone but is increasing on $[0,-2\gamma/\beta)$ and decreasing thereafter. This typically corresponds to a non-monotone $r(\|X\|^2)$ in the Baranchik-type reprepresentation of $\delta_\pi(X)$.

For simplicity, in the following examples, we assume $\sigma^2=1$.

Example 4.4. (Student-*t* priors) In this example we take $\psi(v)$ as in Examples 4.2 and 4.3 with the specific choice $\alpha=(m-p+4)/2\le 0$, $\beta=(m(1-\varphi)+2)/2$, and $\gamma=-\frac{m\varphi}{2}\le 0$, where $m\ge 1$. In this case $h(v)=Cv^{-\frac{m+2}{2}}e^{-\frac{m\varphi}{2v}}$, an inverse gamma distribution, and hence, as is well known, $\pi(\theta)$ is a multivariate-*t* distribution with *m*-degrees of freedom and scale parameter φ if m is an integer (see e.g. Muirhead p.33 or Robert p.174). If $\sigma^2\ne 1$, the scale of the *t*-distribution is φ_σ.

According to different values of m and φ, either the conditions of Example 4.2 or the conditions of Example 4.3 apply. Both examples require $\alpha=m-p+4)/2\le 0$, or equivalently $1\le m\le p-4$ (so that $p\ge s$), and $\gamma=-m\varphi/2\le 0$.

Example 4.2 requires $\beta=\big(m(1-\varphi)+2\big)/2<0$, that is, $\varphi\ge(m+2)/m$. The condition for minimaxity $2-p<\alpha\le 0$ is satisfied since it is equivalent to $m>-p$. Furthermore the condition for proper Bayes minimaxity $2-\frac{p}{2}<\alpha\le 0$ is satisfied as well since it reduces to $m>0$. Hence, if the scale parameter φ is greater than or equal to $(m+2)/m$, the scaled p-variate t prior distribution leads to a proper Bayes minimax estimator for $p \geq 5$ and $n \leq p-4$.

On the other hand, when $\varphi < (m+2)/m$, that is, $\beta > 0$, conditions of Example 4.3 are applicable. Considering the proper Bayes case only, the condition for minimaxity of the Bayes estimator is

$$2 - \frac{p}{2} < \alpha = \frac{m-p+4}{2} \leq \frac{\beta^2}{2\gamma} = \frac{\big(m(1-\varphi)+2\big)^2}{-4m\varphi}.$$

The first inequality is satisfied by the fact that $m > 0$. The second inequality can be satisfied only for certain φ since, when φ goes to 0, the last expression tends to $-\infty$. A straightforward calculation shows that the second inequality can hold only if

$$\varphi \geq \frac{p-2}{m}\left[1 - \sqrt{1 - (\frac{m+2}{p-2})^2}\right] > 0.$$

In particular if $\varphi = 1$ (the standard multivariate t) the condition becomes $2 - p/2 < \frac{m-p+4}{2} \leq -\frac{1}{m}$. As $m \geq 1$ this is equivalent to $m + \frac{2}{m} \leq p-4$, which requires $p \geq 7$ for $m = 1$ or 2, and $p \geq m+5$ for $m \geq 3$.

An alternative approach to the results of this section can be made using the techniques of Subsection 3.4.2 applied to Baranchik-type estimators of the form $\left(1 - \frac{a\, r(\|X\|^2}{\|X\|^2}\right)X$. Indeed any spherically symmetric prior distribution will lead to an estimator of the form $\phi(\|X\|^2)X$. More to the point, for prior distributions of the form studied in this section, the r function is closely connected to the function $-(V+1)h'(V)/h(V)$. To see this note that

$$\begin{aligned}\delta_\pi(X) &= X + \sigma^2 \frac{\nabla m(X)}{m(X)} \\ &= \left(1 - \frac{I_{\frac{p}{2}+1}(y)}{I_{\frac{p}{2}}(y)}\right) X\end{aligned}$$

(from (4.2) with $y = \frac{\|X\|^2}{2\sigma^2}$)

$$= \left(1 - \frac{1}{y}\left(\frac{p-2}{2} - \frac{\int_0^\infty e^{-\frac{y}{1+v}}(1+v)^{-\frac{p}{2}}[(v+1)h'(v)/h(V)]\,dv - e^{-y}h(0)}{I_{\frac{p}{2}}}\right)\right) X$$

(from (4.4))

$$= \left(1 - \frac{2\sigma^2}{\|X\|^2}\left(\frac{p-2}{2} + E_{\frac{p}{2}}^{\frac{\|X\|^2}{2\sigma^2}}\left[-\frac{(V+1)h'(V)}{h(V)}\right] - \frac{e^{-\frac{\|X\|^2}{2\sigma^2}}h(0)}{I_{\frac{p}{2}}(\frac{\|X\|^2}{2\sigma^2})}\right)\right) X$$

(where $E_k^y(f)$ is as in the proof of Theorem 4.1).

Hence the Bayes estimator is of Baranchik form with

$$ar(\|X\|^2) = 2\left(\frac{p-2}{2} + E_{\frac{p}{2}}^{\frac{\|X\|^2}{2\sigma^2}}\left[-\frac{(V+1)h'(V)}{h(V)}\right] - \frac{e^{-\frac{\|X\|^2}{2\sigma^2}}h(0)}{I_{\frac{p}{2}}(\frac{\|X\|^2}{2\sigma^2})}\right). \quad \square$$

Recall, as the proof of Theorem 4.1, that the density $g_k^y(V)$ has monotone decreasing likelihood ratio in k but notice also that it has monotone increasing likelihood ratio (actually an exponential family) in y. Hence if $-\frac{(V+1)h'(V)}{h(V)}$ is non decreasing, it follows that r is non decreasing since $e^{-y}/I_{p/2}(y)$ is also non decreasing. Then the following corollary is immediately from Theorem 4.3.

Corollary 4.4. *Suppose the prior is of the form (4.4) where $-(v+1)\frac{h'(v)}{h(v)}$ is non decreasing and bounded above by $A > 0$. Then the generalized Bayes estimator is minimax provided $A \leq \frac{p-2}{2}$.*

Proof. As noted r is non decreasing and, by (2.21.a), is bounded above by $p-2+2A \leq 2(p-2)$. □

Corollary 4.3 yields an alternative proof for the minimaxity of the generalized Bayes estimator in Example 4.1.

Finally, as indicated earlier in this section, an alternative parametrization has often been used in minimaxity proofs for mixture of normal priors, namely $\lambda = \frac{1}{1+v}$, or equivalently $v = \frac{1-\lambda}{\lambda}$.

Perhaps the easiest way to proceed is to reconsider the prior distribution as a hierarchical prior as discussed in Section 1.7. Here the distribution of $\theta \mid v \sim N(0, v\sigma^2 X)$ and the unconditional density of v is the mixing density $h(v)$. The conditional distribution of θ given X and v is $N(\frac{v}{1+v}X, \frac{V}{1+v}\sigma^2 I)$. The Bayes estimator is

$$\begin{aligned}
\delta_\pi(X) &= E(\theta \mid X) \\
&= E[E(\theta \mid X, V) \mid X] \\
&= E[\tfrac{v}{1+v}X \mid X] \\
&= (1 - E[\tfrac{1}{1+v} \mid X])X \\
&= (1 - E[\lambda \mid X])X.
\end{aligned}$$

Note also that the Bayes estimator for the first stage prior

$$\theta \mid \lambda \sim N(0, \frac{1-\lambda}{\lambda}\sigma^2 I)$$

is $(1-\lambda)X$. Therefore in terms of the λ parametrization one may think of $E[\lambda \mid X]$ as the posterior mean of the shrinkage factor and of the (mixing) distribution on λ as the distribution of the shrinkage factor.

[Find a reference to Morris ??]

In particular for the prior distribution of Example 5.1, i.e. $h(V) = C(1+V)^{-\alpha - \frac{p-2}{2}}$, the corresponding mixture density on λ is given by $g(\lambda) = C\lambda^{\alpha + \frac{p-2}{2} - 2} = C\lambda^{\beta}$, and

($\beta = \alpha + \frac{p}{2} - 3$) the resulting prior is proper Bayes minimax if $2 - \frac{p}{2} < \alpha \leq 0$ or equivalently $-1 < \beta \leq \frac{p}{2} - 3$ (and $p \geq 5$). Note that, if $p \geq 6$, $\beta = 0$ satisfies the conditions and hence the mixing prior $g(\lambda) \equiv 1$ on $0 \leq \lambda \leq 1$, i.e. the uniform prior on the shrinkage factor λ, gives a proper Bayes minimax estimator.

To formalize the above discussion further we present a version of Theorem 4.3 in terms of the mixing distribution on λ. The proof follows from Theorem 4.3 and the change of variable $\lambda = \frac{1}{1+V}$.

Corollary 4.5. *Let θ have the hierarchical prior $\theta \mid \lambda \sim N_p(0, \frac{1-\lambda}{\lambda}\sigma^2 I_p)$ where $\lambda \sim g(\lambda)$ for $0 \leq \lambda \leq 1$. Assume that $\lim_{\lambda \to 0} g(\lambda)\lambda^{\frac{p}{2}+1} = 0$ and that $\int_0^1 e^{-\lambda}\lambda^{p/2} g(\lambda) d\lambda < \infty$. Suppose $\lambda g'(\lambda)/g(\lambda)$ can be decomposed as $l_1^*(\lambda) + l_2^*(\lambda)$ where $l_1^*(\lambda)$ is monotone nonincreasing and $l_1^*(\lambda) \leq A^*$ and where $0 \leq l_2^*(\lambda) \leq B^*$ with $A^* + 2B^* \leq \frac{p}{2} - 3$. Then the generalized Bayes estimator is minimax. Furthermore if $\int_0^1 g(\lambda) d\lambda < \infty$ the estimator is also proper Bayes.*

Example 4.5. (Beta priors) Suppose the prior $g(\lambda)$ on λ is a Beta (a,b) distribution, i.e. $g(\lambda) = K\lambda^{a-1}(1-\lambda)^{b-1}$. Note that the Strawderman [1971] prior is of this form if $b = 1$. An easy calculation shows $\frac{\lambda g'(\lambda)}{g(\lambda)} = a - 1 - (b-1)\frac{\lambda}{1-\lambda}$. Letting $l_1^*(\lambda) = \frac{\lambda g'(\lambda)}{g(\lambda)}$ and $l_2^*(\lambda) \equiv 0$, we see that the resulting proper Bayes estimator is minimax for $0 < a \leq \frac{p}{2} - 2$ and $b \geq 1$.

It is clear that our proof fails for $0 < b < 1$ since in this case $\lambda g'(\lambda)/g(\lambda)$ is not bounded from above (and is also monotone increasing). Maruyama [1998] shows using a different proof technique involving properties of confluent hypergeometric functions that the generalized Bayes estimator is minimax (in our notation) for

$-\frac{p}{2} < a \leq \frac{p}{2} - 2$ and $b \geq \frac{p+2a+2}{\frac{3p}{2}+a}$. This bound in b is in $(0,1)$ for $a < \frac{p}{2} - 2$ and hence certain Beta distributions with $0 < b < 1$ also give proper Bayes minimax estimators. The generalized Bayes minimax estimators of Alam [1973] are also in Maruyama's class.

4.1.3 Multiple shrinkage estimators

In this subsection we consider a class of estimators that adaptively choose a point (or subspace) toward which to shrink. George (1986 a, b) originated work in this area and the results in this section are largely due to him. The basic result upon which the results rely is that a mixture of superharmonic functions is superharmonic (see the discussion in the Appendix), that is, if $m_\alpha(x)$ is superharmonic for each α, then $\int m_\alpha(x)dG(\alpha)$ is superharmonic if $G(\cdot)$ is a positive measure such that $\int m_\alpha(x)dG(\alpha) < \infty$. Using this property we have immediately the following result from Corollary 4.1.

Theorem 4.4. *Let $m_\alpha(x)$ be a family of twice weakly differentiable non-negative superharmonic functions and $G(x)$ a positive measure such that $m(x) = \int m_\alpha(x)dG(x) < \infty$, forall $x \in R^p$. Then the (generalized, proper, or pseudo) Bayes estimator*

$$\delta(X) = X + \sigma^2 \frac{\nabla m(X)}{m(X)}$$

is minimax provided $E[\frac{\|\nabla m\|^2}{m^2}] < \infty$.

The following corollary for finite mixtures is useful.

Corollary 4.6. *Suppose $m_i(x)$ is superharmonic and $E[\frac{\|\nabla m_i(X)\|^2}{m_i^2(X)}] < \infty$ for $i = 1, \ldots, n$. Then if $m(x) = \sum_{i=1}^n m_i(x)$ the (generalized, proper, or pseudo) Bayes estimator*

$$\begin{aligned}\delta(X) &= X + \sigma^2 \tfrac{\nabla m(X)}{m(X)} \\ &= \textstyle\sum_{i=1}^n (X + \sigma^2 \tfrac{\nabla m_i(X)}{m_i(X)}) W_i(X)\end{aligned}$$

where $W_i(X) = \frac{m_i(X)}{\sum_{i=1}^n m_i(X)} (0 < W_i(X) < 1, \sum_{i=1}^n W_i(X) = 1)$ is minimax. (Note that $E_\theta[\frac{\|\nabla m(X)\|^2}{m^2(X)}] < \sum_{i=1}^n E[\frac{\|\nabla m_i(X)\|^2}{m^2(X_i)}] < \infty$.)

Example 4.6. (1) Multiple shrinkage James-Stein. Suppose we have several possible points $X_1, X_2, \ldots, X_n$ toward which to shrink. Recall that $m_i(x) = \left(\frac{1}{\|x-X_i\|^2}\right)^{\frac{p-2}{2}}$ is superharmonic if $p \geq 3$ and the corresponding pseudo-Bayes estimator is $\delta_i(X) = X_i + \left(1 - \frac{(p-2)\sigma^2}{\|X-X_i\|^2}\right)(X - X_i)$. Hence if $m(x) = \sum_{i=1}^n m_i(x)$ the resulting minimax pseudo Bayes estimator is given by

$$\delta(X) = \sum_{i=1}^n \left[X_i + (1 - \frac{(p-2)\sigma^2}{\|X - X_i\|^2})\right](X - X_i) W_i(X)$$

where $W_i(X) \propto \left(\frac{1}{\|X-X_i\|^2}\right)^{\frac{p-2}{2}}$ and $\sum_{i=1}^n W_i(X) = 1$. Note that $W_i(X)$ is large when X is close to X_i and the estimator is seen to adaptively shrink toward X_i.

(2) Multiple shrinkage positive-part James-Stein estimators. Another possible choice for the $m_i(x)$ (leading to a positive-part James Stein estimator) is

$$m_i(x) = \begin{cases} C\, e^{\frac{\|x-X_i\|^2}{2\sigma^2}} & \text{if } \|x - X_i\|^2 < (p-2)\sigma^2 \\ \left(\frac{1}{\|x-X_i\|^2}\right) & \text{if } \|x - X_i\|^2 \geq (p-2)\sigma^2 \end{cases}$$

where $C = \left(\frac{1}{(p-2)\sigma^2}\right)^{\frac{p-2}{2}} e^{\frac{p-2}{2}}$ so that $m_i(x)$ is continuous. This gives

$$\delta_i(X) = X_i + \left(1 - \frac{(p-2)\sigma^2}{\|X - X_i\|^2}\right)_+ (X - X_i)$$

since

$$\left(\frac{\nabla m_i(X)}{m_i(X)} = \begin{cases} -\frac{X-X_i}{\sigma^2} \text{ if } \|X-X_i\|^2 < (p-2)\sigma^2 \\ -\frac{(p-2)}{\|X-X_i\|^2} \end{cases} \right)$$

The adaptive combination is again minimax by the corollary and inherits the usual advantages of the positive-part estimator over the James-Stein estimator.

A smooth alternative to the above is $m_i(x) = \left(\frac{1}{b+\|x-X_i\|^2}\right)^{\frac{p-2}{2}}$ for some $b > 0$.

In each of the above examples we may replace $\frac{p-2}{2}$ in the exponent by $a/2$ where $0 \leq a \leq p-2$ (and where $0 \leq \|x-X_i\|^2 < (p-2)\sigma^2$ is replaced by $0 \leq \|x-X_i\|^2 < a\sigma^2$ for the positive-part estimator). The choice of $p-2$ as an upper bound for a ensures superharmonicity of $m_i(x)$. A choice of a in the range of $p-2 < a \leq 2(p-2)$ seems also quite natural since $\sqrt{m_i(x)}$ is superharmonic (but $m_i(x)$ is not) for a in this range so that each $\delta_i(X)$ is minimax. Unfortunately minimaxity of $\delta(X) = \sum_{i=1}^n W_i(X)\delta_i(X)$ does not follow from Corollary 4.3 for $p-2 < a \leq 2(p-2)$ since it need not be true that $\sqrt{\sum_{i=1}^n m_i(x)}$ is superharmonic even though $\sqrt{m_i(x)}$ is superharmonic for each i.

(3) A generalized Bayes multiple shrinkage estimator. If $\pi_i(\theta)$ is superharmonic and $\pi(\theta) = \sum_{i=1}^n \pi_i(\theta)$, then $\pi(\theta)$ is also superharmonic as is $m(x) = \sum_{i=1}^n m_i(x)$.

For example $\pi_i(\theta) = \left(\frac{1}{b+\|\theta-X_i\|^2}\right)^{a/2}$ for $b \geq 0$ and $0 \leq a \leq p-2$ is a suitable prior. Interestingly, according to a heuristic of Brown [1971] $m(x)$ in this case should behave for large $\|x\|^2$ as $\sum_{i=1}^n \left(\frac{1}{b+\|x-X_i\|^2}\right)^{a/2}$, the "smooth" version of the adaptive positive-part multiple shrinkage pseudo-marginal in part (2) of this example.

By obvious modifications of the above, multiple shrinkage estimators may be constructed which shrink adaptively toward subspaces. Further examples can be found in George [68, 67], Ki and Tsui [84] and Wither [147].

4.2 The Bayes estimate with unknown variance

In this section we extend the results of the previous section, using the ideas in Wells and Zhou [2008], to consider the point estimation for the mean of a multivariate normal when the variance is unknown. Specifically, we assume the following model

$$X \sim N_p(\theta, \sigma^2 I), \quad S \sim \sigma^2 \chi^2_m, \tag{4.13}$$

where S is independent of X. We are interested in constructing generalized Bayes minimax estimators of θ under the scaled squared loss function

$$L(\delta(X), \theta) = \frac{\|\delta(X) - \theta\|^2}{\sigma^2}. \tag{4.14}$$

In particular, we consider the following class of generalized hierarchical prior distributions

$$\theta \sim N(0, v\sigma^2), \quad \sigma^2 {\sim} \sigma^{-2B}, \quad v \sim h(v), \tag{4.15}$$

where, B is a positive constant, and $h(v)$ is a continuously differentiable probability density function on $[0, \infty)$. We develop sufficient conditions on m, p, and $h(v)$ such that the generalized Bayes estimators with respect to the class of priors (4.15) are minimax under the invariant loss function in (4.14). Wells and Zhou [2008] were able to obtain such sufficient conditions by applying the bounds and monotonicity results of Baranchik [1970] and Efron [1976].

In order to derive the (formal) Bayes estimator we reparameterize the model in (4.13) by replacing σ by η^{-1}, the model then becomes

$$\begin{aligned} &X{\sim}N_p(\theta,\eta^{-2}I_p), \quad S \sim S^{\frac{m}{2}-1}\eta^m e^{\frac{-1}{2}S\eta^2}, \\ &\theta{\sim}N_p(0,\nu\eta^{-2}I_p), \quad \nu{\sim}h(\nu), \quad \eta \sim \eta^{-2K}, \end{aligned} \tag{4.16}$$

where K is a positive constant. Under this model, the prior for θ is a scaled mixture of normal distributions.

Before we derive the formula for the generalized Bayes estimator under the model (4.15), we need to impose three regularity conditions on the parameters of priors. These conditions are easily satisfied by many hierarchial priors. These three conditions are assumed throughout this article.

C1: $A = -K + \frac{p}{2} + \frac{m}{2} + \frac{3}{2} > 1$;
C2: $\int_0^1 \lambda^{\frac{p}{2}-2} h(\frac{1-\lambda}{\lambda})\, d\lambda < \infty$; and
C3: $\lim_{\nu\to\infty} \frac{h(\nu)}{(1+\nu)^{\frac{p}{2}-1}} = 0$.

Lemma 4.2. *Under the model in (4.16), the generalized Bayes estimator can be written as*

$$\delta(X,S) = X - R(F)X = X - \frac{r(F)}{F}X, \tag{4.17}$$

where $F = ||X||^2/S$,

$$R(F) = \frac{\int_0^1 \lambda^{\frac{p}{2}-1}(1+\lambda F)^{-A} h(\frac{1-\lambda}{\lambda})\, d\lambda}{\int_0^1 \lambda^{\frac{p}{2}-2}(1+\lambda F)^{-A} h(\frac{1-\lambda}{\lambda})\, d\lambda}, \tag{4.18}$$

and

$$r(F) = FR(F). \tag{4.19}$$

Proof. Under the loss function (4.14), the generalized Bayes estimator for the model (4.16) is

$$\begin{aligned}
\delta(X,S) &= \frac{E(\frac{\theta}{\sigma^2}|X,S)}{E(\frac{1}{\sigma^2}|X,S)} \\
&= \frac{\int_0^\infty h(\nu) \int_0^\infty [(\eta^2)^{A-\frac{1}{2}} e^{-\frac{1}{2}\eta^2 S} \int_{R^p} (\frac{1}{2\pi\nu\eta^{-2}})^{\frac{p}{2}} \theta e^{-\frac{1}{2}\eta^2(\frac{||\theta||^2}{\nu}+||X-\theta||^2)} d\theta] d\eta\, d\nu}{\int_0^\infty h(\nu) \int_0^\infty [(\eta^2)^{A-\frac{1}{2}} e^{-\frac{1}{2}\eta^2 S} \int_{R^p} (\frac{1}{2\pi\nu\eta^{-2}})^{\frac{p}{2}} e^{-\frac{1}{2}\eta^2(\frac{||\theta||^2}{\nu}+||X-\theta||^2)} d\theta] d\eta d\nu} \\
&= \left(1 - \frac{\int_0^\infty [(\frac{1}{1+\nu}) h(\nu) (\frac{1}{1+\nu})^{\frac{p}{2}} \int_0^\infty (\eta^2)^{A-\frac{1}{2}} e^{-\frac{1}{2}\eta^2(S+\frac{l}{1+\nu})} d\eta] d\nu}{\int_0^\infty [h(\nu)(\frac{1}{1+\nu})^{\frac{p}{2}} \int_0^\infty (\eta^2)^{A-\frac{1}{2}} e^{-\frac{1}{2}\eta^2(S+\frac{l}{1+\nu})} d\eta] d\nu}\right) X \\
&= \left(1 - \frac{\int_0^\infty (\frac{1}{1+\nu}) h(\nu) (\frac{1}{1+\nu})^{\frac{p}{2}} (1+\frac{F}{1+\nu})^{-A} d\nu}{\int_0^\infty h(\nu) (\frac{1}{1+\nu})^{\frac{p}{2}} (1+\frac{F}{1+\nu})^{-A} d\nu}\right) X,
\end{aligned}$$

where $l = ||X||^2$. Letting $\lambda = (1+\nu)^{-1}$, then $\delta(X,S) = (1-R(F))X$, which gives the form of the generalized Bayes estimator. □

Recall from Stein [1981] that when σ^2 is known the Bayes estimator under squared error loss and corresponding to a prior $\pi(\theta)$ is given by

$$\delta^\pi(X) = X + \sigma^2 \frac{\bigtriangledown m(X)}{m(X)}. \tag{4.20}$$

The form of the Bayes estimator given in (4.17) gives an analogous form with the unknown variance replaced by the usual unbiased estimator. Furthermore, define

$$\mathbf{M}(x,s) = \int\int f_X(x) \frac{f_S(s)}{s^{\frac{m}{2}-1}} \pi(\theta,\sigma^2) d\theta d\sigma^2,$$

where

$$f_X(x) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{p}{2}} e^{-\frac{1}{2\sigma^2}||x-\theta||^2},$$

and

$$f_S(s) = \frac{1}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})} s^{\frac{m}{2}-1}(\sigma^2)^{-\frac{m}{2}} e^{-\frac{s}{2\sigma^2}}.$$

It can be shown that

$$\mathbf{M}(x,s) \propto \int_0^\infty h(\nu) \int_0^\infty [(\eta^2)^{A-\frac{3}{2}} e^{-\frac{1}{2}\eta^2 s} \int_{\mathbb{R}^p} (\frac{1}{2\pi\nu\eta^{-2}})^{\frac{p}{2}} e^{-\frac{1}{2}\eta^2(\frac{||\theta||^2}{\nu}+||x-\theta||^2)} d\theta] d\eta d\nu.$$

It is interesting to note the unknown variance analog of (4.20) is

$$\delta(X,S) = X - \frac{1}{2}\frac{\nabla_X \mathbf{M}(X,S)}{\nabla_S \mathbf{M}(X,S)}.$$

4.2.1 Preliminary results

The minimax property of generalized Bayes estimator is closely related to the behavior of $r(F)$ and $R(F)$ function, which is in turn closely related to the behavior of

$$g(\nu) = -(\nu+1)\frac{h'(\nu)}{h(\nu)}. \tag{4.21}$$

Fourdrinier, Strawderman and Wells [1998] also gave a detailed analysis of the type of function in (4.21) however their argument was deduced from the superharmonicity of the square root of a marginal condition. Baranchik[1970] and Efron [1976] gave certain regularity conditions on the shrinkage function $r(\cdot)$ such that an estimator

$$\widehat{\theta}(X,S) = X - \frac{r(F)}{F}X \tag{4.22}$$

is minimax under the loss function (4.14) for the model (4.13). Both results require an upper bound on $r(F)$ and a condition on how fast $R(F) = \frac{r(F)}{F}$ decreases with

F. In this section, we derive a bound for the function $r(F)$ in (4.19). The following results are due to Baranchik [1970] and Efron [1976], respectively.

Theorem 4.5. *Assume that $r(F)$ is increasing in F, and $0 \le r(F) \le 2\frac{p-2}{m+2}$, then any point estimator of the form (4.22) is minimax.*

Theorem 4.6. *Define $c_m = \frac{p-2}{m+2}$. Assume that $0 \le r(F) \le 2c_m$, that for all F with $r(F) < 2c_m$,*

$$\frac{F^{\frac{p}{2}-1}r(F)}{(2-\frac{r(F)}{c_m})^{1+2c_m}} \text{ is increasing in } F, \tag{4.23}$$

and that if an F_0 exists such that $r(F_0) = 2c_m$, then $r(F) = 2c_m$ for all $F \ge F_0$. With the above assumptions, the estimator $\widehat{\theta}(X,S) = X - \frac{r(F)}{F}X$ in (4.22) is minimax.

Consequently, to apply these results one has to establish an upper bound for $r(F)$ in (4.19) and the monotonicity property for some variants of $r(F)$, and the candidate we use is $\widetilde{r}(F) = F^c r(F)$ with a constant c. This bound and the monotonicity property is used in proving the minimaxity of the generalized Bayes estimator.

We need certain bounds in order to apply Theorems 4.5 and 4.6. Before we derive the bound for $r(F)$ in (4.19), first note that if $h(\nu)$ is a continuously differentiable function on $[0,\infty)$, and regularity conditions C1, C2 and C3 hold, then that the integration by parts used in Lemmas 4.3 and 4.4 are valid.

Lemma 4.3. *With the regularity conditions C1, C2 and C3, and assume that $g(\nu) \le M$, where M is a positive constant and $g(\nu)$ is defined as in (4.21), then for the $r(F)$ function in (4.19), we have*

$$0 \le r(F) \le \frac{\frac{p}{2}-1+M}{A-\frac{p}{2}-M}.$$

Proof. By the definition in (4.18), $R(F) \geq 0$, then $r(F) = FR(F) \geq 0$. Note that

$$r(F) = F\frac{\int_0^1 \lambda^{\frac{p}{2}-1}(1+\lambda F)^{-A}h(\frac{1-\lambda}{\lambda})\,d\lambda}{\int_0^1 \lambda^{\frac{p}{2}-2}(1+\lambda F)^{-A}h(\frac{1-\lambda}{\lambda})\,d\lambda} = F\frac{I_{\frac{p}{2}-1,A,h}(F)}{I_{\frac{p}{2}-2,A,h}(F)},$$

where we are using the notation

$$I_{\alpha,A,h}(F) = \int_0^1 \lambda^{\alpha}(1+\lambda F)^{-A}h(\frac{1-\lambda}{\lambda})\,d\lambda.$$

Using integration by parts, we obtain

$$\begin{aligned} FI_{\frac{p}{2}-1,A,h}(F) &= \int_0^1 \lambda^{p/2-1}h\left(\frac{1-\lambda}{\lambda}\right)d\left[\frac{(1+\lambda F)^{1-A}}{1-A}\right] \\ &= \lambda^{\frac{p}{2}-1}h\left(\frac{1-\lambda}{\lambda}\right)\frac{(1+\lambda F)^{1-A}}{1-A}|_0^1 + \frac{1}{A-1}\int_0^1 (1+\lambda F)^{-A}(1+\lambda F) \\ &\left[\left(\frac{p}{2}-1\right)\lambda^{\frac{p}{2}-2}h\left(\frac{1-\lambda}{\lambda}\right) - \frac{1}{\lambda^2}\lambda^{\frac{p}{2}-1}h'\left(\frac{1-\lambda}{\lambda}\right)\right]d\lambda. \end{aligned}$$

By C1 and C3, we know that the first term of the RHS is nonpositive. The second term of the RHS can be written as $N_1 + N_2 + N_3 + N_4$ where

$$N_1 = \frac{1}{A-1}\int_0^1 (1+\lambda F)^{-A}\left(\frac{p}{2}-1\right)\lambda^{\frac{p}{2}-2}h\left(\frac{1-\lambda}{\lambda}\right)d\lambda = \frac{\frac{p}{2}-1}{A-1}I_{\frac{p}{2}-2,A,h}(F),$$

$$\begin{aligned} N_2 &= \frac{1}{A-1}\int_0^1 (1+\lambda F)^{-A}\lambda^{\frac{p}{2}-2}h'\left(\frac{1-\lambda}{\lambda}\right)\left(\frac{-\lambda}{\lambda^2}\right)d\lambda \\ &= \frac{I_{\frac{p}{2}-2,A,h}(F)}{A-1}\frac{\int_0^1 \lambda^{\frac{p}{2}-2}(1+\lambda F)^{-A}g(\frac{1-\lambda}{\lambda})h(\frac{1-\lambda}{\lambda})\,d\lambda}{\int_0^1 \lambda^{\frac{p}{2}-2}(1+\lambda F)^{-A}h(\frac{1-\lambda}{\lambda})\,d\lambda} \\ &\leq \frac{M}{A-1}I_{\frac{p}{2}-2,A,h}(F), \end{aligned}$$

$$N_3 = \frac{\frac{p}{2}-1}{A-1}FI_{\frac{p}{2}-1,A,h}(F) = \frac{(\frac{p}{2}-1)r(F)}{A-1}I_{\frac{p}{2}-2,A,h}(F),$$

and

$$\begin{aligned}N_4 &= \frac{I_{\frac{p}{2}-2,A,h}(F)}{A-1}\frac{F\int_0^1 \lambda^{\frac{p}{2}-1}(1+\lambda F)^{-A}h'(\frac{1-\lambda}{\lambda})(\frac{-1}{\lambda})d\lambda}{I_{\frac{p}{2}-2,A,h}(F)}\\ &= \frac{I_{\frac{p}{2}-2,A,h}(F)}{A-1}\frac{F\int_0^1 (1+\lambda F)^{-A}\lambda^{\frac{p}{2}-1}g(\frac{1-\lambda}{\lambda})h(\frac{1-\lambda}{\lambda})d\lambda}{I_{\frac{p}{2}-2,A,h}(F)}\\ &\le \frac{Mr(F)}{A-1}I_{\frac{p}{2}-2,A,h}(F).\end{aligned}$$

Combining all the terms, we get the following inequality

$$(A-1)r(F) \le \left(\frac{p}{2}-1\right)+M+\left(\frac{p}{2}-1\right)r(F)+Mr(F) \Rightarrow r(F) \le \frac{\frac{p}{2}-1+M}{A-\frac{p}{2}-M}.$$

Therefore we have the needed bound on the $r(F)$ function. □

We will now show that under certain regularity conditions on $g(\nu)$, we have the monotonicity property for $\tilde{r}(F) = F^c r(F)$ with a constant c. This monotonicity property enables us to the minimaxity of the generalized Bayes estimator.

Lemma 4.4. *If $g(\nu) = -(\nu+1)\frac{h'(\nu)}{h(\nu)} = l_1(\nu)+l_2(\nu)$ such that $l_1(\nu)$ is increasing in ν and $0 \le l_2(\nu) \le c$, then $\widetilde{r}(F) = F^c r(F)$ is nondecreasing.*

Proof. By taking derivative, we only need to show

$$0 \le (1+c)R(F)+FR'(F), \tag{4.24}$$

which is equivalent to

$$0 \le (1+c)\frac{I_{\frac{p}{2}-1,A,h}(F)}{I_{\frac{p}{2}-2,A,h}(F)}+F\frac{I'_{\frac{p}{2}-1,A,h}(F)I_{\frac{p}{2}-2,A,h}(F)-I'_{\frac{p}{2}-2,A,h}(F)I_{\frac{p}{2}-1,A,h}(F)}{I^2_{\frac{p}{2}-2,A,h}(F)},$$

which is in turn equivalent to

$$\begin{aligned}&-FI'_{\frac{p}{2}-1,A,h}(F)I_{\frac{p}{2}-2,A,h}(F)\\ &\le (1+c)I_{\frac{p}{2}-2,A,h}(F)I_{\frac{p}{2}-1,A,h}(F)FI'_{\frac{p}{2}-2,A,h}(F)I_{\frac{p}{2}-1,A,h}(F).\end{aligned} \tag{4.25}$$

Now note that

$$-FI'_{a,A,h}(F) = \int_0^1 \lambda^a (1+\lambda F)^{-A} h\left(\frac{1-\lambda}{\lambda}\right) \frac{A\lambda F}{1+\lambda F} d\lambda.$$

Letting

$$J_a\left(\frac{Au}{1+u} h\left(\frac{F-u}{u}\right)\right) = \int_0^F u^a (1+u)^{-A} \frac{Au}{1+u} h\left(\frac{F-u}{u}\right) du,$$

and

$$J_a\left(h\left(\frac{F-u}{u}\right)\right) = \int_0^F u^a (1+u)^{-A} h\left(\frac{F-u}{u}\right) du.$$

Also note that

$$J_a\left(\frac{Au}{1+u} h\left(\frac{F-u}{u}\right)\right) = F^{a+1} \int_0^1 \lambda^a (1+\lambda F)^{-A} h\left(\frac{1-\lambda}{\lambda}\right) \frac{A\lambda F}{1+\lambda F} d\lambda,$$

and

$$J_a\left(h\left(\frac{F-u}{u}\right)\right) = F^{a+1} I_{a,A,h}(F).$$

Then it follows that (4.25) is equivalent to

$$\frac{J_{\frac{p}{2}-1}(\frac{Au}{1+u} h(\frac{F-u}{u}))}{J_{\frac{p}{2}-1}(h(\frac{F-u}{u}))} \leq (1+c) + \frac{J_{\frac{p}{2}-2}(\frac{Au}{1+u} h(\frac{F-u}{u}))}{J_{\frac{p}{2}-2}(h(\frac{F-u}{u}))}. \tag{4.26}$$

By an integration by parts, we have

$$\begin{aligned} J_a\left(\frac{Au}{1+u} h\left(\frac{F-u}{u}\right)\right) &= \int_0^F u^a (1+u)^{-A} h\left(\frac{F-u}{u}\right) \frac{Au}{1+u} du \\ &= -u^{a+1} h\left(\frac{F-u}{u}\right) (1+u)^{-A} |_0^F \\ &+ \int_0^F (1+u)^{-A} \left[(a+1) u^a h\left(\frac{F-u}{u}\right) + u^{a+1} h'\left(\frac{F-u}{u}\right)\left(\frac{-F}{u^2}\right)\right] du. \end{aligned}$$

Hence (4.26) is equivalent to

$$
\begin{aligned}
&\frac{-F^{\frac{p}{2}}h(0)(1+F)^{-A}}{J_{\frac{p}{2}-1}(h(\frac{F-u}{u}))}+\Big(\frac{p}{2}\Big)\\
&+\frac{\int_0^F u^{\frac{p}{2}-1}(1+u)^{-A}h(\frac{F-u}{u})\left[\frac{h'(\frac{F-u}{u})}{h(\frac{F-u}{u})}(\frac{-F}{u})\right]du}{\int_0^F u^{\frac{p}{2}-1}(1+u)^{-A}h(\frac{F-u}{u})\,du}\\
&\le 1+c+\frac{-F^{\frac{p}{2}-1}h(0)(1+F)^{-A}+0}{J_{\frac{p}{2}-2}(h(\frac{F-u}{u}))}+\Big(\frac{p}{2}-1\Big)\\
&+\frac{\int_0^F u^{\frac{p}{2}-2}(1+u)^{-A}h(\frac{F-u}{u})\left[\frac{h'(\frac{F-u}{u})}{h(\frac{F-u}{u})}(\frac{-F}{u})\right]du}{\int_0^F u^{\frac{p}{2}-2}(1+u)^{-A}h(\frac{F-u}{u})\,du},
\end{aligned}
\tag{4.27}
$$

which in turn is equivalent to

$$
\begin{aligned}
&\frac{-h(0)(1+F)^{-A}}{I_{\frac{p}{2}-1,A,h}(F)}+\frac{J_{\frac{p}{2}-1}(h(\frac{F-u}{u})l_1(\frac{F-u}{u}))}{J_{\frac{p}{2}-1}(h(\frac{F-u}{u}))}+\frac{J_{\frac{p}{2}-1}(h(\frac{F-u}{u})l_2(\frac{F-u}{u}))}{J_{\frac{p}{2}-1}(h(\frac{F-u}{u}))}\\
&\le c+\frac{-h(0)(1+F)^{-A}}{I_{\frac{p}{2}-2,A,h}(F)}+\frac{J_{\frac{p}{2}-2}(h(\frac{F-u}{u})l_1(\frac{F-u}{u}))}{J_{\frac{p}{2}-2}(h(\frac{F-u}{u}))}+\frac{J_{\frac{p}{2}-2}(h(\frac{F-u}{u})l_2(\frac{F-u}{u}))}{J_{\frac{p}{2}-2}(h(\frac{F-u}{u}))}
\end{aligned}
\tag{4.28}
$$

Therefore it is clear that $I_{\frac{p}{2}-1,A,h}(F)\le I_{\frac{p}{2}-2,A,h}(F)$, so that we then have

$$
\frac{-h(0)(1+F)^{-A}}{I_{\frac{p}{2}-1,A,h}(F)}\le\frac{-h(0)(1+F)^{-A}}{I_{\frac{p}{2}-2,A,h}(F)}.
$$

Note also that $l_1(v)$ is increasing in v implies that for all F fixed, $l_1(\frac{F-u}{u})$ is decreasing in u. When $t<u$, we have

$$
\frac{(1+u)^{-A}u^{\frac{p}{2}-2}h(\frac{F-u}{u})\,\mathbb{1}\{u\le F\}}{(1+t)^{-A}t^{\frac{p}{2}-2}h(\frac{F-t}{t})\,\mathbb{1}\{t\le F\}}\le\frac{(1+u)^{-A}u^{\frac{p}{2}-1}h(\frac{F-u}{u})\,\mathbb{1}\{u\le F\}}{(1+t)^{-A}t^{\frac{p}{2}-1}h(\frac{F-t}{t})\,\mathbb{1}\{t\le F\}}.
$$

By a monotone likelihood argument, we have

$$
\begin{aligned}
\frac{J_{\frac{p}{2}-1}(h(\frac{F-u}{u})l_1(\frac{F-u}{u}))}{J_{\frac{p}{2}-1}(h(\frac{F-u}{u}))}&=\frac{\int_0^F u^{\frac{p}{2}-1}(1+u)^{-A}h(\frac{F-u}{u})l_1(\frac{F-u}{u})}{\int_0^F u^{\frac{p}{2}-1}(1+u)^{-A}h(\frac{F-u}{u})\,du}\\
\le\frac{\int_0^F u^{\frac{p}{2}-2}(1+u)^{-A}h(\frac{F-u}{u})l_1(\frac{F-u}{u})\,du}{\int_0^F u^{\frac{p}{2}-2}(1+u)^{-A}h(\frac{F-u}{u})\,du}&=\frac{J_{\frac{p}{2}-2}(h(\frac{F-u}{u})l_1(\frac{F-u}{u}))}{J_{\frac{p}{2}-2}(h(\frac{F-u}{u}))}.
\end{aligned}
$$

Finally, we note that

$$0 \leq \frac{J_{\frac{p}{2}-2}(l_2(\frac{F-u}{u})h(\frac{F-u}{u}))}{J_{\frac{p}{2}-2}(h(\frac{F-u}{u}))} \leq c$$

and

$$0 \leq \frac{J_{\frac{p}{2}-1}(l_2(\frac{F-u}{u})h(\frac{F-u}{u}))}{J_{\frac{p}{2}-1}(h(\frac{F-u}{u}))} \leq c.$$

Therefore we established the inequality (4.28) and the proof is completed. □

4.2.2 Minimaxity of the generalized Bayes estimators

In this subsection, we state the result from Wells and Zhou [2008] that uses Lemmas 4.2, 4.3, 4.4 and the results of Baranchik [1970] and [1976] to show minimaxity of the generalized Bayes estimator (4.17).

Theorem 4.7. *Assume that $g(\nu)$ is increasing in ν, $g(\nu) \leq M$, where M is a positive constant, and $\frac{p-2+2M}{m+3-2K-2M} \leq 2\frac{p-2}{m+2}$, then $\delta(X,S)$ in (4.17) is minimax.*

Proof. Let $l_2(\nu) = 0$ and $l_1(\nu) = g(\nu)$. By applying Lemma 4.4 to the case $c = 0$, we have $r(F)$ increasing in F. Applying the bound in Lemma 4.3, we can get $0 \leq r(F) \leq 2\frac{p-2}{m+2}$. Therefore, by Lemma 4.2, $\delta(X,S)$ is minimax. □

It is interesting to make connections to the result in Faith [1978]. Faith [1978] considered generalized Bayes estimator for $N_p(\theta, I_p)$ and showed that when $g(\nu)$ is increasing in ν, and $M \leq \frac{p-2}{2}$, the generalized Bayes estimator $\delta(X)$ would be minimax. By taking $m \to \infty$, we deduce the same conditions as Faith [1978]. The next lemma is a variant of Alam [1973] for known variance case.

Lemma 4.5. *Define* $c_m = \frac{p-2}{m+2}$, *if there exists* $b \in (0,1]$, *and* $c = \frac{b(p-2)}{4+4(2-b)c_m}$, *such that* $0 \le r(F) \le (2-b)c_m$, *and* $F^c r(F)$ *is increasing in* F, *then the generalized Bayes estimator* $\delta(X,S)$ *in (4.17) is minimax.*

Proof. By taking derivative, (4.23) can be satisfied by requiring

$$0 \le 2\left(\frac{p}{2}-1\right)R(F)\left(2-\frac{r(F)}{c_m}\right)+4r'(F)(1+r(F)). \tag{4.29}$$

Since $r(F) \le (2-b)c_m$, then at the point where $r'(F) \ge 0$, (4.29) is satisfied. At the point where $r'(F) < 0$, since $r(F) \le (2-b)c_m$, then

$$4r'(F)(1+\beta) \le 4r'(F)(1+r(F)), \tag{4.30}$$

where $\beta = (2-b)c_m$. We now have

$$\begin{aligned} 0 &\le (4+4\beta)(cR(F)+R(F)+FR'(F)) \\ &= 2b\left(\frac{p}{2}-1\right)R(F)+4r'(F)(1+\beta) \\ &\le 2\left(\frac{p}{2}-1\right)R(F)\left(2-\frac{r(F)}{c_m}\right)+4r'(F)(1+r(F)) \end{aligned}$$

since $F^c r(F)$ is increasing in F. Thus for all values of F, we proved (4.29), and combining with the bound on $r(F)$ function, we proved the minimaxity of the generalized Bayes estimator. □

It is interesting to observe that by requiring a tighter upper bound on $r(F)$, we can relax the monotonicity requirement on the $r(F)$. The tighter the upper bound, the more flexible $r(F)$ can be. This result enriches the class of priors whose generalized Bayes estimators are minimax. Direct application of Lemmas 4.2, 4.3, 4.4, 4.5 will get us the following theorem.

Theorem 4.8. *If there exists $b \in (0,1]$ such that $g(\nu) = l_1(\nu) + l_2(\nu) \leq M$, and $l_1(\nu)$ is increasing in ν, $0 \leq l_2(\nu) \leq c = \frac{b(p-2)}{4+4(2-b)\frac{p-2}{m+2}}$, and $\frac{p-2+2M}{m+3-2K-2M} \leq \frac{(2-b)(p-2)}{m+2}$. Then the generalized Bayes estimator $\delta(X,S)$ in (4.17) is minimax.*

4.2.3 Examples

In this subsection we will give several examples on which our results can be applied and make some connection to the existing literature found in Maruyama and Strawderman [2005] and Fourdrinier [1998].

Example 4.7. (Maruyama and Strawderman (2005) priors.) Consider the priors with $h(\nu) \propto \nu^b(1+\nu)^{-a-b-2}$. Maruyama and Strawderman [2005] showed that if $b \geq 0$, $\frac{m}{2} + e > a > -\frac{p}{2} - 1$ and $0 \leq \frac{\frac{p}{2}+a+1}{\frac{m}{2}+e-a} \leq 2c_m$, where $c_m = \frac{p-2}{m+2}$, then the resulting generalized Bayes estimator is minimax. Note that their parameter e and our parameter K are related by the equality $-K = e + 1/2$.

Condition C1 is equivalent to the condition that $\frac{m}{2} + e > -\frac{p}{2} - 1$. C2 and C3 are equivalent here, and both are equivalent to the condition that $a + \frac{p}{2} + 1 > 0$ in this case. Then using Theorem 4.7, we have $g(\nu) = a + 2 - b\nu^{-1}$. The condition that $g(\nu)$ is increasing in ν is equivalent to the condition that $b \geq 0$. Clearly, we can let $M = a + 2$. To satisfy the condition $0 \leq r(F) \leq 2c_m$, the conditions $\frac{m}{2} + e > a$ and $0 \leq \frac{\frac{p}{2}+a+1}{\frac{m}{2}+e-a} \leq 2c_m$ are sufficient since $r(F) \leq \frac{p-2+2a+4}{m+3+2e+1-2a-4} = \frac{\frac{p}{2}+a+1}{\frac{m}{2}+e-a}$.

A close examination of the Maruyama and Strawderman [2005] proof shows that their upper bound on $r(F)$ is sharp, and this implies that our bound in Lemma 4.3 can not be relaxed.

Example 4.8. (Generalized Student priors.) Let

$$h(v) = c(v+1)^{\beta-\alpha-\gamma-\frac{p-2}{2}} v^{\gamma-\beta} e^{\frac{\gamma}{v}}.$$

We shall consider two cases here. The first case where $\alpha \le 0$, $\beta \le 0$ and $\gamma < 0$ involves the construction of a monotonic $r(\cdot)$ function. The second case where $\alpha \le 0$, $\beta > 0$ and $\gamma < 0$ does not require the $r(\cdot)$ function to be monotonic. In all cases,

$$\ln h(v) = (\beta - \alpha - \gamma - \frac{p-2}{2}) \ln(1+v) + (\gamma - \beta) \ln v + \frac{\gamma}{v}$$

and

$$g(v) = \left(\frac{p-2}{2} + \alpha + \gamma - \beta\right) + \frac{(1+v)(\beta-\gamma)}{v} + \frac{\gamma(1+v)}{v^2} = \frac{p-2}{2} + \alpha + \frac{\beta}{v} + \frac{\gamma}{v^2}.$$

Clearly $g(v)$ is monotonic in the first case, and we can get minimaxity of the generalized Bayes estimator by having

$$0 \le \frac{p-2+\alpha}{\frac{m}{2} + \frac{1}{2} - K - \frac{p}{2} - \alpha} \le \frac{p-2}{\frac{m}{2}+1}$$

in addition to conditions C1, C2, and C3. In the limiting case where $m \to \infty$, C1 holds trivially. Both C2 and C3 can be satisfied by $\alpha > 2-p$. The upper bound on $R(F)$ can be satisfied by any $\alpha \le 0$. Consequently, our conditions reduce to those of Fourdrinier [1998] for the case of known variance.

Now we consider spherical multivariate Student-t priors with n degree of freedom and a scale parameter τ, with $\alpha = \frac{n-p+4}{2}$, $\beta = \frac{n(1-\tau)+2}{2}$ and $\gamma = -\frac{n\tau}{2}$. The case of $\tau = 1$ is of particular interest and could not be derived by an monotonic $r(\cdot)$ function argument. But we can use the result in Theorem 4.8 to show that the generalized Bayes estimator is minimax under the following conditions: $n \leq p-4$, there exists a constant $b \in (0,1]$ such that

$$\begin{aligned} \frac{p+n+\frac{1}{n}}{m+1-2K-n-\frac{1}{n}} &\leq (2-b)\frac{p-2}{m+2}, \\ \frac{1}{2n} &\leq c = \frac{b(p-2)}{4+4(2-b)\frac{p-2}{m+2}}. \end{aligned} \tag{4.31}$$

Condition (4.31) can be established by observing that in this case,

$$g(\nu) = \frac{p-2}{2} + \alpha + \frac{\beta}{\nu} + \frac{\gamma}{\nu^2} = \frac{n}{2} + 1 + \frac{1}{\nu} - \frac{n}{2\nu^2}$$

and clearly is nonmonotonic. We then let $M = \frac{n}{2} + 1 + \frac{1}{2n}$ and apply Lemma 4.3 to get the upper bound on $r(\cdot)$. We define

$$l_1(\nu) = g(\nu) - \frac{1}{2n} \text{ when } \nu \leq n, \text{ and } l_1(\nu) = \frac{n}{2} + 1 \text{ otherwise.}$$

We also define

$$l_2(\nu) = \frac{1}{2n} \text{ when } \nu \leq n, \text{ and } l_2(\nu) = \frac{1}{\nu} - \frac{n}{2\nu^2} \text{ otherwise.}$$

By applying Lemma 4.4, we get condition (4.31).

We give an explicit computation by considering the case $n = 1$, which corresponds to spherical multivariate Cauchy prior. We assume that $m = O(p)$, and $-2K = 3$, then condition (4.31) reduces to $p \geq 5$, $\frac{p+2}{m+2} \leq (2-b)\frac{p-2}{m+2}$, and $\frac{1}{2} \leq$

$\frac{b(p-2)}{4+8-4b}$. We set $b = 5/9$, then $p = 11$ is sufficient for the generalized Bayes estimator to be minimax.

4.3 Results for known Σ and general quadratic loss

4.3.1 Results for the diagonal case

Much of this section is based on the discussion in Strawderman [2003]. We begin with a discussion of the multivariate normal case where Σ is diagonal. Let

$$X \sim N_p(\theta, \Sigma) \tag{4.32}$$

where Σ is diagonal, $\Sigma = diag(\sigma^2_{1,...,}\sigma^2_p)$ and loss equal to a weighted sum of squared errors loss

$$L(\theta, \delta) = (\delta - \theta)' D(\delta - \theta) = \sum_{i=1}^{p} (\delta_i - \theta_i)^2 d_i \tag{4.33}$$

The results of the previous section extend by the use of Stein's Lemma in a straightforward way to give the following basic theorem.

Theorem 4.9. *Let X have the distribution (4.32) and let the loss be given by (4.33).*

(1) *If $\delta(X) = X + \Sigma g(X)$, where $g(X)$ is weakly differentiable and $E||g||^2 < \infty$, then the risk of δ is*

$$R(\delta, \theta) = E_\theta((\delta - \theta)' Q(\delta - \theta)) = tr\Sigma D + E_\theta \left[\sum_{i=1}^{p} \sigma_i^4 d_i \left(g_i^2(X) + 2\frac{\partial g_i(X)}{\partial X_i} \right) \right].$$

(2) *If $\theta \sim \pi(\theta)$, then the Bayes estimator of θ is $\delta_\Pi(X) = X + \Sigma \frac{\nabla m(X)}{m(X)}$, where $m(X)$ is the marginal distribution of X.*

(3) *If $\theta \sim \pi(\theta)$, then the risk of a proper (generalized, pseudo-) Bayes estimator of the form $\delta_m(X) = X + \Sigma \frac{\nabla m(X)}{m(X)}$ is given by*

$$R(\delta_m, \theta) = tr\Sigma D + E_\theta \left[\frac{2m(X) \sum_{i=1}^{p} \sigma_i^4 d_i \partial m^2(X)/\partial^2 X_i}{m^2(X)} - \frac{\sum_{i=1}^{p} \sigma_i^4 d_i \left(\partial m^2(X)/\partial^2 X_i\right)^2}{m^2(X)} \right]$$

$$= tr\Sigma D + 4E_\theta \left[\frac{\sum_{i=1}^{p} \sigma_i^4 d_i \partial^2 \sqrt{m(X)}/\partial^2 X_i}{\sqrt{m(X)}} \right].$$

(4) *If $\frac{\sum_{i=1}^{p} \sigma_i^4 d_i \partial^2 \sqrt{m(X)}/\partial^2 X_i}{\sqrt{m(X)}}$ is non-positive, the proper (generalized, pseudo-Bayes $\delta_m(X)$ is minimax).*

The proof follows closely that of corresponding results in Section 4.3. The result is basically from Stein [1981].

A key observation that allows us to construct Bayes minimax procedures for this situation, based on the procedures for the case $\Sigma = D = I$, is the following.

Lemma 4.6. *Suppose $\eta(X)$ is such that $\nabla^2 \eta(X) = \sum_{i=1}^{p} \partial^2 \eta(X)/\partial^2 X_i^2 \leq 0$ (i.e. $\eta(X)$ is superharmonic) then $\eta^*(X) = \eta(\Sigma^{-1} D^{-1/2} X)$ is such that $\sum_{i=1}^{p} \sigma_i^4 d_i \partial^2 \eta^*(X)/\partial^2 X_i \leq$* 0.

The proof is a straightforward calculation. Details can be found in Strawderman [2003].

Note, also, that for any scalar, a, if $\eta(X)$ is superharmonic, then so is $\eta(aX)$. This all leads to the following main result.

Theorem 4.10. *Suppose X has distribution (4.32) and loss is given by (4.33).*

(1) *Suppose $\sqrt{m(X)}$ is superharmonic ($m(X)$ is a proper, generalized, or pseudo-marginal for the case $\Sigma = D = I$). Then*

$$\delta_m(X) = X + \Sigma \left(\frac{\nabla m(\Sigma^{-1} D^{-1/2} X)}{m(\Sigma^{-1} D^{-1/2} X)} \right)$$

is a minimax estimator.

(2) *If $\sqrt{m(\|X\|^2)}$ is spherically symmetric and superharmonic, then*

$$\delta_m(X) = X + \frac{2m'(X'\Sigma^{-1} D^{-1} \Sigma^{-1} X) D^{-1} \Sigma^{-1} X}{m(X'\Sigma^{-1} D^{-1} \Sigma^{-1} X)}$$

is minimax.

(3) *Suppose the prior distribution $\pi(\theta)$ has the hierarchical structure $\theta|\lambda \sim MVN_p(0, A_\lambda)$ for $\lambda \sim h(\lambda)$, $0 < \lambda < 1$, where $A_\lambda = (c/\lambda)\Sigma D \Sigma - \Sigma$ where c is such that A_1 is positive definite and $h(\lambda)$ satisfies the conditions Theorem 4.10. Then*

$$\delta_\pi(X) = X + \Sigma \frac{\nabla m(X)}{m(X)}$$

is minimax.

(4) *Suppose $m_i(X), i = 1, 2 \ldots.. k$ are superharmonic, then the multiple shrinkage estimator*

$$\delta_m(X) = X + \Sigma \left[\frac{\sum_{i=1}^{k} \nabla m_i(\Sigma^{-1} D^{-1/2} X)}{\sum_{i=1}^{k} m_i(\Sigma^{-1} D^{-1/2} X)} \right]$$

is a minimax multiple shrinkage estimator.

Proof. (1) This follows directly from Theorem 4.9 parts (3) and (4) and Lemma 4.6.

(2) This follows from part (1) and Theorem 4.9 part (2) with a straightforward calculation.

(3) First note that $\theta|\lambda \sim MVN_p(0, A_\lambda)$ and $X - \Theta|\lambda \sim MVN_p(0, \Sigma)$. Thus, $X - \theta$ and θ are therefore conditionally independent given λ. Hence $X|\lambda \sim MVN_p(0, A_\lambda + \Sigma)$. It follows that

$$m(X)\alpha \int_0^1 \lambda^{p/2} \exp\left[-\frac{\lambda}{c}\left(X'\Sigma^{-1}D^{-1}\Sigma^{-1}X\right)\right] h(\lambda) d\lambda$$

but $m(X) = \eta\left(X'\Sigma^{-1}D^{-1}\Sigma^{-1}X/c\right)$, where $\sqrt{\eta\left(X'X\right)}$ is superharmonic by Theorem 4.9. Hence by part (2) $\delta_\pi(X)$ is minimax (and proper or generalized Bayes depending on whether $h(\lambda)$ is integrable or not).

(4) Since superharmonicity of $\eta(X)$ implies that of $\sqrt{\eta\left(X\right)}$, part (4) follows from part (1) and superharmonicity of mixtures of superharmonic functions. □

Example 4.9. (*Pseudo-Bayes minimax estimators.*) When $\Sigma = D = \sigma^2 I$, we saw in Section 4.3 that by choosing $m(X) = \frac{1}{\|X\|^{2b}}$, the pseudo-Bayes estimator was the James-Stein estimator $\delta_m(X) = (1 - \frac{2b\sigma^2}{\|X\|^2})X$. It now follows from this and part (2) of Theorem 4.10 that $m(X'\Sigma^{-1}D^{-1}\Sigma^{-1}X) = (1/X'\Sigma^{-1}D^{-1}\Sigma^{-1}X)^b$ has associated with it the pseudo-Bayes estimator $\delta_m(X) = (1 - \frac{2\mathbf{b}\mathbf{D}^{-1}\Sigma^{-1}}{(\mathbf{X}'\Sigma^{-1}\mathbf{D}^{-1}\Sigma^{-1}\mathbf{X})})X$. This estimator is minimax for $0 < b \leq 2(p-2)$.

Example 4.10. (*Hierarchical proper Bayes minimax estimator.*) As suggested by Berger [1976] suppose the prior distribution has the hierarchical structure $\theta|\lambda \sim MVN_p(0, A_\lambda)$ where $A_\lambda = c\Sigma D\Sigma - \Sigma$, $c > (1/\min(\sigma_i^2 d_i)$ and $h(\lambda) = (1+b)\lambda^b$, $0 < \lambda < 1$ for $-1 < b \leq \frac{(p-6)}{2}$. The resulting proper Bayes estimator will be minimax for $p \geq 5$ by Theorem 4.10 part (3) and Example 4.9. For $p \geq 3$ the estimator

$\delta_\pi(X)$ given in part (3) of Theorem 4.10 is a generalized Bayes minimax provided $-\frac{(p+2)}{2} < b \leq \frac{(p-6)}{2}$.

It can be shown to be admissible if the lower bound is replaced by -2, by the results of Brown [1971] (See also Berger and Strawderman [1996] and Kubokawa and Strawderman [2005]).

Example 4.11. (*Multiple shrinkage minimax estimator.*) It follows from Example 4.9 and Theorem 4.10 that $m(X) = \sum_{i=1}^{k} \left[\frac{1}{(X-\nu_i)'\Sigma^{-1}D^{-1}\Sigma^{-1}(X-\nu_i)}\right]^b$ satisfies the conditions of Theorem 4.10 (4) for $0 < b \leq (p-2)/2$ and hence that

$$\delta_m(X) = X - \frac{2b \sum_{i=1}^{k} \left[D^{-1}\Sigma^{-1}(X-\nu_i)\right] \Big/ \left[(X-\nu_i)'\Sigma^{-1}D^{-1}\Sigma^{-1}(X-\nu_i)\right]^{b+1}}{\sum_{i=1}^{k} 1 \Big/ \left[(X-\nu_i)'\Sigma^{-1}D^{-1}\Sigma^{-1}(X-\nu_i)\right]^{b}} \tag{4.34}$$

is a minimax multiple shrinkage (pseudo-Bayes) estimator.

If, as in Example 4.11 we used the generalized prior

$$\pi(\theta) = \sum_{i=1}^{k} \left[\frac{1}{(\theta-\nu_i)'\Sigma^{-1}D^{-1}\Sigma^{-1}(\theta-\nu_i)}\right]^b,$$

the resulting generalized Bayes (as opposed to pseudo-Bayes) estimators would be minimax for $0 < b \leq (p-2)/2$.

4.3.2 General Σ and general quadratic loss

In this section we generalize the above results to the case of

$$X \sim MVN_p(\theta, \Sigma), \tag{4.35}$$

where Σ is a general positive definite covariance matrix, and

$$L(\theta,\delta) = (\delta-\theta)'Q(\delta-\theta), \tag{4.36}$$

where Q is a general positive definite matrix. We will see that this case can be reduced to the canonical form $\Sigma = I$ and $Q = diag(d_1, d_2 \ldots.. d_p) = D$. We continue to follow the development in Strawderman [2003].

The following well known fact will be used repeatedly to obtain the desired generalization.

Lemma 4.7. *For any pair of positive definite matrices, Σ and Q, there exits a non-singular matrix A such that $A\Sigma A' = I$ and $(A')^{-1}QA^{-1} = D$ where D is diagonal.*

Using this fact we can now present the canonical form of the estimation problem.

Theorem 4.11. *Let $X \sim MVN_p(\theta,\Sigma)$ and suppose that the loss is $L_1(\delta,\theta) = (\delta-\theta)'Q(\delta-\theta)$. Let A and D be as in Lemma 4.7, and let $Y = AX \sim MVN_p(v,I)$, where $v = A\theta$, and $L_2(\delta,v) = (\delta-v)'D(\delta-v)$.*

(1) *If $\delta_1(X)$ is an estimator with risk function $R_1(\delta_1,\theta) = E_\theta L_1(\delta_1(X),\theta)$, then the estimator $\delta_2(Y) = A\delta_1(A^{-1}Y)$ has risk function $R_2(\delta_2,v) = R_1(\delta_1,\theta) = E_\theta L_2(\delta_2(Y),v)$.*

(2) *$\delta_1(X)$ is proper or generalized Bayes with respect to the proper prior distribution $\pi_1(\theta)$ (or pseudo-Bayes with respect to the pseudo-marginal $m_1(X)$) under loss L_1 if and only if $\delta_2(Y) = A\delta_1(A^{-1}Y)$ is proper or generalized Bayes with respect to $\pi_2(v) = \pi_1(A^{-1}v)$ (or pseudo-Bayes with respect to the pseudo-marginal $m_2(Y) = m_1(A^{-1}Y)$).*

(3) *$\delta_1(X)$ is admissible (or minimax or dominates $\delta_1^*(X)$) under L_1 if and only if $\delta_2(Y) = A\delta_1(A^{-1}Y)$ is admissible (or minimax or dominates $\delta_2^*(Y) = A\delta_1^*(A^{-1}Y)$ under L_2).*

Proof. (1) The risk function

$$\begin{aligned} R_2(\delta_2, v) &= E_\theta L_2[\delta_2(Y), v] = E_\theta[(\delta_2(Y) - v)' D(\delta_2(Y) - v)] \\ &= E_\theta[(A\delta_1(A^{-1}(AX)) - A\theta)' D(A\delta_1(A^{-1}(AX)) - A\theta)] \\ &= E_\theta[(\delta_1((X) - \theta)' A' D A(\delta_1(X) - \theta)] \\ &= E_\theta[(\delta_1((X) - \theta)' Q(\delta_1(X) - \theta)] = R_1(\delta_1, \theta). \end{aligned}$$

(2) The Bayes estimator for any quadratic loss is the posterior mean. Hence, since $\theta \sim \pi_1(\theta)$ and $v = A\theta \sim \pi_2(v) = \pi_1(A^{-1}v)$ (ignoring constants) then

$$\delta_2(Y) \ = E[v|Y] = E[A\theta|Y] = E[A\theta|AX] = A\, E[\theta|X] = A\ \delta_1(X) = A\delta_1(A^{-1}Y).$$

(3) This follows directly from part (1).

Note: If $\Sigma^{1/2}$ is the positive definite square root of Σ and $A = P\Sigma^{-1/2}$ where P is orthogonal and diagonalizes $\Sigma^{1/2}Q\Sigma^{1/2}$, then this A and $D = P\Sigma^{1/2}Q\Sigma^{1/2}P'$ satisfy the requirements of the theorem.

Example 4.12. Proceeding as we did in Example **??** and applying Theorem 4.4.5, $m(X'\Sigma^{-1}Q^{-1}\Sigma^{-1}X) = (X'\Sigma^{-1}Q^{-1}\Sigma^{-1}X)^{-b}$ has associated with it, for $0 < b \leq 2(p-2)$, the pseudo-Bayes minimax James-Stein estimators is $\delta_m(X) = (1 - \frac{2bQ^{-1}\Sigma^{-1}}{(X'\Sigma^{-1}Q^{-1}\Sigma^{-1}X)})X$.

Generalizations of Example **??** to hierarchical Bayes minimax estimators and Example 4.11 to multiple shrinkage estimators are straightforward. We omit the details.

4.4 Estimation of a predictive density

Consider a parametric model $(\mathscr{Y},(\mathscr{P'}_{\mu})_{\mu\in\vartheta})$ where $\mathscr{Y}$ is the sample space, ϑ is the parameter space and $P'=\{p(y|\mu):\mu\in\vartheta\}$ is a class of densities of P'_{μ} with respect to a measure σ -finite a measure. In addition, suppose that a sample x, comes from a random variable X described by a similar model, index by the same parameter space, is given by $(\mathscr{X},(\mathscr{P}_{\mu})_{\mu\in\vartheta})$. The problem is the prediction (estimate) of the true density of a random variable Y independent of X, either the density $p(.|\mu)\in P$ Would be from the observation x.

Let the estimate of the true density $p(y|\mu)$, based on the observed data x, be a density (estimator) $\hat{q}(y|x)$ (belonging to some class of models $\mathscr{C}\supset p'$). The objective of this section is to assess the goodness of $\hat{q}(y|x)$ as an estimate of the true density. Aitchison [1975] proposed using the divergence of the Kullback Leibler [1951], defined in (4.37) below, as a loss function for estimating $p(y|\mu)$.

The class of estimates $\mathscr{C}$ can be identical to the class P', for example, for any $y\in\mathscr{Y}$

$$\hat{q}(y|x)=p(y|\mu=\hat{\mu}(x))$$

where $\hat{\mu}$ is a possible estimate of μ. this density estimator is called the plug-in estimate associated with the estimator $\hat{\mu}$. Alternatively, if one takes as an estimator of the unknown density the density, defined for any $y \in \mathscr{Y}$ by

$$\hat{q}(y|x) = \int_{\vartheta} p(y|\mu)\, p(\mu|x) d\mu$$

where $p(\mu|x)$ may be a weight function, or a *a posteriori* density associated with *a priori* density $\pi(\mu)$. In this case the class $\mathscr{C}$ will be broader than the class of the model P'. Aitchison [1975] showed that the predictive method method is preferable to plug-in, for several families of probability distributions, by comparing their risks induced the Kullback Leibler divergence.

4.4.1 The Kullback Leibler divergence

First recall the definition of the Kullback Leibler divergence and some of its properties.

Lemma 4.8. *The relative entropy (or the Kullback Leibler divergence)* $D_{KL}(p,q)$ *between two densities p and q is defined by*

$$D_{\ KL}(p,q) = E_p \left[\log \frac{p}{q}\ , \right] = \int \log \left[\frac{p(x)}{q(x)} \right] p(x)\, dx \geq 0 \tag{4.37}$$

and equality is achieved if and only if $p = q$ *p -almost surely.*

Note that the divergence is finite if and only if the support of the density p is contained in the support of the density q and by convention we define $0 \log \frac{0}{0} = 0$.

Proof. By definition of the the Kullback Leibler divergence we can write

$$\begin{aligned}
-D_{\mathrm{KL}}(p,q) &= \int \log\left[\frac{q(x)}{p(x)}\right] p(x)\,dx \\
&\leq \log\left[\int \frac{q(x)}{p(x)} p(x)\,dx\right] \text{ (by Jensen's inequality)} \\
&= \log\left[\int q(x)\,dx\right] \\
&= 0.
\end{aligned}$$

We have equal if and only if equality occurs in Jensen's inequality, which occurs if and only if $p = Q\ p$ -almost surely. Note that, the lemma is true if it takes q as sub-density (mass less than or equal to 1). □

The Kullback Leibler divergence is not a true distance, it is not symmetric and it does not satisfy the triangular inequality. But it appears as the natural distance in the information theory. One of these important properties, given in the lemma next, is that it is strictly convex.

Lemma 4.9. *The Kullback Leibler divergence is strictly convex, that is to say if* (p_1, p_2) *and* (q_1, q_2) *are two pairs of densities, then for any* $0 \leq \lambda \leq 1$,

$$D_{KL}(\lambda p_1 + (1-\lambda)p_2, \lambda q_1 + (1-\lambda)q_2) \leq \lambda D_{KL}(p, q_1) + (1-\lambda)D_{KL}(p_2, q_2). \tag{4.38}$$

Proof. For $f(t) = t\log(t)$. it is clear that f is strictly convex on $(0,\infty)$. Let

$$\alpha_1 = \frac{\lambda q_1}{\lambda q_1 + (1-\lambda)q_2}, \alpha_2 = \frac{(1-\lambda)q_2}{\lambda q_1 + (1-\lambda)q_2}, t_1 = \frac{p_1}{q_1} \text{ and } t_2 = \frac{p_2}{q_2}.$$

From the convexity of the function f it follows that

$$f(\alpha_1 t_1 + \alpha_2 t_2) \leq \alpha_1 f(t_1) + \alpha_2 f(t_2)$$

and consequently

$$(\alpha_1 t_1 + \alpha_2 t_2)\log(\alpha_1 t_1 + \alpha_2 t_2) \leq t_1 \alpha_1 \log(t_1) + t_2 \alpha_2 \log(t_2).$$

Substituting the values of α_1, $\alpha 2 t_1$ and t_2 we obtain

$$(\lambda p_1 + (1-\lambda)p_2)\log\frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2} \leq \lambda p_1 \log\frac{p_1}{q_1} + (1-\lambda)p_2 \log\frac{p_2}{q_2}.$$

Finally, by integrating the latter term, (4.38) and the strict Convexity follows from the strict convexity of the function f. □

4.4.2 The Bayesian predictive density

For any estimator of the density, define the Kullback Leibler loss by

$$\mathrm{KL}(\mu, \hat{p}(.|x)) = \int p(y|\mu)\log\left[\frac{p(y|\mu)}{\hat{p}(y|x)}\right] dy \tag{4.39}$$

and its corresponding risk as

$$\mathscr{R}_{\mathrm{KL}}(\mu, \hat{p}) = \int p(x|\mu)\left[\int p(y|\mu)\log\left[\frac{p(y|\mu)}{\hat{p}(y|x)}\right] dy\right] dx. \tag{4.40}$$

We say that the density estimate $\hat{p}_2$ is dominated the density estimate $\hat{p}_1$ if we, for any $\mu \in \vartheta,\ \mathscr{R}_{\mathrm{KL}}(\mu, \hat{p}_1) - \mathscr{R}_{\mathrm{KL}}(\mu, \hat{p}_2) \geq 0$, where the inequality is strict for at least value of μ.

When the inferential framework is Bayes we will compare estimates using Bayes risk. We will consider the class, more general than Aitchison, of all sub-densities.

$$\mathscr{D} = \left\{ q(.|X)| \text{ for all } x, \int q(y|x)\, dy \leq 1 \right\}.$$

Lemma 4.10. *(Aitchison, [1975]) The Bayes risk $r_\pi(\hat{p}) = \int \mathscr{R}_{KL}(\mu, \hat{p})\, \pi(\mu)\, d\mu$ is minimized by*

$$\hat{p}_\pi(y|x) = \int p(y|\mu)\, p(\mu|x)\, d\mu = \frac{\int p(y|\mu)\, p(x|\mu) \pi(\mu)\, d\mu}{\int p(x|\mu)\, \pi(\mu)\, d\mu}. \tag{4.41}$$

We call $\hat{p}_\pi$ the Bayesian predictive density.

Proof. The difference between the Bayes risks of $\hat{p}_\pi$ and another competing density $\hat{q}$ equals

$$\begin{aligned} r_\pi(\hat{q}) - r_\pi(\hat{p}_\pi) &= \int_\Theta \left[\int_{\mathscr{X}} \left[\int_{\mathscr{Y}} p(y|\mu) \log \frac{\hat{p}_\pi(y|x)}{\hat{q}(y|x)}\, dy \right] p(x|\mu)\, dx \right] \pi(\mu)\, d\mu \\ &= \int_\Theta \left[\int_{\mathscr{X}} \left[\int_{\mathscr{Y}} p(y|\mu) \log \frac{\hat{p}_\pi(y|x)}{\hat{q}(y|x)}\, dy \right] p(x|\mu) \pi(\mu)\, dx \right] d\mu \\ &= \int_\Theta \left[\int_{\mathscr{X}} \left[\int_{\mathscr{Y}} p(y|\mu) \log \frac{\hat{p}_\pi(y|x)}{\hat{q}(y|x)}\, dy \right] p(\mu|x) m(x)\, dx \right] d\mu. \end{aligned}$$

By applying Fubini's Theorem and reversing the order of integration gives

$$\begin{aligned} r_\pi(\hat{q}) - r_\pi(\hat{r}) &= \int_{\mathscr{X}} \left[\int_{\mathscr{Y}} \left[\int_\Theta p(\mu|x)\, p(y|\mu)\, d\mu \right] \log \frac{\hat{p}_\pi(y|x)}{\hat{q}(y|x)}\, dy \right] m(x)\, dx \\ &= \int_{\mathscr{X}} \left[\int_{\mathscr{Y}} \hat{p}_\pi(y|x) \log \frac{\hat{p}_\pi(y|x)}{\hat{q}(y|x)}\, dy \right] m(x)\, dx \\ &= \int_{\mathscr{X}} D_{\mathrm{KL}}(\hat{p}_\pi(.|x), \hat{q}(.|x))\, m(x)\, dx \geq 0. \end{aligned}$$

□

4.4.3 The sufficiency reduction

Let $X_{(n)} = (X_1, \ldots, X_n)$ and $Y_{(m)} = (Y_1, \ldots, Y_m)$ be samples from i.i.d p-dimensional normal distributions $\mathscr{N}_p(\mu, \Sigma_1)$ and $\mathscr{N}_p(\mu, \Sigma_2)$ with unknown common mean μ and known positive definite covariance matrices Σ_1 and Σ_2. On the basis of a observation $x_{(n)} = (x_1, \dots, x_n)$ from $X_{(n)}$, consider the problem of estimating the true predictive density of $y_{(m)} = (y_1, \dots, y_m)$ from $Y_{(m)}$, $p(y_{(m)}|\mu)$, under the Kullback Leibler loss.

For a prior density $\pi(\mu)$, the Bayesian predictive density is given by

$$\hat{p}_\pi(y_{(m)}|x_{(n)}) = \frac{\int p(y_{(m)}|\mu) p(x_{(n)}|\mu)\, \pi(\mu)\, d\mu}{\int p(x_{(n)}|\mu)\, \pi(\mu)\, d\mu}. \tag{4.42}$$

For simplicity sake, we consider the case where $\Sigma_1 = \Sigma_2 = I_p$.

Komaki [2001] showed that for Bayesian predictive densities satisfy

$$\int p(y_{(m)}|\mu) \log \frac{p(y_{(m)}|\mu)}{\hat{p}_\pi(y_{(m)}|x_{(n)})} dy_{(m)} = \int p(\bar{y}_m|\mu) \log \frac{p(\bar{y}_m|\mu)}{\hat{p}_\pi(\bar{y}_m|\bar{x}_m)} d\bar{y}_m \tag{4.43}$$

where

$$\bar{x}_m = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y}_m = \frac{1}{m}\sum_{j=1}^{m} y_j, \quad v_x = \frac{1}{n}, \quad v_y = \frac{1}{m}.$$

Using the fact that

$$\sum_{i=1}^{m} \left[\|y_i - \mu\|^2\right] = \sum_{i=1}^{m} \left[\|y_i - \bar{y}_m\|^2\right] + m\,(\|\bar{y}_m - \mu\|)^2,$$

we can express $p(y_{(m)}|\mu)$ as

$$\begin{aligned} p(y_{(m)}|\mu) &= (2\pi)^{-mp/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{m} \|y_i - \bar{y}_m\|^2\right) \exp\left(-\frac{1}{2} m\,(\|\bar{y}_m - \mu\|)^2\right) \\ &= \frac{m^{p/2}}{(2\pi)^{(m-1)p/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{m} \|y_i - \bar{y}_m\|^2\right) p(\bar{y}_m|\mu). \end{aligned} \tag{4.44}$$

Similarly, it follows that

$$p(x_{(n)}|\mu) = (2\pi)^{-np/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\|x_i-\bar{x}_m\|^2\right) \exp\left(-\frac{n}{2}(\|\bar{x}_m-\mu\|)^2\right)$$
$$= n^{p/2}\,(2\pi)^{-(n-1)p/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\|x_i-\bar{x}_m\|^2\right) p(\bar{x}_m|\mu).$$

By replacing these expressions in the form of the predictive density in (4.42), we get

$$\hat{p}_\pi(y_{(m)}|x_{(n)}) = \left\{\frac{m^{p/2}}{(2\pi)^{(m-1)p/2}}\exp(-\frac{1}{2}\sum_{i=1}^{m}\|y_i-\bar{y}_m\|^2)\right\}\frac{\int p(\bar{y}_m|\mu)\,p(\bar{x}_m|\mu)\pi(\mu)\,d\mu}{\int p(\bar{x}_m|\mu)\,\pi(\mu)\,d\mu}$$
$$= \left\{\frac{m^{p/2}}{(2\pi)^{(m-1)p/2}}\exp(-\frac{1}{2}\sum_{i=1}^{m}\|y_i-\bar{y}_m\|^2)\right\}\hat{p}_\pi(\bar{y}_m|\bar{x}_m) \qquad (4.45)$$

Finally, for (4.44) and (4.45) it follows that

$$\int p(y_{(m)}|\mu)\log\frac{p(y_{(m)}|\mu)}{\hat{p}(y_{(m)}|x_{(n)})}dy_{(m)} = \int p(y_{(m)}|\mu)\log\frac{p(\bar{y}_m|\mu)}{\hat{p}(\bar{y}_m|\bar{x}_m)}dy_{(m)}$$
$$= \int p(\bar{y}_m|\mu)\log\frac{p(\bar{y}_m|\mu)}{\hat{p}(\bar{y}_m|\bar{x}_m)}d\bar{y}_m$$

Therefore, for any prior π, the risk of the Bayesian density estimator is equal to the risk of the Bayesian density associated to π in the reduced model $X \sim \mathcal{N}(\mu, \frac{1}{n}I)$ and $Y \sim \mathcal{N}(\mu, \frac{1}{m}I)$. Thus, for the Bayesian predictive densities, it is sufficient to consider the reduced model.

Now we will compare two plug-in density estimators, $\hat{p}_1$ and $\hat{p}_2$ associated with the two estimator of μ δ_1 and δ_2. That is, for $i = 1.2$, define

$$\hat{p}_i(y_{(m)}|x_{(n)}) = p(y_{(m)}|\mu = \delta_i(x_{(n)})) \qquad (4.46)$$

The difference in risk between $\hat{p}_1$ and $\hat{p}_2$ is given by

$$\begin{aligned}
\Delta\mathscr{R}_{\mathrm{KL}}(\hat{p}_2,\hat{p}_1) &= \mathscr{R}_{\mathrm{KL}}(\mu,\hat{p}_2)-\mathscr{R}_{\mathrm{KL}}(\mu,\hat{p}_1)\\
&= \int p(x_{(n)}|\mu)\int p(y_{(m)}|\mu)\log\frac{\hat{p}_1(y_{(m)}|x_{(n)})}{\hat{p}_2(y_{(m)}|x_{(n)})}dy_{(m)}dx_{(n)}\\
&= \int p(x_{(n)}|\mu)\int p(y_{(m)}|\mu)\bigg(\frac{1}{2}\sum_{i=1}^{m}\|\delta_2(x_{(n)})-y_i\|^2\\
&\quad -\frac{1}{2}\sum_{i=1}^{m}\|\delta_1(x_{(n)})-y_i\|^2\bigg)dy_{(m)}dx_{(n)}.
\end{aligned}$$

Now, in this case, we write

$$\begin{aligned}
\Delta\mathscr{R}_{\mathrm{KL}}(\hat{p}_2,\hat{p}_1) &= \frac{1}{2}\sum_{i=1}^{m}E_{X_{(n)},Y_{(m)}}\Big(\|\delta_2(X_{(n)})-\mu+\mu-Y_i\|^2-\|\delta_1(X_{(n)})-\mu+\mu-Y_i\|^2\Big)\\
&= \frac{m}{2}E_{X_{(n)},Y_{(m)}}\left[\|\delta_2(X_{(n)})-\mu\|^2-\|\delta_1(X_{(n)})-\mu\|^2\right]\\
&\quad+\sum_{i=1}^{m}E_{X_{(n)},Y_{(m)}}\Big(\left[(\delta_2(X_{(n)})-\mu).(\mu-Y_i)\right]-\left[(\delta_1(X_{(n)})-\mu).(\mu-Y_i)\right]\Big)\\
&= \frac{m}{2}\left(E_{X_{(n)}}\left[\|\delta_2(X_{(n)})-\mu\|^2\right]-E_{X_{(n)}}\left[\|\delta_1(X_{(n)})-\mu\|^2\right]\right).\\
&= \frac{m}{2}\left[\mathscr{R}_Q(\delta_2,\mu)-\mathscr{R}_Q(\delta_1,\mu)\right].
\end{aligned}$$

By completeness of statistics $\bar{X}_n$ and $\bar{Y}_m$, the estimate of μ under quadratic loss can be reduced by sufficiency.

4.4.4 The best invariant density $\hat{p}_U$

In this subsection we will use the fact that we are assuming a location parameter model. We will see $X\sim p(x|\mu)=p(x-\mu)$ and $Y\sim p'(y|\mu)=p'(y-\mu)$, where p and p' are two known densities. A density $\hat{q}$ is called invariant (equivariant) wit respect to a location parameter if, for any $a\in\mathbb{R}^p$, $x\in\mathbb{R}^p$, and $y\in\mathbb{R}^p$ $q(y|x+a)=q(y-a|x)$. This is equivalent to $q(y+a|x+a)=q(y|x)$.

Lemma 4.11. *The invariant predictive densities with respect to the group of translations have of constant risk.*

Proof. By the property of invariance, the risk of a invariant density $\hat{q}$ is equal to

$$\begin{aligned}\mathscr{R}(\mu,\hat{q}) &= \int \log \frac{p'(y-\mu)}{\hat{q}(y|x)}\, p(x-\mu)\, p'(y-\mu)\, dy dx \\ &= \int \log \frac{p'(y-\mu)}{\hat{q}(y-\mu|x-\mu)}\, p(x-\mu)\, p'(y-\mu)\, dy dx \\ &= \int \log \frac{p(z')}{\hat{q}(z'|z)}\, p(z)\, p'(z')\, dz'\, dz, \qquad (4.47)\end{aligned}$$

by the change of variables $z = x-\mu$ and $z' = z-\mu$. Therefore, the risk $\mathscr{R}(\mu,\hat{q})$ does not depend on μ and it is constant. Any predictive density which minimizes the risk and is invariant by report to the group of translations is known as the best invariant density. □

4.4.5 A Bayesian property

Lemma 4.12. *The best predictive invariant density is the Bayesian predictive density $\hat{p}_U$ associated with Lesbegue measure on $\mathbb{R}^p$, $\pi(\mu) = 1$, is*

$$\hat{p}_U(y|x) = \frac{\int_{\mathbb{R}^p} p'(y|\mu)\, p(x|\mu)\, d\mu}{\int_{\mathbb{R}^p} p(x|\mu)\, d\mu} \qquad (4.48)$$

Proof. Let $Z = X-\mu$, $Z' = y-\mu$ and $T = Y-X = z'-Z$. We will show that $\hat{p}(T)$, the density of T, which is independent of μ, is the best invariant density. In effect, the invariance property of $\hat{q}$ with $a = -z$ is obtained for any other competing density $\hat{q}$,

$$\begin{aligned}
\mathscr{R}(\mu,\hat{q})-\mathscr{R}(\mu,\hat{p}) &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left[\log\frac{\hat{p}(y-x)}{\hat{q}(y-x)}\right]p(x-\mu)p'(y-\mu)\,dx\,dy\\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\left[\log\frac{\hat{p}(z'-z)}{\hat{q}(z'-z)}\right]p(z)p'(z')\,dz\,dz'\\
&= \int_{\mathbb{R}^p}\left[\log\frac{\hat{p}(t)}{\hat{q}(t)}\right]\hat{p}(t)\,dt \qquad (4.49)
\end{aligned}$$

which is always positive by the inequality of the information (4.37). The result of the lemma follows from the fact that $\hat{p}(t)=\hat{p}(y-x)=\hat{p}_U(y|x)$, that is,

$$\begin{aligned}
\hat{p}(t) &= \int_{\mathbb{R}^p} p(z)\,p'(z+t)\,dz\\
&= \int_{\mathbb{R}^p} p(z)\,p'(z+y-x)\,dz\\
&= \int_{\mathbb{R}^p} p(x-\mu)\,p'(y-\mu)\,d\mu\\
&= \frac{\int_{\mathbb{R}^p} p'(y|\mu)\,p(x|\mu)\,d\mu}{\int_{\mathbb{R}^p} p(x|\mu)\,d\mu}, \qquad (4.50)
\end{aligned}$$

which is the expression of $\hat{p}_U$ given in (4.41) with $\pi(\mu)=1$. □

Murray [1977] showed that $\hat{p}_U$ is the best invariant density under the action of translations and of linear transformations for a gaussian model. In general location families, Ng [1980] has generalized this result. Liang and Barron [2004], without the hypothesis of independence between X and Y, for the estimation of $p'(y|x,\mu)$ showed that $\hat{p}_U = \frac{\int_{\mathbb{R}^p} p'(y|x,\mu)\,p(x|\mu)\,d\mu}{\int_{\mathbb{R}^p} p(x|\mu)\,d\mu}$ is the best invariant density.

4.4.6 Minimaxity of $\hat{p}_U$

We will now show that $\hat{p}_U$ is minimax in location problems.

Lemma 4.13. *Let $X \sim p(x|\mu) = p(x-\mu)$ and $Y \sim p(y|\mu) = p'(y-\mu)$, with unknown location parameter $\mu \in \mathbb{R}^p$. Assuming that $E_0\left[\,|X\,|^2\right] < \infty$, then the best predictive invariant density $\hat{p}_U$ is minimax.*

Proof. Consider a sequence $\{\pi_k\}$ of normal laws $N_p(0, kI_p)$ priors. The difference of Bayes risk between $\hat{p}_U$ and $\hat{p}_{\pi_k}$, is given by

$$\begin{aligned} r(\hat{p}_U, \pi_k) - r(\hat{p}_{\pi_k}, \pi_k) &= \int_{\mathbb{R}^p} \left[\mathscr{R}(\mu, \hat{p}_U) - \mathscr{R}(\mu, \hat{p}_{\pi_k})\right] \pi_k(\mu)\, d\mu \\ &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} \log \frac{\hat{p}_{\pi_k}(y|x)}{\hat{p}_U(y|x)} p(y|\mu)\, p(x|\mu) \pi_k(\mu)\, dy\, dx\, d\mu \\ &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} \log \frac{\hat{p}_{\pi_k}(y|x)}{\hat{p}_U(y|x)} \left[\int_{\mathbb{R}^p} p(y|\mu)\, p(x|\mu) \pi_k(\mu)\, d\mu\right] dy\, dx \\ &= E^{X,Y}_{\pi_k} \log \frac{\hat{p}_{\pi_k}(Y|X)}{\hat{p}_U(Y|X)} \end{aligned}$$

where $E^{x,y}_{\pi_k}$ denotes the expectation with respect to the joint marginal of (X,Y),

$$m_{\pi_k}(x,y) \;=\; \int_{\mathbb{R}^p} p(y|\mu)\, p(x|\mu) \pi_k(\mu)\, d\mu.$$

By simplifying one gets

$$\begin{aligned} & r(\hat{p}_U, \pi_k) - r(\hat{p}_{\pi_k}, \pi_k) \\ &= E^{X,Y}_{\pi_k}\left[\log\left(\frac{\int p(x,y|\mu)\, \pi_k(\mu)\, d\mu}{\int p(x|\mu)\, \pi_k(\mu)\, d\mu} \frac{1}{\int p(x,y|\mu)\, d\mu}\right)\right] \\ &= E^{X,Y}_{\pi_k}\left[-\log \frac{\int p(x,y|\mu)\, \pi_k(\mu) \frac{1}{\pi_k(\mu)}\, d\mu}{\int p(x,y|\mu)\, \pi_k(\mu)\, d\mu} - \log\left(\int p(x|\mu)\, \pi_k(\mu)\, d\mu\right)\right] \\ &= E^{X,Y}_{\pi_k}\left[-\log E_{\mu|X,Y} \frac{1}{\pi_k(\mu)} - \log\left(\int p(x|\mu)\, \pi_k(\mu)\, d\mu\right)\right] \end{aligned}$$

where $E_{\mu|X,Y}$ denotes the expectation with respect to posterior of μ given (X,Y). An application of Jensen's inequality gives

$$\begin{aligned}&r(\hat{p}_U,\pi_k)-r(\hat{p}_{\pi_k},\pi_k)\\&\leq E^{X,Y}_{\pi_k}E_{\mu|X,Y}\log\pi_k(\mu)-E^{X,Y}_{\pi_k}\left[\int p(X|\mu)\,\log\pi_k(\mu)\,d\mu\right]\end{aligned}\tag{4.51}$$

By developing the expectations it follows that

$$\begin{aligned}E^{X,Y}_{\pi_k}E_{\mu|X,Y}\log\pi_k(\mu) &= \iint m_{\pi_k}(x,y)\frac{\int p(x,y|\mu)\pi_k(\mu)\log(\pi_k(\mu))d\mu}{m_{\pi_k}(x,y)}dxdy\\&=\iiint \pi_k(\mu)\log(\pi_k(\mu))\,d\mu\,dxdy\\&=\int\pi_k(\mu)\log(\pi_k(\mu))d\mu.\end{aligned}\tag{4.52}$$

Similarly by integrating with respect to y, and by interchanging between μ and μ' we have

$$\begin{aligned}&E^{X,Y}_{\pi_k}\left[\int p(X|\mu)\,\log\pi_k(\mu)\,d\mu\right]\\&=\iiiint p(x|\mu')p(y|\mu')\pi_k(\mu')p(x|\mu)\log\pi_k(\mu)\,d\mu' d\mu dxdy\\&=\iiint\pi_k(\mu')p(x|\mu)p(x|\mu')\log\pi_k(\mu)d\mu'\,dx\,d\mu\\&=\iiint\pi_k(\mu)p(x|\mu)p(x|\mu')\log\pi_k(\mu')d\mu\,dxd\mu'.\end{aligned}\tag{4.53}$$

By grouping the expressions (4.51), (4.53) and (4.54) and making the changes in variables $z=x-\mu$ and $z'=x-\mu'$ it follows that

$$\begin{aligned}&r(\hat{p}_U,\pi_k)-r(\hat{p}_{\pi_k},\pi_k)\\&\leq\iiint p(x|\mu)p(x|\mu')\pi_k(\mu)\left[\log(\pi_k(\mu))-\log(\pi_k(\mu'))\right]d\mu d\mu'\,dx\\&=\iiint\pi_k(\mu)p(x-\mu)p(x-\mu')\log\left(\frac{\pi_k(\mu)}{\pi_k(\mu')}\right)d\mu\,dzdz'\\&=\iiint\pi_k(\mu)p(z)p(z')\log\left(\frac{\pi_k(\mu)}{\pi_k(\mu+z-z')}\right)d\mu\,dzdz',\end{aligned}\tag{4.54}$$

where $<,>$ denotes the scalar product in $\mathbb{R}^p$. In view of the form $\pi_k(\mu)$, the term of right in (4.54) is written as

$$\begin{aligned} E_{\pi_k} E_{Z,Z'} \log\left(\frac{\pi_k(\mu)}{\pi_k(\mu+Z-Z')}\right) &= E_{\pi_k} E_{Z,Z'} \frac{1}{2k}\left(\|\mu+Z-Z'\| - \|\mu^2\|\right) \\ &= E_{\pi_k} E_{Z,Z'}\left[\frac{1}{2k}\left(\|Z\|^2 + \|Z'\|^2 + 2 < \mu, Z-Z' >\right)\right] \\ &= E_{Z,Z'}\left[\frac{1}{2k}\left(\|Z\|^2 + \|Z'\|^2\right)\right] \end{aligned}$$

Since $E(Z) = E(Z') = E_0(X)$ (here, $E_{Z,Z'}$ denotes the expectation with respect to $p(z,z') = p(z)p(z')$). We then see that the limit of the difference of Bayes risks tends toward zero when $k \to \infty$. Therefore $\hat{p}_U$ is minimax since it is a limit of Bayes rules. □

This result is in Liang and Barron [2004], a more direct proof for the Gaussian case can be found in George et al. [2006].

4.4.7 An explicit expression of $\hat{p}_U$

We now give an explicit expression of $\hat{p}_U$ in the Gaussian setting given described in (4.4.3).

Lemma 4.14. *The Bayesian predictive density associated with the uniform prior on $\mathbb{R}^p$, $\pi(\mu) \equiv 1$, is given by the following expression*

$$\hat{p}_U(y|x) = \frac{1}{[2\pi(v_x+v_y)]^{p/2}} \exp\left[-\frac{1}{2}\frac{\|y-x\|^2}{v_x+v_y}\right]. \tag{4.55}$$

Proof. For $W = (v_y X + v_x Y)/(v_x + v_y)$, $v_w = (v_x v_y)/(v_x + v_y)$ it is clear that, by independence of X and Y, $W \sim N_p(\mu, v_w I_p)$. Further note that,

$$\frac{\|x-\mu\|^2}{2v_x} + \frac{\|y-\mu\|^2}{2v_y} = \frac{\|\mu - w\|^2}{2v_w} + \frac{\|y-x\|^2}{2(v_x+v_y)}. \tag{4.56}$$

By definition, and through the previous representation, it follows that

$$\begin{aligned}
\hat{p}_U(y|x) &= \frac{\int_{\mathbb{R}^p} p(y|\mu, v_y)\, p(x|\mu, v_x)\, d\mu}{\int_{\mathbb{R}^p} p(x|\mu, v_x)\, d\mu} \\
&= \int_{\mathbb{R}^p} \frac{1}{(2\pi)^p (v_y v_x)^{p/2}} \exp\left(-\frac{\|x-\mu\|^2}{2v_x} - \frac{\|y-\mu\|^2}{2v_y}\right) d\mu \\
&= \int_{\mathbb{R}^p} \frac{1}{(2\pi)^p (v_y v_x)^{p/2}} \exp\left(-\frac{\|\mu-w\|^2}{2v_w}\right) \exp\left(-\frac{\|y-x\|^2}{2(v_x+v_y)}\right) d\mu \\
&= \frac{(2\pi v_w)^{p/2}}{(2\pi)^p (v_y v_x)^{p/2}} \exp\left(-\frac{\|y-x\|^2}{2(v_x+v_y)}\right).
\end{aligned}$$

□

The risk of $\hat{p}_U$ is constant, as we have previously seen for invariant densities. We will exploit this fact in determining, in advance, the loss incurred by $\hat{p}_U(.|x)$ after having observed x then that the parameter equals μ. Therefore we can now write

$$\begin{aligned}
&\mathrm{KL}(\hat{p}_U(.|x), \mu) \\
&= \int p(y|\mu, v_y) \log \frac{p(y|\mu, v)}{\hat{p}_U(y|x)}\, dy \\
&= E^Y\left[\log \frac{p(Y|\mu, v)}{\hat{p}_U(Y|x)}\right] \\
&= E^Y\left[-\frac{p}{2}\log\frac{v_y}{v_x+v_y} - \frac{1}{2v_y}\|Y-\mu\|^2 + \frac{1}{2(v_x+v_y)}\|Y-x\|^2\right] \\
&= -\frac{p}{2}\log\frac{v_y}{v_x+v_y} - \frac{p}{2} + E^Y\left[\frac{1}{2(v_x+v_y)}\left(\|Y-\mu\|^2 + \|\mu-x\|^2\right)\right] \\
&= \left[-\frac{p}{2}\log\frac{v_y}{v_x+v_y} - \frac{p}{2} + \frac{pv_y}{2(v_x+v_y)}\right] + \frac{1}{2(v_x+v_y)}\|\mu-x\|^2.
\end{aligned} \tag{4.57}$$

Hence we can conclude that the risk of $\hat{p}_U$ is

$$\begin{aligned}
\mathscr{R}_{\mathrm{KL}}(\hat{p}_U,\mu) &= E^X\left[\mathrm{KL}(\hat{p}_U,\mu,X)\right] \\
&= \left[-\frac{p}{2}\log\frac{v_y}{v_x+v_y}-\frac{p}{2}+\frac{pv_y}{2(v_x+v_y)}\right]+\frac{pv_x}{2(v_x+v_y)} \\
&= -\frac{p}{2}\log\left(\frac{v_y}{v_x+v_y}\right)=\frac{p}{2}\log\left(1+\frac{v_x}{v_y}\right). \qquad (4.58)
\end{aligned}$$

Note that we can express the risk, in the framework of the i.i.d sampling model presented in (4.4.3) as

$$\mathscr{R}_{\mathrm{KL}}(\hat{p}_U,\mu)=\frac{p}{2}\log\left(1+\frac{m}{n}\right).$$

A predictive density is called the plug-in relative to an estimator Δ if it has the form

$$\hat{p}_1(y|x)=\frac{1}{[2\pi v_y]^{p/2}}\exp\left[-\frac{1}{2}\frac{|y-\Delta(x)\ |^2}{v_y}\right].$$

The predictive plug-in density which corresponds to the standard estimator of the mean μ $\delta_0(X)=X$ is

$$\hat{p}_1(y|x)=\frac{1}{[2\pi v_y]^{p/2}}\exp\left[-\frac{1}{2}\frac{|y-x\ |^2}{v_y}\right].$$

We can directly verify that the predictive density $\hat{p}_U$ dominates the plug-in density $\hat{p}_1$ for any $\mu\in\mathbb{R}^p$. In effect their difference in risk equals

$$\begin{aligned}
\triangle\mathscr{R}_{\mathrm{KL}}(\hat{p}_U,\hat{p}_1) &= E^{X,Y}\left[\log\frac{\hat{p}_U(Y|X)}{\hat{p}_1(Y|X)}\right] \\
&= -\frac{p}{2}\log\left(\frac{v_x+v_y}{v_y}\right)-\frac{1}{2}\left[\frac{1}{v_x+v_y}-\frac{1}{v_y}\right]E^{X,Y}\left(\|Y-X\|^2\right).
\end{aligned}$$

Since $E^{X,Y}\left(\|Y-X\|^2\right)$ equals

$$E^{X,Y}\left(\|Y-\mu\|^2\right)+E^{X,Y}\left(\|X-\mu\|^2\right)-2\left\langle E^{X,Y}(Y-\mu),E^{X,Y}(X-\mu)\right\rangle = p(v_x+v_y),$$

we have then

$$\triangle\mathscr{R}_{\text{KL}}(\hat{p}_U,\hat{p}_1) = -\frac{p}{2}\left[\log\left(1+\frac{v_x}{v_y}\right)-\frac{v_x}{v_y}\right] > 0.$$

4.4.8 Domination of $\hat{p}_U$

Surprisingly the predictive density $\hat{p}_U$ has similar properties as the standard estimator $\delta_0(X)=X$, for the estimation of the mean under quadratic loss. Komaki [2001] showed that the density $\hat{p}_U$ is dominated by the Bayesian density using the harmonic prior, $\pi(\mu) = \ ||\mu\ ||^{2-p}$.

Lemma 4.15. *George et al. [2006] (Lemma 2) For $W=(v_yX+v_xY)/(v_x+v_y)$, $v_w=(v_xv_y)/(v_x+v_y)$, and $m_\pi(W;v_w)$ and $m_\pi(X;v_x)$ the marginal W and X respectively compared to the* a priori π. *We have*

$$\hat{p}_\pi(y|X) = \frac{m_\pi(W;v_w)}{m_\pi(X;v_x)}\ \hat{p}_U(y|X)\,, \tag{4.59}$$

where $\hat{p}_U(\cdot|X)$ is the Bayes estimator associated with uniform prior on $\mathbb{R}^p$ given by (4.55). In addition, for any prior measure π, the risk difference between the Kullback Leibler and $\hat{p}_U(\cdot|x)$ and the Bayesian predictive density $\hat{p}_\pi(\cdot|x)$ is given by

$$\mathscr{R}_{KL}(\mu,\hat{p}_U)-\mathscr{R}_{KL}(\mu,\hat{p}_\pi) = E_{\mu,v_w}\left[\log m_\pi(W;v_w)\right]-E_{\mu,v_x}\left[\log m_\pi(X;v_x)\right] \tag{4.60}$$

where $E_{\mu,v}$ denotes the expectation with respect to the normal $N_p(\mu,vI_p)$ law.

Proof. The marginal density of (X,Y) associated with π is equal to

$$\begin{aligned}\hat{p}_\pi(x,y) &= \int_{\mathbb{R}^p} p(x|\mu,v_x)\, p(y|\mu,v_y)\, \pi(\mu)\, d\mu \\ &= \int_{\mathbb{R}^p} \frac{1}{(2\pi v_x)^{p/2}} \exp\left(-\frac{\|x-\mu\|^2}{2v_x}\right) \frac{1}{(2\pi v_y)^{p/2}} \exp\left(-\frac{\|y-\mu\|^2}{2v_y}\right) \pi(\mu)\, d\mu.\end{aligned}$$

using (4.60) it follows that

$$\begin{aligned}\hat{p}_\pi(x,y) &= \frac{1}{(2\pi)^p\,(v_x v_y)^{p/2}} \int_{\mathbb{R}^p} \exp\left(-\frac{\|y-x\|^2}{2(v_x+v_y)}\right) \exp\left(-\frac{\|\mu-w\|^2}{2v_w}\right) \pi(\mu)\, d\mu \\ &= \frac{(2\pi v_w)^{p/2}}{(2\pi)^p\,(v_x v_y)^{p/2}} \exp\left(-\frac{\|y-x\|^2}{2(v_x+v_y)}\right) m_\pi(w;v_w) \\ &= \hat{p}_U(y|x)\, m_\pi(w;v_w).\end{aligned}$$

Hence, we can then write the risk difference as

$$\begin{aligned}\mathscr{R}_{\mathrm{KL}}(\mu,\hat{p}_U) - \mathscr{R}_{\mathrm{KL}}(\mu,\hat{p}_\pi) &= \int\int p(x|\mu,v_x)\, p(y|\mu,v_y) \log \frac{\hat{p}_\pi(y|x)}{\hat{p}_U(y|x)}\, dy\, dx \\ &= \int\int p(x|\mu,v_x)\, p(y|\mu,v_y) \log \frac{m_\pi(W(x,y);v_w)}{m_\pi(x;v_x)}\, dy\, dx \\ &= ibnE^{X,Y} \log m_\pi(W(X,Y);v_w) - E^{X,Y} \log m_\pi(X;v_x) \\ &= E^W \log m_\pi(W(X,Y);v_w) - E^m \log m_\pi(X;v_x)\end{aligned}$$

□

Using this lemma, George [2006] give a simple proof of the result of Liang and Barron [2004] for the Gaussian setting. By taking the same $\{\pi_k\} = N_p(0,kI_p)$ sequence, the difference of the Bayes risk equals

$$
\begin{aligned}
r(\pi_k, \hat{p}_U) - r(\pi_k, \hat{p}_{\pi_k}) &= \int \pi_k(\mu) \left[E_{\mu, v_w} \log m_{\pi_k}(W, v_w) - E_{\mu, v_x} \log m_{\pi_k}(X, v_x) \right] d\mu \\
&= \int \pi_k(\mu) \Bigg[E_{\mu, v_w} \log \left\{ (2\pi(v_w + k))^{-p/2} \exp\left(-\frac{\|W\|^2}{2(v_w + k)} \right) \right\} \\
&- E_{\mu, v_x} \log \left\{ (2\pi(v_x + k))^{-p/2} \exp\left(-\frac{\|X\|^2}{2(v_x + k)} \right) \right\} \Bigg] d\mu \\
&= \int \pi_k(\mu) \Bigg[-p/2 \log(2\pi(v_w + k)) - \frac{p v_w}{2(v_w + k)} \\
&+ p/2 \log(2\pi(v_x + k)) + \frac{p v_x}{2(v_x + k)} \Bigg] d\mu \\
&= -\frac{p}{2} \log \frac{v_w + k}{v_x + k} - \frac{p v_w}{2(v_w + k)} + \frac{p v_x}{2(v_x + k)}.
\end{aligned}
$$

Hence we see that $\lim_{k \to \infty} r(\pi_k, \hat{p}_U) - r(\pi_k, \hat{p}_{\pi_k}) = 0$ and so that $\hat{p}_U$ is minimax. George [2006] also show that the best predictive invariant density is dominated by any Bayesian predictive density relative to the surharmonic prior. This result is parallels the result of Stein for the estimation of the mean under quadratic loss. The lemma following allows us to give sufficient conditions for domination, we use Stein's identity in the proof.

Lemma 4.16. *George et al [2006] If $m_\pi(z; v_x)$ is finite for any z, then for any $v_w \le v \le v_x$ the marginal $m_\pi(z; v)$ is finite. In addition,*

$$
\begin{aligned}
\frac{\partial}{\partial v} E \log m_\pi(z; v) &= E_{\mu, v} \left[\frac{\Delta m_\pi(Z; v)}{m_\pi(Z; v)} - \frac{1}{2} \|\nabla \log m_\pi(Z; v)\|^2 \right] \\
&= E_{\mu, v} \left[2 \frac{\Delta \sqrt{m_\pi(Z; v)}}{\sqrt{m_\pi(Z; v)}} \right]. \qquad (4.61)
\end{aligned}
$$

Proof. For any $v_w \le v \le v_x$,

$$\begin{aligned}
m_\pi(z;v) &= \int_{\mathbb{R}^p} \frac{1}{(2\pi v)^p/2} \exp\left(-\frac{\|z-\mu\|^2}{2v}\right) d\mu \\
&= (\frac{v_x}{v})^{p/2} \int_{\mathbb{R}^p} \frac{1}{(2\pi v_x)^p/2} \exp\left(-\frac{v_x}{v}\frac{\|z-\mu\|^2}{2v_x}\right) d\mu \\
&\leq (\frac{v_x}{v})^{p/2} m_\pi(z;v_x) < \infty.
\end{aligned}$$

Hence the marginal m_π is finite. Setting $Z^* = (Z-\mu)/\sqrt{v} \sim (0,I)$,

$$\begin{aligned}
\frac{\partial}{\partial v} E_{\mu,v} \log m_\pi(Z;v) &= \frac{\partial}{\partial v} \int p(z|\mu,v) \log\left(m_\pi(z;v)\, dz\right) \\
&= \frac{\partial}{\partial v} \int p(z'|0,1) \log\left(m_\pi(\sqrt{v}z'+\mu;v)\right) dz' \\
&= E_{Z'} \frac{(\partial/\partial v) m_\pi(\sqrt{v}Z'+\mu;v)}{m_\pi(\sqrt{v}Z'+\mu;v)} \qquad (4.62)
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial}{\partial v} m_\pi(\sqrt{v}z'+\mu;v) &= \frac{\partial}{\partial v} \int \frac{1}{(2\pi v)^{p/2}} \exp\left\{\frac{\|\sqrt{v}z'+\mu-\mu'\|^2}{2v}\right\} \pi(\mu')\, d\mu' \\
&= \int \left(-\frac{p}{2v} + \frac{\|z-\mu'\|^2}{2v^2} - \frac{\|z'\|^2}{2v} - \frac{< z', \mu-\mu' >}{2v^{3/2}}\right) p(z|\mu')\pi(\mu') d\mu' \\
&= \frac{\partial}{\partial v} m_\pi(z;v) - \int \frac{< z-\mu, z-\mu' >}{2v^2} p(z|\mu')\pi(\mu') d\mu'. \qquad (4.63)
\end{aligned}$$

Note that

$$\nabla_z m_\pi(z,v) = \int \frac{-(z-\mu)}{v} p(z|\mu)\pi(\mu) d\mu \qquad (4.64)$$

and

$$\begin{aligned}
\Delta_z m_\pi(z,v) &= \int \left[\frac{-p}{v} + \frac{\|z-\mu\|^2}{v^2}\right] p(z|\mu)\pi(\mu) d\mu \\
&= 2\frac{\partial}{\partial v} m_\pi(z;v). \qquad (4.65)
\end{aligned}$$

It can be shown that

$$E_{Z'} \frac{(\partial/\partial v) m_\pi(\sqrt{v}Z'+\mu;v)}{m_\pi(\sqrt{v}Z'+\mu;v)} = E_{\mu,v} \left(\frac{1}{2}\frac{\Delta m_\pi(Z;v)}{m_\pi(Z;v)} + \frac{< Z-\mu, \nabla \log m_\pi(Z;v) >}{2v}\right)$$

Now, using Stein's identity it follows that

$$
\begin{aligned}
E_{\mu,v}\left[\frac{< Z-\mu, \nabla \log m_\pi(Z;v) >}{2v}\right] &= E_{\mu,v}\left[\frac{1}{2}\Delta \log m_\pi(Z;v)\right] \\
&= E_{\mu,v}\left[\frac{1}{2}\left(\frac{\Delta m_\pi(Z;v)}{m_\pi(Z;v)} - \|\nabla \log m_\pi(Z;v)\|^2\right)\right].
\end{aligned}
$$

which is the desired result. □

Lemma 4.16 can be used to prove the following result regarding minimaxity and domination.

Theorem 4.12. *George [2006] Assuming that $m_\pi(z;v_x)$ is finite for any z in $\mathbb{R}^p$. If $\Delta m_\pi \leq 0$ for all $v_w \leq v \leq v_x$, then the Bayesian predictive density $\hat{p}_\pi(y|x)$ is minimax and she dominated $\hat{p}_U$ (when π is not the uniform itself). If $\Delta\pi \leq 0$, then the Bayesian predictive density $\hat{p}_\pi(y|x)$ is minimax and she dominated $\hat{p}_U$ (when π is not the uniform it even).*

4.4.9 Links with estimates under quadratic loss

The next result illuminates the link between the two problems, the estimate of predictive densities under the Kullback Leibler loss and the estimation of the mean under quadratic loss. This theorem sets out the link in terms of risk differences.

Theorem 4.13. *Brown [2008] Suppose the prior $\pi(\mu)$ is such that the marginal $m_\pi(z;v)$ is finite for any $z \in \mathbb{R}^p$. Then*

$$
\mathscr{R}_{KL}(\mu,\hat{p}_U) - \mathscr{R}_{KL}(\mu,\hat{p}_\pi) = \frac{1}{2}\int_{v_w}^{v_x} \frac{1}{v^2}\left(\mathscr{R}_Q^v(\mu,X) - \mathscr{R}_Q^v(\mu,\hat{\mu}_{\pi,v})\right) dv. \tag{4.66}
$$

Proof. From (4.60) and (4.61) it follows

$$\begin{aligned}\mathscr{R}_{\mathrm{KL}}(\mu,\hat{p}_U)-\mathscr{R}_{\mathrm{KL}}(\mu,\hat{p}_\pi) &= \int_{v_w}^{v_x} -\frac{\partial}{\partial v}E_{\mu,v}\log m_\pi(Z;v)\,dv \\ &= \int_{v_w}^{v_x} E_{\mu,v}\left[2\frac{\Delta\sqrt{m_\pi(Z;v)}}{\sqrt{m_\pi(Z;v)}}\right]dv. \qquad (4.67)\end{aligned}$$

On the other hand, Stein [1981] showed that

$$\mathscr{R}_Q^v(\mu,X)-\mathscr{R}_Q^v(\mu,\hat{\mu}_{\pi,v}) = -4v^2E_{\mu,v}\frac{\Delta\sqrt{m_\pi(Z;v)}}{\sqrt{m_\pi(Z;v)}}. \qquad (4.68)$$

Finally, replacing (4.68) in the integral (4.67) we find the statement. □

Note that

$$\begin{aligned}\frac{1}{2}\int_{v_w}^{v_x}\frac{1}{v^2}\left(\mathscr{R}_Q^v(\mu,X)\right)dv &= \frac{1}{2}\int_{v_w}^{v_x}\frac{p}{v}\,dv \\ &= p\log\frac{v_x}{v_w} = \frac{p}{2}\log\left(1+\frac{v_x}{v_y}\right). \\ &= \mathscr{R}_{\mathrm{KL}}(\mu,\hat{p}_U) \qquad (4.69)\end{aligned}$$

This is same as the risk of $\hat{p}_U$ in (4.65). Then from (4.68) and (4.69) we have for any for any prior π,

$$\mathscr{R}_{\mathrm{KL}}(\mu,\hat{p}_\pi) = \frac{1}{2}\int_{v_w}^{v_x}\frac{1}{v^2}\mathscr{R}_Q^v(\mu,\hat{\mu}_{\pi,v})\,dv. \qquad (4.70)$$

4.5 Admissibility of Bayes estimators

Recall from Section 2.4 that an admissible estimator is one that cannot be dominated in risk, i.e. $\delta(X)$ is admissible if there does not exist an estimator $\delta'(X)$ such that $R(\theta,\delta')\leq R(\theta,\delta)$ for all θ, with strict inequality for some θ. We have already derived classes of minimax estimators in the previous sections. In this section we study their possible admissibility or inadmissibility. One reason that admissibility of these minimax estimators is interesting is that, as we have already seen, the usual estimator $\delta_0(X) = X$ is minimax but inadmissible if $p \geq 3$. Actually we have seen that it is possible to dominate X with a minimax estimator (e.g., $\delta^{JS}_{(p-2)}(X)$) that has a substantially smaller risk at $\theta = 0$. Hence it is of interest to know if a particular (dominating) estimator is admissible.

Note that a unique proper Bayes estimator is automatically admissible (see Lemma 2.6), so we already have examples of admissible minimax estimators for $p \geq 5$.

We note that the class of generalized Bayes estimators contains all admissible estimators if loss is quadratic (i.e., it is a complete class; see e.g., Sacks [1963], Brown [1971], and Berger and Srinivasan [1978]. It follows that if an estimator is not generalized Bayes, it is not admissible. Further, in order to be generalized Bayes, an estimator must be everywhere differentiable by properties of the Laplace Transform. In particular, the James-Stein estimators and the positive-part James-Stein estimators (for $a \neq 0$) are not generalized Bayes and therefore not admissible.

Brown [1971] gave an important general result which is almost a characterization of admissible generalized Bayes estimators in the multivariate normal case. We give below (without proof) a version of Brown's result (with $\sigma^2 = 1$ for convenience and without essential loss of generality). Let $X \sim N_p(\theta, I)$, and loss be $L(\theta, \delta) = \|\delta - \theta\|^2$ and let $\pi(\theta)$ be a generalized prior distribution for θ with associated generalized marginal density $m(x)$ and generalized Bayes estimator $\delta(X) = X + \frac{\nabla m(X)}{m(X)}$.

Define, for $r = \|x\|$,

$$\bar{m}(r) = \int m(x) dU_r(x) \tag{4.71}$$

and

$$\underline{m}(r) = \int (1/m(x)) dU_r(x) \tag{4.72}$$

where $U_r(x)$ is the uniform distribution on the sphere of radius r. Brown's (1971) result is the following.

Theorem 4.14. (1) (Admissibility) *If* $\|\delta(X) - X\|$ *is uniformly bounded and*

$$\int_c^\infty (r^{p-1}\bar{m}(r))^{-1} \, dr = \infty$$

for some $c > 0$*, then* $\delta(X)$ *is admissible.*

(2) (Inadmissibility) *If* $\int_c^\infty r^{1-p}\underline{m}(r)\, dr < \infty$ *for some* $c > 0$*, then* $\delta(X)$ *is inadmissible.*

Brown's result has a wide applicability. See for example Berger and Strawderman [1996] and Berger, Strawderman and Tang [2005] and Kubokawa and Strawderman [2005] for application to classes of hierarchical prior distributions.

In this section we will study admissibility of estimators corresponding to priors which are variance mixtures of normal distributions as in Subsection 4.1.2. In particular we consider prior densities of the form

$$\pi(\theta) = \int_0^\infty \exp\left\{ -\frac{\|\theta\|^2}{2(1+v)} \right\} (1+v)^{-\frac{p}{2}} \, h(v) \, dv \tag{4.73}$$

and establish a connection between admissibility and the behavior of the mixing (generalized) density $h(v)$ at infinity. An Abelian Theorem (see, e.g., Wider [1946], Corollary 1.a, p. 182) along with Brown's theorem are our main tools. We use the notation $f(X) \sim g(X)$ as $X \to a$ to mean $\lim_{X\to a} \frac{f(X)}{g(X)} = 1$. Here is an adaptation of the Abelian Theorem in Widder that meets our needs.

Theorem 4.15. *Assume $g(t)$ from $\mathbb{R}^+$ into $\mathbb{R}$ has a Laplace transform $f(s) = \int_0^\infty g(t)e^{-st}\,dt$ which is finite for $s \geq 0$. If $g(t) \sim t^\gamma$ as $t \to 0_+$ for some $\gamma > -1$, then $f(s) \sim s^{-(\gamma+1)}\Gamma(\gamma+1)$ as $s \to \infty$.*

The proof is essentially as in Widder but the assumption of finiteness of the Laplace transform at $s = 0$ allows the extension from $\gamma \geq 0$ to $\gamma > -1$.

We first give a lemma which relates the tail behavior of the mixing density $h(v)$ to the tail behavior of $\pi(\|\theta\|^2)$ and $m(\|x\|^2)$ and also shows that $\|\delta(x) - x\|$ is bounded whenever $h(v)$ has polynomial tail behavior.

Lemma 4.17. *Suppose $X \sim N_p(\theta, I_p), L(\theta, \delta) = \|\delta - \theta\|^2$ and $\pi(\theta)$ is given by (2.24) where $h(v) \sim Kv^a$ as $v \to \infty$ with $a < \frac{p-2}{2}$ and $v^{-\frac{p}{2}}h(v)$ is integrable in a neighborhood of 0. Then*

(1) *$\pi(\theta) \sim K(\|\theta\|^2)^{a-\frac{p-2}{2}}\Gamma\left(\frac{p-2}{2} - a\right)$ as $\|\theta\|^2 \to \infty$, $m(x) \sim K(\|x\|^2)^{a-\frac{p-2}{2}}\Gamma\left(\frac{p-2}{2} - a\right)$ as $\|x\|^2 \to \infty$, and therefore $\pi(\|x\|^2) \sim m(\|x\|^2)$ as $\|x\|^2 \to \infty$, and*

(2) *$\|\delta(x) - x\|$ is uniformly bounded.*

Proof. First note that (with $t = 1/v$)

$$\pi(\theta) = \pi^*(\|\theta\|^2) = \int_0^\infty \exp\left\{-\frac{\|\theta\|^2}{2}t\right\} t^{\frac{p}{2}-2} h(1/t)\,dt$$

and $g(t) = t^{\frac{p}{2}-2}h(1/t) \sim Kt^{\frac{p-4}{2}-a}$ as $t \to 0_+$. Therefore, by Theorem 4.9, $\pi(\theta) \sim K(\|\theta\|^2)^{a-\frac{p-2}{2}}\Gamma\left(\frac{p-2}{2}-a\right)$ as $\|\theta\|^2 \to \infty$. Similarly

$$\begin{aligned} m(x) &= \int_0^\infty e^{-\frac{\|\theta\|^2}{2(1+v)}}(1+v)^{-\frac{p}{2}}h(v)\,dv \left(\text{for } t = \frac{1}{1+v}\right) \\ &= \int_1^\infty e^{-\frac{\|\theta\|^2}{2}t}t^{\frac{p}{2}-2}h\left(\frac{1-t}{t}\right)dt. \end{aligned}$$

We note as $t \to 0_+$, $t^{\frac{p}{2}-2}h\left(\frac{1-t}{t}\right) \sim t^{\frac{p-4}{2}}\left(\frac{1-t}{t}\right)^a \sim t^{\frac{p-4}{2}-a}$. Thus, again by Theorem 4.9,

$$m(x) \sim K(\|x\|^2)^{a-\frac{p-2}{2}}\Gamma\left(\frac{p-2}{2}-a\right) \text{ as } \|x\|^2 \to \infty,$$

and part (1) follows.

To prove part (2) note that

$$\begin{aligned} \delta(x) - x &= \frac{\nabla m(x)}{m(x)} \\ &= -\frac{-\int_0^\infty \exp\left\{-\frac{\|x\|^2}{2(1+v)}\right\}(1+v)^{-(\frac{p}{2}+1)}h(v)\,dv}{\int_0^\infty \exp\left\{-\frac{\|x\|^2}{2(1+v)}\right\}(1+v)^{\frac{p}{2}}h(v)\,dv}x. \end{aligned}$$

The above argument applied to the numerator and denominator shows

$$\begin{aligned} \|\delta(x) - x\|^2 &\sim \left[\frac{(\|x\|^2)^{a-\frac{p}{2}}\Gamma(\frac{p}{2}-a)}{\|x\|^2)^{a-\frac{p-2}{2}}\Gamma(\frac{p-2}{2}-a)}\right]^2\|x\|^2 \\ &\sim \left(\frac{p-2}{2}-a\right)^2\frac{1}{\|x\|^2} \text{ as } \|x\|^2 \to \infty. \end{aligned}$$

Since $\delta(x) - x$ is in C_∞ and tends to zero as $\|x\|^2 \to \infty$, the function is uniformly bounded.

The following result characterizes admissibility and inadmissibility for generalized Bayes estimators when the mixing density $h(v) \sim v^a$ as $v \to \infty$.

Theorem 4.16. *For the set-up of Lemma 4.2 the generalized Bayes estimator corresponding to $\pi(\theta)$, with mixing density $h(v) \sim v^a$ as $v \to \infty$, is admissible if and only if $a \leq 0$.*

Proof. (Admissibility if $a \leq 0$) By Lemma 4.6 a with $m(x) = m^*(\|x\|^2)$ and $\bar{m}(r) = m^*(r^2) \sim K^*(r^2)^{a-\frac{p-2}{2}}$. Thus, for any $\varepsilon > 0$, there is an $r_0 > 0$ such that, for $r > r_0$, $\bar{m}(r) \leq (1+\varepsilon)K^* r^{2a-(p-2)}$. Since $\|\delta(x) - x\|$ is uniformly bounded, and

$$\int_{r_0}^{\infty} (r^{p-1}\bar{m}(r))^{-1}\, dr \geq (K^*(1+\varepsilon))^{-1} \int_{r_0}^{\infty} r^{-(2a+1)}\, dr = \infty$$

if $a \geq 0$. Hence $\delta(x)$ is admissible if $a \leq 0$, by Theorem 4.4.

Now for inadmissibility, similarly

$$\underline{m}(r) = \frac{1}{m^*(r^2)} \sim \frac{1}{K^*}(r^2)^{\frac{p-2}{2}-a},$$

for $r \geq r_0$

$$\underline{m}(r) \leq \frac{1}{(1-\varepsilon)K^*} r^{p-2-2a},$$

and

$$\int_0^{\infty} r^{1-p}\underline{m}(r)\, dr \leq \frac{1}{K^*} \int_{r_0}^{\infty} r^{-(1+2a)}\, dr < \infty$$

if $a > 0$.

Thus $\delta(x)$ is inadmissible if $a > 0$. □

Example 4.13. (Continued) Recall for the Strawderman prior that $h(v) = C(1+v)^{-\alpha-(\frac{p-2}{2})} \sim v^a$ for $a = -(\alpha + \frac{p-2}{2})$ as $v \to \infty$.

The above theorem implies that the generalized Bayes estimator is admissible if and only if $\alpha + \frac{p-2}{2} \geq 0$ or $1 - \frac{p}{2} \leq \alpha$. We previously established minimaxity when $2 - p < \alpha \leq 0$ for $p \geq 3$ and propriety of the prior when if $2 - \frac{p}{2} < \alpha \leq 0$ for $p \geq 5$.

Note in general that for a mixing distribution of the form $h(v) \sim Kv^a$ as $v \to \infty$, the prior distribution $\pi(\theta)$ will be proper if and only if $a < -1$ by the same argument as in the proof of Theorem 4.6. Hence the bound for admissibility, $a \leq 0$, differs from the bound for propriety, $a < -1$ by 1.

Chapter 5

Estimation of location parameter for the spherically symmetric case I

5.1 Introduction

In the previous chapter, we studied improved estimators over the "usual" estimator of the location vector for the case of a normal distribution. In this chapter, we extend the discussion to the spherically symmetric case. The reader is referred to the discussion in Section 1.3 for basic concepts. Section 5.2 is devoted to a discussion of basic domination results for Baranchik type estimators while Section 5.3 examines more general estimators. Section 5.4 discusses Bayes minimax estimation. Finally, Section 5.5 discusses admissibility issues.

We close this introductory section by extending the discussion of Subsection 2.2 on the empirical Bayes justification of the James-Stein estimator to the general multivariate (but not necessarily normal) case.

Suppose X has a p-variate distribution with density $f(x-\theta)$, unknown mean vector θ and known covariance matrix $\sigma^2 I_p$. The problem is to estimate θ under loss $L(\theta,\delta)=\|\delta-\theta\|^2$. Let the prior distribution on θ be given by $\pi(\theta)=f^{*n}(\theta)$, the n-fold convolution of the density $f(\cdot)$ with itself.

The Bayes estimator of θ is given by

$$\delta_n(X)=\frac{n}{n+1}X=\left(1-\frac{1}{n+1}\right)X.$$

To see this, note that $X=(X-\theta)+\theta$. But since $Y_0=X-\theta\sim f(Y_0)$ and is independent of $\theta\sim\sum_{i=1}^{n}Y_i$, where Y_i are i.i.d. $\sim f(Y_i)$, it follows that $X=\sum_{i=0}^{n}Y_i\sim f^{*(n+1)}(X)$. Therefore the Bayes estimator is given by

$$\begin{aligned}\delta_n(X)&=E(\theta|X)=X+E[(\theta-X)|X]\\&=X-E[Y_0|X]\\&=X-E\left[\frac{1}{n+1}\sum_{i=0}^{n+1}Y_i\middle|X\right]\quad(\text{since } E[Y_i|X]=E[Y_j|X])\\&=X-E\left[\frac{X}{n+1}\middle|X\right]\\&=\left(1-\frac{1}{n+1}\right)X\\&=\frac{n}{n+1}X.\end{aligned}$$

Assume now that n is unknown. Since

$$\begin{aligned}E(X'X)&=E\left(\sum_{i=0}^{n}Y_i'Y_i\right)\\&=(n+1)\,E(Y_0'Y_0)\\&=(n+1)\,(\operatorname{tr}\sigma^2 I)\\&=(n+1)\,p\,\sigma^2,\end{aligned}$$

an unbiased estimator of $n+1$ is $X'X/(p\sigma^2)$, and so $p\sigma^2/(X'X)$ is a reasonable estimator of $1/(n+1)$. Substituting $p\sigma^2/(X'X)$ for $1/(n+1)$ in the Bayes estimator, we have that

$$\delta^{EB}(X) = \left(1 - \frac{p\,\sigma^2}{X'X}\right) X$$

can be viewed as an empirical Bayes estimator of θ without any assumption on the form of the density (and in fact there is not even any need to assume there is a density). Hence a Stein-like estimator can be viewed as a reasonable alternative to X from an empirical Bayes perspective regardless of the form of the underlying distribution.

Note that Diaconis and Freedman introduced the prior $f^{*n}(\theta)$ as a reasonable conjugate prior for location families since it gives linear Bayes estimators. Strawderman [1992] gave the above empirical Bayes argument. In the normal case the sequence of priors corresponds to that in Subsection 3.2.3 with $\tau^2 = n\sigma$. The shrinkage factor $p\sigma^2$ in the present argument differs from $(p-2)\sigma^2$ in the normal case since in this general case we use a "plug-in" estimator of $1/(n+1)$ as opposed to the unbiased estimator (in the normal case) of $1/(\sigma^2+\tau^2)$.

5.2 Baranchik type estimators

In this section, assuming that X has a spherically symmetric distribution with mean vector θ and that loss is $L(\theta,\delta) = \|\delta - \theta\|^2$, we consider estimators of the

Baranchik type for different families of densities. In Section 5.3, we consider results for general estimators of the form $X + g(X)$.

5.2.1 Variance mixtures of normal distributions

We first consider spherically symmetric densities which are scale mixtures of normal distributions. Suppose

$$f(\|x-\theta\|^2) = \frac{1}{(2\pi)^{p/2}} \int_0^\infty \frac{1}{v^{p/2}} \exp\left\{-\frac{\|x-\theta\|^2}{2v}\right\} dG(v), \tag{5.1}$$

where $G(\cdot)$ is a probability distribution on $(0,\infty)$, i.e., a mixture of $N_p(\theta, vI)$ distributions with mixing distribution $G(\cdot)$.

Our first result gives a domination result for Baranchik type estimators for such distributions. This result is analogous to Theorem **??** in the normal case.

Theorem 5.1. *Let X have density of the form (5.1) and let*

$$\delta^B_{a,r}(X) = \left(1 - a\frac{r(\|X\|^2)}{\|X\|^2}\right)X,$$

where the function $r(\cdot)$ is absolutely continuous. Assume the expectations $E[V]$ and $E[V^{-1}]$ are finite where V has distribution G. Then $\delta^B_{a,r}(X)$ is minimax for loss $L(\theta,\delta) = \|\delta-\theta\|^2$ provided

(1) $0 \le a \le 2(p-2)/E[V^{-1}]$,

(2) $0 \le r(t) \le 1$ *for any* $t \ge 0$,

(3) $r(t)$ *is nondecreasing in* t *and*

(4) *$r(t)/t$ is nonincreasing in t.*

Furthermore, $\delta^B_{a,r}(X)$ dominates X provided the inequalities in (1) or (2) (on a set of positive measure) are strict or $r'(t)$ is strictly increasing on a set of positive measure.

Proof. The proof proceeds by calculating the conditional risk given $V = v$, noting that the distribution of $X|V = v$ is normal $N(\theta, vI_p)$. First note that $E[V] < \infty$ is equivalent to $E_0[\|X\|^2] < \infty$ so that the risk of X is finite. Similarly, it can be seen that $E[V^{-1}] < \infty$ if and only if $E_0[\|X\|^{-2}] < \infty$. Then, thanks to (2), we have $E_0[r^2(\|X\|^2)\|X\|^{-2}] < \infty$. Actually, we will see below that, for any θ, $E_\theta[\|X\|^{-2}] \leq E_0[\|X\|^{-2}]$, and hence, $E_\theta[r^2(\|X\|^2)\|X\|^{-2}] < \infty$ which guarantees that the risk of $\int^{\mathbb{R}}_{a,r}$ is finite. Note that, conditionally on V, $\|X\|^2/V$ has a noncentral chi-square distribution with p degrees of freedom and noncentrality parameter $\|\theta\|^2/V$. Hence, since the family of noncentral chi-square distributions have monotone (increasing) likelihood ratios (and therefore are stochastically increasing) in the noncentrality parameter.

Hence

$$E_\theta\left[\frac{1}{\|X\|^2/V}\right] \leq E_0\left[\frac{1}{\|X\|^2/V}\right]$$

and, as a result,

$$\begin{aligned}
E_\theta\left[\frac{1}{\|X\|^2}\right] &= E\left[E_\theta\left[\frac{1}{\|X\|^2|V}\right]\right] \\
&= E\left[\frac{1}{V}E_\theta\left[\frac{1}{\|X\|^2/V}|V\right]\right] \\
&\leqq E\left[\frac{1}{V}E_0\left[\frac{1}{\|X\|^2/V}\right]\right] \\
&= E_0\left[\frac{1}{\|X\|^2}\right].
\end{aligned}$$

We deal now with the main result of the theorem. Using Corollary 3.1 and Theorem 3.3, we have

$$\begin{aligned}
R(\theta,\delta^B_{a,r}) &= E\{E[\|\delta^B_{a,r}(X)-\theta\|^2\,|V]\}\\
&= E\left\{E\left[\|X-\theta\|^2+V^2\left(\frac{a^2r^2(\|X\|^2)}{V^2\|X\|^2}-\frac{2a(p-2)}{V}\frac{r(\|X\|^2)}{\|X\|^2}\right)\right.\right.\\
&\qquad\qquad \left.\left.-4\,aV\,r'(\|X\|^2)|V\right]\right\}\\
&\le R(\theta,X)+E\left\{aE\left[\frac{r(\|X\|^2)}{\|X\|^2/V}\middle|V\right]\left(\frac{a}{V}-2(p-2)\right)\right\}, \qquad (5.2)
\end{aligned}$$

since $r^2(\|X\|^2)\le r(\|X\|^2)$ and $r'(\|X\|^2)\ge 0$. Now, as a consequence of the above monotone likelihood property, $\|X\|^2/V$ is stochastically decreasing in V. It follows that the conditional expection in (5.2) is nondecreasing in V since, if $v_1<v_2$, we have

$$\begin{aligned}
E\left[\frac{r(\|X\|^2)}{\|X\|^2/V}\middle|V=v_1\right] &= E\left[\frac{r\left(v_1\frac{\|X\|^2}{V}\right)}{\|X\|^2/V}\middle|V=v_1\right]\\
&\le E\left[\frac{r\left(v_2\frac{\|X\|^2}{V}\right)}{\|X\|^2/V}\middle|V=v_1\right]\\
&\quad \text{(since } r(\|X\|^2) \text{ is nondecreasing)}\\
&\le E\left[\frac{r\left(v_2\frac{\|X\|^2}{V}\right)}{\|X\|^2/V}\middle|V=v_2\right]\\
&\quad \text{(since } r(t)/t \text{ is nonincreasing and}\\
&\quad \|X\|^2/V \text{ is stochastically decreasing in} V)\\
&= E\left[\frac{r(\|X\|^2)}{\|X\|^2/V}\middle|V=v_2\right].
\end{aligned}$$

Finally, using the fact that $aV^{-1}-2(p-2)$ is decreasing in V, and the fact that $E[g(Y)h(Y)]\le E[g(Y)]E[h(Y)]$ if g and h are monotone in opposite directions, it follows that

$$
\begin{aligned}
R(\theta, \delta^B_{a,r}) &\le R(\theta, X) + aE\left[\frac{V r(\|X\|^2)}{\|X\|^2}\right] E\left[\frac{a}{V} - 2(p-2)\right] \\
&\le R(\theta, X) \qquad (5.3)
\end{aligned}
$$

by assumption (a). Hence $\delta^B_{a,r}(X)$ is minimax, since X is minimax.

The dominance result follows since the inequality in (5.2) is strict if there is strict inequality in (2) or if $r'(\cdot)$ is strictly positive on a set of positive measure and the inequality in (5.3) is strict if the inequalities in (1) are strict. □

Example 5.1. Student-t distributions: If V has the an inverse Gamma $(v/2, v/2)$ distribution, then the distribution of X is a multivariate Student-t distribution with ν degrees of freedom. Since $E[V] = \{v/(v-2)\}^{v/2+1}$ provided $\nu > 2$, and $E[V^{-1}] = \{v/(v-2)\}^{v/2+1}$, the condition on the constant a of the theorem requires $0 \le a \le 2(p-2)\{(v+2)/v\}^{v/2+1}$ and $v > 2$. the condition of the theorem requires $0 \le a \le 2(p-2)$ and $v > 2$.

Example 5.2. Examples of Function $r(t)$: The James-Stein estimator has

$r(t) \equiv 1$ and hence satisfies conditions (2), (3) and (4) of Theorem 5.1. Also $r(t) = t/(t+b)$ satisfies these conditions. Similarly, the positive-part James-Stein estimator $\left(1 - a/X'X\right)_+ X$ is such that

$$
r(t) = \begin{cases} t/a & \text{for } 0 \le t \le a \\ 1 & \text{for } t \ge a \end{cases}
$$

and

$$
\frac{r(t)}{t} = \begin{cases} 1/a & \text{for } 0 \le t \le a \\ 1/t & \text{for } t \ge a \end{cases}
$$

and hence also satisfies the conditions (2), (3) and (4).

It is worth noting, and easy to see, that if the sampling distribution is $N(\theta, I_p)$ and the prior distribution is any scale mixture of normal distributions as in (**??**), in the Baranchik representation of the Bayes estimator (see Corollary 4.1), the function $r(t)/t$ is always nonincreasing. This fact leads to the following observation on the (sampling distribution) robustness of Bayes minimax estimators for a normal sampling distribution. If $\delta^{\pi}(X) = \big(1 - a\, r(\|X\|^2)/\|X\|^2\big)X$ is a Bayes minimax estimator with respect to a scale mixture of normal priors for a $N(\theta, I_p)$ sampling distribution, and if $r(t)$ is nondecreasing, this Bayes minimax estimator remains minimax for a multivariate-*t* sampling distribution in Example 5.1 as long as the degrees of freedom is greater than two.

It is also interesting to note that, in general, there will be no uniformly optimal choice of the shrinkage constant "*a*" in the James-Stein estimator if the mixing distribution $G(\cdot)$ is nondegenerate. The optimal choice will typically depend on $\|\theta\|^2$. This is in contrast to the normal sampling distribution case ($G(\cdot)$ is degenerate), where the optimal choice is $a = (p-2)\sigma^2$.

5.2.2 Densities with tails flatter than the normal

In this section we consider the sub class of spherically symmetric densities $f(\|x - \theta\|^2)$ such that, for any $t \geq 0$ for which $f(t) > 0$,

$$\frac{F(t)}{f(t)} \geq c > 0 \tag{5.4}$$

for some fixed positive c, where

$$F(t) = \frac{1}{2}\int_t^\infty f(u)du. \tag{5.5}$$

This class was introduced in Berger [1975] (without the constant 1/2 multiplier).

This class of densities contains a large subclass of variance mixtures of normal densities but also many others. The following lemma gives some conditions which guarantee inclusion or exclusion from the class satisfying (5.4) and (5.5).

Lemma 5.1. *Suppose X has density $f(\|x-\theta\|^2)$.*

(1) *(Mixture of normals). If, for some distribution G on $(0,\infty)$,*

$$f(\|x-\theta\|^2) = \left(\frac{1}{\sqrt{2\pi}}\right)^p \int_0^\infty v^{-p/2} \exp\left\{-\frac{\|x-\theta\|^2}{2v}\right\} dG(v)$$

where $E[V^{-p/2}]$ is finite, E denoting the expectation with respect to G, then $f(\cdot)$ is in the class (5.4) with $c = E[V^{-p/2+1}]/E[V^{-p/2}]$ for $p \geq 3$.

(2) *If $f(t) = h(t)e^{-at}$ with $h(t)$ nondecreasing, then $f(\cdot)$ is in the class (5.4).*

(3) *If $f(t) = e^{-atg(t)}$ where $g(t)$ is nondecreasing and $\lim_{t\to\infty} g(t) = \infty$, then $f(t)$ is not in the class (5.4).*

Proof. (1) We have

$$\begin{aligned} F(t) &= \frac{1}{2}\int_t^\infty f(u)du \\ &= \frac{1}{2(\sqrt{2\pi})^p}\int_t^\infty \int_0^\infty v^{-p/2}\exp\{-u/2v\}\, dG(v)du \\ &= \frac{1}{(\sqrt{2\pi})^p}\int_0^\infty v^{-p/2+1}\exp\{-t/2v\}\, dG(v). \end{aligned}$$

Hence

$$\begin{aligned}\frac{F(t)}{f(t)} &= \frac{\int_0^\infty v^{-p/2+1}\exp\{-t/2v\}\,dG(v)}{\int_0^\infty v^{-p/2}\exp\{-t/2v\}\,dG(v)}\\ &\geq \frac{\int_0^\infty v^{-p/2+1}dG(v)}{\int_0^\infty v^{-p/2}dG(v)}\\ &= \frac{E[V^{-p/2+1}]}{E[V^{-p/2}]}.\end{aligned} \tag{5.6}$$

The inequality follows since the family of densities proportional to $v^{-p/2}\exp\{-t/2v\}$ have monotone (increasing) likelihood ratio in the parameter t. Note that if $p \geq 3$, $E[V^{-p/2}] < \infty$ implies $E[V^{-p/2+1}] < \infty$. This completes the proof of (1).

(2) In this case

$$\begin{aligned}\frac{F(t)}{f(t)} &= \frac{\frac{1}{2}\int_t^\infty h(u)e^{-au}du}{h(t)e^{-at}}\\ &\geq \frac{1}{2}\int_t^\infty e^{-a(u-t)}du\\ &= \frac{1}{2a}.\end{aligned}$$

Hence (5.4) is satisfied with $c = 1/2a$, which proves (2).

(3) In this case

$$\begin{aligned}\lim_{t\to\infty}\frac{F(t)}{f(t)} &= \lim_{t\to\infty}\frac{\int_t^\infty \exp\{-aug(u)\}\,du}{\exp\{-atg(t)\}}\\ &= \lim_{t\to\infty}\int_t^\infty \exp\{-aug(u)+atg(t)\}\,du\\ &= \lim_{t\to\infty}\int_0^\infty \exp\{-a(u+t)g(u+t)+atg(t)\}\,du\\ &\leq \lim_{t\to\infty}\int_0^\infty \exp\{-aug(t)\}\,du\\ &= \lim_{t\to\infty}\frac{1}{ag(t)}\\ &= 0.\end{aligned}$$

Hence $f(t)$ is not in the class (5.4). This shows (c). □

Part (2) of the lemma shows that densities with tails flatter than the normal (and including the normal) are in the class (5.4), while densities with tails "sufficiently lighter" than the normal are not included. Also the condition in part (2) is stronger than necessary in that it suffices that condition hold only for all t larger than some positive K. See Berger [1975] for the details.

Example 5.3. Some specific examples in the class (5.4) include (see Berger [1975] for more details)

$$(1) \qquad f(t) = K/\cosh t \qquad (c \approx 1/2)$$

$$(2) \qquad f(t) = Kt(1+t^2)^{-m} \text{ with } m > p/4 \qquad (c = m/2)$$

$$(3) \qquad f(t) = Ke^{-\alpha t-\beta}/(1+e^{-\alpha t-\beta})^2 \qquad (c = \alpha/2)$$

$$(4) \qquad f(t) = Kt^n e^{-t/2} \text{ for } n \geq 0 \qquad (c = 1).$$

The latter two distributions are known as the logistic type and Kotz, respectively.

The following lemma plays the role of Stein's lemma (Theorem **??**) for this family of distributions.

Lemma 5.2. *Let X have density $f(\|x-\theta\|^2)$ and let $g(X)$ be a weakly differentiable function such that $E_\theta[|(X-\theta)'g(X)|] < \infty$. Then*

$$E_\theta[(X-\theta)'g(X)] = E_\theta\left[\frac{F(\|X-\theta\|^2)}{f(\|X-\theta\|^2)}\operatorname{div} g(X)\right]$$

where $F(t)$ is defined as in (5.5).

Proof. Note that the existence of the expectation in Lemma 5.2 will be guaranteed for function $g(x)$ such that $E_\theta[\|y(x)\|^2] < \infty$ as soon as $E_0[\|X\|^2] < \infty$. The proof will follow that of Theorem **??** up to the application of the divergence (Stokes') theorem, (see Appendix ?).

$$\begin{aligned}
E[(X-\theta)'g(X)] &= \int_{R^p} (x-\theta)'g(x)f(\|x-\theta\|^2)\,dx \\
&= \int_0^\infty \int_{S_{R,\theta}} (x-\theta)'g(x)\,f(\|x-\theta\|^2)\,d\sigma_{R,\theta}(x)\,dR \quad (1.4) \\
&= \int_0^\infty \int_{S_{R,\theta}} \left(\frac{x-\theta}{\|x-\theta\|}\right)' d\sigma_{R,\theta}(x)\,R\,f(R^2)\,dR \\
&= \int_0^\infty \int_{B_{R,\theta}} \operatorname{div} g(x)\,dxRf(R^2)\,dR \qquad (\text{Stokes}'\text{ theorem}) \\
&= \int_{R^p} \operatorname{div} g(x) \int_{\|x-\theta\|}^\infty Rf(R^2)\,dR\,dx \qquad (\text{Fubini}'\text{s theorem}) \\
&= \int_{R^p} \operatorname{div} g(x) \frac{1}{2} \int_{\|x-\theta\|^2}^\infty f(t)\,dt\,dx \\
&= \int_{R^p} \operatorname{div} g(x)\,F(\|x-\theta\|^2)\,dx \\
&= E_\theta\left[\operatorname{div} g(x) \frac{F(\|x-\theta\|^2)}{f(\|x-\theta\|^2)}\right].
\end{aligned}$$

□

Note that the following result gives conditions for minimaxity of estimators of the Baranchik type.

Theorem 5.2. *Let X have density $f(\|x-\theta\|^2)$ which satisfies (5.4) for some $0 < c < \infty$. Assume also that $E_0[\|X\|^2] < \infty$ and $E_0[\|X\|^{-2}] < \infty$. Let*

$$\delta^B_{a,r}(X) = \left(1 - \frac{a\,r(\|X\|^2)}{\|X\|^2}\right)X$$

where $r(\cdot)$ is absolutely continuous. Then $\delta^B_{a,r}(X)$ is minimax for $p \geq 3$ provided

(1) $0 < a \leq 2c\,(p-2)$,

(2) $0 \leq r(t) \leq 1$,

(3) *$r(t)$ is nondecreasing.*

Furthermore it dominates X provided both inequalities are strict in (1) or in (2) on a set of positive measure or if $r'(\cdot)$ is strictly positive on a set of positive measure.

Proof. We note that the conditions ensure finiteness of the risk so that Lemma 5.2 is applicable. Hence we have

$$\begin{aligned} R(\theta,\delta^B_{a,r}) &= E\left[\|X-\theta\|^2 + \frac{a^2r^2(\|X\|^2)}{\|X\|^2} - 2\frac{a,r(\|X\|^2)X'(X-\theta)}{\|X\|^2}\right] \\ &= R(\theta,X) + aE\left[\frac{ar^2(\|X\|^2)}{\|X\|^2} - \operatorname{div}\left(\frac{r(\|X\|^2)X}{\|X\|^2}\right) 2\frac{F(\|X-\theta\|^2)}{f(\|X-\theta\|^2)}\right] \end{aligned}$$

by Lemma 5.2. Therefore the risk difference between $\delta^B_{a,r}(X)$ and X equals

$$\begin{aligned} \Delta_\theta &= aE\left[\frac{ar^2(\|X\|^2)}{\|X\|^2} - \left(\frac{2(p-2)r(\|X\|^2)}{\|X\|^2} + 4r'(\|X\|^2)\right)\frac{F(\|X-\theta\|^2)}{f(\|X-\theta\|^2)}\right] \\ &\le aE\left[\frac{r(\|X\|^2)}{\|X\|^2}\left(a - 2(p-2)\frac{F(\|X-\theta\|^2)}{f(\|X-\theta\|^2)}\right)\right] \\ &\le aE\left[\frac{r(\|X\|^2)}{\|X\|^2}\big(a-2(p-2)c\big)\right] \\ &\le 0. \end{aligned}$$

The domination part follows as in Theorem 5.1. □

Theorem 5.2 applies for certain densities for which Theorem 5.1 is not applicable and additionally lifts the restriction that $r(t)/t$ is nonincreasing. However, if the density is a mixture of normals, and both theorems apply, the shrinkage constant "a" given by Theorem 5.1 ($a = 2(p-2)/E[V^{-1}]$) is strictly larger than that for Theorem 5.2 ($a = 2(p-2)c$) whenever the mixing distribution $G(\cdot)$ is not degenerate. To see this note that

$$\frac{1}{E[V^{-1}]} > c = \frac{E[V^{-p/2+1}]}{E[V^{-p/2}]}$$

or equivalently

$$E[V^{-p/2}] > E[V^{-1}]E[V^{-p/2+1}]$$

whenever the positive random variable V is non-degenerate. Note also that $E[V^{-1}] < \infty$ whenever $E[V^{-p/2}] < \infty$ and $p \geq 3$.

Example 5.4. Student-t: Suppose X has a p-variate Student-t distribution with ν degrees of freedom as in Example 5.1, so that V has the distribution of a $yg(\nu/1, \nu/2)$. In this case

$$E[V^{-p/2}] = \frac{2^{p/2}\Gamma\left(\frac{p+\nu}{2}\right)}{\nu^{p/2}\Gamma\left(\frac{\nu}{2}\right)}$$

which is finite for all $\nu > 0$ and $p > 0$.

The bound on the shrinkage constant, “a”, in Theorem 5.1 is $2(p-2)$ as shown in Example 5.1, while the bound on “a”, in Theorem 5.2, as indicated above, is given by

$$2(p-2)\frac{E[V^{-p/2+1}]}{E[V^{-p/2}]} = 2(p-2)\left(\frac{\nu}{\nu+p-2}\right) < 2(p-2).$$

Hence, for large p, the bound on the shrinkage factor “a” can be substantially less for Theorem 5.2 than for Theorem 5.1 in the case of a multivariate-t sampling distribution. Note that, for fixed p, as ν tends to infinity the smaller bound tends to the larger one (and the t-distribution tends to the normal).

Example 5.5. Example 5.3 (Continued) All of the distributions in Example 5.3 satisfy the assumptions of Theorem 5.2 (under suitable moment conditions for the second density). It is interesting to note that for the Kotz distribution, the value of c ($= 1$), as in (5.4), doesn’t depend on the parameter $n > 0$. Hence the bound on the shrinkage factor “a” is $2(p-2)$ and is also independent of n, indicating a certain distributional robustness of the minimaxity property of Baranchik type estimators with $a < 2(p-2)$.

With additional assumptions on the function $F(t)/f(t)$ in (5.4) (i.e. it is either monotone increasing or monotone decreasing), theorems analogous to Theorem 5.2 can be developed which further improve the bounds on the shrinkage factor "a". These typically may involve additional assumptions on the function $r(\cdot)$. We will see examples of this type in the next section.

5.3 More general minimax estimators

In this section, we consider minimaxity of general estimators of the form $X + a\,g(X)$. The initial results rely on Lemma 5.2. The first result follows immediately from this lemma and gives an expression for the risk.

Corollary 5.1. *Let X have a density $f(\|x-\theta\|^2)$ such that $E_0[\|X\|^2] < \infty$ and let $g(X)$ be weakly differentiable and be such that $E_\theta[\|g(X)\|^2] < \infty$. Then, for loss $L(\theta,\delta) = \|\delta-\theta\|^2$, the risk of $X + a\,g(X)$ can be expressed as*

$$R(\theta, X + a\,g(X)) = R(\theta, X) + E_\theta\left[a^2\,\|g(X)\|^2 + 2\,a\,Q(\|X-\theta\|^2)\,\operatorname{div} g(X)\right] \tag{5.7}$$

where

$$Q(\|X-\theta\|^2) = \frac{F(\|X-\theta\|^2)}{f(\|X-\theta\|^2)} \tag{5.8}$$

and where $F(\|X-\theta\|^2)$ is defined in (5.5).

An immediate consequence of Corollary 5.1 when the density of f satisfies (5.4), i.e. $Q(t) \geq c > 0$ for some constant c, is the following.

Corollary 5.2. *Under the assumptions of Corollary 5.1, assume that, for some* $c > 0$*, we have* $Q(t) \geq c$ *for any* $t \geq 0$*. Then* $X + g(X)$ *is minimax and dominates* X *provided, for any* $x \in R^p$*,*

$$\|g(x)\|^2 + 2c \operatorname{div} g(x) \leq 0$$

with strict inequality on a set of positive measure.

In the following two theorems, we establish minimaxity results under the assumption that $Q(t)$ is monotone.

Theorem 5.3. *Suppose* X *has density* $f(\|x-\theta\|^2)$ *such that* $E_0[\|X\|^2] < \infty$ *and that* $Q(t)$ *in (5.8) is nonincreasing. Suppose there exists a nonpositive function* $h(U)$ *such that* $E_{R,\theta}[h(U)]$ *is nondecreasing where* $U \sim U_{R,\theta}$ *(the uniform distribution on the sphere of radius* R *centered at* θ*) and such that* $E_\theta[-h(x)] < \infty$*. Furthermore suppose that* $g(X)$ *is weakly differentiable and also satisfies*

(1) $\operatorname{div} g(X) \leq h(X)$*,*

(2) $\|g(X)\|^2 + 2h(X) \leq 0$

and

(3) $0 \leq a \leq E_0(\|X\|^2)/p$*.*

Then $\delta(X) = X + ag(X)$ *is minimax. Also* $\delta(X)$ *dominates* X *provided* $g(\cdot)$ *is nonzero with positive probability and strict inequality holds with positive probability in* (1) *or* (2)*, or both inequalities are strict in* (3)*.*

Proof. Note that $g(x)$ satisfies the conditions of Corollary 5.1. Then we have

$$\begin{aligned} R(\theta,\delta) &= R(\theta,X) + aE[a\|g(X)\|^2 + 2Q(\|X-\theta\|^2)\operatorname{div} g(X)] \\ &= R(\theta,X) + aE[E_{R,\theta}[a\|g(X)\|^2 + 2Q(R^2)\operatorname{div} g(X)]] \end{aligned}$$

where $E_{R,\theta}$ is as above and E denotes the expectation with respect to the radial distribution. Now, using (1) and (2), we have

$$\begin{aligned} R(\theta,\delta) &\leq R(\theta,X) + aE[E_{R,\theta}[-2ah(X) + 2Q(R^2)h(X)]] \\ &= R(\theta,X) + 2aE[(a - Q(R^2))E_{R,\theta}[-h(X)]] \\ &\leq R(\theta,X) + 2aE[a - Q(R^2)]E_\theta[-h(X)] \end{aligned}$$

by the monotonicity assumptions on $E_{R,\theta}[h(\cdot)]$ and $Q(t)$ as well as the covariance inequality.

Hence, since $-h(X) \geq 0$, we have $R(\theta,\delta) \leq R(\theta,X)$, provided $0 \leq a \leq E[Q(R^2)]$. Now $E[Q(R^2)] = E_0[\|X\|^2]/p$ by Lemma 5.3 below, hence δ is minimax. The domination result follows since the additional conditions imply strict inequality between the risks. □

Lemma 5.3. *For any $k > -p$ such that $E[R^{k+2}] < \infty$, we have*

$$E[R^k Q(R^2)] = \frac{1}{p+k} E[R^{k+2}].$$

In particular, we have

$$E[Q(R^2)] = \frac{1}{p} E[R^2] = \frac{1}{p} E_0[\|X\|^2]$$

and, for $p \geq 3$,

$$E\left[\frac{Q(R^2)}{R^2}\right] = \frac{1}{p-2}.$$

Proof. Recall that the radial density $\varphi(r)$ of $R = \|X - \theta\|$ can be expressed as $\varphi(r) = \sigma(S) r^{p-1} f(r^2)$ where $\sigma(S)$ is the area of the unit sphere S in R^p. By (5.8) and (5.5), we have

$$\begin{aligned}
E[R^k Q(R^2)] &= \frac{1}{2}\int_{R^p} \|x\|^k \int_{\|x\|^2}^{\infty} f(t)\,dt\,dx \\
&= \frac{1}{2}\int_0^{\infty}\int_{B_{\sqrt{t}}} \|x\|^k dx\, f(t)\,dt \quad \text{by Fubini's theorem} \\
&= \frac{1}{2}\int_0^{\infty}\int_0^{\sqrt{t}} \sigma(S)\, r^{k+p-1} dr\, f(t)\,dt \quad \text{by Lemma 1.4} \\
&= \frac{1}{2}\int_0^{\infty} \sigma(S)\,\frac{t^{(k+p)/2}}{k+p}\, f(t)\,dt \\
&= \frac{1}{k+p}\int_0^{\infty} r^{k+2}\varphi(r)\,dr \quad \text{by the change of variable} t = r^2 \\
&= \frac{1}{k+p} E[R^{k+2}].
\end{aligned}$$

Note that positivity of integrands and $E[R^{k+2}] < \infty$ implies $E[R^k Q(R^2)] < \infty$. □

The next theorem reverses the monotonicity assumption on $Q(\cdot)$ and changes the condition on the function $h(X)$ which, in turn, bounds the divergence of $g(X)$.

Theorem 5.4. *Suppose X has a density $f(\|x-\theta\|^2)$ such that $E_0[\|X\|^2] < \infty$ and $E_0[1/\|X\|^2] < \infty$ and such that $Q(t)$ in (5.8) is nondecreasing. Suppose there exists a nonpositive function $h(X)$ such that $E_{R,\theta}\left[R^2 h(U)\right]$ is nonincreasing where $U \sim U_{R,\theta}$ and such that $E_\theta[-h(X)] < \infty$.*

Furthermore suppose that $g(X)$ is weakly differentiable and also satisfies

(1) $\operatorname{div} g(X) \le h(X)$,

(2) $\|g(X)\|^2 + 2\,h(X) \le 0$,

and

(3) $0 \le a \le \frac{1}{(p-2)E_0(1/\|X\|^2)}$.

Then $\delta(X) = X + a\,g(X)$ *is minimax. Also* $\delta(X)$ *dominates* X *provided* $g(\cdot)$ *is nonzero with positive probability and strict inequality holds with positive probability in* (1) *or* (2)*, or both inequalities are strict in* (3)*.*

Proof. As in the proof of Theorem 5.3, we have

$$\begin{aligned} R(\theta,\delta) &\le R(\theta,X) + 2\,a\,E[(a - Q(R^2))\,E_{R,\theta}[-h(X)]] \\ &= R(\theta,X) + 2\,a\,E\left[\left(\frac{a}{R^2} - \frac{Q(R^2)}{R^2}\right) E_{R,\theta}[-R^2\,h(X)]\right] \\ &\le R(\theta,X) + 2\,a\,E\left[\frac{a}{R^2} - \frac{Q(R^2)}{R^2}\right] E_{R_0,\theta}[-R_0^2\,h(X)] \end{aligned}$$

where R_0 is a point such that $a - Q(R_0^2) = 0$, provided such a point exists. Here we have used the version of the covariance inequality that states

$$Ef(X)g(X) \le Ef(X)g(X_0)$$

provided that $g(X)$ is nondecreasing (respectively, nonincreasing) and $f(X)$ changes sign once from $+$ to $-$ (respectively, $-$ to $+$) at X_0. But such a point R_0 does exist provided

$$E\left[\frac{a}{R^2} - \frac{Q(R^2)}{R^2}\right] \le 0$$

since $Q(R^2)$ is nondecreasing.

It follows that $R(\theta,\delta) \le R(\theta,X)$ provided that $aE[\frac{1}{R^2}] \le E[\frac{Q(R^2)}{R^2}]$. However $E[\frac{Q(R^2)}{R^2}] = \frac{1}{p-2}$ by Lemma 5.3 and hence the result follows as in Theorem 5.3. □

Note that the bound on "a" in both of these theorems is strictly larger than the bound in Theorem 5.2 provided $Q(R^2)$ is not constant. This is so since the bound in Theorem 5.2 is based on $c = \inf Q(R^2)$ while, in these results, the bound is equal to a (possibly weighted) average of $Q(R^2)$.

We indicate the utility of these two results by applying them to the James-Stein estimator.

Corollary 5.3. *Let $X \sim f(\|x-\theta\|^2)$ for $p \geq 4$ and let $\delta_b^{JS}(X) = (1 - b/\|X\|^2)X$. Assume also that $E_0[\|X\|^2] < \infty$ and $E_0[1/\|X\|^2] < \infty$. Then $\delta_b^{JS}(X)$ is minimax and dominates X provided either*

(1) *$Q(R^2)$ is nonincreasing and*

$$0 < b < 2(p-2)\frac{E_0\|X\|^2}{p},$$

or

(2) *$Q(R^2)$ is nondecreasing and*

$$0 < b < \frac{2}{E_0(1/\|X\|^2)}.$$

Proof. We apply Theorems 5.3 and 5.5 with $g(X) = -[2\,(p-2)/\|X\|^2]X$, $\operatorname{div} g(X) = -2\,(p-2)^2/\|X\|^2 = h(X)$. It follows from Theorem 00 in the Appendix **[see file 1]** that when $p \geq 4$, $E_{\theta,R}[h(U)]$ is nondecreasing in R and $E_{\theta,R}[R^2 h(U)]$ is nonincreasing in R. Hence, if $Q(R^2)$ is nonincreasing, Theorem 5.3 implies that

$$\delta_a(X) = X - \frac{2\,(p-2)\,a}{\|X\|^2}X = \delta^{JS}_{2\,(p-2)\,a}(X)$$

is minimax and dominates X provided $0 < a < E_0[\|X\|^2]/p$ or equivalently $0 < 2\,(p-2)\,a < 2\,(p-2)\,E_0(\|X\|^2)/p$ which is (1) with $b = 2\,(p-2)\,a$. Similarly, applying Theorem 5.5 when $Q(R^2)$ is nondecreasing, we find that $\delta_a(X)$ is minimax and dominates X if

$$0 < a < \frac{1}{(p-2)E_0(1/\|X\|^2)}$$

which is (2). □

Example 5.6 (Densities with increasing and decreasing $Q(R^2)$). Note first that variance mixtures of normal distributions have increasing $Q(R^2)$ since, by (5.6) and (5.8), $Q(R^2)$ may be viewed as the expected value of V with respect to a family of distributions with monotone increasing likelihood ratio in $t = R^2$. Note also that the bound for the shrinkage constant "a" in a James-Stein estimator is the same in Corollary 5.3 as it is in Theorem 5.1 for mixtures of normals.

We note that, if we consider $f(t)$ to be proportional to a density of a positive random variable, then $2\,Q(t)$ is the reciprocal of the hazard rate. There is a large literature on increasing and decreasing hazard rates (see, for example, Barlow and Proschan [1981]).

We note that the monotonicity of $Q(t)$ may be determined in many cases by studying the log-convexity or the log-concavity of $f(t)$. In particular, if $\ln f(t)$ is convex (concave), then $Q(t)$ is nondecreasing (nonincreasing). To see this, note that

$$Q(t) = \frac{1}{2}\frac{\int_t^\infty f(u)\,du}{f(t)} = \frac{1}{2}\int_0^\infty \frac{f(s+t)}{f(t)}\,ds$$

and hence $Q(t)$ will be nondecreasing (nonincreasing) if $\frac{f(s+t)}{f(t)}$ is nondecreasing (nonincreasing) in t for each $s > 0$. But, assuming for simplicity that f is differentiable, for any $t \geq 0$ such that $f(t) > 0$,

$$\begin{aligned} \frac{d}{dt}\left(\frac{f(s+t)}{f(t)}\right) &= \frac{f(t)f'(s+t) - f(s+t)f'(t)}{f^2(t)} \\ &= \frac{f(s+t)}{f(t)}\left[\frac{f'(s+t)}{f(s+t)} - \frac{f'(t)}{f(t)}\right] \\ &= \frac{f(s+t)}{f(t)}\left[\frac{d}{dt}\ln f(s+t) - d/dt \ln f(t)\right]. \end{aligned}$$

This is positive or negative respectively when $\ln f(s+t)$ is convex or concave in t. For example if X has a Kotz distribution with parameter n, $f(t) \propto t^n e^{-t/2}$. Then $\ln f(t) = K + n \ln t - \frac{t}{2}$ which is concave if $n \geq 0$ and convex if $n \leq 0$. Hence $Q(t)$ is decreasing if $n > 0$ and increasing if $n < 0$. Of course the log-convexity (log-concavity) of $f(t)$ is not a necessary condition for the nondecreasing (nonincreasing) monotonicity of $Q(t)$. Thus, it is easy to check that $f(t) \alpha \exp(-t^2) \exp[-1/2 \int_0^t \exp(-u^2)\, du]$ leads to $Q(t) = \exp(t^2)$, but is not log [this ran off the page here].

An important class of distributions is covered by the following corollary.

Corollary 5.4. *Let* $X \sim f(\|x-\theta\|^2)$ *for* $p \geq 4$ *with* $f(t) \propto \exp(-\beta t^\alpha)$ *where* $\alpha > 0$ *and* $\beta > 0$. *Then* $\delta_b^{JS}(X) = (1 - b/\|X\|^2)X$ *is minimax and dominates* X *provided either*

(1) $\alpha \leq 1$ *and* $0 < b < \frac{2}{\beta^{1/\alpha}} \frac{p-2}{p} \frac{\Gamma((p+2)/2\alpha)}{\Gamma(p/2\alpha)}$ *or*

(1) $\alpha \leq 1$ *and* $0 < b < \frac{2}{\beta^{1/\alpha}} \frac{\Gamma(p/2\alpha)}{\Gamma((p-2)/2\alpha)}$.

Proof. By the above discussion, $Q(R^2)$ is nonincreasing (nondecreasing) for $\alpha \geq 1$ ($\alpha \leq 1$). Then the result follows from Corollary 5.3 and the fact that

$$E_0[\|X\|^k] = \frac{1}{\beta^{k/2\alpha}} \frac{\Gamma(\frac{p+k}{2\alpha})}{\Gamma(\frac{p}{2\alpha})}$$

for $k > -p$. □

The final theorem of this section gives conditions for minimaxity of estimators of the form $X + a\,g(X)$ for general spherically symmetric distributions. Note that no density is needed for this result which relies on the radial distribution defined in Theorem 1.1.

We first need the following lemma which will play the role of the Stein lemma in the theorem.

Lemma 5.4. *Let X have a spherically symmetric distribution around θ, and let $g(X)$ be a weakly differentiable function such that $E_\theta[\,|(X-\theta)'g(X)|\,] < \infty$. Then*

$$E_\theta[(X-\theta)'g(X)] = \frac{1}{p} E\left[R^2 \int_{B_{R,\theta}} \operatorname{div} g(X)\, dV_{R,\theta}(X)\right]$$

where E denotes the expectation with respect to the radial distribution and where $V_{R,\theta}(\cdot)$ is the uniform distribution on $B_{R,\theta}$, the ball of radius R centered at θ.

Proof. Denoting by ρ the radial distribution and according to Theorem 1.1, we have

$$\begin{aligned}
E[(X-\theta)'g(X)] &= \int_{\mathbb{R}_+}\int_{S_{R,\theta}} (x-\theta)'g(x)\,d\mathscr{U}_{R,\theta}(x)\,d\rho(R)\\
&= \int_{\mathbb{R}_+} \frac{1}{\sigma_{R,\theta}(S_{R,\theta})}\int_{S_{R,\theta}} \frac{(x-\theta)'}{\|x-\theta\|}\,g(x)\,d\sigma_{R,\theta}(x)\,R\,d\rho(R)\\
&= \int_{\mathbb{R}_+} \frac{1}{\sigma_{R,\theta}(S_{R,\theta})}\int_{B_{R,\theta}} \operatorname{div} g(x)\,dx\,R\,d\rho(R) \quad \text{by Stokes' theorem}\\
&= \frac{1}{p}\int_{\mathbb{R}_+}\int_{B_{R,\theta}} \operatorname{div} g(x)\,dV_{R,\theta}(x)\,R^2 d\rho(R)
\end{aligned}$$

since the volume of $B_{R,\theta}$ equals

$$\lambda(B_{R,\theta}) = \frac{R}{p}\sigma_{R,\theta}(S_{R,\theta}).$$

□

Theorem 5.5. *Let X have a spherically symmetric distribution around θ, and suppose $E_0[\|X\|^2] < \infty$ and $E_0[1/\|X\|^2] < \infty$. Suppose there exists a nonpositive function $h(\cdot)$ such that $h(X)$ is subharmonic and $E_{R,\theta}[R^2\,h(U)]$ is nonincreasing where $U \sim \mathscr{U}_{R,\theta}$ and such that $E_\theta[-h(x)] < \infty$. Furthermore suppose that $g(X)$ is weakly differentiable and also satisfies*

(1) $\operatorname{div} g(X) \le h(X)$,

(2) $\|g(X)\|^2 + 2\,h(X) \le 0$

and

(3) $0 \le a \le \frac{1}{pE_0(1/\|X\|^2)}$.

Then $\delta(X) = X + a\,g(X)$ is minimax. Also $\delta(X)$ dominates X provided $g(\cdot)$ is nonzero with positive probability and strict inequality holds with positive probability in (1) *or* (2)*, or both inequalities are strict in* (3)*.*

Proof. Using Lemma 5.4 and Conditions (1) and (2), we have

$$
\begin{aligned}
R(\theta,\delta) &= R(\theta,X)+aE_\theta\left[a\|g(X)\|^2+2(X-\theta)'g(X)\right]\\
&\le R(\theta,X)+2aE_\theta\left[-ah(X)+(X-\theta)'g(X)\right]\\
&= R(\theta,X)+2a\left\{E_\theta\left[-ah(X)\right]+\frac{1}{p}E\left[R^2\int_{B_{R,\theta}}\operatorname{div} g(X)\,dV_{R,\theta}(X)\right]\right\}\\
&\le R(\theta,X)+2a\left\{E_\theta\left[-ah(X)\right]+\frac{1}{p}E\left[R^2\int_{B_{R,\theta}}h(X)\,dV_{R,\theta}(X)\right]\right\}.
\end{aligned}
$$

By subharmonicity of h (see Lemma 000 in the Appendix),

$$
\int_{B_{R,\theta}} h(X)dV_{R,\theta}(X)\le\int_{S_{R,\theta}} h(X)dU_{R,\theta}(X).
$$

Hence,

$$
\begin{aligned}
R(\theta,\delta) &\le R(\theta,X)+2a\left\{E_\theta\left[-ah(X)\right]+\frac{1}{p}E\left[R^2\int_{S_{R,\theta}}h(X)\,dU_{R,\theta}(X)\right]\right\}\\
&= R(\theta,X)+2aE\left[\left(\frac{a}{R^2}-\frac{1}{p}\right)\cdot\left(-R^2\int_{S_{R,\theta}}h(X)dU_{R,\theta}(X)\right)\right]\\
&= R(\theta,X)+2aE\left[\left(\frac{a}{R^2}-\frac{1}{p}\right)\left(-E_{R,\theta}[R^2h(X)]\right)\right]\\
&\le R(\theta,X)+2aE\left[\left(\frac{a}{R^2}-\frac{1}{p}\right)\right]E\left[-E_{R,\theta}[R^2h(X)]\right]\\
&\qquad(\text{by monotonicity of} E_{R,\theta}[R^2h(X)]\text{and the covariance inequality}).
\end{aligned}
$$

Hence $R(\theta,\delta)\le R(\theta,X)$ when $E\left[a/R^2-1/p\right]\le 0$ which is equivalent to (3). The domination part follows as before. □

We note that the shrinkage constant in the above result $1/\{pE_0[1/\|X\|^2]\}$ is somewhat smaller than the constant in Theorem 5.4 ($a=1/\{(p-2)E_0[1/\|X\|^2]\}$), but Theorem 5.5 has essentially no restrictions on the distribution of X aside from moment conditions (which coincide in Theorems 5.4 and 5.5). In particular we do not even assume that a density exists! However there is an additional assumption of subharmonicity of h.

A useful corollary gives minimaxity for James-Stein estimators in dimension $p \geq 4$ for all spherically symmetric distributions with finite $E_0[\|X\|^2]$ and $E_0[1/\|X\|^2]$.

Corollary 5.5. *Let X have a spherically symmetric distribution with $p \geq 4$, and let $E_0[\|X\|^2] < \infty$ and $E_0[1/\|X\|^2] < \infty$. Then*

$$\delta_a^{JS}(X) = \left(1 - \frac{a}{\|X\|^2}\right) X$$

is minimax and dominates X provided

$$0 < a < \frac{1}{pE_0(1/\|X\|^2)}.$$

Proof. Here $g(X) = -X/\|X\|^2$ and is weakly differentiable for $p \geq 3$. Then div $g(X) = -(p-2)/\|X\|^2$ and $\|g(X)\|^2 = 1/\|X\|^2$ so that Conditions (1) and (2) of Theorem 5.5 are satisfied with $h(X) = -\alpha/\|X\|^2$ where $1/2 \leq \alpha \leq p-2$. Now the subharmonicity of $h(X)$ and its monotonicity condition hold since it is shown in the appendix that, for $p \geq 4, 1/\|X\|^2$ is super-harmonic (so that $E_{R,\theta}[1/\|X\|^2]$ is nonincreasing in R) and that $R^2 E_{R,\theta}[1/\|U\|^2]$ is nondecreasing in R.

Furthermore, it is worth noting that $E_{R,\theta}[1/\|U\|^2]$ is nonincreasing in $\|\theta\|$. Hence, for any $\theta \in \mathbb{R}^p$, we have $E_\theta[-h(X)] < \infty$ since

$$E_{R,\theta}[1/\|X\|^2] \leq E_{R,\theta}[1/\|X^2\|$$

so that

$$E_\theta[1/\|X\|^2] \leq E_\theta[1/\|X\|^2] < \infty,$$

by assumption. □

Example 5.7. Nonspherical minimax estimators: In Section **??**, we considered estimators which shrink toward a subspace. Theorem 5.5 allows us to show that estimators of this type are minimax for general spherically symmetric distributions. To be specific, suppose V is a $s < p$ dimensional subspace and let

$$\delta_a(X) = P_V X + \left(1 - \frac{a}{\|X - P_V X\|^2}\right)(X - P_V X).$$

As in the proof of Theorem 3.6, it can be shown that the risk of $\delta_a(X)$ equals

$$R(\theta, \delta_a(X)) = E_{\nu_1}[\|Y_1 - \nu_1\|^2] + E_{\nu_2}\left[\left\|\left(1 - \frac{a}{\|Y_2\|^2}\right)Y_2 - \nu_2\right\|^2\right], \tag{5.9}$$

where Y_1, Y_2, ν_1 and ν_2 are as in Theorem 3.6.

In the present case, Y_2 has a spherically symmetric distribution about ν_2 of dimension $p - 5$. Hence, by Theorem 5.5,

$$\begin{aligned} E(\theta, \delta_a(X)) &\le E_{\nu_1}[\|Y_1 - V_1\|^2] + E_{\nu_2}[\|Y_2 - V_2\|^2] \\ &= E_\theta \|X - \theta\|^2 \\ &= R(\theta, X), \end{aligned}$$

provided $p - s \ge 4$ and

$$0 < a < \frac{1}{(p-s)\, E_0[1/\|X - P_V X\|^2]}.$$

5.4 Bayes estimators

In this section, we consider (generalized) Bayes estimators of the location vector $\theta \in \mathbb{R}^p$ of a spherically symmetric distribution. More specifically let X be a random vector in $\mathbb{R}^p$ with density $f(\|x-\theta\|^2)$ and let $\pi(\theta)$ be a prior density. Under quadratic loss $\|\delta-\theta\|^2$, the (generalized) Bayes estimator of θ is the posterior mean given by

$$\delta_\pi(X) = X + \frac{1}{m(X)} \int_{\mathbb{R}^p} (\theta - X)\, f(\|X-\theta\|^2)\, \pi(\theta)\, d\theta \tag{5.10}$$

where $m(x)$ is the marginal

$$m(x) = \int_{\mathbb{R}^p} f(\|x-\theta\|^2) \pi(\theta)\, d\theta. \tag{5.11}$$

Recall from Subsection 4.1.1 that, in the normal case (that is, $f(t) \propto \exp(-t/2\sigma^2)$ with σ^2 known) the superharmonicity of $\sqrt{m(x)}$ is a sufficient condition for minimaxity of $\delta_\pi(X)$. This superharmonicity is implied by that of $m(x)$ and in turn by that of $\pi(\theta)$. While in the nonnormal case minimaxity has been studied by many authors (for example, see Strawderman [1974], Berger [1975], Brandwein and Strawderman [1978], Brandwein and Strawderman [1991] relatively few results on minimaxity of Bayes estimators are known. The primary technique to establish minimaxity is through a Baranchik representation of the form $(1 - a\, r(\|X\|^2)/\|X\|^2)X$. The minimaxity conditions are essentially those developed in Theorem 5.3 and Theorem 5.4 and most of the derivations are in the context of variance mixtures of normals. See Strawderman [1974], Maruyama [2003] and Fourdrinier, Kortbi and Strawderman [2008].

The main difficulty in using Theorem 5.1 with mixtures of normals densities for the sampling distribution is to prove the monotonicity (and boundedness) properties of the function r. Maruyama [2003] and Fourdrinier, Kortbi and Strawderman [2008] consider priors which are mixtures of normals as well. Their main condition for obtaining minimaxity of the corresponding Bayes estimator is that the mixing density g of the sampling distribution has monotone non decreasing likelihood ratio when considered as a scale parameter family. In Fourdrinier, Kortbi and Strawderman [2008], explicit use is made of that monotone likelihood ratio property for the mixing (possibly improper) density h of the prior distribution.

The main result of Fourdrinier, Kortbi and Strawderman [2008] is the following. Consult that paper for the technical proof.

Theorem 5.6. *Let X be a random vector in $\mathbb{R}^p$ $(p \geq 3)$ distributed as a variance mixture of multivariate normal distributions with density*

$$f(x) = \int_0^\infty \frac{1}{(2\pi v)^{p/2}} \exp\left(-\frac{1}{2}\frac{\|x-\theta\|^2}{v}\right) g(v)\, dv \tag{5.12}$$

where g is the density of a known nonnegative random variable V. Let π be a (generalized) prior with density of the form

$$\pi(\theta) = \int_0^\infty \frac{1}{(2\pi t)^{p/2}} \exp\left(-\frac{1}{2}\frac{\|\theta\|^2}{t}\right) h(t)\, dt \tag{5.13}$$

where h is a function from $\mathbb{R}_+$ into $\mathbb{R}_+$ such that this integral exists.

Assume that the mixing density g is such that

$$E[V] = \int_0^\infty v\, g(v)\, dv < \infty \text{ and } E[V^{-p/2}] = \int_0^\infty v^{-p/2} g(v)\, dv < \infty. \tag{5.14}$$

Assume also that the mixing function h of the (possibly improper) prior density π is absolutely continuous and satisfies

$$\lim_{t\to\infty} \frac{h(t)}{t^\beta} = c \tag{5.15}$$

for some $\beta < p/2 - 1$ and some $0 < c < \infty$. Assume finally that h and g have monotone increasing likelihood ratio when considered as a scale parameter family.

Then, if there exist $K > 0$, $t_0 > 0$ and $\alpha < 1$ such that

$$h(t) \le K t^{-\alpha} \quad \text{for } 0 < t < t_0, \tag{5.16}$$

the (generalized or proper) Bayes estimator δ_h with respect to the prior distribution corresponding to the mixing function h is minimax provided that β satisfies

$$-(p-2)\left[\frac{E[V^{-p/2+1}]}{E[V]E[V^{-p/2}]} - \frac{1}{2}\right] \le \beta. \tag{5.17}$$

For priors with mixing distribution h satisfying (5.16) and (5.17) an argument as in Maruyama [2003] using Brown [1979] and a Tauberian theorem suggests that the resulting generalized Bayes estimator is admissible if $\beta \le 0$. Recently, Maruyama and Takemura [2006] have verified this under additional conditions which imply, in the setting of Theorem 5.6, that $E_\theta[\|X\|^3] < \infty$.

Assume that the sampling distribution is a p-variate Student t with n_0 degrees of freedom. It corresponds to the inverse gamma mixing density $(n_0/2, n_0/2)$, that is, to $g(v) \propto v^{-(n_0+2)/2} \exp(-n_0/2v)$. Let also the prior be a Student t distribution with n degrees of freedom, that is, with mixing density $h(t) \propto t^{-(n+2)/2} \exp(-n/2t)$. It is clear that Conditions (5.14) and (5.15) are satisfied with $n_0 \ge 7$. It is also clear that

Condition (5.16) holds for any $\alpha < 1$. Finally a simple calculation shows that

$$\frac{E[V^{-p/2+1}]}{E[V]E[V^{-p/2}]} = \frac{n_0 - 2}{p + n_0 - 2}$$

so that Condition (5.17) reduces to

$$n \leq (p-2)\left[\frac{2(n_0-2)}{p+n_0-2} - 1\right] - 2.$$

Note that, as $n > 0$, this condition can hold if and only if $p \geq 5$ and

$$n_0 \geq 3 + p\frac{p}{p-4}.$$

Other examples (including generalized priors) can be found in Fourdrinier, Kortbi and Strawderman [2008].

In the following, we consider broader classes of spherically symmetric distributions which are not restricted to variance mixtures of normals. Minimaxity of generalized Bayes estimators is obtained for unimodal spherically symmetric superharmonic priors $\pi(\|\theta\|^2)$ under the additional assumption that the Laplacian of $\pi(\|\theta\|^2)$ is a nondecreasing function of $\|\theta\|^2$. The results presented below are derived in Fourdrinier and Strawderman [2008]. An interesting feature is that their approach does not rely on the Baranchik representation used in Maruyama [2003] and Fourdrinier, Kortbi and Strawderman [2008]. Note however that the superharmonicity property of the priors implies that the corresponding Bayes estimators cannot be proper.

First note that, for any prior $\pi(\theta)$, the Bayes estimator in (5.10) can be written as

$$\delta_\pi(X) = X + \frac{\nabla M(X)}{m(X)}$$

where, for any $X \in \mathbb{R}^p$,

$$M(x) = \int_{\mathbb{R}^p} F(\|x-\theta\|^2)\, \pi(\theta)\, d\theta$$

with F given in (5.5). Thus $\delta_\pi(X)$ has the general form $\delta_\pi(X) = X + g(X)$ (with $g(X) = \nabla M(X)/m(X)$) and, if the density $f(\|x-\theta\|^2)$ is as in Subsection **??**, that is, such $F(t)/f(t) \geq c > 0$ for some fixed positive constant c, then Corollary 5.2 applies and $\delta_\pi(X) = X + g(X) = X + \nabla M(X)/m(X)$ is minimax provided, for any $x \in \mathbb{R}^p$,

$$2\,c\,\operatorname{div} g(x) + \|g(x)\|^2 \leq 0.$$

In particular, it follows that if

$$2\,c\,\frac{\Delta M(x)}{m(x)} - 2\,c\,\frac{\nabla M(x) \cdot \nabla m(x)}{m^2(x)} + \frac{\|\nabla M(x)\|^2}{m^2(x)} \leq 0 \tag{5.18}$$

and

$$E_\theta\left[\left\|\frac{\nabla M(X)}{m(X)}\right\|^2\right] < \infty,$$

δ_π is minimax.

For a spherically symmetric prior $\pi(\|\theta\|^2)$, the main result of Fourdrinier and Strawderman [2008] is the following.

Theorem 5.7. *Assume that X has a spherically symmetric distribution in $\mathbb{R}^p$ with density $f(\|x-\theta\|^2)$. Assume that $\theta \in \mathbb{R}^p$ has a superharmonic prior $\pi(\|\theta\|^2)$ such that $\pi(\|\theta\|^2)$ is nonincreasing and $\Delta\pi(\|\theta\|^2)$ is nondecreasing in $\|\theta\|^2$. Assume also that*

$$E_\theta\left[\left\|\frac{\nabla M(X)}{m(X)}\right\|^2\right] < \infty.$$

Then the Bayes estimator δ_π is minimax under quadratic loss provided that $f(t)$ is log-convex, $c = \frac{F(0)}{f(0)} > 0$ and

$$\int_0^\infty f(t) t^{p/2} dt \leq 4c \int_0^\infty -f'(t) t^{p/2} dt < \infty. \tag{5.19}$$

To prove Theorem 5.7 we need preliminary lemmata whose proofs are given in the Appendix. Note first that it follows from the spherical symmetry of π that, for any $x \in \mathbb{R}^p$, $m(x)$ and $M(x)$ are functions of $t = \|x\|^2$. Then, setting

$$m(x) = m(t) \quad \text{and} \quad M(x) = M(t),$$

we have

$$\nabla m(x) = 2m'(t)\, x \quad \text{and} \quad \nabla M(x) = 2M'(t)\, x. \tag{5.20}$$

Lemma 5.5. *Assume that $\pi'(t) \leq 0$, for any $t \geq 0$. Then we have $M'(t) \leq 0$, for any $t \geq 0$.*

Lemma 5.6. *For any $x \in \mathbb{R}^p$,*

$$x \cdot \nabla m(x) = -2 \int_0^\infty H(u,t)\, u^{p/2} f'(u)\, du$$

and

$$x \cdot \nabla M(x) = \int_0^\infty H(u,t)\, u^{p/2} f(u)\, du$$

where, for $u \geq 0$ and for $t \geq 0$,

$$H(u,t) = \lambda(B) \int_{B_{\sqrt{u},x}} x \cdot \theta\, \pi'(\|\theta\|^2)\, dV_{\sqrt{u},x}(\theta) \tag{5.21}$$

and $V_{\sqrt{u},x}$ is the uniform distribution on the ball $B_{\sqrt{u},x}$ of radius $\sqrt{u}$ centered at x and $\lambda(B)$ is the volume of the unit ball.

Lemma 5.7. *For any $t \geq 0$, the function $H(u,t)$ in (5.21) is nondecreasing in u provided that $\Delta\pi(\|\theta\|^2)$ is nondecreasing in $\|\theta\|^2$.*

Lemma 5.8. *Let $h(\|\theta - x\|^2)$ be a unimodal density and let $\psi(\theta)$ be a symmetric function. Then*

$$\int_{\mathbb{R}^p} x \cdot \theta \, \psi(\theta) \, h(\|\theta - x\|^2) \, d\theta \geq 0$$

as soon as ψ is nonnegative.

Proof. (Proof of Theorem 5.7.) By the superharmonicity of $\pi(\|\theta\|^2)$, we have $\Delta M(x) \leq 0$ for all $x \in \mathbb{R}^p$ so that by (5.18), it suffices to prove that

$$-2\,c\,\nabla M(x) \cdot \nabla m(x) + \|\nabla M(x)\|^2 \leq 0 \tag{5.22}$$

for all $x \in \mathbb{R}^p$. Since m and M are spherically symmetric, by (5.20), (5.22) reduces to

$$-2cM'(t)m'(t) + \big(M'(t)\big)^2 \leq 0$$

where $t = \|x\|^2$. Since $M'(t) \leq 0$ by Lemma 5.5, Inequality (5.22) reduces to

$$-2cm'(t) + M'(t) \geq 0$$

or, by (5.20), to

$$-2\,c\,x \cdot \nabla m(x) + x \cdot \nabla M(x) \geq 0$$

or, by Lemma 5.6, to

$$4cE\left[H(u,t)\frac{f'(u)}{f(u)}\right] + E[h(u,t)] \geq 0, \tag{5.23}$$

where E denotes the expectation with respect to the density proportional to $u^{p/2} f(u)$. Since, by assumption, $\Delta\pi(\|\theta\|^2)$ is non decreasing in $\|\theta\|^2$, $H(u,t)$ is nondecreasing in u by Lemma 5.7. Furthermore $f'(u)/f(u)$ is non decreasing by log-convexity of f so that Inequality (5.16) is satisfied as soon as

$$4\,c\,E[H(u,t)]\,E\left[\frac{f'(u)}{f(u)}\right] + E[H(u,t)] \geq 0. \tag{5.24}$$

Finally, as $\pi'(\|\theta\|^2) \leq 0$ by assumption, Lemma 5.2 guarantees that $H(u,t) \leq 0$ (note that $V_{\sqrt{u},x}$ has a unimodal density) and hence (5.24) reduces to

$$4cE\left[\frac{f'(u)}{f(u)}\right] + 1 \leq 0$$

which is equivalent to (5.19). □

5.5 Shrinkage estimators for concave loss

In this section we consider improved shrinkage estimators for loss functions that are concave functions of squared error loss. The basic results are due to Brandwein and Strawderman [1980, 1991, 1991] and we largely follow the method of proof in the later paper. The general nature of the main result is that (under mild conditions) if an estimator can be shown to dominate X under squared error loss then the same estimator, with a suitably altered shrinkage constant, will dominate X for a loss which is a concave function of squared error loss.

Let X have a spherically symmetric distribution around θ, and let $g(X)$ be a weakly differentiable function. The estimators considered are of the form

$$\delta(X) = X + ag(X) \tag{5.25}$$

The loss functions are of the form

$$L(\theta, \delta) = f(||\delta - \theta||^2), \tag{5.26}$$

where $f(\cdot)$ is a differentiable non-negative, non-decreasing concave function (so that, in particular $f'(\cdot) \geq 0$).

One basic tool needed for the main result is Lemma 3.2.3 , and the other is the basic property of the concave function $f(\cdot)$ that $f(t+a) \leq f(t) + af'(t)$.

The following result shows that shrinkage estimators that improve on X for squared error loss also improve on X for concave loss provided the shrinkage constant is adjusted properly.

Theorem 5.8. *Let X have a spherically symmetric distribution around θ, let $g(X)$ be a weakly differentiable function, and let the loss be given by (5.26).*

Suppose there exists a function $h(\cdot)$ such that $h(X)$ is subharmonic and $E_{\theta,R}[R^2 h(U)]$ is non increasing where $U \sim \mathscr{U}_{R,\theta}$. Furthermore suppose that $g(X)$ satisfies $E^*_\theta[||g(X)||^2] < \infty$ and also satisfies

(1) div $g(X) \leq h(X)$,

(2) $\|g(X)\|^2 + 2h(X) \leq 0$ *and*

(3) $0 \leq a \leq \frac{1}{pE^*_0(1/\|X\|^2)}$,

where E^*_θ refers to the expectation with respect to the distribution whose Radon-Nikodyn derivative with respect to the distribution of X is proportional to $f'(||X - \theta||^2)$. Then, $\delta(X) = X + ag(X)$ is minimax. Also $\delta(X)$ dominates X provided $g(\cdot)$

*is non-zero with positive probability and strict inequality holds with positive probability in (1) or (2), or both inequalities are strict in (3). (Here, positive probability is with respect to the * distribution)*

Proof. Note, by concavity of $f(\cdot)$ (and the usual identity

$$||X+ag(X)-\theta||^2||X-\theta||^2+a^2||g(X)||^2+2(X-\theta)'g(X))R(\theta,\delta)$$

$$=E_\theta[f(||\delta(X)-\theta||^2)]\leq E_\theta[f(||X-\theta||^2)]$$

$$+\,E_\theta[f'(||X-\theta||^2)(a^2||g(X)||^2+2a(X-\theta)'g(X))],$$

so that the difference in risk, $R(\theta,\delta)-R(\theta,X)$ is bounded by

$$R(\theta,\delta)-R(\theta,X)\leq E_\theta[f'(||X-\theta||^2)(a^2||g(X)||^2+2a(X-\theta)'g(X))]$$

$$=E^*_\theta[(a^2||g(X)||^2+2a(X-\theta)'g(X))]\leq 0,$$

by Lemma 5.5 applied to the distribution corresponding to E^*_θ. This completes the proof. □

5.6 Shrinkage estimators for concave loss in the presence of a residual vector

In this section, we consider the case of concave loss and illustrate that certain classes of shrinkage estimators which properly use the residual vector have the strong robustness property of dominating the usual unbiased estimator uniformly over the class of spherically symmetric distributions, simultaneously for a broad class of concave loss functions. It extends and broadens the results of Section 5.5 to the residual vector case. We follow closely the development in Strawderman [2001].

Specifically, let (X,U) be a $p+k$ dimensional vector with mean vector $(\theta,0)$, where the dimensions of X and θ are equal to p and the dimensions of the residual vector U and its mean vector, 0, are equal to k. We suppose also that the random vector (X,U) has a spherically symmetric distribution about its mean. Notation for this model is

$$(X,U)SS_{p+k}(\theta,0). \tag{5.27}$$

The loss function we consider is

$$L(\theta,\delta)=f(||\theta-\delta||^2), \tag{5.28}$$

for $f(t)$ a non-negative concave monotone non-decreasing function.

The estimators we consider will be of the now familiar form

$$\delta(X,U)=X+a(S/(k+2))g(X), \tag{5.29}$$

where $S=||U||^2$, and $g(.)$ maps R^p into R^p.

The following result, extracted from the development in Theorem **??** and due to Brandwein and Strawderman [1991] is basic to the development of this section.

Lemma 5.9. *Brandwein and Strawderman ([1991]). Let X $SS_p(\theta)$, for $p\geq 4$ and let $g(X)$ map R^p into R^p be weakly differentiable, and such that*

(1) $||g(X)||^2/2\leq -h(X)\leq -\nabla' g(X)$,

(2) $-h(X)$ *is superharmonic and* $E_\theta[R^2h(W)|R]$ *is a nondecreasing function of R, where W has a uniform distribution on the sphere of radius r centered at* θ.

Then

$$E_\theta[||X+ag(X)-\theta||^2-||X-\theta| \le E[(-2a^2/r^2+2a/p)E_\theta[r^2h(W)|r^2]]\,,$$

where $r^2=||X-\theta||^2$.

We will also need the following well known result.

Lemma 5.10. *Suppose* (X,U) *has distribution (5.27). Then the random variable* $\beta=||X-\theta||^2/(||X-\theta||^2+S)$ *has a* $Beta(p/2,k/2)$ *distribution, independent of* $R^2=||X-\theta||^2+S$, *where* $S=||U||^2$.

The main result is the following:

Theorem 5.9. *Suppose* (X,U) *is distributed according to model (5.27), that loss is given by loss (5.26) and that the estimator* $\delta(X,S)$ *is given by (5.29). Then* $\delta(X,S)$ *dominates the unbiased estimator* X, *provided that* $(p-2-2\alpha)>0$, *and*

(1) $g(X)$ *satisfies assumptions (1) and (2) of Lemma 5.9,*

(2) *the concave nondecreasing function* $f(t)$ *also satisfies* $t^\alpha f'(t)$ *is nondecreasing, and*

(3) $0\le a\le(p-2-2\alpha)/p$.

Note first, by concavity of $f(.)$, that $f(t)\le f(y)+(y-a)f'(y)$. Hence the risk satisfies

$$\begin{aligned}
R(\theta,\delta)&=E[f(||X+\frac{aSg(X)}{k+2}-\theta||^2)]\\
&\le E[f(||X-\theta||^2)+f'(||X-\theta||^2)(||X+\frac{aSg(X)}{k+2}-\theta||^2-||X-\theta||^2)]\\
&=R(\theta,X)+E[f'(||X-\theta||^2)(||X+\frac{aSg(X)}{k+2}-\theta||^2-||X-\theta||^2)].
\end{aligned}$$

It suffices to prove the second term in the above expression is negative. Now, let $r^2 = ||X-\theta||^2, R^2 = ||X-\theta||^2 + S$ (where $S = ||U||^2 = R^2 - r^2$), and note that the conditional distribution of X given r, and R is $SS_p(\theta)$. Then it follows, using Lemma 5.9 that

$$\begin{aligned}
&E[f'(||X-\theta||^2)(||X+\frac{aSg(X)}{k+2}-\theta||^2-||X-\theta||^2)])\\
&=E[f'(r^2)E[||X+\frac{aSg(X)}{k+2}-\theta||^2-||X-\theta||^2|R,r]]\\
&\le E[f'(r^2)E[(2(\frac{aS}{(k+2)r})^2)-2\frac{aS}{(k+2)p})E_\theta[-r^2h(W)|r^2]|R,r]].
\end{aligned}$$

Now using Lemma 5.10, this last expression may be written as

$$\begin{aligned}
&2E[f'(\beta R^2)((\frac{a(1-\beta)R^2)}{(k+2)})^2\frac{1}{\beta R^2}-\frac{a(1-\beta)R^2}{(k+2))p})E_\theta[-\beta R^2h(W)|\beta R^2]|R]]=\\
&\frac{2a}{(k+2)}E[(R^2(\beta R^2)^\alpha f'(\beta R^2)(\beta R^2)^{-\alpha}(1-\beta)(\frac{(1-\beta)a}{\beta(k+2)}-\frac{1}{p})E_\theta[-\beta R^2h(W)|\beta R^2]|R]].
\end{aligned}$$

Next, using assumption (5.28), note that for fixed R, $\beta^\alpha f'(\beta R^2)$ is nonnegative and nondecreasing in β, and by assumption (2) of Lemma 5.9, so is $E_\theta[-\beta R^2h(W)|\beta R^2]$. Also $(1-\beta)/\beta$ is decreasing in β. Hence it follows from the covariance inequality (and independence of β and R) that the previous expression is less than or equal to

$$\frac{2a}{k+2}E\left[E_\theta[-\beta R^2h(W)|\beta R^2]R^2(R^2\beta)^\alpha f'(\beta R^2)|R]E[(\beta^\alpha(1-\beta))\left(\frac{a(1-\beta)}{\beta)(k+2)}-\frac{1}{p}\right)\right]$$

Since the first expectation in this term is nonnegative, it suffices that the second expectation is negative. But this is equivalent to

$$0\le a\le\frac{k+2}{p}E[((\beta^\alpha(1-\beta))]/E[(\beta^\alpha(1-\beta)^2)/\beta]=(p-2-2\alpha)/p,$$

which completes the proof.

For the loss $L(\theta,\delta) = ||\theta - \delta||^q, f(t) = t^{q/2}$, and it follows that $t^\alpha f'(t) = (q/2)t^{\alpha+q/2-1}$ is nondecreasing for $\alpha \geq 1 - q/2$. Thus, the following Corollary is immediate.

Corollary 5.6. *Under the loss* $L(\theta,\delta) = ||\theta - \delta||^q$, *for* $p > 4$ *and* $0 < q \leq 2$, *the estimator in 5.9 dominates X for* $0 < a \leq (p-4+2q)/p$ *simultaneously for all spherically symmetric distributions with finite second moment. It does so simultaneously for all such losses for* $0 < a \leq (p-4)/p$.

Note that the range of shrinkage constants for which domination holds includes $a = 1/2$ as soon as $p \geq 8$. For the usual James-Stein estimator,

$$\delta(X) = (1 - a(2(p-2)S/((k+2)||X||^2))X, \tag{5.30}$$

the uniformly optimal constant for quadratic loss ($f(.) = 1$) is $a = 1/2$ and hence this optimal estimator improves for all such l_q losses simultaneously for $p \geq 8$.

Chapter 6

Estimation of location parameter for the spherically symmetric case II

6.1 The general linear model case with residual vector

In this chapter, we consider the general linear model introduced in Section 2.6 when a residual vector U is available. Recall that (X,U) is a random vector around $(\theta,\mathbf{0})$ (such that $\dim X = \dim \theta = p$ and $\dim U = \dim \mathbf{0} = k$) with a spherically symmetric distribution. Estimation of θ under quadratic loss $\|\delta - \theta\|^2$ parallels the normal situation presented in Section 3.3 where $X \sim N_p(\theta, \sigma^2 I_p)$ (with σ^2 known) and the estimators of θ are of the form $\delta(X) = X + \sigma^2 g(X)$. Here the estimators will be of the form

$$\delta(X) = X + \frac{\|U\|^2}{k+2}\, g(X) \tag{6.1}$$

for some function g from $\mathbb{R}^p$ into $\mathbb{R}^p$. In this section,

$$\sigma^2 = \mathrm{Var}(X_i) = \mathrm{Var}(U_i) = \frac{1}{p} E_\theta[\|X-\theta\|^2] = \frac{1}{k} E_\theta[\|U\|^2] = \frac{1}{p+k} E[R^2]$$

can be considered as known or unknown. When σ^2 is unknown, $\|U\|^2/k$ is an unbiased estimator of σ^2. Also, when σ^2 is unknown, it is perhaps preferable to use the invariant loss $\|\delta - \theta\|^2/\sigma^2$ since the estimator X has constant risk p and is minimax for this loss provided the variance of X is finite, while the minimax risk for the loss $\|\delta - \theta\|^2$ is infinite.

When σ^2 is known, estimators of the form $\delta(X) = X = \sigma^2 g(X)$ can be used and we will contrast these estimators with estimators (6.1) in the next section. One advantage of the estimators in (6.1) is that they share a striking robustness property, namely that, if $\|g(X)\|^2 + 2\,\mathrm{div}\, g(X) \leq 0$, then $X + \|U\|^2/(k+2)\, g(X)$ dominates X whatever the distribution of (X,U). In particular, the form of the density may not be known and indeed there is no need that a density exists. The results of this section can be found in Cellier and Fourdrinier [37].

Assuming the risk of X is finite (i.e., $E_\theta[\|X - \theta\|^2] = E_0[\|X\|^2] < \infty$) the risk of $\delta(X)$ is finite if and only if $E_\theta[\|U\|^4 \|g(X)\|^2] < \infty$ and the difference in risk between $\delta(X)$ and X is

$$\mathit{\Delta}(\theta) = E_\theta\left[2\,(X-\theta)' g(X)\,\frac{\|U\|^2}{k+2} + \|g(X)\|^2 \frac{\|U\|^4}{(k+2)^2}\right]. \qquad (6.2)$$

The cross product term, that is, the first term in the right-hand side of (6.2) will be analyzed as in the normal case. Here follows an adaptation of Stein's identity.

Lemma 6.1. *(Stein type lemma for the general linear model) Assume that* $(X,U) \sim ss(\theta, \mathbf{0})$ *where* $\dim X = \dim \theta = p$ *and* $\dim U = \dim \mathbf{0} = k$*. Then, for any weakly differentiable function g from* $\mathbb{R}^p$ *into* $\mathbb{R}^p$ *such that*

$$E_\theta\left[|(X-\theta)' g(X)|\right] < \infty,$$

we have

$$E_\theta\left[(X-\theta)' g(X)\,\|U\|^2\right] = E_\theta\left[\operatorname{div} g(X)\,\frac{\|U\|^4}{k+2}\right]. \tag{6.3}$$

Proof. We will show that, conditionally on the radius $R = \|X-\theta\|^2 + \|U\|^2$, Equality (6.3) holds. First, conditionally on R, the left-hand side of (6.3) is expressed as

$$\begin{aligned} E_{R,\theta}\left[(X-\theta)' g(X)\,\|U\|^2\right] &= \int_{S_{R,\theta}} (x-\theta)' g(x)\,(R^2 - \|x-\theta\|^2)\, dU_{R,\theta}(x) \quad (6.4) \\ &= \int_{B_{R,\theta}} (x-\theta)' g(x)\, C_R^{p,k} (R^2 - \|x-\theta\|^2)^{k/2}\, dx \end{aligned}$$

since, according to (1.9), X given R has density

$$\psi_{R,\theta}(x) = C_R^{p,k} (R^2 - \|x-\theta\|^2)^{k/2-1}\, \mathbb{1}_{B_{R,\theta}}(x)$$

with

$$C_R^{p,k} = \frac{\Gamma((p+k)/2)}{\Gamma(k/2)}\,\frac{R^{2-(p+k)}}{\pi^{p/2}}.$$

Now, note that

$$(R^2 - \|x-\theta\|^2)^{k/2}\,(x-\theta) = \nabla\gamma(x)$$

where

$$\gamma(x) = \frac{-(R^2 - \|x-\theta\|^2)^{k/2+1}}{k+2}.$$

Hence, using the classical identity

$$\big(\nabla\gamma(x)\big)' g(x) = \operatorname{div}\big(\gamma(x)\, g(x)\big) - \gamma(x) \operatorname{div} g(x),$$

it follows from (6.4) that

$$E_{R,\theta}\left[(X-\theta)' g(X)\,\|U\|^2\right] = A + B \tag{6.5}$$

where

$$A = C_R^{p,k} \int_{B_{R,\theta}} \operatorname{div}\big(\gamma(x)\, g(x)\big)\, dx \tag{6.6}$$

and

$$B = C_R^{p,k} \int_{B_{R,\theta}} -\gamma(x) \operatorname{div} g(x) dx. \tag{6.7}$$

Applying Stokes' theorem to the integral in (6.6) gives

$$A = C_{S_R^{p,k}} \int_{S_{R,\theta}} \gamma(x)\, g(x)\, \frac{x-\theta}{\|x-\theta\|}\, d\sigma_{R,\theta}(x) = 0 \tag{6.8}$$

since, for any $x \in S_{R,\theta}$, $\gamma(x) = 0$. The term B in (6.7) can be expressed as

$$B = \int_{B_{R,\theta}} \operatorname{div} g(x)\, \frac{(R^2 - \|x-\theta\|^2)^2}{k+2}\, \psi_{R,\theta}(x)\, dx = E_{R,\theta}\left[\operatorname{div} g(X)\, \frac{\|U\|^4}{k+2}\right]$$

and, finally, the lemma follows from (6.4), (6.5) and (6.8). □

As a consequence of Lemma 6.1, we can derive a sufficient condition of domination of $\delta(X) = X + \|U\|^2/(k+2) g(X)$ over X.

Theorem 6.1. *Assume that $E_\theta[\|X\|^2] < \infty$ and $E_\theta[\|U\|^4\, \|g(X)\|^2] < \infty$. Then an unbiased estimator of the risk difference $\Delta(\theta)$ in (6.2) between $\delta(X) = X + \|U\|^2/(k+2)\, g(X)$ and X is*

$$[2 \operatorname{div} g(X) + \|g(X)\|^2]\, \frac{\|U\|^4}{(k+2)^2}. \tag{6.9}$$

A sufficient condition for domination of $\delta(X)$ over X is that, for any $x \in \mathbb{R}^p$,

$$2 \operatorname{div} g(x) + \|g(x)\|^2 \leq 0 \tag{6.10}$$

with strict inequality on a set a positive measure on $\mathbb{R}^p$.

Proof. The proof of (6.9) follows immediately from (6.3) and (6.2). The domination condition (6.10) is a direct consequence of (6.9). □

The inclusion of the residual term U in the estimate yields an interesting and strong property. Note that the hypotheses in Theorem (6.1) are independent of the radial distribution and are consequently valid for any spherically symmetric distribution. This is in contrast with the results of Section 6.2 which require conditions on the radial distribution.

The improved estimators in Section 5.3 require two critical hypotheses. The first one is the superharmonicity condition on an auxillary function h such that $\|g\|^2/2 \leq -h \leq -\operatorname{div} g$. Secondly these estimators require the assumption that the function $R \to R^2 E_{R,\theta}[h]$ is nonincreasing. In contrast, the conditions for improvement of the improved estimator with the residual term included share the same set of hypotheses as the general Stein type estimators in the normal case (see Section 3.3). As a result, estimators which dominate X (through the differential inequality) in the normal case dominate X simultaneously for all spherically symmetric distributions. At this point, we will focus on the so-called robust James-Stein estimators rather than discussing general examples as in Section 3.3.

Consider

$$\delta^a_{RJS}(X) = \left(1 - \frac{a}{\|X\|^2} \frac{\|U\|^2}{k+2}\right) X$$

where a is a positive constant which is of the form (6.1) with $g(X) = -aX/\|X\|^2$. Note this is the shrinkage in the basic James-Stein estimator in (3.6) with $\sigma^2 = 1$. Using the divergence calculation of this $g(X)$ from (3.9), the unbiased estimator of risk implied by (6.9) is,

$$\left(a^2 - 2a(p-2)\right) \frac{1}{\|X\|^2} \frac{\|U\|^4}{(k+2)^2},$$

and so it follows that the optimal constant a (i.e., with minimum risk) is $a = p - 2$. Note that this optimal a is independent of the sampling distribution and yields improvement on X for any spherically symmetric distribution. Hence the best a has a nice robust optimality property.

An alternative approach to the results of this section can be based on the approach used in Lemma 5.2. Thus a straightforward adaptation of the proof of Lemma 5.2 leads to

$$E_\theta\left[(X-\theta)^t g(X)\|U\|^2\right] = E_\theta\left[\frac{F(\|X-\theta\|^2+\|U\|^2)}{f(\|X-\theta\|^2+\|U\|^2)} \operatorname{div}_X g(X)\,\|U\|^2\right].$$

Similarly

$$\begin{aligned} E_\theta\left[\|g(X)\|^2\|U\|^4\right] &= E_\theta\left[U^t\left(U\|U\|^2\,\|g(X)\|^2\right)\right] \\ &= E_\theta\left[\frac{F(\|X-\theta\|^2+\|U\|^2)}{f(\|X-\theta\|^2+\|U\|^2)} \operatorname{div}_U\left(U\|U\|^2\right)\|g(X)\|^2\right] \\ &= E_\theta\left[\frac{F(\|X-\theta\|^2+\|U\|^2)}{f(\|X-\theta\|^2+\|U\|^2)} (k+2)\,\|U\|^2\,\|g(X)\|^2\right]. \end{aligned}$$

Hence the difference in risk between $X + \|U\|^2/(k+2)g(X)$ and X can be written as

$$C E^*_\theta\left[\left(2\operatorname{div} g(X) + \|g(X)\|^2\right)\frac{\|U\|^2}{k+2}\right] \tag{6.11}$$

where E^*_θ denotes the expectation with respect to the density proportional to $F\left(\|x-\theta\|^2+\|u\|^2\right)$ and C is the normalizing constant

$$C = \int_{\mathbb{R}^p \times \mathbb{R}^k} F(\|x-\theta\|^2 + \|u\|^2)\, dx\, du\,.$$

Note that a straightforward application of the Fubini theorem shows that

$$C = \frac{1}{p+k} \int_0^\infty r^2 h(r) dr$$

where $h(r)$ is the radial density. Thus C is the common variance of each coordinate of (X,U). Therefore it follows from (6.11) that condition (6.10) is sufficient for the minimaxity of the estimator $X + \|U\|^2/(k+2)\, g(X)$, provided we treat the density $f(\cdot)$ as fixed and known, which implies implicitely that σ^2 is known. Alternatively,

6.2 A paradox concerning shrinkage estimators

In this section, we contrast the result of the previous section and (6.2). We continue our study of the problem of estimating the mean vector θ of spherically symmetric distribution when the scale σ^2 is known but when a residual vector U is available.

An important class of improved estimators is the class of James-Stein estimators $\delta^a_{JS}(X) = \big(1 - a\sigma^2/\|X\|^2\big)X$. The previous section provided an alternative class of robust James-Stein estimators $\delta^a_{RJS}(X,U) = \big(1 - a/\|X\|^2\, \|U\|^2/(k+2)\big)X$. We show that there often exist situations where $\delta^a_{RJS}(X,U)$ dominates $\delta^a_{JS}(X)$ and hence that the use of the residual vector U to estimate σ^2 may be superior to using its known value. This phenomenon seems paradoxical in the sense that the risk behavior of an estimator may be improved by substituting an estimate for a known quantity. This phenomenon adds to the attractiveness of the robust James-Stein class by demonstrating not only domination of the usual estimator X simultaneously for all spherically symmetric distributions, but also domination of the usual James-Stein estimators in many cases. A similar paradox was found in the context of goodness of fit testing by [144]. The results of this section are from [55] and [64].

Note that the paradox cannot occur in the case of a normal distribution since by the Rao-Blackwell theorem, when σ^2 is known in the normal case, X is a complete sufficient statistic so that the conditional expectation of $\delta^a_{RJS}(X,U)$ given X reduces to $\delta^{ak/(k+2)}_{JS}(X)$ which dominates $\delta^a_{RJS}(X,U)$. Note also that, if the paradox holds for one value of σ^2 for a particular family, it holds for all values of σ^2 by the scale equivariance of $\delta^a_{RJS}(X,U)$ and, therefore, holds for any scale mixture. Hence, as the normal distribution arises as a mixture of uniform distributions on spheres, and also as a mixture of uniform distributions on balls, the paradox cannot occur for these distributions as well.

For easy presentation, it is convenient to define the general estimator $\delta^a_\alpha(X,U) = \left(1-a\|U\|^{2\alpha}/\|X\|^2\right)X$ for $\alpha = 0$ or 1. Note that, for $\alpha = 0$, $\delta^a_0 = \delta^a_{JS}$ and, for $\alpha = 1$, $\delta^a_1 = \delta^{a(k+2)}_{RJS}$. As in Section 6.1, we assume the finiteness of the risk of X (i.e., $E_0[\|X\|^2] < \infty$) and it is clear that the finiteness of the risk of $\delta^a_\alpha(X,U)$ is guaranteed as soon as $E_\theta\left[\|U\|^{2\alpha}/\|X\|^2\right] < \infty$. Under that condition, the following proposition yields the risk of δ^a_α.

Proposition 6.1. *The risk of δ^a_α equals*

$$R(\delta^a_\alpha,\theta) = E_0[\|X\|^2] + a^2\,E_\theta\left[\frac{\|U\|^{4\alpha}}{\|X\|^2}\right] - 2\,a\,\frac{p-2}{k+2\alpha}\,E_\theta\left[\frac{\|U\|^{2(\alpha+1)}}{\|X\|^2}\right].$$

Proof. The risk calculation is a straightforward extension of the one in Section 5.2 as in Lemma 5.2 for the normal case, with

$$g(x,s) = \frac{s^{\alpha}}{\|x\|^2} x.$$

□

It is easy to deduce from Lemma 6.1 that, for any $\theta \in \mathbb{R}^p$, the constant a for which the risk of δ_α^a is minimum is

$$a(\theta) = \frac{p-2}{k+2\alpha} \frac{E_\theta\left[\frac{\|U\|^{2(\alpha+1)}}{\|X\|^2}\right]}{E_\theta\left[\frac{\|U\|^{4\alpha}}{\|X\|^2}\right]}.$$

The corresponding risk is

$$R\big(\delta_\alpha^{a(\theta)}, \theta\big) = E_0\big[\|X\|^2\big] - \left(\frac{p-2}{k+2\alpha}\right)^2 \frac{\left(E_\theta\left[\frac{\|U\|^{2(\alpha+1)}}{\|X\|^2}\right]\right)^2}{E_\theta\left[\frac{\|U\|^{4\alpha}}{\|X\|^2}\right]}. \tag{6.12}$$

We already noticed in Section 6.1 that, for $\alpha = 1$, the optimal a does not depend on θ and equals $\frac{p-2}{k+2}$, which can be easily seen from the above expression. For $\alpha = 0$, the optimal a depends on θ and equals

$$a(\theta) = \frac{p-2}{k} \frac{E_\theta\left[\frac{\|U\|^2}{\|X\|^2}\right]}{E_\theta\left[\frac{1}{\|X\|^2}\right]}. \tag{6.13}$$

Then the paradox will occur if, for any $a \geq 0$,

$$R\big(\delta_1^{(p-2)/(k+2)}, \theta\big) < R\big(\delta_0^a, \theta\big),$$

and it will certainly occur if

$$R\big(\delta_1^{(p-2)/(k+2)}, \theta\big) < R\big(\delta_0^{a(\theta)}, \theta\big)$$

with $a(\theta)$ as in (6.13) and with the corresponding risks given by (6.12). This is equivalent to

$$\left(\frac{p-2}{k}\right)^2 \frac{\left(E_\theta\left[\frac{\|U\|^2}{\|X\|^2}\right]\right)^2}{E_\theta\left[\frac{1}{\|X\|^2}\right]} < \left(\frac{p-2}{k+2}\right)^2 E_\theta\left[\frac{\|U\|^4}{\|X\|^2}\right],$$

that is, to

$$\frac{\left(E_\theta\left[\frac{\|U\|^2}{\|X\|^2}\right]\right)^2}{E_\theta\left[\frac{\|U\|^4}{\|X\|^2}\right]E_\theta\left[\frac{1}{\|X\|^2}\right]} < \left(\frac{k}{k+2}\right)^2. \tag{6.14}$$

Expression (6.14) is a general condition for the paradox to occur. [55] develop a series of bounds for the quantities in the left-hand side of (6.14). However the resulting sufficient condition was complex and could be verified in a limited number of cases, the primary example being the Student case. Subsequently [64] developed a new approach to deal with the expectations in (6.14) for the case of mixtures of normals.

Assume that (X,U) has a scale mixture of normals distribution with the representation

$$(X,U)|Z = z \sim N_{p+k}\big((\theta,0), zI_{p+k}\big) \tag{6.15}$$

where Z is a positive random variable. For model (6.15), expressions of the expectations in (6.14) are given by the following lemma.

Lemma 6.2. *Assume that (X,U) is a scale mixture of normals as in (6.15) and that $p \geq 3$. Let $q > -k/2$ and assume that $E[Z^{q-1}] < \infty$. Then we have*

$$E_\theta\left[\frac{\|U\|^{2q}}{\|X\|^2}\right] = 2^q \frac{\Gamma(k/2+q)}{\Gamma(k/2)} E\left[Z^{q-1} f_p\left(\frac{\|\theta\|^2}{Z}\right)\right]$$

where $f_p(\gamma) = E[Y^{-1}]$ for a random variable Y having a noncentral chi-square distribution with p degrees of freedom and noncentrality parameter γ.

Proof. Note that conditionally on Z, X and U are independent and $(\|U\|^2/Z)|Z \sim \chi^2_k(0)$ and $(\|X\|^2/Z)|Z \sim \chi^2_p\big(\|\theta\|^2/Z\big)$. Hence we can write

$$\begin{aligned}
E_\theta\left[\frac{\|U\|^{2q}}{\|X\|^2}\,\middle|\, Z\right] &= E[\|U\|^{2q}|Z]E_\theta\left[\frac{1}{\|X\|^2}\,\middle|\, Z\right]\\
&= Z^{q-1}E\left[\left(\frac{\|U\|^2}{Z}\right)^q\,\middle|\, Z\right]E_\theta\left[\frac{Z}{\|X\|^2}\,\middle|\, Z\right]\\
&= Z^{q-1}2^q\,\frac{\Gamma(k/2+q)}{\Gamma(k/2)}\,f_p\left(\frac{\|\theta\|^2}{Z}\right)
\end{aligned}$$

since $q > -k/2$. Now use the fact that f_p is bounded if $p \geq 3$ and $E[Z^{q-1}] < \infty$ and uncondition to complete the proof. □

It follows directly from Lemma 6.2 for $q = 0, 1, 2$ that (6.14) is equivalent to

$$H_Z(\lambda) = \frac{\big(E\big[f_p(\lambda^2/Z)\big]\big)^2}{E\big[Zf_p(\lambda^2/Z)\big]E\big[Z^{-1}f_p(\lambda^2/Z)\big]} < \frac{k}{k+2} \tag{6.16}$$

for all $\lambda = \|\theta\| \geq 0$.

Alternatively note that

$$H_Z(\lambda) = \big(E_\lambda[W]E_\lambda[W^{-1}]\big)^{-1} \tag{6.17}$$

where W is a positive random variable with density

$$h_\lambda(w) = c(\lambda)f_p(\lambda^2 w)g(w)$$

where g is a density of $V = Z^{-1}$ and $c(\lambda)$ is a normalizing constant. Then (6.14) can also be expressed as

$$E_\lambda[W]E_\lambda[W^{-1}] > 1 + \frac{2}{k} \tag{6.18}$$

for all $\lambda \geq 0$.

The following main result shows that the paradox occurs for any nondegenerate mixture of normals.

Theorem 6.2. *Assume that (X,U) is a scale mixture of normals as in (6.15), with Z nondegenerate, $E[Z] < \infty$ and $E[Z^{-1}] < \infty$. Then, for any $p \geq 3$, there exists a positive integer k_0 such that, for any integer $k \geq k_0$, the optimal robust James-Stein estimator $\delta_{RJS}^{(p-2)}$ simultaneously dominates all James-Stein δ_{JS}^{a}.*

Proof. Setting $\bar{H} = \sup_{\lambda \geq 0} H_Z(\lambda)$, Condition (6.16) reduces to $k > 2\frac{\bar{H}}{1-\bar{H}}$. From (6.17) we know (by covariance inequality) that $H_Z(\lambda) \leq 1$ with equality if and only if W is degenerate, that is, if and only if Z is degenerate, which corresponds to the normal case. Then $\bar{H} \leq 1$ and we only need to show that $\bar{H} < 1$ since H_Z, and hence $\bar{H}$ does not depend on k.

Now it can be shown (see Lemma 3 in [64]) that

$$\begin{aligned}
\lim_{\lambda \to \infty} H_Z(\lambda) &= \left(\lim_{\lambda \to \infty} E_\lambda[W] \lim_{\lambda \to \infty} E_\lambda[W^{-1}] \right)^{-1} \\
&= \left(\frac{1}{E[Z]} \cdot \frac{E[Z^2]}{E[Z]} \right)^{-1} \\
&= \frac{(E[Z])^2}{E[Z^2]} \\
&< 1,
\end{aligned}$$

for $p \geq 3$ and nondegenerate Z. □

The necessity of nondegeneracy of Z is explicit in the proof of Theorem 6.2. Therefore the paradox occurs only in the case of nondegenerate mixtures of normals and not in the normal case.

Outside the class of mixtures of normals little is known. In the case where the radial distribution is concentrated on two points, [55] show that the paradox can occur for suitable weights. Showing the existence of the paradox in other families of spherically symmetric distributions is an open question.

6.3 The general linear model with unknown covariance matrix

In this section, consider the general linear model with an unknown non-singular scale matrix. Most of the material of this section is taken from [63]. As introduced in Section 2.6, in the canonical form of this model, $X, V_1, \ldots, V_{n-1}$ are n random vectors in $\mathbb{R}^p$ with joint density of the form

$$f(x-\theta)'\Sigma^{-1}(x-\theta) + \sum_{j=1}^{n-1} V_j'\Sigma^{-1}V_j\Big) \tag{6.19}$$

where the $p \times 1$ location vector θ and the $p \times p$ scale matrix Σ are unknown. Note that we have absorbed the normalizing factor $|\Sigma^{-1}|^{n/2}$ in the function f.

We consider the problem of estimating θ with the invariant loss

$$L(\theta, \delta) = (\delta - \theta)'\Sigma^{-1}(\delta - \theta). \tag{6.20}$$

Recall that the usual estimator $\delta_0(X) = X$ is minimax provided $E_{0,I}[\|X\|^2] < \infty$ (where $E_{\theta,\Sigma}$ denotes the expectation with respect to the density in (6.19)). We make this assumption throughout this section. Note that, when Σ is a covariance matrix, this expectation is necessarily finite and equal to p. Moreover X is typically admissible when $p \leq 2$ and inadmissible when $p \geq 3$.

We concentrate on the case $p \geq 3$ and construct a class of estimators, depending on the sufficient statistics (X,S), of the form

$$\delta(X,S) = X + g(X,S), \tag{6.21}$$

where $S = \sum_{i=1}^{n-1} V_i V_i'$, which dominate $\delta_0(X) = X$ under loss (6.20), for the entire class of distributions defined in (6.19) such that $E_{0,I}[\|X\|^2] < \infty$. Note that, although the loss in (6.20) is invariant, the estimate in (6.21) may not be equivariant (except for $\delta_0(X)$).

The risk difference $\Delta_{\theta,\Sigma}$ between $\delta(X,S)$ given in (6.21) and $\delta_0(X) = X$ equals

$$\begin{aligned} \Delta_{\theta,\Sigma} &= R\big(\theta,\delta(X,S)\big) - R\big(\theta,\delta_0(X)\big) \qquad (6.22)\\ &= E_{\theta,\Sigma}\big[2g'(X,S)\Sigma^{-1}(X-\theta)\big] + E_{\theta,\Sigma}\big[g'(X,S)\Sigma^{-1}g(X,S)\big], \end{aligned}$$

provided $E_{\theta,\Sigma}\big[g'(X,S)\Sigma^{-1}g(X,S)\big] < \infty$.

We first give a lemma which expresses the two terms in the last expression of (6.22) as expectations $E^*_{\theta,\Sigma}$ with respect to the distribution

$$C^{-1}F\left((x-\theta)'\Sigma^{-1}(x-\theta) + \sum_{j=1}^{n-1} V_j'\Sigma^{-1}V_j\right)$$

where F and C are defined as

$$F(t) = \frac{1}{2}\int_t^\infty f(s)ds$$

and

$$C = \int_{\mathbb{R}^p \times \cdots \times \mathbb{R}^p} F\left((x-\theta)'\Sigma^{-1}(x-\theta) + \sum_{j=1}^{n-1} V_j'\Sigma^{-1}V_j\right) dx\,dv_1 \cdots dv_{n-1}.$$

Lemma 6.3. (1) *Suppose $g(x,s)$ is a weakly differentiable function in x for each s such that the expectation $E_{\theta,\Sigma}\left[g'(X,S)\Sigma^{-1}(X-\theta)\right]$ exists. Then*

$$E_{\theta,\Sigma}\left[g'(X,S)\Sigma^{-1}(X-\theta)\right] = CE^*_{\theta,\Sigma}\left[\operatorname{div}_X g(X,S)\right] \tag{6.23}$$

where $\operatorname{div}_x g(x,s)$ *is the divergence of $g(x,s)$ with respect to x.*

(2) *Suppose $T(x,s)$ is a $p \times p$ matrix function weakly differentiable in s for any x such that the expectation $E_{\theta,\Sigma}\left[\operatorname{tr}\left(T(X,S)\right)\Sigma^{-1}\right]$ exists. Then*

$$E_{\theta,\Sigma}\left[\operatorname{tr}\left(T(X,S)\Sigma^{-1}\right)\right] = CE^*_{\theta,\Sigma}\left[2D^*_{1/2}T(X,S) + (n-p-2)\operatorname{tr}(S^{-1}T)\right] \tag{6.24}$$

where

$$D^*_{1/2}T(x,s) = \sum_{i=1}^{p} \frac{\partial T_{ii}(x,s)}{\partial s_{ii}} + \frac{1}{2}\sum_{i \neq j} \frac{\partial T_{ij}(x,s)}{\partial s_{ij}}. \tag{6.25}$$

The proof of Lemma 6.3 is given at the end of this section.

Note that, when X, $V_1, \ldots, V_{n-1}$ are independent normal vectors with covariance Σ, then $f = F$ and therefore $E_{\theta,\Sigma}[\quad] = E^*_{\theta,\Sigma}[\quad]$. Hence Lemma 6.3. (1) essentially reduces to Stein's lemma 1981 (cf. [134]) and Lemma 6.3. (2) corresponds to a result of [71].

Applying part (1) to the first term in (6.22) and part (2) to the second term in (6.22) with $T(x,s) = g(x,s)g'(x,s)$ (noting that $g'(x,s)\Sigma^{-1}g(x,s) = \operatorname{tr}\left(g(x,s)g'(x,s)\Sigma^{-1}\right)$) gives immediately the following theorem.

Theorem 6.3. *Assume that $g(x,s)$ and $T(x,s) = g(x,s)g'(x,s)$ satisfy the assumptions of Lemma 6.3. Assume also that $E_{0,\Sigma}[\|X\|^2] < \infty$ and $E_{\theta,\Sigma}\left[g'(X,S)\Sigma^{-1}g(X,S)\right] < \infty$. Then the risk difference $\Delta_{\theta,\Sigma}$ in (6.22) between $\delta(X,S) = X + g(X,S)$ and $\delta_0(X) = X$ equals*

$$C E^*_{\theta,\Sigma}\left[2\,\mathrm{div}_X g(X,S) + (n-p-2)\,g'(X,S)S^{-1}g(X,S) + 2D^*_{1/2}\big(g(X,S)g'(X,S)\big)\right]. \tag{6.26}$$

A sufficient condition for $\delta(X,S)$ to be minimax is

$$2\,\mathrm{div}_x g(x,s) + (n-p-2)\,g'(x,s)s^{-1}g(x,s) + 2D^*_{1/2}\big(g(x,s)g'(x,s)\big) \leq 0 \tag{6.27}$$

for all x and s. Furthermore $\delta(X,S)$ dominates $\delta_0(X)$ as soon as (6.27) is satisfied with strict inequality on a set of positive measure.

Note that, as $E^*_{\theta,\Sigma}[\quad] = E_{\theta,\Sigma}[\quad]$ in the normal case, the left-hand side of (6.27) is an unbiased estimator of the risk difference between $\delta(X,S)$ and $\delta_0(X)$. Perhaps, most importantly, observe that the theorem leads to an extremely strong robustness property for estimators satisfying (6.27). Namely, any such estimator is minimax and, as soon as strict inequality occurs on a set of positive measure in (6.27), dominates $\delta_0(X)$ for the entire class of distributions (6.19). This property is analogous to the robustness property mentioned in Section 6.1 in the case of spherically symmetric distributions.

The following corollary gives a general class of examples of minimax estimates which dominate $\delta_0(X)$ uniformly for densities of the form (6.19).

Corollary 6.1. *Assume that* $E_{0,\Sigma}[\|X\|^2] < \infty$ *and* $E_{\theta,\Sigma}\left[\frac{\|X\|^2}{(X'S^{-1}X)^2}\right] < \infty$*. Let* $\delta(X,S) = (1 - r(X'S^{-1}X)/X'S^{-1}X)X$ *where* r *is a nondecreasing function bounded between 0 and* $2(p-2)/(n-p+2)$*. Then* $\delta(X,S)$ *is minimax for any density of the form (6.19). Furthermore* $\delta(X,S)$ *dominates* $\delta_0(X)$ *as soon as either* r *is strictly increasing or bounded away for 0 and* $\frac{2(p-2)}{n-p+2}$ *on a set of positive measure.*

Proof. Setting

$$g(x,s) = -\frac{r(x's^{-1}x)}{x's^{-1}x}\,x,$$

we have

$$\mathrm{div}_s g(x,s) = -\left[(p-2)\frac{r(x's^{-1}x)}{xs^{-1}x} + 2r'(x's^{-1}x)\right]$$

by routine calculations. Now we have

$$\begin{aligned}
&d^*_{1/2}\big(g(x,s)g'(x,s)\big)\\
&= \sum_{i=1}^{p}\frac{\partial}{\partial s_{ii}}\left[\frac{r^2(x's^{-1}x)}{(x's^{-1}x)^2}\right]x_i^2 + \frac{1}{2}\sum_{i\neq j}\frac{\partial}{\partial s_{ij}}\left[\frac{r^2(x's^{-1}x)}{(x's^{-1}x)^2}\right]x_i x_j\\
&= \frac{2(x's^{-1}x)^2 r(x's^{-1}x)r'(x;s^{-1}x) - 2(x's^{-1}x)r^2(x's^{-1}x)}{(x's^{-1}x)^4}\\
&\qquad\times\left\{\sum_{i=1}^{p}\frac{\partial}{\partial s_{ii}}(x's^{-1}x)X_i^2 + \frac{1}{2}\sum_{i\neq j}\frac{\partial}{\partial s_{ij}}(x's^{-1}x)x_i x_j\right\}. \qquad (6.28)
\end{aligned}$$

Using the fact that

$$\frac{\partial}{\partial s_{ij}}(x's^{-1}x) = -(2-\delta_{ij})(x's^{-1})_i(x's^{-1})_j$$

it follows that the bracketed term in (6.28) equals

$$
\begin{aligned}
&-\left\{\sum_{i=1}^{p}(x's^{-1})_i^2 x_i^2+\frac{1}{2}\sum_{i\neq j}2(x's^{-1})_i(x's^{-1})_j x_i x_j\right\}\\
&=-\sum_{1\leq i,j\leq p}(x's^{-1})_i(x's^{-1})_j x_j\\
&=-\left(\sum_{i=1}^{p}(x's^{-1})_i X_i\right)^2\\
&=-(x's^{-1}x)^2
\end{aligned}
$$

and hence

$$
d^*_{1/2}\big(g(x,s)g'(x,s)\big)=-2\left\{r(x's^{-1}x)r'(x'sx)-\frac{r^2(x's^{-1}x)}{x's^{-1}x}\right\}.
$$

Finally it is clear that

$$
g'(x,s)s^{-1}g(x,s)=\frac{r^2(x's^{-1}x)}{x's^{-1}x}
$$

so that the left-hand side of (6.27) equals

$$
\begin{aligned}
&-2\left\{(p-2)\frac{r(x's^{-1}x)}{x's^{-1}x}+2r'(x's^{-1}x)\right\}+(n-p-2)\\
&-4\big\{r(x's^{-1}x)r'(x's^{-1}x)-\frac{r^2(x's^{-1}x)}{x's^{-1}x}\big\}\\
&=\frac{r(x's^{-1}x)}{}x's^{-1}x\big\{-2(p-2)+(n-p+2)r(x's^{-1}x)\big\}\\
&\quad-4r'(x's^{-1}x)\big\{1+r(x's^{-1}x)\big\}\\
&\leq 0 \qquad (6.29)
\end{aligned}
$$

according to the assumptions on r.

Hence the minimaxity of $\delta(X,S)$ follows. The domination result follows as well since strict inequality in (6.29) holds on a set of positive measure under the additional assumptions. □

6.3.1 Proof of Lemma 6.3

(1) By definition, we have

$$E_\theta \left[g(X,S)'\Sigma^{-1}(X-\theta) \right] = \int_{\mathbb{R}^p \times \cdots \times \mathbb{R}^p} \int_{\mathbb{R}^p} g(x,s)'\Sigma^{-1}(x-\theta)$$
$$f\left((x-\theta)'\Sigma^{-1}(x-\theta) + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j \right) dx\, dv_1 \ldots dv_{n-1}.$$

Now applying the integration-by-slice in Appendix AAA with $\varphi(x) = \sqrt{(x-\theta)'\Sigma^{-1}(x-\theta)}$ to the inner most integral gives

$$\nabla\varphi(x) = \frac{\Sigma^{-1}(x-\theta)}{\sqrt{(x-\theta)'\Sigma^{-1}(x-\theta)}}$$

and

$$\begin{aligned}
&\int_{\mathbb{R}^p} g(x,s)'\Sigma^{-1}(x-\theta)\, f\left((x-\theta)'\Sigma^{-1}(x-\theta) + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j \right) dx \\
&= \int_0^\infty f\left(R^2 + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j \right) \int_{[\varphi=R]} \frac{g(x,s)'\Sigma^{-1}(x-\theta)}{\|\nabla\varphi(x)\|} d\sigma_R(x)\, dR \\
&= \int_0^\infty f\left(R^2 + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j \right) \int_{[\varphi=R]} g(x,s)' \sqrt{(x-\theta)'\Sigma^{-1}(x-\theta)}\, \frac{\nabla\varphi(x)}{\|\nabla\varphi(x)\|}\, d\sigma_R(x)\, dR \\
&= \int_0^\infty R\, f\left(R^2 + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j \right) \int_{[\varphi=R]} g(x,s)\, \frac{\nabla\varphi(x)}{\|\nabla\varphi(x)\|}\, d\sigma_R(x)\, dR \\
&= \int_0^\infty R\, f\left(R^2 + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j \right) \int_{[\varphi\le R]} \text{div}_x\, g(x,s)\, dx\, dR \\
&= \int_{\mathbb{R}^p} \text{div}_x g(x,s) \int_{\sqrt{(x-\theta)'\Sigma^{-1}(x-\theta)}}^\infty R\, f\left(R^2 + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j \right) dR\, dx \\
&= \int_{\mathbb{R}^p} \text{div}_x\, g(x,s)\, \frac{1}{2} \int_{(x-\theta)'\Sigma^{-1}(x-\theta)}^\infty f\left(r + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j \right) dr\, dx \\
&= \int_{\mathbb{R}^p} \text{div}_x\, g(x,s) F\left((x-\theta)'\Sigma^{-1}(x-\theta) + \sum_{j=1}^{n-1} v_j'\Sigma^{-1}v_j \right) dx. \qquad (6.30)
\end{aligned}$$

Finally integrating (6.30) with respect to the v_j gives an expression for the expectation $E_\theta[g(X,S)'\Sigma^{-1}(X-\theta)]$ and yields (1). □

(2) First note that

$$\begin{aligned}
\operatorname{tr}\left(T(X,S)\Sigma^{-1}\right) &= \operatorname{tr}\left(T(X,S)\Sigma^{-1}SS^{-1}\right)\\
&= \operatorname{tr}\left(T(X,S)\Sigma^{-1}\sum_{i=1}^{n-1}V_iV_i'S^{-1}\right)\\
&= \sum_{i=1}^{n-1}\operatorname{tr}\left(V_i'S^{-1}T(X,S)\Sigma^{-1}V_i\right)\\
&= \sum_{i=1}^{n-1}V_i'S^{-1}T(X,S)\Sigma^{-1}V_i.
\end{aligned}$$

Then, by the argument in (1) above,

$$\begin{aligned}
E_\theta\left[\operatorname{tr}\left(T(X,S)\Sigma^{-1}\right)\right] &= C\sum_{i=1}^{n-1}E_\theta^*\left[\operatorname{div}_{V_i}T(X,S)S^{-1}V_i\right]\\
&= C\sum_{i=1}^{n-1}E_\theta^*\left[\sum_{j=1}^{p}\frac{\partial}{\partial V_{ij}}\left(\sum_{m=1}^{p}\sum_{\ell=1}^{p}T_{j\ell}(X,S)S^{\ell m}V_{im}\right)\right]\\
&= C\,E_\theta^*(A_1+A_2+A_3)
\end{aligned}$$

where

$$A_1 = \sum_{i=1}^{n-1}\sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{\ell=1}^{p}\left(\frac{\partial}{\partial V_{ij}}V_{im}\right)T_{j\ell}(X,S)S^{\ell m},$$

$$A_2 = \sum_{i=1}^{n-1}\sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{\ell=1}^{p}V_{im}\left(\frac{\partial}{\partial V_{ij}}T_{j\ell}(X,S)\right)S^{\ell m}$$

and

$$A_3 = \sum_{i=1}^{n-1}\sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{\ell=1}^{p}V_{im}T_{j\ell}(X,S)\left(\frac{\partial}{\partial V_{ij}}S^{\ell m}\right).$$

First it is easy to see that

$$\begin{aligned} A_1 &= \sum_{i=1}^{n-1}\sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{\ell=1}^{p}\delta_{jm}T_{j\ell}(X,S)S^{\ell m} \\ &= (n-1)\sum_{j=1}^{p}\sum_{\ell=1}^{p}T_{j\ell}(X,S)S^{\ell j} \\ &= (n-1)\operatorname{tr}\left(T(X,S)S^{-1}\right). \end{aligned}$$

Now since S is symmetric we have

$$A_2 = \sum_{i=1}^{n-1}\sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{\ell=1}^{p}V_{im}S^{\ell m}\left[\sum_{q\le r}\frac{\partial T_{j\ell}(X,S)}{\partial S_{qr}}\frac{\partial S_{qr}}{\partial V_{ij}}\right].$$

By definition of S, the last derivative is

$$\frac{\partial S_{qr}}{\partial V_{ij}} = \frac{\partial}{\partial V_{ij}}\left(V_{iq}V_{ir}\right) = V_{iq}\delta_{jr} + V_{ir}\delta_{jq}.$$

Multiplying the last expression by V_{im} and summing on i we obtain

$$A_2 = \sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{\ell=1}^{p}S^{\ell m}\left[\sum_{q\le r}\frac{\partial T_{j\ell}(X,S)}{\partial S_{qr}}\left(S_{mq}\delta_{jr} + S_{mr}\delta_{jq}\right)\right]. \tag{6.31}$$

Summing on m and using the fact that $\sum_{m=1}^{p}S^{am}S_{mb} = \delta_{ab}$ it follows that

$$\begin{aligned} A_2 &= \sum_{j=1}^{p}\sum_{\ell=1}^{p}\sum_{q\le r}\frac{\partial T_{j\ell}(X,S)}{\partial S_{qr}}\left(\delta_{\ell q}\delta_{jr} + \delta_{\ell r}\delta_{jq}\right) \\ &= \sum_{j=1}^{p}\sum_{\ell=1}^{p}\sum_{q\le r}\left(\frac{\partial T_{j\ell}(X,S)}{\partial S_{\ell j}}\delta_{\ell q}\delta_{jr} + \frac{\partial T_{j\ell}(X,S)}{\partial S_{j\ell}}\delta_{\ell r}\delta_{jq}\right) \\ &= \sum_{j=1}^{p}\sum_{\ell=1}^{p}\sum_{q\le r}\frac{\partial T_{j\ell}(X,S)}{\partial S_{\ell j}}\left(\delta_{\ell q}\delta_{jr} + \delta_{\ell r}\delta_{jq}\right). \end{aligned}$$

Note that

$$\sum_{q\le r}\left(\delta_{\ell q}\delta_{jr} + \delta_{\ell r}\delta_{jq}\right) = \begin{cases} 2 \text{ if } j = \ell \\ 1 \text{ if } j \ne \ell. \end{cases}$$

Hence

$$\begin{aligned}A_2 &= 2\left[\sum_{j=1}^{p}\frac{\partial T_{jj}(X,S)}{\partial S_{jj}}+\frac{1}{2}\sum_{j\neq\ell}\frac{\partial T_{j\ell}(X,S)}{\partial S_{j\ell}}\right]\\ &= 2D^{*}_{1/2}T.\end{aligned}$$

We now treat the term A_3. Using the same argument which led to (6.31), we can write A_3 as

$$A_3=\sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{\ell=1}^{p}T_{j\ell}(X,S)\sum_{q\leq r}\frac{\partial S^{\ell m}}{\partial S_{qr}}\left(S_{mq}\delta_{jr}+S_{mr}\delta_{jq}\right).$$

Using the fact that

$$\frac{\partial S^{\ell m}}{\partial S_{qr}}=\begin{cases}-S^{\ell q}S^{rm}-S^{mq}S^{r\ell} & \text{if } q\neq r\\ -S^{\ell q}S^{qm} & \text{if } q=r\end{cases}$$

we have that A_3 is expressed as

$$\begin{aligned}&-\sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{\ell=1}^{p}T_{j\ell}(X,S)\left\{\sum_{q=r}2S_{mq}\delta_{jq}S^{\ell q}S^{qm}+\sum_{q<r}\left(S_{mq}\delta_{jr}+S_{mr}\delta_{jq}\right)\left(S^{\ell q}S^{rm}+S^{mq}S^{r\ell}\right)\right\}\\ &\quad=-\sum_{j=1}^{p}\sum_{m=1}^{p}\sum_{\ell=1}^{p}T_{j\ell}(X,S)\left\{2S_{mj}S^{\ell j}S^{jm}+\sum_{q<r}\left(S_{mq}\delta_{jr}+S_{mr}\delta_{jq}\right)\left(S^{\ell q}S^{rm}+S^{mq}S^{r\ell}\right)\right\}.\end{aligned}$$

Summing on m, using the fact that $\sum_{m=1}^{p}S_{ma}S^{mb}=\delta_{ab}$ and the symmetry of S and S^{-1}, we obtain

$$\begin{aligned}A_3&=-\sum_{j=1}^{p}\sum_{\ell=1}^{p}T_{j\ell}(X,S)\left\{2\,S^{\ell j}+\sum_{q<r}\left(\delta_{jr}S^{r\ell}+\delta_{jq}S^{\ell q}\right)\right\}\\ &=-\sum_{j=1}^{p}\sum_{\ell=1}^{p}T_{j\ell}(X,S)\,S^{\ell j}\left\{2+\sum_{q<r}\left(\delta_{jr}+\delta_{jq}\right)\right\}.\end{aligned}$$

Using the fact that, for any fixed j, $\sum_{q<r}(\delta_{jr}+\delta_{jq})=p-1$ we have $A_3=-(p+1)\operatorname{tr}(TS^{-1})$. Finally, summing $A_1+A_2+A_3$, the lemma follows. □

6.4 Bayes estimators

Let (X,U) be a random vector in $\mathbb{R}^p \times \mathbb{R}^k$ with density

$$\frac{1}{\sigma^{p+k}}\, f\left(\frac{\|x-\theta\|^2+\|u\|^2}{\sigma^2}\right), \tag{6.32}$$

where $\theta \in \mathbb{R}^p$ and $\sigma \in \mathbb{R}_+ \backslash \{0\}$ are unknown. We assume throughout that $p \geq 3$.

We consider generalized Bayes estimators of θ for priors of the form

$$\pi(\|\theta\|^2)\, \eta^b\,, \tag{6.33}$$

where $\eta = 1/\sigma^2$, under the quadratic loss

$$\eta\, \|\delta-\theta\|^2\,. \tag{6.34}$$

We first show that, under weak moment conditions, such generalized Bayes estimators are robust in the sense that they do not depend on the underlying density f. Furthermore, we exhibit a large class of superharmonic priors π for which these generalized Bayes estimators dominate the usual minimax estimator X for the entire class of densities (6.32). Hence this subclass of estimators has the extended robustness property of being simultaneously generalized Bayes and minimax for the entire class of spherically symmetric distributions.

Note that, paralleling Section 2.6, the above model arises as the canonical form of the general linear model $Y = V\beta + \varepsilon$ where V is a $(p+k) \times p$ design matrix, β

is a $p \times 1$ vector of unknown regression coefficients, and ε is an $(p+k) \times 1$ error vector with spherically symmetric density $f(\|\varepsilon\|^2/\sigma^2)/\sigma^{p+k}$.

In the following, for a real valued function $g(x,u)$, we denote by $\bigtriangledown_x g(x,u)$ and $\bigtriangleup_x g(x,u)$ the gradient and the Laplacian of $g(x,u)$ with respect to the variable x. Analogous notations hold with respect to the variable u. When $g(x,u)$ is a vector valued function, $\mathrm{div}_x g(x,u)$ is the divergence with respect to x (here $\dim g(x,u) = \dim x$).

[134] shows that, when the density in (6.32) is normal with known scale, the generalized Bayes estimator corresponding to a prior $\pi(\theta)$, for which the square root of the marginal density $m(x)$ is superharmonic, is minimax under the loss (6.34). Fundamental to this result is the development of an unbiased estimator of risk based on a differential expression involving $m(x)$ which has become a basic tool to prove minimaxity. This differential expression has been extended to non normal models such as (6.32) by several authors (see, for example, [38], [39], [24], [25], [37], [63], [65], [90], [107], [57] and [58]).

A notable aspect of many of the papers dealing with model (6.32), in particular in the presence of a residual vector U, is the development of robust estimators in the sense that they are minimax for a wide class of spherically symmetric distributions (see particularly, for example, [38], [37], [65]).

Another line of research pertinent to this paper is the development of Bayes and generalized Bayes minimax estimators. In the case of a normal distribution with known scale, see Section 4.1, When the scale is unknown, see Section 4.5. For variance mixture of normals with known scale and with no residual vector, see Section 4.5 and for general spherically symmetric distributions.

[108] showed that, for spherically symmetric distributions with a residual vector U and unknown scale parameter, the generalized Bayes estimator with respect to a prior on θ and η proportioned to $\|\theta\|^{2-p}$ (i.e. the fundamental harmonic) is independent of the density f and is minimax under weak moment conditions (see also [112], [109] and [110]).

The goal of this section is to extend the phenomenon in [108] to a broader class of priors of the form $\pi(\|\theta\|^2)\,\eta^b$ with $\pi(\|\theta\|^2)$ superharmonic. In particular, in Subsection 6.4.2, we show that the generalized Bayes estimators do not depend on the density f under weak moment conditions and, in Subsection 6.4.3, we prove that these generalized Bayes estimators are minimax provided the prior $\pi(\|\theta\|^2)$ is superharmonic and its Laplacian $\Delta\pi(\|\theta\|^2)$ is a nondecreasing function of $\|\theta\|^2$, under conditions on b, p and k.

In the case of a known scale parameter in model (1), [57] studied the same class of priors $\pi(\|\theta\|^2)$ and proved minimaxity of generalized Bayes estimators for a large subclass of unimodal densities. We rely strongly on the techniques of that paper.

6.4.1 Risk Considerations

Any estimator $\delta = \delta(X,U)$ of θ is evaluated by its risk associated to the loss (6.34), that is, by

$$R(\theta,\eta,\delta) = E_{\theta,\eta}\left[\eta \,\|\delta(X,U)-\theta\|^2\right], \tag{6.35}$$

where $E_{\theta,\eta}$ denotes the expectation with respect to the density (6.32) with $\eta = 1/\sigma^2$. For the rest of this section, we assume

$$E_{\theta,\eta}\left[\|X-\theta\|^2\right] < \infty, \tag{6.36}$$

which guarantees that the standard estimator X has finite risk and is minimax. As $\delta(X,U)$ can be written as $\delta(X,U) = X + g(X,U)$, the finiteness of its risk is guaranteed by

$$E_{\theta,\eta}\left[\|g(X,U)\|^2\right] < \infty. \tag{6.37}$$

To express the risk difference between $\delta(X,U)$ and X, we introduce first the function F defined, for any $t > 0$, by

$$F(t) = \frac{1}{2}\int_t^\infty f(u)\,du. \tag{6.38}$$

Note that, according to (6.36), we have

$$c = \int_{\mathbb{R}^{p+k}} F(\|x\|^2 + \|u\|^2)\,dx\,du < \infty.$$

A version of the following lemma can be found in [63].

Lemma 6.4. *Assume that the function $g(x,u)$ is weakly differentiable from $\mathbb{R}^{p+k}$ into $\mathbb{R}^p$. Then*

$$\eta E_{\theta,\eta}\left[(X-\theta)'g(X,U)\right] = c E^*_{\theta,\eta}\left[\mathrm{div}_X g(X,U)\right],$$

where $E^*_{\theta,\eta}$ is the expectation with respect to the density

$$\frac{\eta^{p+k}}{c} F\big(\eta\ \big(\|x-\theta\|^2+\|u\|^2\big)\big)\,,$$

provided either of the above expectations exists.

Similarly, for any weakly differentiable function h from $\mathbb{R}^{p+k}$ into $\mathbb{R}^p$,

$$\eta E_{\theta,\eta}\left[U'h(X,U)\right] = c\,E^*_{\theta,\eta}\left[\mathrm{div}_U h(X,U)\right],$$

provided either of these expectations exists.

Thanks to Lemma 6.4, an expression of the risk difference between $\delta(X,U)$ and X is given in the following proposition.

Proposition 6.2. *Assume that $E_{\theta,\eta}\left[\|g(X,U)\|^2\right] < \infty$. The risk difference $\triangle_{\theta,\tau}$ between $\delta(X,U) = X + g(X,U)$ and X equals*

$$\triangle_{\theta,\eta} = \mathscr{R}(\theta,\eta,\delta) - \mathscr{R}(\theta,\eta,X) = c\,E^*_{\theta,\eta}\left[\mathscr{O}g(X,U)\right],$$

where

$$\begin{aligned}&\mathscr{O}g(X,U)\\ &= 2\,\mathrm{div}_X g(X,U) + \frac{k-2}{\|U\|^2}\,\|g(X,\|U\|^2)\|^2 + \frac{U'}{\|U\|^2}\nabla_U\|g(X,U)\|^2\,. \qquad (6.39)\end{aligned}$$

Proof. A straightforward calculation gives

$$\begin{aligned}\triangle_{\theta,\eta} &= E_{\theta,\eta}\left[2\,(X-\theta)'g(X,U) + \|g(X,U)\|^2\right]\\ &= \eta\,E_{\theta,\eta}\left[2\,(X-\theta)'g(X,U) + U'\frac{U}{\|U\|^2}\|g(X,U)\|^2\right].\end{aligned}$$

Using Lemma 6.4 on each term in the brackets, we obtain

$$\begin{aligned}\triangle_{\theta,\eta} &= c\,E^*_{\theta,\eta}\left[2\,\mathrm{div}_X g(X,U) + \mathrm{div}\left(\frac{U}{\|U\|^2}\,\|g(X,U)\|^2\right)\right]\\ &= c\,E^*_{\theta,\eta}\left[2\,\mathrm{div}_X g(X,U) + \frac{k-2}{\|U\|^2}\,\|g(X,U)\|^2 + \frac{U'}{\|U\|^2}\,\nabla_U \|g(X,U)\|^2\right]\end{aligned}$$

by the divergence formula. □

6.4.2 Form of the Bayes Estimators

We will see that, for priors of the form (6.33), the generalized Bayes estimators do not depend on the density (6.32); more precisely their expressions depend only on π and b provided that

$$\int_0^\infty f(\tau)\,\tau^{(p+k)/2+b+1}\,d\tau < \infty, \tag{6.40}$$

which is equivalent to

$$E_{0,1}\left[(\|X\|^2 + \|U\|^2)^{2(b+2)}\right] < \infty.$$

Proposition 6.3. *For a prior of the form (6.33), the generalized Bayes estimator $\delta(X,U) = X + g(X,U)$ is such that, for any $(x,u) \in \mathbb{R}^p \times \mathbb{R}^k$,*

$$g(x,u) = \frac{\int_{\mathbb{R}^p} \frac{\theta - x}{(\|x-\theta\|^2 + \|u\|^2)^{(p+k)/2+b+2}}\,\pi(\|\theta\|^2)\,d\theta}{\int_{\mathbb{R}^p} \frac{1}{(\|x-\theta\|^2 + \|u\|^2)^{(p+k)/2+b+2}}\,\pi(\|\theta\|^2)\,d\theta}, \tag{6.41}$$

provided (6.40) holds and (6.41) exists.

Note that $g(x,u)$ in (6.41) depends on u only through $\|u\|^2$ so that we write $g(x,u) = g(x,\|u\|^2)$. Note also that it arises as

$$\frac{\nabla_x M(x, \|u\|^2)}{m(x, \|u\|^2)} \, ,$$

where $m(x, \|u\|^2)$ is the marginal associated to π and the density

$$\varphi\left(\|x-\theta\|^2 + \|u\|^2\right) \propto \frac{1}{(\|x-\theta\|^2 + \|u\|^2)^{(p+k)/2+b+2}} \, , \tag{6.42}$$

and M is the marginal associated to ϕ with

$$\phi(t) = \frac{1}{2} \int_t^\infty \varphi(v)\, dv \, . \tag{6.43}$$

Therefore, for each fixed u, $\delta(X,u) = X + g(X,u)$ with $g(X,u)$ in (6.41) can be interpreted as the Bayes estimator of θ under the density φ and the prior π for fixed scale parameter $\|u\|$. This observation will be important in the next subsection since it will allow us to use results in [57] which are developed for the case of known scale parameter.

Finally, note that existence of (6.41) will be guaranteed by the stronger finiteness risk condition developed in the proof of Theorem 6.4. More generally, it suffices that π be locally integrable and have tails that do not grow too fast at infinity. In particular, superharmonic priors are locally integrable and have bounded tails.

Proof of Proposition 6.3

The Bayes estimator under loss (6.34) is

$$\delta(X,U) = \frac{E[\eta\, \theta | X, U]}{E[\eta | X, U]} = X + g(X, \|U\|^2),$$

with, for any $(x,u) \in \mathbb{R}^p \times \mathbb{R}^k$,

$$
\begin{aligned}
g(x,\|u\|^2) &= \frac{E[\eta\,(\theta-x)\,|\,x,u]}{E[\eta\,|x,u]} \\
&= \frac{\int_0^\infty \int_{\mathbb{R}^p} \eta\,(\theta-x)\,\eta^{(p+k)/2} f(\eta\,(\|x-\theta\|^2+\|u\|^2))\,\pi(\|\theta\|^2)\,\eta^b\,d\theta\,d\eta}{\int_0^\infty \int_{\mathbb{R}^p} \eta^{(p+k)/2+1} f(\eta\,(\|x-\theta\|^2+\|u\|^2))\,\pi(\|\theta\|^2)\eta^b\,d\theta\,d\eta} \\
&= \frac{\int_{\mathbb{R}^p} \int_0^\infty \eta^{(p+k)/2+b+1} f(\eta\,(\|x-\theta\|^2+\|u\|^2))\,d\eta\,(\theta-x)\,\pi(\|\theta\|^2)\,d\theta}{\int_{\mathbb{R}^p} \int_0^\infty \eta^{(p+k)/2+b+1} f(\eta\,(\|x-\theta\|^2+\|u\|^2))\,d\eta\,\pi(\|\theta\|^2)\,d\theta}\,,
\end{aligned}
$$

by Fubini's theorem. Now, through the change of variable $\tau=\eta\,(\|x-\theta\|^2+\|u\|^2)$ in the innermost integrals, we obtain

$$
\begin{aligned}
g(x,\|u\|^2) &= \frac{\int_{\mathbb{R}^p} \int_0^\infty \tau^{(p+k)/2+b+1} f(\tau)\,d\tau\,\frac{(\theta-x)\,\pi(\|\theta\|^2)}{(\|x-\theta\|^2+\|u\|^2)^{(p+k)/2+b+2}}\,d\theta}{\int_{\mathbb{R}^p} \int_0^\infty \tau^{(p+k)/2+b+1} f(\tau)\,d\tau\,\frac{\pi(\|\theta\|^2)}{(\|x-\theta\|^2+\|u\|^2)^{(p+k)/2+b+2}}\,d\theta} \\
&= \frac{\int_{\mathbb{R}^p} \frac{(\theta-x)\,\pi(\|\theta\|^2)}{(\|x-\theta\|^2+\|u\|^2)^{(p+k)/2+b+2}}\,d\theta}{\int_{\mathbb{R}^p} \frac{\pi(\|\theta\|^2)}{(\|x-\theta\|^2+\|u\|^2)^{(p+k)/2+b+2}}\,d\theta}
\end{aligned}
$$

thanks to (6.40). □

6.4.3 Minimaxity of generalized Bayes estimators

According to the expression of $g(X,U)$ above, we give an expression of the differential operator $\mathscr{O}g(X,U)$ in (6.39). The proof of Proposition 6.4 follows from straightforward calculations.

Proposition 6.4. *For $g(X,\|U\|^2)=\frac{\nabla_X M(X,\|U\|^2)}{m(X,\|U\|^2)}$, (6.39) can be expressed as*

$$
\begin{aligned}
\mathscr{O}g(X,\|U\|^2) = {}& 2\,\frac{\Delta_X M(X,\|U\|^2)}{m(X,\|U\|^2)} - 2\,\frac{\nabla_X m(X,\|U\|^2)'\nabla_X M(X,\|U\|^2)}{m^2(X,\|U\|^2)} \qquad (6.44) \\
& + \frac{k-2}{\|U\|^2}\left\|\frac{\nabla_X M(X,\|U\|^2)}{m(X,\|U\|^2)}\right\|^2 + 2\,\frac{\partial}{\partial s}\left\|\frac{\nabla_X M(X,s)}{m(X,s)}\right\|^2\Bigg|_{s=\|U\|^2},
\end{aligned}
$$

where, for any $(x,u)\in\mathbb{R}^p\times\mathbb{R}^k$,

$$m(x,\|u\|^2) = \int_{\mathbb{R}^p} \varphi(\|x-\theta\|^2 + \|u\|^2)\, \pi(\|\theta\|^2)\, d\theta, \tag{6.45}$$

and

$$M(x,\|u\|^2) = \int_{\mathbb{R}^p} \phi(\|x-\theta\|^2 + \|u\|^2)\, \pi(\|\theta\|^2)\, d\theta \tag{6.46}$$

with φ and ϕ given by (6.42) and (6.43).

[57] studied Bayes minimax estimation of a location vector in the case of spherically symmetric distributions with known scale parameter. For a subclass of spherically symmetric densities, they proved minimaxity of generalized Bayes estimators for spherically symmetric priors of the form $\pi(\|\theta\|^2)$ under the following assumptions.

Assumption 1

(1) $\pi'(\|\theta\|^2) \;\; \leq 0$ i.e. $\pi(\|\theta\|^2)$ is unimodal;

(2) $\Delta\pi(\|\theta\|^2) \;\; \leq 0$ i.e. $\pi(\|\theta\|^2)$ is superharmonic;

(3) $\Delta\pi(\|\theta\|^2)$ is non decreasing in $\|\theta\|^2$.

Note that Condition (2) in fact implies Condition (1) by the mean value property of superharmonic functions.

Our main result below is that a generalized Bayes estimator of θ for a density (6.32), a prior (6.33) and the loss (6.34) is minimax under weak moment conditions and conditions on b, provided the prior satisfies the Assumptions above. We remind the reader that, according to Proposition 6.3, the generalized Bayes estimator is independent of the sampling density, f, provided the assumption (6.40) holds. Hence, each such estimator is simultaneously generalized Bayes and minimax for the entire class of spherically symmetric distributions.

Before developing our minimaxity result, we give a theorem which guarantees the risk finiteness of the generalized Bayes estimators.

Theorem 6.4. *Assume that* π *satisfies Assumption 1.b and that* $b > -k/2+1)$. *Then the generalized Bayes estimator associated to* π *has finite risk.*

Proof. According to (6.41), the risk finiteness condition (6.35) is satisfied as soon as

$$E_{\theta,\eta}\left[\left\|\frac{\displaystyle\int_{R^p}(\theta-X)\frac{\pi(\|\theta\|^2)}{(\|X-\theta\|^2+\|U\|^2)^{(p+k)/2+b+2}}d\theta}{\displaystyle\int_{R^p}\frac{\pi(\|\theta\|^2)}{(\|X-\theta\|^2+\|U\|^2)^{(p+k)/2+b+2}}d\theta}\right\|^2\right]$$

$$\begin{aligned}&\leq E_{\theta,\eta}\left[\frac{\displaystyle\int_{R^p}\|\theta-X\|^2\frac{\pi(\|\theta\|^2)}{(\|X-\theta\|^2+\|U\|^2)^{(p+k)/2+b+2}}d\theta}{\displaystyle\int_{R^p}\frac{\pi(\|\theta\|^2)}{(\|X-\theta\|^2+\|U\|^2)^{(p+k)/2+b+2}}d\theta}\right]\\&<\infty. \qquad (1)\end{aligned}$$

Note that, for any $(x,u)\in\mathbb{R}^p\times\mathbb{R}^k$ and for any nonnegative function h on $\mathbb{R}_+\times\mathbb{R}_+$,

$$\begin{aligned}&\int_{R^p}\pi(\|\theta\|^2)\,h(\|x-\theta\|^2,\|u\|^2)\,d\theta=\\&\int_0^\infty\int_{S_{R,x}}\pi(\|\theta\|^2)\,d\mathscr{U}_{R,x}(\theta)\,\sigma(S)\,R^{p-1}h(R^2,\|u\|^2)\,dR, \qquad (2)\end{aligned}$$

where $\mathscr{U}_{R,x}$ is the uniform distribution on the sphere $S_{R,x}$ of radius R and centered at x and $\sigma(S)$ is the area of the unit sphere. Through the change of variable $R = \sqrt{v}$, the right hand side of (2) can be written as

$$\int_0^\infty \mathscr{S}_\pi(\sqrt{v},x)\, v^{p/2-1}\, h(v,\|U\|^2)\, dv,$$

where

$$\mathscr{S}_\pi(\sqrt{v},x) = \frac{\sigma(S)}{2}\int_{S_{\sqrt{v},x}} \pi(\|\theta\|^2)\, d\mathscr{U}_{\sqrt{v},x}(\theta)$$

is non increasing in v by the superharmonicity of $\pi(\|\theta\|^2)$.

Now we can express the last quantity in brackets in (1) as

$$\begin{aligned}\frac{\displaystyle\int_o^\infty \mathscr{S}_\pi(\sqrt{v},x)\ \frac{v^{p/2}}{(v+\|u\|^2)^{(p+k)/2+b+2}dv}}{\displaystyle\int_o^\infty \mathscr{S}_\pi(\sqrt{v},x)\ \frac{v^{p/2-1}}{(v+\|u\|^2)^{(p+k)/2+b+2}}dv}\\
&= E_1[v]\\
&\leq E_2[v], \qquad (3)\end{aligned}$$

where E_1 is the expectation with respect to the density $f_1(v)$ proportional to

$$\mathscr{S}_\pi(\sqrt{v},x)\,\frac{v^{p/2-1}}{(v+\|u\|^2)^{(p+k)/2+b+2}},$$

and E_2 is the expectation with respect to the density $f_2(v)$ proportional to

$$\frac{v^{p/2-1}}{(v+\|u\|^2)^{(p+k)/2+b+2}}.$$

Indeed the ratio $f_2(v)/f_1(v)$ is nondecreasing by the monotonicity of $\mathscr{S}_\pi(\sqrt{v},x)$. In (3), $E_2[v]$ is

$$E_2[v] = \frac{\int_0^\infty \frac{v^{p/2}}{(v+\|u\|^2)^{(p+k)/2+b+2}}\,dv}{\int_0^\infty \frac{v^{p/2-1}}{(v+\|u\|^2)^{(p+k)/2+b+2}}\,dv}$$

$$= \|u\|^2 \;\frac{\int_0^\infty \frac{v^{p/2}}{(v+1)^{(p+k)/2+b+2}}\,dv}{\int_0^\infty \frac{v^{p/2-1}}{(v+1)^{(p+k)/2+b+2}}\,dv}$$

$$= \|u\|^2 \;\frac{B(p/2+1, k/2+b+1)}{B(p/2, k/2+b+2)},$$

which is finite for $k/2+b+1>0$.

Finally the expectations in (1) are bounded above by $K\,E_{\theta,\eta}[\|U\|^2]$ where K is a constant, and hence are finite. □

We will need the following result which essentially gathers results in Lemma 3.1–3.3 of [57].

Lemma 6.5. *Let $m(x,\|u\|^2)$ and $M(x,\|u\|^2)$ be as defined in (6.45) and (6.46) and let $\cdot$ be the inner product in $\mathbb{R}^p$. Then we have*

(1)

$$x\cdot\nabla_x m(x,\|u\|^2) = -2\int_0^\infty H(v,\|x\|^2)\,v^{p/2}\,\varphi'(v+\|u\|^2)\,dv,$$

and

$$x\cdot\nabla_x M(x,\|u\|^2) = \int_0^\infty H(v,\|x\|^2)\,v^{p/2}\,\varphi(v+\|u\|^2)\,dv,$$

where, for $v>0$,

$$H(v,\|x\|^2) = \lambda(B)\int_{B_{\sqrt{v},x}} x\cdot\theta\,\pi'(\|\theta\|^2)\,dV_{\sqrt{v},x}(\theta) \qquad (6.47)$$

and $V_{\sqrt{v},x}$ is the uniform distribution on the ball $B_{\sqrt{v},x}$ of radius $\sqrt{v}$ centered at x and $\lambda(B)$ is the volume of the unit ball;

(2) *For any $x \in \mathbb{R}^p$, the function $H(v, \|x\|^2)$ in (6.47) is nondecreasing in v provided that $\Delta\pi(\|\theta\|^2)$ is nondecreasing in $\|\theta\|^2$. (Assumption 1.c);*

(3) *For any $v > 0$ and any $x \in \mathbb{R}^p$, the function $H(v, \|x\|^2)$ in (6.47) is non positive provided $\pi'(\|\theta\|^2) \leq 0$. (Assumption 1.a).*

Given these preliminaries, we present our main result.

Theorem 6.5. *Suppose that π satisfies Assumptions 1. Then the generalized Bayes estimator associated to $\pi(\|\theta\|^2)\,\eta^b$ is minimax provided that $b \geq \frac{2p-k-2}{4}$ and the assumptions of Theorem 6.4 are satisfied.*

Proof. It suffices to show that $\mathscr{O}g(X,U)$ in (6.43), with $m(X, \|U\|^2)$ and $M(X, \|U\|^2)$ given respectively by (6.44) and (6.46), is non positive since the assumptions guarantee that the generalized Bayes estimator δ is of the form $\delta(X,U) = X + \nabla_X M(X, \|U\|^2)/m(X, \|U\|^2)$ and has finite risk.

Due to the superharmonicity of $\pi(\|\theta\|^2)$, for any $(x,u) \in \mathbb{R}^p \times \mathbb{R}^k$, we have $\Delta_x M(x\|u\|^2) \leq 0$ so that

$$\mathscr{O}g(x, \|u\|^2) \leq -2\, \frac{\nabla_x m(x, \|u\|^2)' \nabla_x M(x, \|u\|^2)}{m^2(x, \|u\|^2)} + \frac{k-2}{\|u\|^2} \left\| \frac{\nabla_x M(x, \|u\|^2)}{m(x, \|u\|^2)} \right\|^2 + 2 \frac{\partial}{\partial s} \left\| \frac{\nabla_x M(x,s)}{m(x,s)} \right\|^2 \Bigg|_{s=\|u\|^2} .$$

Note that

$$\begin{aligned} m^2(x,s)\, \frac{\partial}{\partial s} \left\| \frac{\nabla_x M(x,s)}{m(x,s)} \right\|^2 &= \frac{\partial}{\partial s}\, \|\nabla_x M(x,s)\|^2 + \|\nabla_x M(x,s)\|^2\, m^2(x,s)\, \frac{\partial}{\partial s} \frac{1}{m^2(x,s)} \\ &\leq \frac{\partial}{\partial s}\, \|\nabla_x M(x,s)\|^2 + (p+k+2b+4)\, \frac{1}{s}\, \|\nabla_x M(x,s)\|^2 , \end{aligned}$$

since

$$\begin{aligned}
\frac{\partial}{\partial s}\frac{1}{m^2(x,s)} &= \frac{-2}{m^3(x,s)}\int_{\mathbb{R}^p}\frac{-[(p+k)/2+b+2]}{(\|x-\theta\|^2+s)^{(p+k)/2+b+3}}\,\pi(\|\theta\|^2)\,d\theta \\
&= \frac{p+k+2b+4}{m^3(x,s)}\,\frac{1}{s}\int_{\mathbb{R}^p}\frac{s}{\|x-\theta\|^2+s}\;\frac{1}{(\|x-\theta\|^2+s)^{(p+k)/2+b+2}}\,\pi(\|\theta\|^2)\,d\theta \\
&\leq \frac{p+k+2b+4}{m^2(x,s)}\,\frac{1}{s}\,.
\end{aligned}$$

Therefore

$$\begin{aligned}
m^2(x,s)\,\mathscr{O}g(x,s) \leq & -2\,\nabla_x m(x,s)'\nabla_x M(x,s) \\
& + \frac{k-2+2(p+k+2b+4)}{s}\,\|\nabla_x M(x,s)\|^2 \\
& + 2\frac{\partial}{\partial s}\,\|\nabla_x M(x,s)\|^2\,.
\end{aligned} \tag{6.48}$$

As $m(x,s)$ and $M(x,s)$ depend on x only through $\|x\|^2$, it is easy to check that (as in [57])

$$\nabla_x m(x,s)'\nabla_x M(x,s) = \frac{x'\nabla_x m(x,s)\,x'\nabla_x M(x,s)}{\|x\|^2}$$

and

$$\|\nabla_x M(x,s)\|^2 = \frac{(x'\nabla_x M(x,s))^2}{\|x\|^2}\,.$$

Thus the right hand side of (6.48) will be nonpositive as soon as

$$-2x'\nabla_x m(x,s) + \frac{2p+3k+4b+6}{s}\,x'\nabla_x M(x,s) + 4\frac{\partial}{\partial s}x'\nabla_x M(x,s) \geq 0, \tag{6.49}$$

since, according to Lemma 6.5, the common factor $x'\nabla_x M(x,s)$ is nonpositive. Using again Lemma 6.5, the left hand side of (6.49) equals

$$4\int_0^\infty H(v,\|x\|^2)\,v^{p/2}\,\varphi'(v+s)\,dv + \frac{2p+3k+4b+6}{s}\int_0^\infty H(v,\|x\|^2)\,v^{p/2}\,\varphi(v+s)\,dv$$
$$+4\int_o^\infty H(v,\|x\|^2)v^{p/2}\varphi'(v+s)dv$$
$$=\int_0^\infty v^{p/2}\,\varphi(v+s)\,dv\left\{8E\left[H(v,\|x\|^2)\,\frac{\varphi'(v+s)}{\varphi(v+s)}\right]\right.$$
$$\left.+\frac{2p+3k+4b+6}{s}E\left[H(v,\|x\|^2)\right]\right\}, \qquad (6.50)$$

where E denotes the expectation with respect to the density proportional to $v \longmapsto v^{p/2}\,\varphi(v+s)$.

As

$$\frac{\varphi'(v+s)}{\varphi(v+s)} = \frac{-((p+k)/2+b+2)}{v+s} \qquad (6.51)$$

is nondecreasing in v and, according to Lemma 6.5, $H(v,\|x\|^2)$ is also nondecreasing in v, the first expectation in (6.50) satisfies

$$E\left[H(v,\|x\|^2)\,\frac{\varphi'(v+s)}{\varphi(v+s)}\right] \geq E\left[H(v,\|x\|^2)\right]E\left[\frac{\varphi'(v+s)}{\varphi(v+s)}\right]$$

by the covariance inequality. Therefore Inequality (6.49) will be satisfied as soon as

$$8E\left[\frac{\varphi'(v+s)}{\varphi(v+s)}\right] + \frac{2p+3k+4b+6}{s} \leq 0, \qquad (6.52)$$

since $H(v,\|x\|^2) \leq 0$ by Lemma 6.5.

From (6.51) we have

$$E\left[\frac{\varphi'(v+s)}{\varphi(v+s)}\right] = -\big((p+k)/2+b+2\big)\,E\left[\frac{1}{v+s}\right] \tag{6.53}$$

$$= -\big((p+k)/2+b+2\big)\,\frac{\int_o^\infty \frac{1}{v+s}\; v^{p/2}\,\frac{1}{(v+s)^{(p+k)/2+b+2}}\,dv}{\int_0^\infty v^{p/2}\,\frac{1}{(v+s)^{(p+k)/2+b+2}}\,dv}$$

$$= -\big((p+k)/2+b+2\big)\,\frac{1}{s}\,\frac{\int_0^\infty \frac{z^{p/2}}{(z+1)^{(p+k)/2+b+3}}\,dz}{\int_0^\infty \frac{z^{p/2}}{(z+1)^{(p+k)/2+b+2}}\,dz}$$

$$= -\big((p+k)/2+b+2\big)\,\frac{1}{s}\,\frac{B(p/2+1,k/2+b+2)}{B(p/2+1,k/2+b+1)},$$

where $B(\alpha,\beta)$ is the beta function with parameters $\alpha > 0$ and $\beta > 0$. Then (6.53) becomes

$$\begin{aligned} E\left[\frac{\varphi'(v+s)}{\varphi(v+s)}\right] &= -\frac{((p+k)/2+b+2)}{s}\,\frac{\Gamma((k/2+b+2))}{\Gamma((p+k)/2+b+3)}\,\frac{\Gamma((p+k)/2+b+2)}{\Gamma(k/2+b+1)} \\ &= \frac{-(k/2+b+1)}{s}. \end{aligned} \tag{6.54}$$

It follows from (6.54) that (6.52) reduces to

$$b \geq \frac{2p-k-2}{4},$$

which is the condition given in the theorem. □

The condition on b in Theorem 6.5 can be alternatively expressed as $k \geq 2p - 4b - 2$ which dictates that the dimension, k, of the residual vector, U, increases with the dimension, p, of θ. This dependence can be (essentially) eliminated provided the generalized Bayes estimator in Proposition 6.3 satisfies the following assumption:

Assumption 2

The function $g(x,u)$ in (6.41) can be expressed as

$$g(x,u) = \frac{\nabla_x M(x,\|u\|^2)}{m(x,\|u\|^2)} = -\frac{r(\|x\|^2,\|u\|^2)\,\|u\|^2}{\|x\|^2}\,x,$$

where $r(\|x\|^2,\|u\|^2)$ is nonnegative and nonincreasing in $\|u\|^2$.

Assumption 2 is satisfied, for example, by the generalized Bayes estimator corresponding to the prior on (θ,η) proportional to $\pi(\|\theta\|^2) = \left(1/\|\theta\|^2\right)^{-b/2}\eta^a$ for $0 < b \leq p-2$ and $a > -\frac{k}{2} - \frac{b}{2} - 2$, in which case the function $r(\|x\|^2,\|u\|^2) = \phi\left(\|x\|^2/\|u\|^2\right)$, where $\phi(t)$ is increasing in t, and hence $r(\|x\|^2,\|u\|^2)$ is decreasing in $\|u\|^2$ (see, [108]).

We have the following corollary.

Corollary 6.2. *Suppose π satisfies Assumptions 1 and the assumptions of Theorem 6.5 and suppose also that the generalized Bayes estimator (which does not depend on the underlying density f) satisfies Assumption 2. Then the generalized Bayes estimator is minimax provided $b \geq -(k+2)/4$.*

Proof. Assumption 2 guarantees that

$$\frac{\partial}{\partial s}\left(\frac{1}{s^2}\left\|\frac{\nabla_x M(x,s)}{m(x,s)}\right\|^2\right) = \frac{\partial}{\partial s}\left(\frac{r^2(\|x\|^2,s)}{\|x\|^2}\right) \leq 0.$$

Since

$$\frac{\partial}{\partial s}\left\|\frac{\nabla_x M(x,s)}{m(x,s)}\right\|^2 = \frac{\partial}{\partial s}\left(\frac{s^2}{s^2}\left\|\frac{\nabla_x M(x,s)}{m(x,s)}\right\|^2\right) = \frac{2}{s}\left\|\frac{\nabla_x M(x,s)}{m(x,s)}\right\|^2 + s^2\frac{\partial}{\partial s}\left(\frac{1}{s^2}\left\|\frac{\nabla_x M(x,s)}{m(x,s)}\right\|^2\right),$$

the inequality for $\mathscr{O}g(X,\|U\|^2)$ in the proof of Theorem 6.5 can be replaced by

$$\mathscr{O}g(x,\|u\|^2) \leq -2\,\frac{\nabla_x m(x,\|u\|^2)'\nabla_x M(x,\|u\|^2)}{m^2(x,\|u\|^2)} + \frac{k+2}{\|u\|^2}\left\|\frac{\nabla_x M(x,\|u\|^2)}{m(x,\|u\|^2)}\right\|^2.$$

It follows that inequality condition (6.49) becomes

$$-2x'\nabla_x m(x,s) + \frac{k+2}{s} x'\nabla_x M(x,s) \geq 0,$$

and that inequality condition (6.52) becomes

$$4E\left[\frac{\varphi'(v+s)}{\varphi(v+s)}\right] + \frac{k+2}{s} \leq 0,$$

which, by (6.54), becomes

$$4\left[-\left(\frac{k/2+b+1}{s}\right)\right] + \frac{k+2}{s} \leq 0,$$

which is equivalent to $b \geq -(k+2)/4$. □

6.4.4 Examples

Several examples of priors which satisfies Assumptions 1 are given in [57]. We briefly summarize these.

Example 6.1. **Priors related to the fundamental harmonic prior**

Let $\pi(\|\theta\|^2) = \left(\dfrac{1}{A+\|\theta\|^2}\right)^c$ with $A \geq 0$ and $0 \leq c \leq \frac{p}{2} - 1$.

Example 6.2. **Mixtures of priors satisfying Assumption 2**

Let $(\pi_\alpha)_{\alpha \varepsilon A}$ be a family of priors such that Assumption 2.1 is satisfied for any $\alpha \in A$. Then any mixture of the form $\int_A \pi_\alpha(\|\theta\|^2)\, dH(\alpha)$ where H is a probability on A satisfies Assumption 1 as well. For instance, Example 6.1 with $c = 1, p \geq 4, A = \alpha$ and the gamma density $\alpha \longmapsto \dfrac{\beta^{1-v}}{\Gamma(1-v)} \alpha^{-v} e^{-\beta\alpha}$ with $\beta > 0$ and $0 < v < 1$ leads to the prior

$$\|\theta\|^{-2-v} e^{\beta\|\theta\|^2} \Gamma(v, \beta\|\theta\|^2),$$

where

$$\Gamma(v,y) = \int_y^\infty e^{-x} x^{v-1}\, dx$$

is the complement of the incomplete gamma function.

Example 6.3. **Variance mixtures of normals**

Let

$$\pi(\|\theta\|^2) = \int_O^\infty \left(\frac{u}{2\pi}\right)^{p/2} \exp\left(\frac{-u\|\theta\|^2}{2}\right) h(u)du$$

a mixture of normals with respect to the inverse of the variance. As soon as, for any $u > 0$,

$$\frac{uh'(u)}{h(u)} \leq -2,$$

the prior $\pi(\|\theta\|^2)$ satisfies Assumptions 1. Note that the priors in Example 6.3 arise as such a mixture with $h(u) \propto \alpha u^{k-p/2-1} \exp(-A/2u)$.

Other examples can be given and a constructive approach is proposed in [57].

Chapter 7

Restricted Parameter Spaces

7.1 Introduction

In this chapter, we will consider the problem of estimating a location vector which is constrained to lie in a convex subset of $\mathbb{R}^P$. Much of the chapter is devoted to one of two types of constraint sets, balls (and spheres), and polyhedral cones. However, Section 7.2 is devoted to general convex constraint sets and more particularly to a striking result of Hartigan which shows that in the normal case, the Bayes estimator of the mean with respect to the uniform prior over any convex set, $\mathscr{C}$, dominates X for all $\theta \in \mathscr{C}$ under the usual quadratic loss $\|\delta - \theta\|^2$.

Section 7.3 will consider estimation of a normal mean vector restricted to a polyhedral cone, $\mathscr{C}$, in the normal case under quadratic loss. Both the cases of known and unknown scale are treated. Special methods need to be developed to handle this restriction because the shrinkage functions considered are not necessarily weakly differentiable. Hence the methods of Chapter 4 are not directly applicable. A version of Stein's lemma is developed for positively homogeneous sets which allows the analysis to proceed.

In general, if the constraint set, $\mathscr{C}$, is convex, a natural alternative to the UMYUE X, is $P_c X$ the projection of X onto $\mathscr{C}$. Our methods lead to Stein type shrinkage estimators that shrink $P_c X$ which dominate $P_c X$, and hence X itself, when $\mathscr{C}$ is a polyhedral cone.

Section 7.4 considers the situation where X is normal with a known scale but the constraint set is a ball, B, of known radius centered at a known point in $\mathbb{R}^p$. Here again, a natural estimator to dominate is the projection onto the Ball $P_B X$. Hartigan's result of course applies and shows that the Bayes estimat corresponding to the uniform prior dominates X, but a finer analysis lead to domination over $P_B X$ (provided the radius of the ball is not too large relative to the dimension) by the Bayes estimator corresponding to the uniform prior on the sphere of the same radius.

Section 7.5 is devoted to the case of a general spherically symmetric distribution with a residual vector when the mean vector is restricted to a polyhedral cone. As in Section 7.3, the potential non differentiability of the shrinkage factors is a complication. We develop a general method that allows the results of Section 7.3 for the normal case to be extended to the general spherically symmetric case as long as a residual vector is available. This method also allows for an alternative development of some of the results of Chapter 4 that rely on an extension of Stein's Lemma to the general spherical case.

Section **??** is devoted to the case of a spherically symmetric distribution where the mean is restricted to lie in a ball. It is shown that some of the results developed in Section 7.4 can be extended to certain scale mixtures of normal distributions including the multivariate-t with sufficiently large degrees of freedom.

7.2 Normal Mean Vector Restricted to a Convex Set

In this Section, we treat the case $X \sim N(\theta, \sigma^2 I)$ where σ^2 is known and where the unknown mean θ is restricted to lie in *a* convex set $\mathscr{C}$ (with non empty interior and sufficiently regular boundary), and where the loss is $L(\theta, \delta) = \|\delta - \theta\|^2$. We show that the (generalized) Bayes estimator with respect to the uniform prior distribution on $\mathscr{C}$, say $\pi(\theta) = \mathbb{1}_{\mathscr{C}}(\theta)$, dominates the usual (unrestricted) estimator $\delta_0(X) = X$. At this level of generality the result is due to [72] although versions of the result (in $\mathbb{R}^1$) date back to [82]. We follow the discussion in [105]

Theorem 7.1. *([72]) Let $X \sim N(\theta, \sigma^2 I)$ with σ^2 known and $\theta \in \mathscr{C}$, a convex set with non-empty interior and sufficiently regular boundary $\partial\mathscr{C}$ ($\partial\mathscr{C}$ Lipshitz (1) suffices). Then the Bayes estimator, $\delta_U(X)$ with respect to the uniform prior on $\mathscr{C}, \pi(\theta) = \mathbb{1}_{\mathscr{C}}(\theta)$, dominates $\delta_0(X) = X$ with respect to quadratic loss.*

Proof. Without loss of generality, assume $\sigma^2 = 1$. Recall (see 2.12 e.g.) that $\delta_U(X) = X + \nabla m(X)/m(X)$ where, for any $x \in \mathbb{R}^p$,

$$m(x) \propto \int_{\mathscr{C}} \exp\left(-\frac{1}{2}\|x - c\|^2\right) d\nu \, .$$

The difference in risk between δ_U and δ_0 is $R(\theta, \delta_U) - R(\theta, \delta_0)$

$$\begin{aligned} R(\theta,\delta_U)-R(\theta,\delta_0) &= E_\theta\left[\left\|X+\frac{\nabla m(X)}{m(X)}-\theta\right\|^2-\|X-\theta\|^2\right] \\ &= E_\theta\left[\frac{\|\nabla m(X)\|^2}{m^2(X)}+2\,\frac{\nabla m(X)'(X-\theta)}{m(X)}\right]. \qquad (7.1) \end{aligned}$$

Hartigan's clever demonstration proceeds by applying Stein's Lemma 3.1 to only half of the cross product term in order to cancel the squared norm term in the above. Indeed, since

$$\begin{aligned} E_\theta\left[(X-\theta)'\left(\frac{\nabla m(X)}{m(X)}\right)\right] &= E_\theta\left[\operatorname{div}\left(\frac{\nabla m(X)}{m(X)}\right)\right] \\ &= E_\theta\left[\frac{\Delta m(X)}{m(X)}-\frac{\|\nabla m(X)\|^2}{m^2(X)}\right], \end{aligned}$$

(7.1) becomes

$$\begin{aligned} R(\theta,\delta_U)-R(\theta,\delta_0) &= E_\theta\left[\frac{\Delta m(X)+(X-\theta)'\nabla m(X)}{m(X)}\right] \\ &= E_\theta\left[\frac{H(X,\theta)}{m(X)}\right] \qquad (7.2) \end{aligned}$$

with $H(x,\theta)=\Delta m(x)+(x-\theta)'\nabla m(x)$. Hence it suffices to show $H(x,\theta)\le 0$ for all $\theta\in\mathscr{C}$ and $x\in\mathbb{R}^p$. Using the facts that

$$\nabla_x \exp\left(-\frac{1}{2}\|x-\nu\|^2\right) = -\nabla_\nu \exp\left(-\frac{1}{2}\|x-\nu\|^2\right)$$

and

$$\Delta_x \exp\left(-\frac{1}{2}\|x-\nu\|^2\right) = \Delta_\nu \exp\left(-\frac{1}{2}\|x-\nu\|^2\right),$$

it follows that

$$\begin{aligned} H(x,\theta) &\propto \Delta_x \int_{\mathscr{C}} \exp\left(-\frac{1}{2}\|x-\nu\|^2\right) d\nu + (x-\theta)'\nabla_x \int_{\mathscr{C}} \exp\left(\frac{1}{2}\|x-\nu\|^2\right) d\nu \\ &= \int_{\mathscr{C}} \left[\Delta_\nu \exp\left(-\frac{1}{2}\|x-\nu\|^2\right) - (x-\theta)'\nabla_\nu \exp\left(-\frac{1}{2}\|x-\nu\|^2\right)\right] d\nu \\ &= \int_{\mathscr{C}} \operatorname{div}_\nu \left[\nabla_\nu \exp\left(-\frac{1}{2}\|x-\nu\|^2\right) - (x-\theta)\exp\left(-\frac{1}{2}\|x-\nu\|^2\right)\right] d\nu \\ &= \int_{\mathscr{C}} \operatorname{div}_\nu \left[(\theta-\nu)\exp\left(-\frac{1}{2}\|x-\nu\|^2\right)\right] d\nu\,. \end{aligned}$$

By Stokes' Theorem (See appendix) this last expression can be expressed as

$$\int_{\partial\mathscr{C}} \eta'(\nu)(\theta-\nu)\exp\left(-\frac{1}{2}\|x-\nu\|^2\right) d\sigma(\nu)$$

where $\eta(\nu)$ is the unit outward normal to $\partial\mathscr{C}$ at ν and σ is the surface area Lebesgue measure on $\partial\mathscr{C}$. Finally, since $\mathscr{C}$ is convex and $\theta \in \mathscr{C}$, the angle between $\eta(\nu)$ and $\theta-\nu$ is obtuse for $\nu \in \partial\mathscr{C}$ and so $\eta'(\nu)(\theta-\nu) \leq 0$, for all $\theta \in \mathscr{C}$ and $\nu \in \partial\mathscr{C}$, which implies the risk difference in (7.2) is nonpositive. □

Note that, if θ is in the interior of $\mathscr{C}, \mathscr{C}^0$, then $\eta'(\nu)(\theta-\nu)$ is strictly negative for all $\nu \in \partial\mathscr{C}$, and hence, $R(\theta,\delta_U) - R(\theta\delta_0) < 0$ for all $\theta \in \mathscr{C}^0$. However, if $\mathscr{C}$ is a pointed cone at θ_0, then $\eta'(\nu)(\theta_0-\nu) \equiv 0$ for all $\nu \in \partial\mathscr{C}$ and $R(\theta_0,\delta_U) = R(\theta_0,\delta_0)$.

Note also that, if $\mathscr{C}$ is compact, the uniform prior on $\mathscr{C}$ is proper, and hence, $\delta_U(X)$ not only dominates $\delta_0(X)$ (on $\mathscr{C}$) but is also admissible for all p.

On the other hand, if $\mathscr{C}$ is not compact, it is often (typically for $p \geq 3$) the case that δ_U is not admissible, and alternative shrinkage estimators may be desirable.

Furthermore, it may be argued in general, that a more natural basic estimator which one should seek to dominate is $P_c X$, the projection of X onto $\mathscr{C}$ which is the MLE. We consider this problem for the case where $\mathscr{C}$ is a cone in Section 7.3 and where $\mathscr{C}$ is a ball in Section 7.4. □

7.3 Normal mean Vector Restricted to a Polyhedral Cone

In this Section, we consider first the case when $X \sim N_p(\theta, \sigma^2 I)$ where σ^2 is known and θ is restricted to a polyhedral cone $\mathscr{C}$. Later in this Section, we will consider the case where σ^2 is unknown and, in Section 7.5, the general spherically symmetric case with a residual vector. The reader is referred to [65] for more details.

A natural estimator in this problem is $\delta_{\mathscr{C}}(X) = P_{\mathscr{C}}X$, the projection of X onto the cone $\mathscr{C}$. The estimator $\delta_{\mathscr{C}}$ is the MLE and dominates X which is itself minimax by .... (?). Our goal will be to dominate $\delta_{\mathscr{C}}$ and therefore also $\delta_0(X) = X$.

We refer the reader to [135] and Robertson et al. (1988) for an extended discussion of polyhedral cones. Here is a brief summary. A polyhedral cone $\mathscr{C}$ is defined as the intersection of a finite number of half spaces, that is,

$$\mathscr{C} = \{x \,/\, a_i' x \leq 0, i = 1, \ldots, n\}$$

for n fixed vectors $a_1, \ldots, a_n$ in $\mathbb{R}^p$.

It is positively homogeneous, closed and convex, and, for each $x \in R^p$, there exists a unique point $P_{\mathscr{C}}x$ in $\mathscr{C}$ such that $\|P_{\mathscr{C}}x - x\| = \inf_{y \in \mathscr{C}} \|y - x\|$.

We assume throughout that $\mathscr{C}$ has non empty interior, $\mathscr{C}^0$ so that $\mathscr{C}$ may be partitioned into $\mathscr{C}_i$, $i = 0, \ldots, n$, where $\mathscr{C}_0 = \mathscr{C}^0$ and $\mathscr{C}_i$, $i = 1, \ldots, n$, are the relative interiors of the proper faces of $\mathscr{C}$. For each set $\mathscr{C}_i$, let $D_i = P_{\mathscr{C}}^{-1}\mathscr{C}_i$ (the pre-image of $\mathscr{C}_i$ under the projection operator $P_{\mathscr{C}}$ and $s_i = \dim \mathscr{C}_i)$. Then $D_i, i = 0, \ldots, n$ form a partition of $\mathbb{R}^p$, where $D_0 = \mathscr{C}_0$.

For each $x \in C_i$, we have $P_{\mathscr{C}} x = P_i x$ where P_i is the orthogonal linear projection onto the s_i-dimensional subspace L_i spanned by $\mathscr{C}_i$. Also for each such x, $P_i^{\perp} x$, the orthogonal projection onto $L_i^{\perp}$, is equal to $P_{\mathscr{C}^*} x$ where $\mathscr{C}^*$ is the polar cone corresponding to $\mathscr{C}$. Additionally, if $x \in C_i$, then $P_i x + P_i^{\perp} \in D_i$ is positively homogeneous in $P_i x$ for each fixed $P_i^{\perp} x$ (see Robertson et al. (1988), Theorem 8.2.7). Hence we may express

$$\delta_{\mathscr{C}}(X) = \sum_{i=0}^{n} \mathbb{1}_{P_i}(X)\, P_i X\,. \tag{7.3}$$

The problem of dominating $\delta_{\mathscr{C}}$ is relatively simple in the case where $\mathscr{C}$ has the form $\mathscr{C} = \mathbb{R}_+^k \oplus \mathbb{R}^{p-k}$ where $\mathbb{R}_+^k = \{(x_1, \ldots, x_k)\,/\, x_i \geq 0,\ i = 1, \ldots, k\}$. In this case,

$$\delta_{\mathscr{C}}(X)_i = \begin{cases} X_i \text{ if } X_i \geq 0 \\ 0 \ \text{ if } X_i < 0 \text{ for } i = 1, \ldots k \text{ and } \delta_{\mathscr{C}}(X)_i = X_i \text{ for } i = k+1, \ldots, p. \end{cases}$$

Furthermore, $\delta_{\mathscr{C}}(X)$ is weakly differentiable and the techniques of Chapter 3 (i.e. Stein's Lemma) are available.

As a simple example, suppose $\mathscr{C} = \mathbb{R}_+^p$, i.e. all coordinates of θ are nonnegative. Then

$$\delta_{\mathscr{C}_i}(X) = X_i + \partial_i(X) \quad i = 1, \ldots, p$$

where

$$\partial_i(X) = \begin{cases} -X_i \text{ if } X_i < 0 \\ \ \ 0 \text{ if } X_i \geq 0 \end{cases}$$

Also, we may rewrite **??** as $X_+ = \sum_{i=1}^{2^p} \mathbb{1}_{O_i}(X)\, P_i X$ where $O_1 = \mathbb{R}_+^p$, and O_j, for $j > 1$, represent the other $2^p - 1$ orthants and P_i is the projection of X onto the space generated by the face of O_1 adjacent to O_i.

Then a James-Stein type shrinkage estimator that dominates X_+ is given by

$$\delta(X) = \sum_{i=1}^{2^p} \left(1 - \frac{c_i}{\|X_+\|^2}\right) X_+ \, \mathbb{1}_{0_i}(X)$$

where $c_i = (s_i - 2)_+$ and s_i is the number of positive coordinates in O_i. Note that shrinkage occurs only in those orthants such that $s_i \geq 3$.

The proof of domination follows essentially by the usual argument of Chapter 3, Section **??**, applied separately to each orthant since X_+ and $X_+/\|X_+\|^2$ are weakly differentiable in O_i and

$$\nabla^{\cdot}_X \frac{X_+}{\|X_+\|^2} \, \mathbb{1}_{O_i}(X) = \frac{s_i - 2}{\|X_+\|^2} \, \mathbb{1}_{O_i}(X) \, ,$$

provided $s_i > 2$. Note also that a_i may take any value between 0 and $2\,(s_i - 2)_+$.

Difficulties arise when the cone $\mathscr{C}$ is not of the form $\mathscr{C} = \mathbb{R}^k_+ \oplus \mathbb{R}^{p-k}$ because the estimator $P_{\mathscr{C}}X$ may not be weakly differentiable (see appendix . . .). In this case, a result of [128] can be used to give an unbiased estimator of the risk. Here is a version of their result.

Lemma 7.1. *([128]) Let $X \sim N(\theta, \sigma^2 I)$ and $\mathscr{C}$ a positively homogeneous set. Then for every absolutely continuous function $h(\cdot)$ from $\mathbb{R}_+$ to $\mathbb{R}$ such that*

$$\forall k \geq 0 \quad \lim_{y \to 0, \infty} h(y)\, y^{k+p/2}\, e^{-y/2} = 0$$

and such that

$$E_\theta[h^2(\|X\|^2)\|X\|^2] < \infty$$

we have

$$\begin{aligned} E_\theta[h(\|X\|^2)\,X'(X-\theta)\,\mathbb{1}_{\mathscr{C}}(X)] &= \sigma^2 E_\theta[2\,\|X\|^2\,h'(\|X\|^2)+p\,h(\|X\|^2)\,\mathbb{1}_{\mathscr{C}}(X)] \\ &= \sigma^2 E_\theta[\mathrm{div}(h(\|X\|^2)\,X)\,\mathbb{1}_{\mathscr{C}}(X)]. \end{aligned}$$

Note that for $\mathscr{C}=\mathbb{R}^p$, Lemma 7.1 follows from Stein's Lemma with $g(X)=h(\|X\|^2)\,X$ provided $E[h(\|X\|^2)\,\|X\|^2]<\infty$. The possible non-weak differentiability of the function $h(\|X\|^2)\,X\,\mathbb{1}_{\mathscr{C}}(X)$ prevents a direct use of Stein's lemma for general $\mathscr{C}$.

PROOF OF LEMMA 7.1 Note first that, if, for any θ, $E_\theta[\|g(X)\|]<\infty$, then

$$E_0\left[g(X)\,e^{X'\theta/\sigma^2}\right]=\sum_{k=0}^{\infty}E_0\left[\frac{g(X)\,X'\theta/\sigma^{2k}}{k!}\right]$$

by the dominated convergence Theorem. Without loss of generality, assume $\sigma^2=1$ and let

$$\begin{aligned} A_\theta &= E_\theta[h(\|X\|^2)\,X'(X-\theta)\,\mathbb{1}_{\mathscr{C}}(X)] \qquad (7.4)\\ &= (2\pi)^{-p/2}\,e^{-\|\theta\|^2/2}\int_{R^p} e^{-\|X\|^2/2}\,e^{X'\theta}\,h(\|X\|^2)\,(\|X\|^2-X'\theta)\,\mathbb{1}_{\mathscr{C}}(X)dx \\ &= (2\pi)^{-p/2}\,e^{-\|\theta\|^2/2}\sum_{k=0}^{\infty}E_0\left[h(\|X\|^2)(\|X\|^2-X'\theta)\frac{(X'\theta)^k}{k!}\,\mathbb{1}_{\mathscr{C}}(X)\right] \\ &= (2\pi)^{-p/2}\,e^{-\|\theta\|^2/2}\sum_{k=0}^{\infty}\frac{1}{k!}E_0\Big[h(\|X\|^2)\,\mathbb{1}_{\mathscr{C}}(X)(X'\theta)^k(\|X\|^2-k)\Big] \\ &= (2\pi)^{-p/2}\,e^{-\|\theta\|^2/2}\sum_{k=0}^{\infty}\frac{1}{k!}E_0\left[h(\|X\|^2)\,\mathbb{1}_{\mathscr{C}}\left(\frac{X'\theta}{\|X\|}\right)^k(\|X\|^{k+2}-k\|X\|^k)\right] \end{aligned}$$

By the positive homogeneity of $\mathscr{C}$, we have $\mathbb{1}_{\mathscr{C}}(X)=\mathbb{1}_{\mathscr{C}}(X/\|X\|)$ and, by the independence of $\|X\|$ and $X/\|X\|$ for $\theta=0$, we have

$$\begin{aligned} A(\theta) = (2\pi)^{-p/2} e^{-\|\theta\|^2/2} \sum_{k=0}^{\infty} \frac{1}{k!} E_0\left[\left(\frac{X'\theta}{\|X\|}\right)^k \mathbb{1}_{\mathscr{C}}\left(\frac{X}{\|X\|}\right)\right] \\ E_0\Big[h(\|X\|^2)\Big(\|X\|^{k+2} - k\,\|X\|^k\Big)\Big] \end{aligned} \tag{7.5}$$

When $\theta = 0$, $\|X\|^2$ has central Chi-square distribution with p degrees of freedom. Thus, with $d = 1/2^{p/2}\,\Gamma(p/2)$, we have

$$\begin{aligned} E_0[h(\|X\|^2)(\|X\|^{k+2} - k\|X\|^k)] &= d\int_0^{\infty} y^{p/2-1}\, h(y)\,(y^{(k+2)/2} - k\, y^{k/2})\, e^{-y/2}\, dy \\ &= d\int_0^{\infty} y^{(p+k)/2}\, h(y)\, e^{-y/2}\, dy - d\, k\int_0^{\infty} y^{(p+k)/2-1}\, h(y)\, e^{-y/2}\, dy \end{aligned}$$

Integrating by parts the first integral gives

$$\begin{aligned} &\int_0^{\infty} y^{(p+k)/2}\, h(y)\, e^{-y/2}\, dy \\ &= 2\left[\int_0^{\infty} \frac{p+k}{2}\, y^{(p+k)/2-1}\, h(y)\, e^{-y/2}\, dy + \int_0^{\infty} y^{(p+k)/2}\, h'(y)\, e^{-y/2}\, dy\right] \end{aligned}$$

and thus combining gives

$$\begin{aligned} E_0[h(\|X\|^2)\,(\|X\|^{k+2} - k\,\|X\|^k)] &= d\int_0^{\infty} y^{k/2}\,[2\,y\,h'(y) + p\,h(y)]\, e^{-y/2}\, dy \\ &= E_0[(2\,\|X\|^2\, h'(\|X\|^2) + p\,h(\|X\|^2))\,\|X\|^k]\,. \end{aligned}$$

Thus (7.5) becomes

$$\begin{aligned} A(\theta) &= (2\pi)^{-p/2}\, e^{-\|\theta\|^2/2} \sum_{k=0}^{\infty} \frac{1}{k!} E_0\left(\frac{X'\theta}{\|X\|}\right)^k \mathbb{1}_{\mathscr{C}}\left(\frac{X'\theta}{\|X\|}\right) \\ &\quad E_0[(2\,\|X\|^2\, h'(\|X\|^2) + p\,h(\|X\|^2))\,\|X\|^k] \\ &= (2\pi)^{-p/2} e^{-\|\theta\|^2/2} \sum_{k=0}^{\infty} \frac{1}{k!} E_0[(X'\theta)^k \{2\,\|X\|^2\, h'(\|X\|^2) + p\,h(\|X\|^2)\}\, \mathbb{1}_{\mathscr{C}}(X)] \\ &= E_0[\{2\,\|X\|^2\, h'(\|X\|^2) + p\,h(\|X\|^2)\}\, \mathbb{1}_{\mathscr{C}}(X)] \end{aligned}$$

where the final identity follows by the dominated convergence theorem. □

General dominating estimators will be obtained by shrinking each P_iX in (7.3) on the set D_i. Recall that each D_i has the property that, if $x \in D_i$, then $a P_i x + P_i^{\perp} x \in C_i$ for all $a > 0$. The next result extends Lemma 7.1 to apply to projections P_i onto sets which have this conditional homogeneity property.

Lemma 7.2. *Let $X \sim N(\theta, \sigma^2 I)$ and P be a linear orthogonal projection of rank s. Further, let D be a set such that, if $x = Px + P^{\perp}x \in D$, then $a Px + P^{\perp}x \in D$ for all $a > 0$. Then, for any absolutely continuous function $h(\cdot)$ on $\mathbb{R}_+$ into $\mathbb{R}$ such that $\lim_{y \to 0,\infty} h(y)\, y^{(j+s)/2}\, e^{-y/2} = 0$ for all $j \geq 0$, we have*

$$\begin{aligned} & E_\theta[(X-\theta)' PX\, h(\|PX\|^2)\, 1\!\!1_D(X)] \\ & = \sigma^2 E_\theta[\{2\, \|PX\|^2\, h'(\|PX\|^2) + s\, h(\|pX\|^2)\}\, 1\!\!1_D(X)]\,. \end{aligned}$$

Proof. By assumption $(Y_1, Y_2) = (PX, P^{\perp}X) \sim N_p((\eta_1, \eta_2), \sigma^2 I)$ where $(P\theta, P^{\perp}\theta) = (\eta_1, \eta_2)$. Also

$$\begin{aligned} A(\theta) &= E_\theta[(X-\theta)' PX h(\|PX\|^2)\, 1\!\!1_D(X)] \\ &= E_\theta[(PX - P\theta)' PX h(\|PX\|^2)\, 1\!\!1_D(X)] \\ &= E_{\eta_1 \eta_2}[(Y_1 - \eta_1)' Y_1 h(\|Y_1\|)\, 1\!\!1_D, (Y_1, \eta_2)\prime] \end{aligned}$$

where

$$D' = \{(Y_1, Y_2) | (Y_1, Y_2) = (PX, P^{\perp}X) \in D)\}\,.$$

On conditioning on Y_2 (which is independent of Y_1), and applying Lemma 7.1 to Y_1, we have

$$
\begin{aligned}
A(\theta) &= E_{\eta_2}[E_{\eta_1}[(Y_1-\eta_1)'Y_1\,h(\|Y_1\|^2)\,\mathbb{1}_{D'(Y_1,Y_2)}|Y_2]] \\
&= \sigma^2\,E_{\eta_2}[E_{\eta_1}[\{2\,\|Y_1\|^2\,h'(\|Y_1\|^2)+s\,h(\|Y_1\|^2)\}\,\mathbb{1}_{D'(Y_1,Y_2)}|Y_2]] \\
&= \sigma^2\,E[\{2\,\|PX\|^2\,h'(\|PX\|^2)+s\,h(\|PX\|^2)\}\,\mathbb{1}_D(X)]
\end{aligned}
$$

□

Now we use Lemma 7.1 to obtain the main domination result of this section.

Theorem 7.2. *Let $X \sim N_p(\theta,\sigma^2 I)$ where σ^2 is known and θ is restricted to lie in the polyhedral cone $\mathscr{C}$ with non-empty interior. Then, under loss $L(\theta,d)=\|d-\theta\|^2$, the estimator*

$$
\delta(X)=\sum_{i=0}^{n}\left(1-\sigma^2\,\frac{r_i(\|P_iX\|^2)\,(s_i-2)_+}{\|P_iX\|^2}\right)P_iX\;\mathbb{1}_D(X) \tag{7.6}
$$

dominates the rule $\delta_{\mathscr{C}}(X)$ given by (7.3) provided a) $0<r_i(t)<2$, b) $r_i(\cdot)$ is absolutely continuous, and c) $r_i'(t)\ge 0$, for each $i=0,1,\ldots,n$.

Proof. The difference in risk between δ and $\delta_{\mathscr{C}}$ can be expressed as

$$
\begin{aligned}
\Delta(\theta) &= R(\theta,\delta)-R(\theta,\delta_{\mathscr{C}}) \\
&= \sum_{i=0}^{n}E_\theta\Bigg[\sigma^4\,\frac{r_i^2(\|P_iX\|^2)\,((s_i-2)_+)^2}{\|P_iX\|^2} \qquad (7.7)\\
&\qquad -2\,\sigma^2\,\frac{r_i(\|P_iX\|^2)\,(s_i-2)_+}{\|P_iX\|^2}\,(P_iX)'(P_iX-\theta)\Bigg]\,\mathbb{1}_{D_i}(X)
\end{aligned}
$$

Now apply Lemma 7.2 (noting that $(P_iX)'(P_iX-\theta)=(P_iX)'(X-\theta)$) to each summand in the second term to get

$$\begin{aligned}\Delta(\theta) = \sigma^4 \sum_{i=0}^{n} E_\theta \Bigg[& \frac{r_i^2(\|P_iX\|^2)\,((s_i-2)_+)^2}{\|P_iX\|^2} \\ & - 2\,\frac{r_i(\|P_iX\|^2)\,(s_i-2)_+}{\|P_iX\|^2} + 4\,r_i'(\|P_iX\|^2)\,(s_i-2)_+ \Bigg] \mathbb{1}_{D_i}(X) \\ \le 0 \end{aligned} \tag{7.8}$$

since each $r_i'(\cdot) \ge 0$ and $0 < r_i(\cdot) < 2$. □

As noted in Chapter 3, the case of an unknown σ^2 is easily handled provided an independent statistic $S \sim \sigma^2 \chi_k^2$ is available. For completeness we give the result for this case in the following theorem.

Theorem 7.3. *Suppose $X \sim N_p(\theta, \sigma^2 I)$ and $S \sim \sigma^2 \chi_k^2$ with X independent of S. Suppose that θ is restricted to the polyhedral cone $\mathscr{C}$ with non-empty interior. Then the estimator*

$$\delta(X,S) = \sum_{i=0}^{n} \left(1 - \left(\frac{S}{k+2}\right) \frac{r_i(\|P_iX\|^2)(s_i-2)_+}{\|P_iX\|^2}\right) P_iX\,\mathbb{1}_{D_i}(X) \tag{7.9}$$

dominates $\delta_{\mathscr{C}}(X)$ given in (7.3) provided $0 < r_i(\cdot) < 2$ and $r_i(\cdot)$ is absolutely continuous with $r_i'(\cdot) \ge 0$, for $i = 0, \ldots, n$.

Many of the classical problems in ordered inference are examples of restrictions to polyhedral cones. Here are a few.

Example 7.1. Orthant Restrictions. Estimation problems where k of the coordinate means are restricted to be greater than (or less than) a given set constants, can be transformed easily into the case where these same components are restricted to be positive. This is essentially the case for $\mathscr{C} = \mathbb{R}_+^k \oplus \mathbb{R}^{p-k}$ mentioned earlier.

Example 7.2. Ordered Means. The restrictions that $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_p$ (or that a subset are so ordered) is a common example in the literature and corresponds to the finite set of half space restrictions

$$\theta_2 \geq \theta_1, \theta_3 \geq \theta_2, \ldots, \theta_p \geq \theta_{p-1}\,.$$

Example 7.3. Umbrella Ordering. The ordering $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_k \geq \theta_{k+1} \geq \theta_{k+2}, \ldots,$ $\theta_{p-1} \geq \theta_p$ corresponds to the polyhedral cone generated by the half space restrictions

$$\theta_2 - \theta_1 \geq 0,\, \theta_3 - \theta_2 \geq 0, \ldots, \theta_k - \theta_{k-1} \geq 0,\, \theta_{k+1} - \theta_k \leq 0, \ldots, \theta_p - \theta_p \leq 0\,.$$

In some examples, such as Example 7.1, it is relatively easy to specify P_i and D_i. In others, such as Example 7.2 and 7.3 it is more complicated. The reader is referred to Robertson et al (1988) and references therein for further discussion of this issue.

7.4 Normal Mean Vector Restricted to a Ball

7.5 Spherically Symmetric Distribution with a Mean Vector Restricted to a Polyhedral Cone

This Section is devoted to proving the extension of Theorem 7.3 to the case of a spherically symmetric distribution when a residual vector is present. Specifically we assume that $(X,U) \sim SS(\theta,0)$ where $\dim X = \dim\theta = p, \dim U = \dim 0 = k$ and where θ is restricted to lie in a polyhedral cone, $\mathscr{C}$, with non-empty interior.

Recall that the shrinkage functions in the estimator (7.9) are not necessarily weakly differentiable because of the presence of the indicator functions $I_{D_i}(X)$. Hence the methods of Chapter 4 are not immediately applicable.

The following theorem develops the required tools for the desired extension of Theorem 7.3. It is the result referred to in Remark 4 of Chapter 4 and allows an alternative development of the results of Section 4 of that chapter.

Theorem 7.4. *(FSW (2006)) Let $(X,U) \sim N_{p+k}((\theta,0),\sigma^2 I)$ and assume $f : \mathbb{R}^p \to \mathbb{R}$ and $g : \mathbb{R}^p \to \mathbb{R}^p$ are such that*

$$E_{\theta,0}[(X-\theta)' g(X)] = \sigma^2 \, E_{\theta,0}[f(X)]$$

where both expected values exist for all $\sigma^2 > 0$. Then, if $(X,U) \sim S^S_{p+k}(\theta,0)$, we have

$$E_{\theta,0}[\|U\|^2 \, (X-\theta)' g(X)] = \frac{1}{k+2} E_{\theta,0}[\|U\|^4 \, f(X)]$$

provided either expected value exists.

Proof. As (X,U) is normal, $X \sim N(\theta,\sigma^2 I)$ and $\|U\|^2 \sim \sigma^2 \chi^2_k$ are independent. Using $E[\|u\|^2] = k\,\sigma^2$ and $E[\|U\|^4] = \sigma^4 \, k\,(k+2)$, we have, for each fixed σ^2,

$$\begin{aligned} E_{\theta,0}[\|U\|^2 \, (X-\theta)' g(X)] &= k\,\sigma^2 \, E_{\theta,0}[(X-\theta)' g(X)] \\ &= k\,\sigma^4 \, E_{\theta,0}[f(X)] \\ &= \frac{1}{k+2} \, E_{\theta,0}[\|U\|^4 \, f(X)] \end{aligned} \tag{7.10}$$

For each θ (considered fixed), $\|X-\theta\|^2 + \|U\|^2$ is a complete sufficient statistic for $(X,U) \sim N_{p+k}((\theta,0),\sigma^2 I)$. Now noting

$$E_{\sigma^2}[E[\|U\|^2(X-\theta)'g(X) \mid \|X-\theta\|^2+\|U\|^2]]$$

and

$$\frac{1}{k+2}E_{\sigma^2}[\|U\|^4 f(X)] = \frac{1}{k+2}E_{\sigma^2}[E[\|U\|^4 f(X) \mid \|X-\theta\|^2+\|U\|^2]]$$

it follows from (7.10) and the completeness of $\|X-\theta\|^2+\|U\|^2$ that

$$\begin{aligned}&E[\|U\|^2(X-\theta)'g(X) \mid \|X-\theta\|^2+\|U\|^2]\\&=\frac{1}{k+2}E[\|U\|^2(X-\theta)'g(X) \mid \|X-\theta\|^2+\|U\|^2]\end{aligned} \tag{7.11}$$

almost everywhere. We show at the end of this section that the functions in (7.11) are both continuous and hence they are in fact, equal everywhere.

Since the conditional distribution of (X,U) conditional on $\|X-\theta\|^2+\|U\|^2=R^2$ is uniform on the sphere centered at $(\theta,0)$ for all spherically symmetric distributions (including the normal), the result follows on integration of (7.11) with respect to the radial distribution of (X,U). □

The main result of this Section results from an application of Theorem 7.2 to the development of the proof of Theorem 7.3.

Theorem 7.5. *Let $(X,U)\sim SS_{p+k}(\theta,0)$ and let θ be restricted to a polyhedral cone with non-empty interior. Then, under loss $L(\theta,d)=\|d-\theta\|^2$, the estimator*

$$\delta(X,U)=\sum_{i=0}^{n}\left(1-\frac{\|U\|^2}{k+2}\,\frac{(s_i-2)_+\,r_i(\|P_iX\|^2)}{\|P_iX\|^2}\right)P_iX\,\mathbb{1}_{D_i}(X) \tag{7.12}$$

*dominates $P_{\mathscr{C}}X=\delta_{\mathscr{C}}(X)$, given in (**??**) provided, $0<r_i(\cdot)<2$, $r_i(\cdot)$ is absolutely continuous and $r_i'(\cdot)\geq 0$ for $i=0,\ldots,n$.*

Proof. The key observation is that, in passing from (7.7) to (7.8) in the proof of Theorem 7.2, we used Lemma 7.2 and the fact that $P_i X'(PX_i - \theta) = p_i X'(X - \theta)$ to establish that

$$E\left[\frac{r_i(\|P_iX\|^2)\,(s_i-2)_+}{\|P_iX\|^2}(P_iX)'(P_iX-\theta)\,1\!\!1_{D_i}(X)\right] =$$
$$\sigma^2 E\left[\frac{r_i(\|P_iX\|^2)\,((s_i-2)_+)^2}{\|P_iX\|^2} + 2\,r_i'(\|P_iX\|^2)\,(s_i-2)_+\,1\!\!1_{D_i}(X)\right].$$

Hence, by Theorem 7.4,

$$E\left[\frac{\|U\|^2}{k+2}\,\frac{r_i(\|P_iX\|^2)\,(s_i-2)_+}{\|P_iX\|^2}(P_iX)'(P_iX-\theta)\,1\!\!1_{D_i}(X)\right] =$$
$$\sigma^2 E\left[\frac{\|U\|^4}{(k+2)^2}\left\{\frac{r_i(\|P_iX\|^2)\,((s_i-2)_+)^2}{\|P_iX\|^2} + 2\,r_i'(\|P_iX\|^2)\,(s_i-2)_+\right\}1\!\!1_{D_i}(X)\right].$$

It follows then, as in the proof of Theorem 7.2,

$$\begin{aligned} R(\theta,\delta(X,U)) - R(\theta,\delta_{\mathscr{C}}) = &\sum_{i=0}^{n} E_\theta\left[\frac{\|U\|^4}{(k+2)^2}\left\{\frac{r_i^2(\|P_iX\|^2)\,((s_i-2)_+)^2}{\|P_iX\|^2}\right.\right. \\ &\left.\left. - \left\{2\,\frac{r_i(\|P_iX\|^2)\,(s_i-2)_+}{\|P_iX\|^2} + 4\,r_i'(\|P_iX\|^2)\,(s_i-2)_+\right\}1\!\!1_{D_i}(X)\right] \\ \le &\,0 \end{aligned} \tag{7.13}$$

□

Theorem 7.4 is an example of a meta result which follows from Theorem 7.5, and states roughly that, if one can find an estimator $X + \sigma^2 g(X)$ that dominates X for each σ^2 using a Stein-type differential equality in the normal case, then $X + \|U\|^2/(k+2)\,g(X)$ will dominate X in the general spherically symmetric case, $(X,U) \sim SS_{1+k}(\theta,U)$, under $L(\theta,\delta) = \|\delta - \theta\|^2$. The “proof” goes as follows.

Suppose one can show also that $E[(X-\theta)'g(X)] = \sigma^2 E[f(X)]$ in the normal case, and also that $\|g(x)\|^2 + 2f(x) \leq 0$, for any $x \in \mathbb{R}^p$. Then, in the normal case,

$$R(\theta, X - \sigma^2 g(X)) - R(\theta, X) = \sigma^4 E[\|g(X)\|^2 + 2f(X)] \leq 0.$$

Using Theorem 7.4 (and assuming finiteness of expectations), it follows in the general case that

$$R\left(\theta, X + \frac{\|U\|^2}{k+2} g(X)\right) - R(\theta, X) = E\left[\frac{\|U\|^4}{(k+2)^2}\{\|g(X)\|^2 + 2f(X)\}\right] \leq 0.$$

In this Section, application of the above meta-result had the additional complication of a separate application (to $P_i X$ instead of X) on each D_i but the basic idea is the same. The results of Chapter 4 which rely on extending a version of Stein's lemma to the general spherically symmetric case can be proved in the same way.

We close this Section with a result that implies the claimed continuity of the conditional expectations in (7.11).

Lemma 7.3. *Let $(X,U) \sim SS_{p+k}(\theta,0)$ and let $\alpha \in N$. Assume $\varphi(\cdot)$ is such that for any $R > 0$, the conditional expectation*

$$f(R) = E_{(\theta,0)}[\|U\|^\alpha \, \varphi(X) \mid \|X-\theta\|^2 + \|U\|^2 = R^2]$$

exists. Then the function f is continuous on $\mathbb{R}_+$.

Proof. Assume without loss of generality that $\theta = 0$ and $\varphi(\cdot) \geq 0$. Since the conditional distribution or (X,U) conditional on $\|X\|^2 + \|U\|^2 = R^2$ is the uniform distribution U_R on the sphere $S_R = \{y \in R^{p+k} / \|y\| = R\}$ centered at 0 with radius R, we have

$$f(R) = \int_{S_R} \|u\|^{\alpha}\, \varphi(x)\, dU_R(x,u)\,.$$

Since $\|u\|^2 = R^2 - \|x\|^2$ for any $(x,u) \in S_R$ and X has distribution concentrated on the ball $B_r = \{x \in \mathbb{R}^p | \|x\| \leq R\}$ in R^p with density proportional to $R^{2-(p+k)}((R^2 - \|x\|^2)^{k/2-1})$ we have that $R^{p+k-2} f(R)$ is proportional to

$$\begin{aligned} g(R) &= \int_{B_R} (R^2 - \|x\|^2)^{(k+\alpha)/2-1}\, \varphi(x)\, dx. \\ &= \int_0^R \int_{S_r} (R^2 - \|x\|^2)^{(k+\alpha)/2-1}\, \varphi(x)\, d\sigma_r(x)\, dr \\ &= \int_0^R (R^2 - r^2)^{(k+\alpha)/2-1}\, H(r)\, dr \end{aligned}$$

where

$$H(r) = \int_{S_r} \varphi(x)\, d\sigma_r(x)$$

and where σ_r is the area measure on the sphere S_r. Since $H(\cdot)$ and $(k+\alpha)/2-1$ are non-negative, the family of integrable functions $r \to K(R,r) = (R^2 - r^2)^{(k+\propto)/2-1} H(r) I^{(}_{[0,R]} r)$, indexed by R, is nondecreasing in R and bounded above for $R < R_0$ by the integrable function $K(R_0, r)$. Then the continuity of $g(R)$, and hence of $f(R)$, is guaranteed by the dominated convergence theorem. □

Chapter 8

Loss Estimation

8.1 Introduction

Suppose X is an observable from a distribution P_θ parameterized by an unknown parameter θ. In classical decision theory, it is usual, after selecting an estimation procedure $\varphi(X)$ of θ, to evaluate it through a criterion, a loss, $L(\theta, \varphi(X)$ which represents the cost incurred by the estimation $\varphi(X)$ when the unknown parameter equals θ. In the long run, as it depends on the particular value of X, this loss cannot be appropriate to assess the performance of the estimator φ. Indeed, to be valid (in the frequentist sense), a global evaluation of such a statistical procedure should be based on all the possible observations. Consequently, it is common to report the risk $R(\theta, \varphi) = E_\theta[L(\theta, \varphi(X)]$ as a gauge of the efficiency of φ (E_θ denotes expectation with respect to P_θ). Thus we have at our disposal a long run performance of $\varphi(X)$ for each value of θ. However, although this notion of risk can effectively be used in comparing $\varphi(X)$ with other estimators, it is inaccessible since θ is unknown. The usual frequentist risk assessment is the maximum risk $\bar{R}_\varphi = \sup_\theta R(\theta, \varphi)$.

By construction, this last report on the estimation procedure is non data-dependent (as we were guided by a global notion of accuracy of $\varphi(X)$). However there exist situations where the fact that the observation X has a particular value x may influence the judgment on a statistical procedure. A particularly edifying example is given by the following simple confidence interval estimation (which can also be seen as a loss estimation problem). Assume that the observable is a pair (X_1,X_2) of independent copies of a random variable X satisfying, for $\theta \in \mathbb{R}$,

$$P[X = \theta - 1] = P[X = \theta + 1] = \frac{1}{2}\,.$$

Then it is clear that the confidence interval for θ defined by

$$I(X_1,X_2) = \left\{\theta \in \mathbb{R} \Bigm/ \left|\frac{X_1+X_2}{2} - \theta\right| < \frac{1}{2}\right\}$$

satisfies

$$\mathbb{1}_{[I(X_1,X_2)\ni\theta]} = \begin{cases} 1 \text{ if } X_1 \neq X_2 \\ \\ 0 \text{ if } X_1 = X_2 \end{cases}$$

so that it suffices to observe $(X_1,X_2) = (x_1,x_2)$ in order to know exactly whether $I(x_1,x_2)$ contains θ or not.

The previous (ad hoc) example indicates that data-dependent reports are relevant. In our estimation context when $X = x$, note that, if it were available (but θ is unknown), it would be the loss $L(\theta,\varphi(x))$ itself which should serve as a perfect measure of the accuracy of φ. It is then natural to estimate $L(\theta,\varphi(x))$ by a data-dependent estimator $\delta(X)$, a new estimator called a loss estimator which will serve as a data-dependent report (instead of $\bar{R}_\varphi$). This is a conditional approach in the sense that accuracy assessment is made on a data-dependent quantity, the loss, instead of the risk.

To evaluate the extend to which $\delta(X)$ successfully estimates $L(\theta, \varphi(X))$, another loss is required and it has become standard to use the squared error

$$L^*(\theta, \varphi(X), \delta(X)) = (\delta(X) - L(\theta, \varphi(X))^2 \,, \tag{8.1}$$

for simplicity. In so far as we are thinking in terms of long-run frequencies, we adopt a frequentist approach to evaluating the performance of L^* by averaging over the sampling distribution of X given θ, that is, by using a new notion of risk

$$\mathscr{R}(\theta, \varphi, \delta) = E_\theta[L^*(\theta, \varphi(X), \delta(X))] = E_\theta[(\delta(X) - L(\theta, \varphi(X))^2] \,. \tag{8.2}$$

As $\bar{R}_\varphi$ reports on the worst situation (the maximum risk), we may expect that a competitive data-dependent report $\delta(X)$ improves on $\bar{R}_\varphi$ under the risk (8.1), that is, for all θ, satisfies

$$\mathscr{R}(\theta, \varphi, \delta) \leq \mathscr{R}(\theta, \varphi, \bar{R}_\varphi) \,. \tag{8.3}$$

More generally, a reference loss estimator δ_0 will be dominated by a competitive estimator δ if, for all θ,

$$\mathscr{R}(\theta, \varphi, \delta) \leq \mathscr{R}(\theta, \varphi, \delta_0) \,, \tag{8.4}$$

with strict inequality for some θ.

Note that, unlike the usual estimation setting where the quantity of interest is a function of the parameter θ, loss estimation involves a function of both θ and X (the data). This feature may make the statistical analysis more difficult but it is clear that the usual notions of minimaxity, admissibility, etc, and their methods of proof can be directly adapted to that situation. Also, although frequentist interpretability was evoked above, in case we would be interested in a Bayesian approach, it is easily seen that this approach would consist of the usual Bayes estimator φ_B of θ and the posterior loss $\delta_B(X) = E[L(\theta, \varphi_B)|X]$.

The problem of estimating a loss function has been considered by Sandved [125] who developed a notion of unbiased estimator of $L(\theta, \varphi(X)$ in various settings. However the underlying conditional approach traces back to Lehmann and Sheffé [96] who estimated the power of a statistical test. Kiefer, in a series of papers ([86], [87], [88]), developed conditional and estimated confidence theories through frequentist interpretability. A subjective Bayesian approach was compared by Berger ([11], [12], [13]) with the frequentist one.

Jonhstone [80] considered (in)admissibility of unbiased estimators of loss for the maximum likelihood estimator $\varphi_0(X) = X$ and for the James-Stein estimator $\varphi^{JS}(X) = \left(1 - \frac{p-2}{||X||^2}\right) X$ of a p-variate normal mean θ. For $\varphi_0(X) = X$, the unbiased estimator of the quadratic loss $L(\theta, \varphi_0(X)) = ||\varphi_0(X) - \theta||^2$, that is, the loss estimator δ_0 which satisfies, for all θ,

$$E_\theta[\delta_0] = E_\theta[L(\theta, \varphi_0(X))] = R(\theta, \varphi_0)\,, \tag{8.5}$$

is $\delta_0 = \bar{R}_\varphi = p$. Johnstone proved that (8.3) is satisfied with the competitive estimator $\delta(X) = p - 2(p-4)/||X||^2$ when $p \geq 5$, the risk difference between δ_0 and δ being expressed as $-4(p-4)^2 E_\theta[1/||X||^4]$.

For the James-Stein estimator φ^{JS}, the unbiased estimator of loss is itself data-dependent and equal to $\delta_0^{JS}(X) = p - (p-2)^2/||X||^2$. Jonhstone showed that improvement on δ_0^{JS} can be obtained with $\delta^{JS}(X) = p - (p-2)^2/||X||^2 + 2p/||X||^2$ when $p \geq 5$, with strict inequality in (8.4) for all θ since the difference in risk between δ^{JS} and δ_0^{JS} equals $-4p^2 E_\theta[1/||X||^2]$.

In Section 8.2, we develop the quadratic loss estimation problem for a p-normal mean. After a review of the basic ideas, a new class of loss estimators is constructed in Subsection 8.2.1. In Subsection 8.2.2, we turn our focus on some interesting and surprising behavior of Bayesian assessments, this paradoxical result is illustrated in a general inadmissibility theorem. Section 8.3 is devoted to the case where the variance is unknown. Extensions to the spherical case are given in Section 8.4. In Subsection 8.4.1, we consider the general case of a spherically symmetric distribution around a fixed vector $\theta \in \mathbb{R}^p$ while, in Subsection 8.4.2, these ideas are then generalized to the case where a residual vector is available. We conclude by mentioning a number of applied and theoretical developments of loss estimation not covered in this overview.

8.2 Estimating the quadratic loss of a p-normal mean with known variance

8.2.1 Dominating unbiased estimators of loss

Let X be a random vector having a p-variate normal distribution $\mathscr{N}(\theta, I_p)$ with unknown mean θ and identity covariance matrix I_p. To estimate θ, the observable X is itself a reference estimator MLE and it is an unbiased estimator of θ) so that it is convenient to write any estimator of θ through X as $\varphi(X) = X + g(X)$, for a certain function g from $\mathbb{R}^p$ into $\mathbb{R}^p$. Under squared error loss $||\varphi(X) - \theta||^2$, the (quadratic) risk of φ is defined by

$$R(\theta, \varphi) = E_\theta[||\varphi(X) - \theta||^2] \tag{8.6}$$

where E_θ denotes the expectation with respect to $\mathscr{N}(\theta, I_p)$.

Clearly, the risk of the MLE X equals p and $\varphi(X)$ will be a reasonable estimator only if its risk is finite. It is easy to see through Schwarz's inequality that this is the case as soon as

$$E_\theta[||g(X)||^2] < \infty, \tag{8.7}$$

which we will assume in the following (it can be also seen that this condition is in fact necessary to guarantee the risk finiteness). Indeed the loss of $\varphi(X)$ can be expanded as

$$||\varphi(X) - \theta||^2 = ||X - \theta||^2 + ||g(X)||^2 + 2(X - \theta)' g(X) \,. \tag{8.8}$$

Now we have $E_\theta[||X-\theta||^2] = p < \infty$. Hence, by Schwarz's inequality, it follows from (8.8) that $|E_\theta[(X-\theta)^t\, g(X)]| < (E_\theta[||X-\theta||^2])^{1/2}\,(E_\theta[||g(X)||^2])^{1/2}$. Therefore, as soon as $E_\theta[||g(X)||^2] < \infty$, we will have $|E_\theta[||\varphi(X)-\theta||^2] < \infty$.

To improve on the MLE X when $p \geq 3$ (that is, to have $R(\theta,\varphi) \leq p$), Stein [134] exhibited (under certain differentiability conditions that we recall below) an unbiased estimator of the risk of $\varphi(X)$, that is, a function $\delta_0(X)$ (depending only on X and not on θ) which verifies

$$R(\theta,\varphi) = E_\theta[\delta_0(X)]\,. \tag{8.9}$$

This statistic yields a natural estimator of the loss $||\varphi(X)-\theta||^2$ since (8.9) expresses that

$$E_\theta[||\varphi(X)-\theta||^2] = E_\theta[\delta_0(X)] \tag{8.10}$$

and hence is an unbiased estimator of the loss. Stein [134] proved more precisely that $\delta_0(X) = p + 2\,\mathrm{div}\,g(X) + ||g(X)||^2$ (where $\mathrm{div}g(X)$ stands for the divergence of $g(X)$, that is, $\mathrm{div}\,g(X) = \sum_{i=1}^{p} \partial_i g_i(X)$). One can see that δ_0 may change sign so that, as an estimator of loss (which is non negative), it cannot be completely satisfactory, and hence, is liable to be improved.

Any competitive loss estimator $\delta(X)$ can be written as $\delta(X) = \delta_0(X) - \gamma(X)$ for a certain function $\gamma(X)$ which can be interpreted as a correction to $\delta_0(X)$. Note that, if the MLE is concerned (that is, if $g(X) = 0$), we may expect that an improvement on $\delta_0(X) = p$ would be obtained with a non negative function $\gamma(X)$ satisfying the requirement expressed by Condition (8.3). Note also that, similarly to the finiteness risk condition (8.7), we will require that

$$E_\theta[\gamma^2(X)] < \infty \tag{8.11}$$

to assure that the risk of $\delta(X)$ is finite.

Using straightforward algebra, the risk difference $\mathscr{D}(\theta, \varphi, \delta) = \mathscr{R}(\theta, \varphi, \delta) - \mathscr{R}(\theta, \varphi, \delta_0)$ simplifies in

$$\mathscr{D}(\theta, \varphi, \delta) = E_\theta[\gamma^2(X) - 2\,\gamma(X)\,\delta_0(X)] + 2\,E_\theta[\gamma(X)\,||\varphi(X) - \theta||^2]\,. \tag{8.12}$$

Conditions for which $\mathscr{D}(\theta, \varphi, \delta) \leq 0$ will be formulated after finding, along the line of Stein's techniques evoked above, an unbiased estimate of the term $\gamma(X)\,||\varphi(X) - \theta||^2$ in the last expectation. We briefly review the context of those techniques.

For a function g from $\mathbb{R}^p$ into $\mathbb{R}^p$, the Stein's identity (see Stein [134]) states that

$$E_\theta[(X - \theta)' g(X)] = E_\theta[\operatorname{div} g(X)] \tag{8.13}$$

provided that these expectations exist. Here Stein specified that g was almost differentiable. Almost differentiability is needed to integrate shrinkage functions $g(X)$, intervening in the James-Stein estimators, of the form $g(X) = -aX\,/||X||^2$ which are not differentiable in the usual sense (such $g(X)$ explode at 0). This notion is equivalent (and it is of more common use in analysis) to the statement that g belongs to the Sobolev space $W^{1,1}_{loc}(\mathbb{R}^p)$ of weakly differentiable functions. That equivalence was noticed by Johnstone [80].

Recall that a locally integrable function γ from $\mathbb{R}^p$ into $\mathbb{R}$ is said to be weakly differentiable if, there exist p functions $h_1, \ldots, h_p$ locally integrable on $\mathbb{R}^p$ such that, for any $i = 1, \ldots, p$

$$\int_{\mathbb{R}^p} \gamma(x)\, \frac{\partial \varphi}{\partial x_i}(x)\, dx = -\int_{\mathbb{R}^p} h_i(x)\, \varphi(x)\, dx \tag{8.14}$$

for any infinitely differentiable function φ on $\mathbb{R}^p$ with compact support. The functions h_i are the i-th partial weak derivatives of γ. Their common notation is $\partial\gamma/\partial x_i$ and the vector $\nabla\gamma = (\partial\gamma/\partial x_1, \dots, \partial\gamma/\partial x_p)^t$ is referred to the weak gradient of γ.

Note that (8.14) usually holds when γ is continuously differentiable; $h_i = \partial\gamma/\partial x_i$, the standard partial derivative, is continuous. Thus, via (8.14), the extension to weak differentiability consists in a propriety of integration by parts with vanishing bracketed term. Naturally a function $g = (g_1, \dots, g_p)$ from $\mathbb{R}^p$ into $\mathbb{R}^p$ is said to be weakly differentiable if each of its components g_j is weakly differentiable. In that case, the function $\operatorname{div} g = \sum_{i=1}^p \partial g_i/\partial x_i$ is referred to the weak divergence of g; this is the operator intervening in the Stein's identity (8.13).

When dealing with an unbiased estimator of a quantity of the form $||X-\theta||^2\,\gamma(X)$ where γ is a function from $\mathbb{R}^p$ into $\mathbb{R}$, writing

$$||X-\theta||^2\,\gamma(X) = (X-\theta)'(X-\theta)\,\gamma(X) \tag{8.15}$$

naturally leads to an iteration of Stein's identity (8.13) and involves twice weak differentiability of γ. This is of course defined through the weak differentiability of all the weak partial derivatives $\partial\gamma/\partial x_i$; these second weak partial derivatives are denoted by $\partial^2\gamma/\partial x_j \partial x_i$. Thus γ belongs to the Sobolev space $W^{2,1}_{loc}(\mathbb{R}^p)$ and $\Delta\gamma = \sum_{i=1}^p \partial^2\gamma/\partial x_i^2$ is referred to as the weak Laplacian of γ.

By (8.15) and (8.13), we have

$$\begin{aligned}E_\theta[||X-\theta||^2\,\gamma(X)] &= E_\theta[\mathrm{div}((X-\theta)'\,\gamma(X))]\\ &= E_\theta[p\,\gamma(X)+(X-\theta)^t\,\nabla\gamma(X)] \qquad (8.16)\end{aligned}$$

by property of the divergence. Then, applying again (8.13) to the last term in (8.16) gives

$$E_\theta[(X-\theta)^t\,\nabla\gamma(X)] = E_\theta[\mathrm{div}(\nabla\gamma(X)] = E_\theta[\Delta\,\gamma(X)] \qquad (8.17)$$

by definition of the Laplacian. Finally, gathering (8.16) and (8.17), we obtain that

$$E_\theta[||X-\theta||^2\,\gamma(X)] = E_\theta[p\,\gamma(X)+\Delta\,\gamma(X)]\,. \qquad (8.18)$$

We are now in a position to provide an unbiased estimator of $\mathscr{D}(\theta,\varphi,\delta)$. Its nonpositivity will be a sufficient condition for $\mathscr{D}(\theta,\varphi,\delta)\le 0$ and hence for δ to improve on δ_0. Indeed we have

$$\begin{aligned}||\varphi(X)-\theta||^2 &= ||X+g(X)-\theta||^2\\ &= ||g(X)||^2+2\,(X-\theta)'\,g(X)+||X-\theta||^2\end{aligned}$$

so that, according to (8.13) and (8.18),

$$E_\theta[||\varphi(X)-\theta||^2\,\gamma(X)] = E_\theta[\gamma(X)\,||g(X)||^2+2\,\mathrm{div}(\gamma(X)\,g(X))+p\,\gamma(X)+\Delta\,\gamma(X)]\,.$$

Therefore, as

$$\mathrm{div}(\gamma(X)\,g(X)) = \gamma(X)\,\mathrm{div}g(X)+\nabla\gamma(X)'\,g(X)$$

and $\delta_0(X)=p+2\,\mathrm{div}\,g(X)+||g(X)||^2$, the risk difference $\mathscr{D}(\theta,\varphi,\delta)$ in (8.12) reduces to

$$\mathscr{D}(\theta,\varphi,\delta) = E_\theta[\gamma^2(X)+4\,\nabla\gamma(X)^t\,g(X)+2\,\Delta\,\gamma(X)]$$

so that a sufficient condition for $\mathscr{D}(\theta,\varphi,\delta)$ to be nonpositive is

$$\gamma^2(x) + 4\,\nabla\gamma(x)'\,g(x) + 2\,\Delta\gamma(x) \leq 0 \tag{8.19}$$

for any $x \in \mathbb{R}^p$.

How can one determine a "best" correction γ satisfying (8.19)? The following theorem provides a way to associate to the function g a suitable correction γ which satisfies (8.19) in the case where $g(x)$ is of the form $g(x) = \nabla m(x)/m(x)$ for a certain nonnegative function m. This is the case when φ is a Bayes estimator of θ related to a prior π, the function m being the corresponding marginal (see Brown [29]). Bock [21] shows that, through the choice of m, such estimators constitute a wide class of estimators of θ (which are called pseudo-Bayes estimators when the function m does not correspond to a true prior π).

Theorem 8.1. *Let m be a non negative function which is also superharmonic (respectively subharmonic) on $\mathbb{R}^p$ such that $\nabla m/m \in W^{1,1}_{loc}(\mathbb{R}^p)$. Let ξ be a real valued function, strictly positive and strictly subharmonic (respectively superharmonic) on $\mathbb{R}^p$, and such that*

$$E_\theta\left[\left(\frac{\Delta\xi(X)}{\xi(X)}\right)^2\right] < \infty. \tag{8.20}$$

Assume also that there exists a constant $K > 0$ such that, for any $x \in \mathbb{R}^p$,

$$m(x) > K\frac{\xi^2(x)}{|\Delta\xi(x)|} \tag{8.21}$$

and let $K_0 = \inf_{x\in\mathbb{R}^p} m(x)\,\frac{|\Delta\xi(x)|}{\xi^2(x)}$.

Then the unbiased loss estimator δ_0 of the estimator φ of θ defined by $\varphi(X) = X + \nabla m(X)/m(X)$ is dominated by the estimator $\delta = \delta_0 - \gamma$, where the correction *term γ is given, for any $x \in \mathbb{R}^p$ such that $m(x) \neq 0$, by*

$$\gamma(x) = -\alpha \, \mathrm{sgn}(\Delta \xi(x)) \, \frac{\xi(x)}{m(x)} \,, \tag{8.22}$$

as soon as $0 < \alpha < 2 K_0$.

Proof. The domination condition will be obtained by proving that the risk difference is less than zero. We only consider the case where m is superharmonic and ξ is strictly subharmonic, the case where m is subharmonic and ξ is strictly superharmonic being similar.

First note that the finiteness risk condition (8.11) is guaranteed by Condition (8.20) and the fact that (8.21) implies that, for any $x \in \mathbb{R}^p$,

$$\gamma^2(x) = \alpha^2 \, \frac{\xi^2(x)}{m^2(x)} \leq \frac{\alpha^2}{K_0^2} \left(\frac{\Delta \xi(x)}{\xi(x)} \right)^2 .$$

Also note that, for a shrinkage function g of the form $g(x) = \nabla m(x)/m(x)$, the left hand side of (8.19) can be expressed as

$$\mathscr{R}\gamma(x) = \gamma^2(x) + 2 \left\{ 2 \, \frac{\Delta(m(x)\, \gamma(x))}{m(x)} - \gamma(x) \, \frac{\Delta m(x)}{m(x)} \right\} \tag{8.23}$$

and hence, for γ in (8.22), as

$$\mathscr{R}\gamma(x) = \alpha^2 \, \frac{\xi^2(x)}{m^2(x)} + 2\, \alpha \left\{ - \frac{\Delta \xi(x)}{m(x)} + \frac{\xi(x)\, \Delta m(x)}{m^2(x)} \right\} . \tag{8.24}$$

Now, since m is superharmonic and ξ is positive, it follows from (8.24) that

$$\mathscr{R}\gamma(x) \leq \frac{\alpha}{m(x)} \left\{ \frac{\alpha \, \xi^2(x)}{m(x)} - 2 \, \Delta \xi(x) \right\}$$

and hence, by subharmonicity of ξ, Inequality (8.21) and definition of K_0, that

$$\mathscr{R}\gamma(x) < \frac{\alpha}{m(x)}\{\alpha - 2K_0\}\,\frac{\xi^2(x)}{m(x)}\,. \tag{8.25}$$

Finally, since $0 < \alpha < 2K_0$, Inequality (8.25) gives $\mathscr{R}\gamma(x) < 0$, which is the desired result. □

As an example, consider $m(x) = 1/||x||^{p-2}$, that is, the fundamental harmonic function which is superharmonic on the entire space $\mathbb{R}^p$ (see Du Plessis [119]). Then $g(x) = -(p-2)/||x||^2$ and $\varphi(X)$ is the James-Stein estimator whose unbiased estimator of loss is $\delta_0(X) = p - (p-2)^2/||X||^2$. Choosing, for any $x \neq 0$, the function $\xi(x) = 1/||x||^p$ gives rise to $\Delta\xi(x) = 2p/||x||^{p+2} > 0$ and hence to

$$\frac{\xi^2(x)}{|\Delta\xi(x)|} = \frac{1}{2p}\frac{1}{||x||^{p-2}}\,,$$

which means that Condition (8.21) is satisfied with $K = 1/2p$. Also we have

$$\left(\frac{\Delta\xi(x)}{\xi(x)}\right)^2 = \frac{4p^2}{||x||^4}$$

which implies that Condition (8.20) is satisfied for $p \geq 5$. Now it is clear that the constant K_0 is equal to $2p$ and that the correction term γ in (8.22) equals, for any $x \neq 0$, $\gamma(x) = -\alpha/||x||^2$. Finally, Theorem 8.1 guarantees that an improved loss estimator over the unbiased estimator of loss $\delta_0(X)$ is $\delta(X) = \delta_0(X) + \alpha/||X||^2$ for $0 < \alpha < 4p$, which is Johnstone's result [80] for the James-Stein estimator.

Similarly Johnstone's result for $\varphi(X) = X$ can be constructed with $m(x) = 1$ (which is both subharmonic and superharmonic) and with the choice of the superharmonic function $\xi(x) = 1/||x||^2$, for which $K_0 = 2(p-4)$, so that $\delta(x) = p - \alpha/||x||^2$ dominates p for $0 < \alpha < 4(p-4)$.

A possible problem with the improved estimator in (8.22) is that it may be negative, which is undesirable since we are estimating a nonnegative quantity. A simple remedy to this problem is to use a positive-part estimator. If we define the positive-part $\delta^+ = \max\{\delta, 0\}$, the loss difference between δ^+ and δ is $(\delta - L(\theta,\varphi))^2 - (\delta^+ - L(\theta,\varphi))^2 = (\delta^2 - 2\delta L(\theta,\varphi))\mathbb{1}_{\delta \leq 0}$, hence it is always nonnegative. Therefore the risk difference is positive, which implies that δ^+ dominates δ. It would be of interest to find an estimator that dominates δ^+.

In the context of variance estimation, despite warnings on its inappropriate behavior (Stein [131] and Brown [27]) the decision theoretic approach to the normal variance estimation is typically based on the standardized quadratic loss function, where overestimation of the variance is much more severely penalized than underestimation, thus leading to presumably too small estimates. Similarly in loss estimation under quadratic loss, the overestimation of the loss is also much more severely penalized than underestimation. A possible alternative to quadratic loss would be a Stein-type loss. Suppose $\varphi(X)$ is an estimator of θ under $\| \theta - \varphi(X) \|^2$ and let $\delta(X)$ be an estimator of $\| \theta - \varphi(X) \|^2$ for $\delta(X) > 0$. Then we can define the Stein-type loss for evaluating $\delta(X)$ as

$$L(\theta, \varphi(X), \delta(X)) = \frac{\| \theta - \varphi(X) \|^2}{\delta(X)} - \log \frac{\| \theta - \varphi(X) \|^2}{\delta(X)} - 1. \tag{8.26}$$

The analysis of the loss estimates under the Stein-type is more challenging but can be carried using the integration-by-parts tools developed in this section.

We have shown that the unbiased estimator of loss can be dominated. Often one may wish to add a frequentist-validity constraint to a loss estimation problem.

Specifically in our problem, the frequentist-validity constraint for some estimator δ would be $E_\theta[\delta(X)] \geq E_\theta[\delta_0(X)]$ for all θ. Kiefer [88] suggested that conditional and estimated confidence assessments should be conservatively biased, that is, the average reported loss should be greater than the average actual loss. Under such a frequentist-validity condition Lu and Berger [103] give improved loss estimators for several of the most important Stein-type estimators. One of their estimators is a generalized Bayes estimator, suggesting that Bayesians and frequentists can potentially agree on a conditional assessment of loss.

8.2.2 Dominating the posterior risk

In the previous sections, we have seen that the unbiased estimator of loss should be often dismissed since it can be dominated. When a (generalized) Bayes estimator of θ is available, incorporating the same prior information for estimating the loss of this Bayesian estimator is coherent, and we may expect that the corresponding Bayes estimator is a good candidate to improve on the unbiased estimator of loss. However, somewhat surprisingly, Fourdrinier and Strawderman [56] found that, in the normal setting, the unbiased estimator often dominates the corresponding generalized Bayes estimator of loss for priors which give minimax estimators in the original point estimation problem. In particular, they give a class of priors for which the generalized Bayes estimator of θ is admissible and minimax but for which the unbiased estimator of loss dominates the generalized Bayes estimator of loss. They also give a general inadmissibility result for a generalized Bayes estimator of loss.

While much of their focus is on pseudo-Bayes estimators, in this section, we essentially present their results on generalized Bayes estimators.

Suppose X is distributed as $\mathcal{N}(\theta, I_p)$ and, estimating θ with the estimator $\varphi(X)$, the loss function is $L(\theta, \varphi(X)) = \|\varphi(X) - \theta\|^2$. For a given generalized prior π, we denote the generalized marginal by m and the generalized Bayes estimator of θ by

$$\varphi_m(X) = X + \frac{\nabla m(X)}{m(X)} \,. \tag{8.27}$$

Then (see Stein [134]) the unbiased estimator of risk of $\varphi_m(X)$ is

$$\delta_0(X) = p + 2\frac{\Delta m(x)}{m(X)} - \frac{\|\nabla m(X)\|^2}{m^2(X)} \tag{8.28}$$

while the posterior risk of $\varphi_m(X)$ is

$$\delta_m(X) = p + \frac{\Delta m(X)}{m(X)} - \frac{\|\nabla m(X)\|^2}{m^2(X)} \,. \tag{8.29}$$

Domination of $\delta_0(X)$ over $\delta_m(X)$ is obtained thanks to the fact that their risk admits $\left(\frac{\Delta m(X)}{m(X)}\right)^2 - 2\,\frac{\Delta^{(2)} m(X)}{m(X)}$ as an unbiased estimator of their risk difference, that is,

$$\mathscr{R}(\theta, \varphi_m, \delta_0) - \mathscr{R}(\theta, \varphi_m, \delta_m) = E_\theta\left[\left(\frac{\Delta m(X)}{m(X)}\right)^2 - 2\,\frac{\Delta^{(2)} m(X)}{m(X)}\right] \tag{8.30}$$

where $\Delta^{(2)} m = \Delta(\Delta m)$ is the bi-Laplacian of m (see [56]). Thus the above domination will occur as soon as

$$\left(\frac{\Delta m(X)}{m(X)}\right)^2 - 2\,\frac{\Delta^{(2)} m(X)}{m(X)} \leq 0 \,. \tag{8.31}$$

Applicability of that last condition is underlined by the remarkable fact that if the prior π satisfies (8.31), that is, if

$$\left(\frac{\Delta\pi(\theta)}{\pi(\theta)}\right)^2 - 2\,\frac{\Delta^{(2)}\pi(\theta)}{\pi(\theta)} \leq 0\,, \tag{8.32}$$

then (8.31) is satisfied for the marginal m.

As an example, [56] consider $\pi(\theta) = \left(\frac{\|\theta\|^2}{2} + a\right)^{-b}$ (where $a \geq 0$ and $b \geq 0$) and show that, if $p \geq 2(b+3)$ then (8.32) holds and hence δ_u dominates δ_m. Since π is integrable if and only if $b > \frac{p}{2}$ (for $a > 0$), the prior π is improper whenever this condition for domination of δ_u over δ_m holds. Of course, whenever π is proper, the Bayes estimator δ_m is admissible provided its Bayes risk is finite.

Inadmissibility of the generalized Bayes loss estimator is not exceptional. Thus, in [56], the following general inadmissibility result is given; its proof is parallel to the proof of Theorem 8.1.

Theorem 8.2. *Under the conditions of Theorem 8.1, δ_m is inadmissible and a class of dominating estimators is given by*

$$\delta_m(X) + \alpha \,\operatorname{sgn}\,(\Delta\xi(X))\,\frac{\xi(X)}{m(X)} \;\textit{for}\; 0 < \alpha < 2\,K_0\,.$$

Note that Theorem 8.2 gives conditions of improvement on δ_m while Theorem 8.1 looks for improvements on δ_0. As we saw that, often, δ_0 dominates δ_m. So it is not surprising that the proof of the two theorems are parallel; more precisely, it suffices to suppress, in the proof of Theorem 8.1), the superharmonicity (or subharmonicity) condition on m to obtain the proof of Theorem 8.2.

In [56], it is suggested that the inadmissibility of the generalized Bayes (or pseudo-Bayes) estimator is due to the fact that the loss function $(\delta(x) - \|\varphi(x) - \theta\|^2)^2$ may be inappropriate. The possible deficiency of this loss is illustrated by the following simple result concerning estimation of the square of a location parameter in $\mathbb{R}$.

Suppose $X \in \mathbb{R} \sim f\left((X-\theta)^2\right)$ such that $E_\theta[X^4] < \infty$. Consider estimation of θ^2 under loss $(\delta - \theta^2)^2$. The generalized Bayes estimator δ_π of θ^2 with respect to the uniform prior $\pi(\theta) \equiv 1$ is given by

$$\delta_\pi(X) = \frac{\int \theta^2 f\left((X-\theta)^2\right) d\theta}{\int f((X-\theta)^2)\, d\theta} = X^2 + E_0\left[X^2\right].$$

Since this estimator has constant bias $2\,E_0[X^2]$, it is dominated by the unbiased estimator $X^2 - E_0[X^2]$ (the risk difference is $4\,(E_0[X^2])^2$). Hence δ_π is inadmissible for any $f(\cdot)$ such that $E_\theta[X^4] < \infty$.

8.2.3 Examples of improved estimators

In this subsection, we give some examples of Theorems 8.1 and 8.2. The only example up to this point of an improved estimator over the unbiased estimator of loss $\delta_0(X)$ is $\delta(X) = \delta_0(X) + \alpha/||x||^2$ for $0 < \alpha < 4\,p$, which is Johnstone's result [80]. Although the shrinkage factor in Theorems 8.1 and 8.2 are the same, in the examples below we will only focus on improvements of posterior risk.

As an application of Theorem 8.2, let $\xi_b(x) = \left(||x||^2 + a\right)^{-b}$ (with $a \geq 0$ and $b \geq 0$). It can be shown that we have $\Delta\xi_b(x) < 0$ for $a \geq 0$ and $0 < 2\,(b+1) < p$. Also $\Delta\xi_b(x) > 0$ if $a = 0$ and $2(b+1) > p$. Furthermore

$$\frac{\xi_b^2(x)}{|\Delta\xi_b(x)|} = \frac{1}{2\,b\,\left|p - 2(b+1)\frac{||x||^2}{||x||^2+a}\right|}\;\frac{1}{(||x||^2+a)^{b-1}}.$$

a) Suppose that $0 < 2\,(b+1) < p$ and $a \geq 0$. Then

$$\frac{\xi_b^2(x)}{|\Delta\,\xi_b(x)|} \leq \frac{1}{2\,b(p-2(b+1))}\;\frac{1}{(\|x\|^2+a)^{b-1}}$$

and $E_\theta\left[\left(\frac{\Delta\,\xi_b(X)}{\xi_b(X)}\right)^2\right] < \infty$ since it is proportional to $E_\theta\left[\frac{1}{(\|X\|^2+a)^2}\right]$ which is finite for $a > 0$ or for $a = 0$ and $p > 4$.

Suppose that $m(x)$ is greater than or equal to some multiple of $\left(\frac{1}{\|x\|^2+a)}\right)^{b-1}$ or equivalently

$$m(x) \geq \frac{k}{2\,b(p-2(b+1))}\;\left(\frac{1}{\|x\|^2+a}\right)^{b-1} \tag{8.33}$$

for some $k > 0$. Theorem 8.2 implies that $\delta_m(X)$ is inadmissible and is dominated by

$$\delta_m(X) - \frac{\alpha}{m(X)(\|X\|^2+a)^b}$$

for

$$0 < \alpha < 4\,b(p-2(b+1))\inf_{x\in\mathbb{R}^p}(m(x)(\|x\|^2+(1)^{b-1}).$$

Alternatively, if $m(x) \geq \frac{k}{(\|x\|^2+a)^c}$ for $0 < c < \frac{p-4}{2}, \delta_m$ is inadmissible and the above gives an explicit improvement upon substituting $c-1$ for b. Note that the improved estimators shrink towards 0.

Suppose, for example, that $m(x) \equiv 1$. Then (8.33) is satisfied for $b \geq 1$. Here $\varphi_m(X) = X$ and $\delta_m(X) = p$. Choosing $b = 1$, an improved class of estimators is given by $p - \frac{\alpha}{\|X\|^2+a}$ for $0 < \alpha < 4\,(p-4)$. The case $a = 0$ is equivalent to Johnstone's result for this marginal.

(2) Suppose that $2\,(b+1) > p > 4$ and $a = 0$. Then

$$\frac{\xi_b^2(x)}{|\Delta \xi_b(x)|} = \frac{1}{2\,b(2(b+1)-p)} \, \frac{1}{\|x\|^{2(b-1)}}.$$

A development similar to the above implies that, when $m(x)$ is greater than or equal to some multiple of $\|x\|^{2(1-b)}$, an improved estimator is

$$\delta_m(X) + \frac{\alpha}{m(X)\|X\|^{2b}}$$

for

$$0 < \alpha < 4\,b(2\,(b+1)-p) \inf_{x \in \mathbb{R}^p} \Big(m(x)\|x\|^{2(b-1)} \Big).$$

Note that, in this case, the correction term is positive and hence the estimators expands away from 0. Note also that this result only works for $a = 0$ and hence applies to pseudo-marginals which are unbounded in a neighborhood of 0. Since all marginals corresponding to a generalized prior π are bounded, this result can never apply to generalized Bayes procedures but only to pseudo-Bayes procedures.

Suppose, for example, that $m(x) = \|x\|^{2-p}$. Here $\varphi_m(X) = \Big(1 - \frac{p-2}{\|X\|^2}\Big) X$ is the James-Stein estimator and $\delta_m(X) = p - \frac{(p-2)^2}{\|X\|^2}$. In particular, the above applies for $b - 1 = \frac{p-2}{2}$, that is, for $b = \frac{p}{2} > \frac{p-2}{2}$. An improved estimator is given by $\delta_m(X) + \frac{\gamma}{\|X\|^2}$ for $0 < \gamma < 4\,p$. This again agrees with Johnstone's result for James-Stein estimators.

8.3 Estimating the quadratic loss of a p-normal mean with unknown variance

In Section 8.2 it was assumed that the covariance matrix was known and equal to the identity matrix I_p. Typically, this covariance is unknown and should be estimated. In the case where it is of the form $\sigma^2 I_p$ with σ^2 unknown, Wan and Zou [143] show that, for the invariant loss $||\varphi(X) - \theta||^2/\sigma^2$, Johnstone's result [80] can be extended when estimating the loss of the James-Stein estimator. In fact, the general framework considered in Section 8.2 can be extended to the case where σ^2 is unknown, and we show that a condition parallel to Condition (8.19) can be found.

Consider an estimator of θ of the form $\varphi(X,S) = X + S\,g(X,S)$ with $E_{\theta,\sigma^2}[S^2\,||g(X,S)||^2] < \infty$, where E_{θ,σ^2} denotes the expectation with respect to the joint distribution of (X,S).

Then, by Theorem 3.5, an unbiased estimator of the loss $||\varphi(X,S) - \theta||^2/\sigma^2$ is

$$\delta_0(X,S) = p + S\left\{(k+2)\,||g(X,S)||^2 + 2\,\mathrm{div}_X g(X,S) + 2\,S\frac{\partial}{\partial S}||g(X,S)||^2\right\}. \tag{8.34}$$

The following theorem provides an extension of results in Section 8.2 to the setting of an unknown variance. The necessary conditions to insure the finiteness of the risks are parallel to the case where the variance σ^2 is known. It should be noticed that the corresponding domination condition of $\delta(X,S)$ over $\delta_0(X,S)$, that is, for any $x \in R^p$ and any $s \in \mathbb{R}_+$, $(n+2)\,||g(x,s)||^2 + 2\,\mathrm{div}_x g(x,s) + 2\,s\frac{\partial}{\partial s}||g(x,s)||^2 \leq 0$, entails that the two conditions $E_{\theta,\sigma^2}[(S\,\mathrm{div} g(X,S))^2] < \infty$ and

$E_{\theta,\sigma^2}\left[\left(S^2\,\frac{\partial}{\partial S}||g(X,S)||\right)^2\right]$ imply the condition $E_{\theta,\sigma^2}[S^2\,||g(X,S)||^4] < \infty$. Also the derivation of the finiteness of $R(\theta,\sigma^2,\varphi)$ follows as in the known variance case.

Theorem 8.3. *Let $X \sim \mathscr{N}(\theta,\sigma^2 I_p)$ where θ and σ^2 are unknown and $p \geq 5$ and let S be a non negative random variable independent of X and such that $S \sim \sigma^2\chi^2_k$.*

For any function $\gamma(X)$ such that $E_{\theta,\sigma^2}[\gamma^2(X)] < \infty$, the risk difference $\mathscr{D}(\theta,\sigma^2,\varphi,\delta) = \mathscr{R}(\theta,\sigma^2,\varphi,\delta) - \mathscr{R}(\theta,\sigma^2,\varphi,\delta_0)$ between the estimators $\delta(X,S) = \delta_0(X,S) - S\,\gamma(X)$ and $\delta_0(X,S)$ is given by

$$E_{\theta,\sigma^2}\left[S^2\left\{\gamma^2(X) + \frac{2}{k+2}\,\Delta\gamma(X) + 4\,g'(X,S)\nabla\gamma(X) + 4\,\gamma(X)\,||g(X,S)||^2\right\}\right], \tag{8.35}$$

so that a sufficient condition for $\mathscr{D}(\theta,\sigma^2,\varphi,\delta)$ to be non positive, and hence for $\delta(X,S)$ to improve on $\delta_0(X,S)$, is

$$\gamma^2(x) + \frac{2}{k+2}\,\Delta\gamma(x) + 4\,g^t(x,s)\,\nabla\gamma(x) + 4\,\gamma(x)\,||g(x,s)||^2 \leq 0 \tag{8.36}$$

for any $x \in \mathbb{R}^p$ and any $s \in \mathbb{R}_+$.

Proof. Consider the finiteness of the risk of the alternative loss estimator $\delta(X,S) = \delta_0(X,S) - S\,\gamma(X)$. It is easily seen that its difference in loss $d(\theta,\sigma^2,X,S)$ with $\delta_0(X,S)$ can be written as

$$\begin{aligned} &d(\theta,\sigma^2,X,S) \\ &= \left(\delta_0(X,S) - \frac{1}{\sigma^2}||\varphi(X)-\theta||^2 - S\,\gamma(X)\right)^2 - \left(\delta_0(X,S) - \frac{1}{\sigma^2}||\varphi(X)-\theta||^2\right)^2 \\ &= S^2\,\gamma^2(X) - 2\,S\,\gamma(X)\,(\delta_0(X,S) - \frac{1}{\sigma^2}||\varphi(X)-\theta||^2)\,. \end{aligned} \tag{8.37}$$

Hence, since $E_{\theta,\sigma^2}[||\varphi(X,S)-\theta||^2/\sigma^2] < \infty$ as the risk of the estimator $\varphi(X,S)$, the condition $E_{\theta,\sigma^2}[\gamma^2(X)] < \infty$ insures that the expectation of the loss in (8.37), that is, the risk difference $\mathscr{D}(\theta,\sigma^2,\varphi,\delta)$ is finite. Then $\mathscr{R}(\theta,\sigma^2,\varphi,\delta) < \infty$ since $\mathscr{R}(\theta,\sigma^2,\varphi,\delta_0) < \infty$.

We now express the risk difference $\mathscr{D}(\theta,\sigma^2,\varphi,\delta) = E_{\theta,\sigma^2}[d(\theta,\sigma^2,X,S)]$. Using (8.34) and expanding $||\varphi(X,S)-\theta||^2/\sigma^2$ give that $d(\theta,\sigma^2,X,S)$ in (8.37) can be written as

$$d(\theta,\sigma^2,X,S) = A(X,S) + B(\theta,\sigma^2,X,S)$$

where

$$\begin{aligned} A(X,S) = S^2\,\gamma^2(X) \;-\; 2\,p\,S\,\gamma(X) - 2\,(k+2)\,S^2\,\gamma(X)\,||g(X,S)||^2 \\ -\; 4S^2\,\gamma(X)\,\mathrm{div}_X g(X,S) - 4S^3\,\gamma(X)\,\frac{\partial}{\partial S}||g(X,S)||^2 \end{aligned} \tag{8.38}$$

and

$$\begin{aligned} B(\theta,\sigma^2,X,S) = 2\,\frac{S^3}{\sigma^2}\,\gamma(X)\,||g(X,S)||^2 + 2\,\frac{S}{\sigma^2}\,\gamma(X)\,||X-\theta||^2 \\ +4\,\frac{S^2}{\sigma^2}\,\gamma(X)\,(X-\theta)^t g(X,S)\,. \end{aligned} \tag{8.39}$$

Through Lemma 3.1 (2) with $h(x,s) = 2\,\frac{s^3}{\sigma^2}\,\gamma(x)\,||g(x,s)||^2$, the expectation of the first term in the right hand side of (8.39) equals

$$\begin{aligned} E_{\theta,\sigma^2}\left[2\,\frac{S^3}{\sigma^2}\,\gamma(X)\,||g(X,S)||^2\right] = E_{\theta,\sigma^2}\big[2\,(k+4)\,S^2\,\gamma(X)\,||g(X,S)||^2 \\ +\; 4S^3\,\gamma(X)\,\frac{\partial}{\partial S}||g(X,S)||^2\Big]\,. \end{aligned} \tag{8.40}$$

Also a reiterated application of Lemma 3.1 (1) to the expectation of the second term in the right hand side of (8.39) allows to write

$$
\begin{aligned}
E_{\theta,\sigma^2}\left[2\,\frac{S}{\sigma^2}\,\gamma(X)\,||X-\theta||^2\right] &= E_{\theta,\sigma^2}[2\,\frac{1}{\sigma^2}\,(X-\theta)^t\,S\,\gamma(X)\,(X-\theta)] \\
&= E_{\theta,\sigma^2}[2\,\mathrm{div}_X\{S\,\gamma(X)\,(X-\theta)\}] \\
&= E_{\theta,\sigma^2}[2\,p\,S\,\gamma(X)+2\,S\,(X-\theta)^t\nabla\gamma(X)] \\
&= E_{\theta,\sigma^2}[2\,p\,S\,\gamma(X)+2\,\sigma^2\,S\,\Delta\,\gamma(X)]
\end{aligned}
$$

which, as $S \sim \sigma^2\chi^2_k$ entails that $E[S^2/(k+2)] = E[\sigma^2\,S]$ and as S is independent of X, gives

$$
E_{\theta,\sigma^2}\left[2\,\frac{S}{\sigma^2}\,\gamma(X)\,||X-\theta||^2\right] = E_{\theta,\sigma^2}\left[2\,p\,S\,\gamma(X)+2\,\frac{S^2}{k+2}\,\Delta\,\gamma(X)\right]. \quad (8.41)
$$

As for the third term in the right hand side of (8.39), its expectation can also also be expressed using Lemma 3.1 (1) as

$$
\begin{aligned}
E_{\theta,\sigma^2}\!\left[4\,\frac{S^2}{\sigma^2}\,\gamma(X)\,(X-\theta)^t g(X,S)\right] &= E_{\theta,\sigma^2}[4\,S^2\,\mathrm{div}_X\{\gamma(X)\,g(X,S)\}] \\
= E_{\theta,\sigma^2}[4\,S^2\,\gamma(X)\,\mathrm{div}_X\{g(X,S)\} &+ 4\,S^2\,g(X,S)^t\,\nabla\gamma(X)] \qquad (8.42)
\end{aligned}
$$

by propriety of the divergence.

Finally, gathering (8.40), (8.41) and (8.42) yields an expression of (8.39) which, with (8.38), gives the integrand term of (8.35), which is the desired result. □

As an example, consider the James-Stein estimator $\varphi^{JS}(X,S) = X - \frac{p-2}{k+2}\,\frac{S}{||X||^2}X$ discussed in Section 3.4. Here the shrinkage factor only depends on X and equals $g(X) = -\frac{p-2}{k+2}\,\frac{X}{||X||^2}$ so that, through routine calculation, the unbiased estimator of loss is $\delta_0(X,S) = p - \frac{(p-2)^2}{k+2}\,\frac{S}{||X||^2}$. For a correction of the form $\gamma(x) = -d/||x||^2$ with $d \geq 0$, it is easy to check that the expression in (8.36) equals

$$d^2 + 4\frac{p-4}{k+2}d - 8\frac{p-2}{k+2}d - 4\left(\frac{p-2}{k+2}\right)^2 d = d\left(d - \frac{4}{k+2}\left[p + \frac{(p-2)^2}{k+2}\right]\right)$$

which is negative for $0 < d < \frac{4}{k+2}\left[p + \frac{(p-2)^2}{k+2}\right]$ and gives domination of $p - \frac{(p-2)^2}{k+2}\frac{S}{||X||^2} + \frac{d}{||x||^2}$ over $p - \frac{(p-2)^2}{k+2}\frac{S}{||X||^2}$. This condition recovers the result of Wan and Zou [143] who considered the case $d = \frac{2}{k+2}\left[p + \frac{(p-2)^2}{k+2}\right]$.

8.4 Extensions to the spherical case

8.4.1 Estimating the quadratic loss of the mean of a spherical distribution

In the previous section the loss estimation problem was considered for the normal distribution setting. The normal distribution has been generalized in two important directions, first as a special case of the exponential family and secondly as a spherically symmetric distribution. In this section we will consider the latter. There are a variety of equivalent definitions and characterizations of the class of spherically symmetric distributions, a comprehensive review is given by [50]. We will use the representation of a random variable from a spherically symmetric distribution, $X = (X_1, \ldots, X_p)'$, as $X \overset{d}{=} RU^{(p)} + \theta$, where $R = ||X - \theta||$ is a random radius, $U^{(p)}$ is a uniform random variable on the p-dimensional unit sphere, where R and $U^{(p)}$ are independent. In such situation, the distribution of X is said spherically symmetric around θ and we write $X \sim SS_p(\theta)$.

We also extend, in Subsection 8.4.2, these results to the case where the distribution of X is spherically symmetric and when a residual vector U is available (which allows an estimation of the variance factor σ^2).

Assume $X \sim SS_p(\theta)$ and suppose we wish to estimate $\theta \in \mathbb{R}^p$ by a decision rule $\delta(X)$ using quadratic loss. Suppose that we also use quadratic loss to assess the accuracy of loss estimate $\delta(X)$, then the risk of this loss estimate is given by (8.2). In [61], the problem of estimating the loss when $\varphi(X) = X$ is the estimate of the location parameter θ is considered. The estimate φ is the least squares estimator and is minimax among the class of spherically symmetric distributions with bounded second moment. Furthermore if one assumes the density of X exists and is unimodal, then φ is also the maximum likelihood estimator.

The unbiased constant estimate of the loss $||X-\theta||^2$ is $\delta_0 = E_\theta[R^2]$. Note that δ_0 is independent of θ, since $E_\theta[||X-\theta||^2] = E_0[||X||^2]$. Fourdrinier and Wells [61] show that the unbiased estimator δ_0 can be dominated by $\delta_0 - \gamma$, where γ is a particular superharmonic function for the case where the sampling distribution is a scale mixture of normals and in a more general spherical case.

The development of the results depends on some interesting extensions of the classical Stein identities in (8.13) and (8.18) to the general spherical setting. Since the distribution of X, say P_θ, is spherically symmetric around θ, for every bounded function f, we have $E_\theta[f] = EE_{R,\theta}[f] = \int_{\mathbb{R}_+} E_{R,\theta}[f]\rho(dR)$, where ρ is the distribution of the radius, namely the distribution of the norm $||X-\theta||$ under P_0 and where E and $E_{R,\theta}$ denotes respectively the expectation with respect to the radial distribution and uniform distribution $U_{R,\theta}$ on the sphere $S_{R,\theta} = \{x \in \mathbb{R}^p |\; ||x-\theta|| = R\}$ of radius

R and center θ. To deduce the various risk domination results it suffices to work conditionally on the radius, that is to say to replace P_θ by $U_{R,\theta}$ in the risk expressions. Denote $\sigma_{R,\theta}$ as the area measure on $S_{R,\theta}$. Therefore, for every Borel measurable set A, $U_{R,\theta}(A) = \sigma_{R,\theta}(A)/\sigma(S_{R,\theta}) = \Gamma(p/2)\sigma_{R,\theta}(A)/2\pi^{p/2}R^{p-1}$. Define the volume measure $\tau_{R,\theta}$ on the ball $B_{R,\theta} = \{x \in E \,|\, ||x-\theta|| \leq R\}$ of radius R and center θ and denote the uniform distribution on $B_{R,\theta}$ as $V_{R,\theta}$. Hence, for every Borel measurable set A, $V_{R,\theta}(A) = \tau_{R,\theta}(A)/\tau_{R,\theta}(B_{R,\theta}) = p\Gamma(p/2)\tau_{R,\theta}(A)/2\pi^{p/2}R^p$. Suppose γ is a weakly differentiable vector valued function, then by applying the Divergence Theorem for weakly differentiable functions to the definition of the expectation we have

$$\begin{aligned} E_\theta[(X-\theta)^t\gamma(X) \mid ||X-\theta|| = R] &= \int_{S_{R,\theta}} (x-\theta)^t \, \gamma(x) U_{R,\theta}\,(dx) \qquad (8.43) \\ &= \frac{R}{\sigma_{R,\theta}(S_{R,\theta})} \int_{B_{R,\theta}} \operatorname{div} \gamma(x)\, dx. \end{aligned}$$

If γ is a real valued function then it follows from (8.43) and the product rule applied to the vector valued function $(x-\theta)\gamma(x)$ that

$$\begin{aligned} &E_\theta[||X-\theta||^2\gamma(X) \mid ||X-\theta|| = R] \\ &= \int_{S_{R,\theta}} (x-\theta)^t\,(x-\theta)\,\gamma(x)\,U_{R,\theta}\,(dx) \\ &= \frac{R}{\sigma_{R,\theta}(S_{R,\theta})} \int_{B_{R,\theta}} [p\,\gamma(x) + (x-\theta)^t\,\nabla\gamma(x)]\,dx. \qquad (8.44) \end{aligned}$$

Our first extension of Theorem 8.1 is to the class of spherically symmetric distributions that are scale mixtures of normal distributions. Well known examples in the class of densities include the double exponential, multivariate t-distribution (hence, the multivariate Cauchy distribution). Let $\phi(x;\theta,I)$ be the probability density function of a random vector X with a normal distribution with mean vector θ and identity covariance matrix. Suppose that there is a probability measure on $\mathbb{R}_+$ such that the probability density function p_θ may be expressed as

$$p_\theta(x|\theta) = \int_0^\infty \phi(x;\theta, I/t) G(dt). \tag{8.45}$$

One can think of T being a random variable with distribution G, the conditional distribution of X given $T = t, X \mid T = t$, is $N_p(\theta, I/t)$. This class contains heavy tailed distributions, possibly with no moments. It is well known (see [50]) that, if a spherical distribution has a density p_θ, it is of the form $p_\theta(x) = g(||x-\theta||^2)$ for a measurable positive function g (called the generating function).

In the scale mixture of normals setting the unbiased estimate, δ_0, of risk equals

$$E[R^2] = E_\theta[||X-\theta||^2] = p\int_0^\infty t^{-1} G(dt).$$

It is easy to see that the risk of the unbiased estimator δ_0 is finite if and only if $E_\theta[||X-\theta||^4] < \infty$, which holds if

$$\int_0^\infty t^{-2} G(dt) < \infty. \tag{8.46}$$

The main theorem in [61] is the following domination result of an improved estimator of loss over the unbiased loss estimator.

Theorem 8.4. *Assume the distribution of X is a scale mixture of a normal random variables as in (8.45) such that (8.46) is satisfied and such that*

$$\int_{\mathbb{R}_+} t^{p/2} G(dt) < \infty. \tag{8.47}$$

Also, assume that the shrinkage function γ is twice weakly differentiable on $\mathbb{R}^p$ and satisfies $E_\theta[\gamma^2] < \infty$, for every $\theta \in \mathbb{R}^p$. Then a sufficient condition for $\delta_0 - \gamma$ to dominate δ_0 is that γ satisfies the differential inequality $k\Delta\gamma + \gamma^2 < 0$ with

$$k = 2\,\frac{\int_{\mathbb{R}_+} t^{p/2}\,G(dt)}{\int_{\mathbb{R}_+} t^{p/2-2}\,G(dt)}\,. \tag{8.48}$$

As an example let $\gamma(x) = c/||x||^2$ where c is a positive constant. Note that γ is only weakly differentiable (but not differentiable in the usual sense) and that is Laplacian exists as a locally integrable function) only when $p > 4$. Then it may be shown that $\Delta\gamma(x) = -2c(p-4)/||x||^4$. Hence $k\Delta(x) + \gamma^2(x) = -2kc(p-4)/||x||^4 + c^2/||x||^4 < 0$ if $-2kc(p-4) + c^2 < 0$, that is, $0 < c < 2k(p-4)$. It is easy to see that the optimal value of c for which this inequality is the most negative equals $k(p-4)$, so an interesting estimate in this class of γ's is $\delta = \delta_0 - k(p-4)/||x||^2 (p > 4)$. This is precisely the estimate proposed by [80] in the normal distribution case $N_p(\theta, I)$ where $k = 2$; recall, in that case $\delta_0 = p$. In this example we have assumed that the dimension p is greater than four. In general we can have domination as long as the assumptions of the theorem are valid. Actually, Blanchard and Fourdrinier [19] show explicitly that, when $p \leq 4$, the only solution γ in $L^2_{\text{loc}}(\mathbb{R}^p)$ of the inequality $k\Delta\gamma + \gamma^2 \leq 0$ is $\gamma \equiv 0$ (a.e., with respect to the Lebesgue measure). In the normal case $N_p(\theta, I/t)$ where $2t^{-2}\Delta\gamma + \gamma^2$ is an unbiased estimator of risk difference, hence, for dimensions four or less, it is impossible to find an estimator $\delta = \delta_0 - \gamma$ whose unbiased estimate of risk is always less that of δ_0 since we cannot have $E_\theta[2t^{-2}\Delta\gamma + \gamma^2] \leq 0$ without having $t^{-2}\Delta\gamma(x) + \gamma^2(x) \leq 0$ for some x.

In the case of scale mixture of normal distributions, the conjecture of admissibility of $\delta_0 - \gamma$ for lower dimensions, although it is probably true, remains open. Indeed, under conditions of Theorem 8.4, $k\Delta\gamma + \gamma^2$ is no longer an unbiased estimator of the risk difference and $E_\theta[k\Delta\gamma + \gamma^2]$ is only its upper bound. The use of Blyth's method would need to specify the distribution of X (that is, the mixture distribution G). It is worth noting that dimension-cutoff also arises through the finiteness of $E_\theta[\gamma^2]$ when using the classical shrinkage function $c/||x||^2$.

In order to prove Theorem 8.4 we need some additional technical results. The first lemma gives some important properties of superharmonic functions and is found in Du Plessis [119] and the second lemma links the integral of the gradient on a ball with the integral of the Laplacian.

Lemma 8.1. *If γ is a real valued superharmonic function then*

(1) $\int_{S_{R,\theta}} \gamma(x) U_{R,\theta}(dx) \leq \int_{B_{R,\theta}} \gamma(x) V_{R,\theta}(dx)$.

(2) *Both of the integrals in (i) are decreasing in R.*

Proof. See Sections 1.3 and 2.5 in [119]. □

Lemma 8.2. *Suppose γ is a twice weakly differentiable function. Then*

$$\int_{B_{R,\theta}} (x-\theta)' \nabla\gamma(x) V_{R,\theta}(dx) = \frac{p\Gamma(p/2)}{2\pi^{p/2}} \frac{1}{R^p} \int_0^R r \int_{B_{r,\theta}} \triangle\gamma(x)\, dx\, dr.$$

Proof. Since the density of the distribution of the radius under $V_{R,\theta}$ is $(p/R^p) r^{p-1}$, we have

$$\int_{B_{R,\theta}} (x-\theta)' \nabla\gamma(x) V_{R,\theta}(dx) = \int_0^R \int_{S_{r,\theta}} (x-\theta)^t \nabla\gamma(x) U_{r,\theta}(dx) \frac{p}{R^p} r^{p-1} dr.$$

The result follows from applying (8.44) to the inner most integral of the right hand side of this equality and by recalling the fact that $\sigma_{r,\theta}(S_{r,\theta}) = (2\pi^{p/2}/\Gamma(p/2))r^{p-1}$.

□

(Proof of Theorem 8.4) The risk difference between δ_0 and $\delta_0 - \gamma$ equals $\alpha(\theta) + \beta(\theta)$ where

$$\alpha(\theta) = 2p \int_{\mathbb{R}_+} \left(\frac{1}{t} - \frac{\delta_0}{p}\right) \int_{\mathbb{R}_+} \int_{S_{R,\theta}} \gamma(x)\, U_{R,\theta}\,(dx)\, \rho_t\,(dt)\, G\,(dt)$$

and

$$\beta(\theta) = \int_{\mathbb{R}_+} \int_{\mathbb{R}^p} (2t^{-2}\nabla\gamma(x) + \gamma^2(x)) \left(\frac{t}{2\pi}\right)^{p/2} \exp\left(-\frac{t}{2}||x-\theta||^2\right) dx\, G(dt)\,.$$

We have from the definition of $V_{R,\theta}$ and an application of Fubini's theorem

$$\begin{aligned}
&\int_{\mathbb{R}_+} R^2 \int_{B_{R,\theta}} \gamma(x) V_{R,\theta}\,(dx) \rho\,(dR) \\
&= p\frac{\Gamma(p/2)}{2\pi^{p/2}} \int_{\mathbb{R}_+} R^{2-p} \int_{B_{R,\theta}} \gamma(x)\, dx \rho\,(dR) \\
&= p\frac{\Gamma(p/2)}{2\pi^{p/2}} \int_{\mathbb{R}^p} \gamma(x) \int_{||x-\theta||}^{+\infty} R^{2-p} \rho\,(dR)\, dx \qquad (8.49)
\end{aligned}$$

Now, for fixed $t \geq 0$, in the normal case $N_p(\theta, I/t)$ the distribution ρ_t of the radius has the density f_t of the form $f_i(R) = \frac{t^{p/2}}{2^{p/2-1}\Gamma(p/2)} R^{p-1} \exp\{-\frac{tR^2}{2}\}$ and $\delta_0 = \frac{p}{t}$. Thus the expression (8.49) becomes

$$\begin{aligned}
&\int_{\mathbb{R}_+} R^2 \int_{B_{R,\theta}} \gamma(x) V_{R,\theta}\,(dx) \rho\,(dR) \\
&= \frac{pt^{p/2}}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} \gamma(x) \int_{||x-\theta||}^{+\infty} R \exp\left\{-\frac{tR^2}{2}\right\} dR\,dx \\
&= \frac{pt^{p/2-1}}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} \gamma(x) \exp\left\{-\frac{t}{2}||x-\theta||^2\right\} dx \\
&= \frac{p}{t} \int_{\mathbb{R}_+} \int_{S_{R,\theta}} \gamma(x) U_{R,\theta}\,(dx) \rho_t\,(dR),
\end{aligned}$$

the last equality holding since $X \overset{D}{=} RU^{(p)}$. Using the mixture representation with mixing distribution G, the expression of $\alpha(\theta)$ is written as

$$\alpha(\theta) = = 2p \int_{\mathbb{R}_+} \left(\frac{1}{t} - \frac{\delta_0}{p}\right) \int_{\mathbb{R}^p} \gamma(x) \left(\frac{t}{2\pi}\right)^{p/2} \exp\left(-\frac{t}{2}||x-\theta||^2\right) dx\,G(dt).$$

Since $\delta_0 = \frac{p}{t}$, the expression for $\alpha(\theta)$ is a covariance with respect to G and is nonpositive by Lemma 8.1.

We can now treat the integral of the expression $\beta(\theta)$ in the same manner. The function $x \to (x-\theta)'\nabla\gamma(x)$ and the function $x \to \nabla\gamma(x)$ taking successively the role of the function γ, we obtain

$$\begin{aligned}
&\int_{\mathbb{R}_+} \frac{R^2}{p} \int_{B_{R,\theta}} (x-\theta)'\nabla\gamma(x) V_{R,\theta}\,(dx) \rho_t\,(dR) \\
&= \frac{1}{t} \int_{\mathbb{R}_+} \int_{S_{R,\theta}} (x-\theta)'\nabla\gamma(x) U_{R,\theta}\,(dx) \rho_t\,(dR) \\
&= \frac{1}{t} \int_{R_+} \frac{R^2}{p} \int_{B_{R,\theta}} \nabla\gamma(x)\,dx \rho_t\,(dR) \\
&= \frac{t^{p/2-2}}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} \nabla\gamma(x) \exp\left\{-\frac{t}{2}||x-\theta||^2\right\} dx
\end{aligned}$$

applying (8.43) for the second equality and remembering that $\triangle\gamma = div(\nabla\gamma)$. Therefore by Fubini Theorem $\beta(\theta)$ can be reexpressed as

$$\beta(\theta) = \int_{\mathbb{R}^p} \left(2 \triangle \gamma(x) \frac{\int_{\mathbb{R}_+} t^{p/2-2} \exp(-t||x-\theta||^2/2) G(dt)}{\int_{\mathbb{R}_+} t^{p/2} \exp(-t||x-\theta||^2/2) G(dt)} + \gamma^2(x) \right)$$
$$\times \int_{\mathbb{R}_+} \left(\frac{t}{2\pi}\right)^{p/2} \exp\left(-\frac{t}{2}||x-\theta||^2\right) G(dt)\, dx.$$

Now, by the superharmonicity condition on γ, assumption (8.48) gives

$$\beta(\theta) \leq \int_{\mathbb{R}^p} (k \triangle \gamma(x) + \gamma^2(x)) \int_{\mathbb{R}_+} \left(\frac{t}{2\pi}\right)^{p/2} \exp\left(-\frac{t}{2}||x-\theta||^2\right) G(dt)\, dx \quad \square$$

The improved loss estimator result in Theorem 8.4 for scale mixture of normal distributions family was extended to the more general family of spherically symmetric distributions in [61]. In this setting the conditions for improvement rest on the generating function g of the spherical density p_θ. A sufficient condition for domination of δ_0 has the usual form $k\nabla\gamma + \gamma^2 \leq 0$.

Theorem 8.5. *Assume the spherical distribution of X with generating function g has finite fourth moment. Assume the function γ is non negative and twice weakly differentiable on $\mathbb{R}^p$ and satisfies $E_\theta[\gamma^2] < \infty$. If, for every $s \geq 0$,*

$$\int_s^\infty g(z)\, dz 2g(s) \leq p\, \delta_0 \tag{8.50}$$

and if there exists a constant k such that, for any $s \geq 0$

$$0 < k < \frac{\int_s^\infty z g(z)\, dz - s \int_s^\infty g(z)\, dz}{2g(s)}. \tag{8.51}$$

Then a sufficient condition for $\delta_0 - \gamma$ to dominate δ_0 is that γ satisfies the differential inequality: $k\Delta\gamma + \gamma^2 < 0$.

We have shown that one can dominate the unbiased constant estimator of loss by a shrinkage-type estimator. As in the normal case one may wish to add a frequentist-validity constraint to the loss estimation problem. It is easy to show that the only frequentist valid estimator of the form δ_0 would be the only frequentist valid loss estimator. The proof of this result follows from a randomization of the origin technique as in Hsieh and Hwang [76].

8.4.2 Estimating the quadratic loss of the mean of a spherical distribution with a residual vector

In this subsection we extend the ideas of the previous sections to a spherically symmetric distribution with a residual vector. We first develop an unbiased estimator of the loss and then construct a dominating shrinkage-type estimator. An important feature of our results is that the proposed loss estimates dominate the unbiased estimates for the entire class of spherically symmetric distributions. That is, the domination results are robust with respect to spherical symmetry.

Let $(X,U) \sim SS(\theta,0)$ where $\dim X = \dim \theta = p$ and $\dim U = \dim 0 = k$ $(p+k = n)$. For convenience representation, here (X,U) and $(\theta,0)$ represent $n \times 1$ vectors Unlike Subsection 8.4.1, the dimension of the observable (X,U) is greater than the dimension of the estimand θ. Recall from Subsection 2.6.1 that this model arises as the canonical form of the following seemingly more general model, the general linear model.

The usual estimator of θ is the orthogonal projector X. A class of competing point estimators which are also considered are of the form $\varphi = X - ||U||^2 g(X)$, g is a measurable function from $\mathbb{R}^p$ into $\mathbb{R}^p$. This class of estimators is closely related to a Stein-like estimators (when estimating the mean of a normal distribution, the square of the residual term $||u||$ is used as an estimate of the unknown variance). Their domination properties are robust with respect to spherical symmetry (cf. [37] and [38]). We will first consider estimation of the loss of the usual least squares estimator X then estimation of the loss of the more general shrinkage estimator φ.

In the spherical case in Section 8.3, the risk of X was constant with respect to θ. Thus this risk provides an unbiased estimator of the loss, that is, $\frac{p}{n}E[R^2]$, which is subject to the knowledge of $E[R^2]$. Its properties, as the properties of any improved estimator, may depend on the specific underlying distribution. An important feature of the results in this subsection is that we propose an unbiased estimator δ_0 of the loss of X which is available for every spherically symmetric distribution (with finite fourth moment), that is, $\delta_0(x) = k||u||^2/(n-k)$. Thus we do not need to know the specific distribution, and we get robustness with an estimator which is no longer constant. Notice δ_0 makes sense because $p < n$.

In this subsection we consider estimation of X. An unbiased estimator of the loss of X of θ is given by $\delta_0 = p||U||^2/k$. The unbiasedness of δ_0 follows from Corollary 8.1 by taking $q = 0$ and $\gamma \equiv 1$. The goal of this subsection is to prove the domination of the unbiased estimator δ_0 by a competing estimator δ of the form

$$\delta = \delta_0 - ||U||^4\,\gamma(X), \tag{8.52}$$

where γ is a positive function. It is important to notice that the "residual term" $||U||$ appears explicitly in the shrinkage function. It has been noted in [37] that the use of this term allows fewer assumptions about the distributions than when it does not appear. Specifically, this including gives a robustness property to the results, since they are valid for the entire class of spherically symmetric distributions.

We require the real-valued function γ to be twice weakly differentiable, in order to include basic examples, which are not twice differentiable.

Theorem 8.6. *Assume that $p \geq 5$, the distribution of (X, U) has a finite fourth moment and the function γ is twice weakly differentiable on $\mathbb{R}^p$. A sufficient condition under which the estimator in (8.52) dominates the unbiased estimator δ_0 is that γ satisfies the differential inequality*

$$\gamma^2 + \frac{2}{(k+4)(k+6)} \triangle \gamma \leq 0 \,. \tag{8.53}$$

The standard example where $\gamma(t) = d/||t||^2$ for all $t \neq 0$ with $d > 0$ satisfies the conditions of the theorem. More precisely it is easy to deduce that $\triangle\gamma(t) = -2d(p-4)/|||t||^4$ and thus the sufficient condition of the theorem is written as $0 < d \leq 4(p-4)/(k+4)(k+6)$, which only occurs when $p \geq 5$. Straightforward calculus shows that the optimal value of d is given by $2(p-4)/(k+4)(k+6)$. The optimal constant in [10] is equal to $2(p-4)$. The extra terms in the denominator compensate for the $||U||^4$ term in our estimator.

We will now consider the estimation of the loss of a class of shrinkage estimators considered in [37], that is for location estimators of the form

$$\varphi_g = X - ||U||^2 \, g(X), \tag{8.54}$$

where g is a weakly differentiable function from $\mathbb{R}^p$ into $\mathbb{R}^p$. In [37] it is shown that, if $||g||^2 \leq 2\ \text{div}g/(n-k+2)$, φ_g dominates X, under quadratic loss for all spherically symmetric distributions with a finite second moment. A general example of a member of this class of estimators is with $g(X) = r(||X||^2)\frac{A(X)}{b(X)}$, where r is a positive differentiable and nondecreasing function, A is a positive definite symmetric matrix and b is a positive definite quadratic form of $\mathbb{R}^p$. When r is equal to some constant a, A is the identity on $\mathbb{R}^p$ and the quadratic form b is the usual norm, g reduces to $a/||X||^2$. It can be shown that the optimal choice of a equals $(p-2)/(k+2)$. A member of the class is $\varphi_r = X - (p-2)\frac{||U||^2}{k+2}\frac{X}{||X||^2}$, the James-Stein estimator used when the variance is unknown as in Section 8.3.

In Proposition **??** of Section 2.3 of [37] it is shown that an unbiased estimator of the loss of the shrinkage estimator φ_g is given by

$$\delta_0^g = \frac{p}{k}||U||^2 + \left(||g(X)||^2 - \frac{2}{k+2}\ \text{div}\ g(X)\right)||U||^4. \tag{8.55}$$

As in Theorem 8.6 above, the unbiased estimator of the loss can be improved by a shrinkage estimator of the loss. Thus the competing estimator we consider is

$$\delta_\gamma^g = \delta_0^g - ||U||^4\gamma(X), \tag{8.56}$$

where γ is a positive function. Note that (8.56) is a true shrinkage estimator, while Johnstone's [80] optimal loss estimate for the normal case is an expanding estimator. This is not contradictory since we are using a different estimator than Johnstone and he is only dealing with the normal case. If $g \equiv 0$ the following result reduces to Theorem 8.6.

Theorem 8.7. *Assume that $p \geq 5$, the distribution of (X,U) has a finite fourth moment and the function γ is twice weakly differentiable on $\mathbb{R}^p$. A sufficient condition under which the estimator δ_γ^g given in (8.56) dominates the unbiased estimator δ_0^s is that γ satisfies the differential inequality*

$$\gamma^2 + \frac{4}{k+2}\gamma \operatorname{div} g - \frac{4}{k+6}\operatorname{div}(\gamma g) + \frac{2}{(k+4)(k+6)} \triangle \gamma \leq 0\,. \tag{8.57}$$

Before proving the theorem, we need some preliminary integration identities.

Lemma 8.3. *For every twice weakly differentiable function $g\,(\mathbb{R}^p \to \mathbb{R}^p)$ and for every function $h(\mathbb{R}_+ \to \mathbb{R})$,*

$$E_{R,\theta}\left[h(||U||^2)(X-\theta)' g(X)\right] = E_{R,\theta}\left[\frac{H(||U||^2)}{(||U||^2)^{\frac{k}{2}-1}} \operatorname{div} g(X)\right], \tag{8.58}$$

where H is the indefinite integral, vanishing at 0, of the function $t \to \frac{1}{2}h(t)t^{\frac{k}{2}-1}$ and provided the expectations exist.

Proof. We have

$$\begin{aligned}
& E_{R,\theta}\left[h(||U||^2)(X-\theta)' g(X)\right] \\
&= C_R^{p,k} \int_{B_{R,\theta}} h(R^2 - \|x-\theta\|^2)(x-\theta)^t g(x) \left(R^2 - \|x-\theta\|^2\right)^{\frac{k}{2}-1} dx \\
&= C_R^{p,k} \int_{B_{R,\theta}} (\nabla H(R^2 - \|x-\theta\|^2))' g(x)\, dx
\end{aligned}$$

since

$$\begin{aligned}
\nabla H(R^2 - \|x-\theta\|^2) &= -2H'(R^2 - \|x-\theta\|^2)(x-\theta) \\
&= h(R^2 - \|x-\theta\|^2)\left(R^2 - \|x-\theta\|^2\right)^{\frac{k}{2}-1}(x-\theta)\,.
\end{aligned}$$

Then, by divergence formula,

$$E_{R,\theta}\left[h(||U||^2)(X-\theta)'g(X)\right] = C_R^{p,k}\int_{B_{R,\theta}} \operatorname{div}\left(H(R^2-\|x-\theta\|^2)g(x)\right)dx$$
$$- C_R^{p,k}\int_{B_{R,\theta}} H(R^2-\|x-\theta\|^2)\operatorname{div} g(x)\,dx$$

Now, if $\sigma_{R,\theta}$ denotes the area measure on the sphere $S_{R,\theta}$, the divergence theorem insures that the first integral equals

$$C_R^{p,k}\int_{S_{R,\theta}} (H(R^2-\|x-\theta\|^2)g(x))'\frac{x-\theta}{\|x-\theta\|}\sigma_{R,\theta}\,(dx)$$

and is null since, for $x \in S_{R,\theta}$, $R^2-\|x-\theta\|^2=0$ and $H(0)=0$. Hence, in terms of expectation, we have

$$E_{R,\theta}\left[h(||U||^2)(X-\theta)'g(X)\right]$$
$$= -C_R^{p,k}\int_{B_{R,\theta}} \frac{H(R^2-\|x-\theta\|^2)}{(R^2-\|x-\theta\|^2)^{\frac{k}{2}}-1}\operatorname{div} g(x)\left(R^2-\|x-\theta\|^2\right)^{\frac{k}{2}-1}dx$$
$$= -E_{R,\theta}\left[\frac{H(||U||^2)}{(||U||^2)^{\frac{k}{2}-1}}\operatorname{div} g(X)\right]$$

which is the desired result. □

Corollary 8.1. *For every twice weakly differentiable function $\gamma(\mathbb{R}^p \to \mathbb{R}_+)$ and for every integer q,*

$$E_{R,\theta}\left[||U||^q\,||X-\theta||^2\,\gamma(X)\right] = \frac{p}{k+q}E_{R,\theta}\left[||U||^{q+2}\,\gamma(X)\right]$$
$$+ \frac{1}{(k+q)(k+q+2)}E_{R,\theta}\left[||U||^{q+4}\,\triangle\,\gamma(X)\right].$$

provided the expectation exists.

Proof. Take $h(t)=t^{q/2}$ and $g(x)=\gamma(x)\,(x-\theta)$ and apply Lemma 8.3 twice. □

(Proof of Theorem 8.7) Since the distribution of (X,U) is spherically symmetric around θ, it suffices to obtain the result working conditionally on the radius. For $R>0$ fixed, we can compute using the uniform distribution $U_{R,\theta}$ on the sphere $S_{R,\theta}$. Hence, the risk of δ_γ^g equals

$$\begin{aligned} E_{R,\theta}\left[(\delta_\gamma^g - ||\varphi-\theta||^2)^2\right] &= E_{R,\theta}\left[(\delta_0^g - ||\varphi-\theta||^2)^2\right] + E_{R,\theta}\left[||U||^8\,\gamma^2(X)\right] \\ &\quad - 2E_{R,\theta}\left[||U||^4\gamma(X)(\delta_0^g - ||\varphi-\theta||^2)\right]. \end{aligned}$$

Applying Lemma 8.3, it follows that

$$2\,E_{R,\theta}\left[||U||^6\,(X-\theta)'\gamma(X)\,g(X)\right] = \frac{2}{k+6}E_{R,\theta}\left[||U||^8\,\mathrm{div}\,(\gamma(X)\,g(X))\right]$$

and expanding the risk and Corollary 8.1 it follows that the risk of δ_γ^g equals

$$\begin{aligned} & E_{R,\theta}\left[(\delta_0^g - ||\varphi-\theta||^2)^2\right] + E_{R,\theta}\left[||U||^8\gamma^2(X)\right] \\ & - \frac{8k}{(k)(k+4)}E_{R,\theta}\left[||U||^6\gamma(X)\right] \\ & + E_{R,\theta}\left[||U||^8\left\{\frac{4}{k+2}\,\gamma(X)\,\mathrm{div}\,g(X) - \frac{4}{k+6}\,\mathrm{div}\,(\gamma(X)\,g(X))\right\}\right] \\ & + \frac{2}{(k+4)(k+6)}E_{R,\theta}\left[||X-X||^8\triangle\gamma(X)\right]. \end{aligned}$$

Since the function γ is positive, the third term on the right-hand side is negative and the $||U||^8$ term occurs in the other expressions; hence, the sufficient condition for domination is

$$\gamma^2 + \frac{4}{k+2}\gamma\,\mathrm{div}\,g - \frac{4}{k+6}\,\mathrm{div}\,(\gamma g) + \frac{2}{(k+4)(k+6)}\,\triangle\,\gamma \le 0$$

in order that the inequality $R(\delta^g,\theta,\varphi) \le R(\delta_0^g,\theta,\varphi)$ holds. □

8.5 Confidence set assessment

In the previous sections, the usual quadratic loss $L(\theta, \varphi(x)) = ||\varphi(x) - \theta||^2$ was considered to evaluate various estimators $\varphi(X)$ of θ. The squared norm $||x-\theta||^2$ was crucial in the derivation of the properties of the loss estimators in conjunction with its role in the normal density or, more generally, in a spherical density. One could imagine other losses, but, to deal with tractable calculations, it matters to keep the Euclidean norm as a component of the loss in use. Hence a natural extension is to consider losses which are function of $||x-\theta||^2$, that is, of the form $c(||x-\theta||^2)$ for a nonnegative function c defined on $\mathbb{R}_+$.

[24], consider a nondecreasing and concave function c of $\|x-\theta\|^2$ in order to compare various estimators $\delta(X)$ of θ. As in the case tackled by [80] and [61] it is still of interest to assess the loss of $\delta(X) = X$, that is, to estimate $c\left(\|x-\theta\|^2\right)$.

Note that estimating $c(||x-\theta||^2)$ can be view as an evaluation of a quantity which is not necessarily a loss. Indeed it includes the problem of estimating the confidence statement of the usual confidence set $\{\theta \in \mathbb{R}^p \,|\, \|x-\theta\|^2 \leq c_\alpha\}$ with confidence coefficient $1-\alpha$: $c(\cdot)$ is the indicator function $1\!\!1_{[0,c_\alpha]}$ (the confidence interval estimation example seen in Section 8.1 has illustrated the necessity of a confidence evaluation depending on the data).

The problem of estimating a function $c(\cdot)$ of $||x-\theta||^2$ was tackled by Fourdrinier and Lepelletier [53] whose we follow the lines and refer to for more details.

Let X be a random vector in $\mathbb{R}^p$ with a spherical density of the form $x \mapsto f(\|x-\theta\|^2)$ where θ is the unknown location parameter. For a given nonnegative function

c on $\mathbb{R}_+$, we are interested in estimating the quantity $c(\|x-\theta\|^2)$ when x has been observed from X. In contrast to the previous sections here, if only X is considered as an estimator of θ, however the function $c(\cdot)$ intervenes in the quantity to estimate. A simple reference estimator is the unbiased estimator $\delta_0 = E_0[c(\|X\|^2)]$. Note that, in the confidence statement estimation problem, $\delta_0 = E_0[\,\mathbb{1}_{[0,c_\alpha]}] = 1-\alpha$ while, in the loss estimation problem of $||x-\theta||^2$ considered in the previous sections, $\delta_0 = E_0[\|X\|^2]$ (that is, p in the normal case).

Since it is a constant estimator, it is natural to search better estimators in the sense of the risk (8.2), that is, estimators $\varDelta$ such that

$$R_c(\delta,\theta) = E_\theta\left[\left(\delta(X) - c\big(\|X-\theta\|^2\big)\right)^2\right] \le E_\theta\left[\left(\delta_0 - c\big(\|X-\theta\|^2\big)\right)^2\right] = R_c(\delta_0,\theta)\,.$$

Improvement on γ_0 will be considered when its own risk is finite, that is, under the condition $E_0[c^2(\|X\|^2)] < \infty$, which also guarantees the existence of δ_0. We assume Condition (8.11) to assure the finiteness of the risk of $\delta(X)$.

Due to the presence of the function $c(\cdot)$, repeated use of Stein's identity is not appropriate any more to deal with the risk difference

$$\begin{aligned}\mathscr{D}_c(\theta,\delta) &= \mathscr{R}_c(\theta,\delta) - \mathscr{R}_c(\theta,\delta_0)\\ &= E_\theta[2\,\{\delta_0 - c(\|X-\theta\|^2)\}\,\gamma(X) + \gamma^2(X)]\end{aligned} \tag{8.59}$$

between $\delta(X)$ and δ_0. As an alternative, the approach in [53] consists in introducing the Laplacian of the correction function γ, say $\varDelta(\gamma)$, under the expectation sign in the right hand side of (8.59) and in developing an upper bound of the risk difference in terms of the expectation of a differential expression of the form $k\,\varDelta\gamma + \gamma^2$ where

k is a constant different from 0. The underlying idea is based on two facts. We know that, in the normal case, $\delta_0 = 1-\alpha$ is admissible for estimating a confidence statement (see [33]) and $\delta_0 = p$ is admissible for estimating the loss $||x-\theta||^2$ (see [80]) when $p \leq 4$ while, for $p \geq 5$, improved estimators are available, mainly through simulations in [120] and formally thanks to the differential inequation $2\Delta(\gamma)+\gamma^2 \leq 0$ in [80]. Now we will see , in Section 8.6, that inequations of the form $k\Delta(\gamma)+\gamma^2 \leq 0$ have no nontrivial solution γ (that is, non equal to almost everywhere) when $p \leq 4$. Therefore it may be reasonable to think that, in (8.59), such operators of the form $k\Delta(\gamma)+\gamma^2$, and hence the Laplacian of γ, should play a role to obtain improved estimators when $p \geq 5$.

Here we illuminate the principle which leads to the role of $\Delta(\gamma)$, assuming that suitable regularity conditions on the various functions in use are satisfied to make valid what it is stated; we will precise the appropriate conditions afterwards. First, it can be checked that, if K is the function depending on f and c defined, for any $t > 0$, by

$$K(t) = \frac{1}{p-2}\int_t^\infty \left[\left(\frac{y}{t}\right)^{p/2-1} - 1\right](\gamma_0 - c(y))\, f(y)\, dy.$$

then, for almost every $x \in \mathbb{R}^p$,

$$\Delta K(\|x-\theta\|^2) = 2(\gamma_0 - c(\|x-\theta\|^2))\, f(\|x-\theta\|^2)$$

and hence the first part of the expectation in (8.59) can be written as

$$E_\theta[2\,(\gamma_0 - c(\|X-\theta\|^2))\,\gamma(X)] = \int_{\mathbb{R}^p} \Delta K(\|x-\theta\|^2)\,\gamma(x)\,dx. \qquad (8.60)$$

Now, through an appropriate Green's formula, the Laplacian in (8.60) can be moved from the function K to the function γ, so that

$$\int_{\mathbb{R}^p} \Delta K(\|x-\theta\|^2)\,\gamma(x)\,dx = \int_{\mathbb{R}^p} K(\|x-\theta\|^2)\,\Delta\gamma(x)\,dx \tag{8.61}$$

and hence (8.60) can be written as

$$E_\theta[2\,(\gamma_0 - c(\|X-\theta\|^2))\,\gamma(X)] = E_\theta\left[\frac{K(\|X-\theta\|^2)}{f(\|X-\theta\|^2)}\,\Delta\gamma(X)\right]. \tag{8.62}$$

Therefore it follows from (8.62) an expression of the risk difference in (8.59) involving $\Delta\gamma(X)$, that is,

$$\mathscr{D}_c(\theta,\Delta) = E_\theta\left[\frac{K(\|X-\theta\|^2)}{f(\|X-\theta\|^2)}\,\Delta\gamma(X) + \gamma^2(X)\right]. \tag{8.63}$$

In [53], under the condition that $\delta_0 - c$ has only one sign change, two cases are considered for a domination result to be obtained: when $\delta_0 - c$ is first negative and then positive, the Laplacian of γ is assumed subharmonic while, when $\delta_0 - c$ is first positive and then negative, the Laplacian of γ is assumed superharmonic. Then, relying on the fact that f is bounded from above by a constant M, it is proved that

$$E_\theta\left[\frac{K(\|X-\theta\|^2)}{f(\|X-\theta\|^2)}\,\Delta s(X)\right] \le E_\theta[k\,\Delta s(X)]$$

with

$$k = \frac{1}{M} E_0[K(\|X\|^2)],$$

so that a sufficient condition for δ to dominate δ_0 is that γ satisfies the partial differential inequality

$$k\,\Delta\gamma(x) + \gamma^2(x) \le 0, \tag{8.64}$$

for any $x \in \mathbb{R}^p$.

Before commenting this result, we specify the regularity conditions (that we will call Conditions ($\mathscr{C}$) in the following) on γ, f and c under which it holds. In addition to the usual requirement that $E_\theta[\gamma^2] < \infty$ and $\gamma \in W^{2,1}_{loc}(\mathbb{R}^p)$, it is assumed that there exists $r > 0$ such that $\gamma \in C^2_b(\mathbb{R}^p \setminus B_r)$ the space of the functions twice continuously differentiable and bounded on $\mathbb{R}^p \setminus B_r$. Also it is supposed that the functions $f(\cdot)$ and $c(\cdot)$ are continuous on $\mathbb{R}^*_+$, except possibly on a finite set T, and that there exists $\varepsilon > 0$ such that f and $f(\cdot)c$ belong to $S^{0,p/2+1+\varepsilon}(\mathbb{R}^*_+ \backslash T)$, the space of the continuous functions v on $\mathbb{R}^*_+ \backslash T$ such that

$$\sup_{x \in \mathbb{R}^*_+ \backslash T; \beta \leq p/2+1+\varepsilon} ||x||^\beta \, |v(x)| < \infty .$$

For more on such spaces, see Appendix.

Typical solutions of (8.64) are functions of the form $\gamma(x) = -\text{sgn}(k)\, d/||x||^2$ with $0 \leq d \leq |k|\,(p-4)$. Intuitively, estimating a loss as $||x-\theta||^2$ (i.e. $c(t) = t$) is different from estimating a confidence statement (i.e. $c(t) = \mathbb{1}_{[0,c_\alpha]}(t)$) : we would like to deal with small losses and with large confidence statements. The two sign change conditions do report on these two situations. Thus, for the first problem, the function $\delta_0 - t = p - t$ is first positive and then negative; this is a case for which it can be shown that $k < 0$ (see [53]), so that a dominating loss estimator is $\delta(X) = \delta_0 - \gamma(X) = p + \text{sgn}(k)\, d/||x||^2 = p - d/||x||^2$ for $0 \leq d \leq -k\,(p-4)$. Now, for the second problem, the function $\delta_0 - \mathbb{1}_{[0,c_\alpha]}(t) = 1 - \alpha - \mathbb{1}_{[0,c_\alpha]}$ is first negative and then positive and it is shown in [53] that $k > 0$, so that a dominating loss estimator is $\delta(X) = \delta_0 - \gamma(X) = 1 - \alpha + \text{sgn}(k)\, d/||x||^2 = 1 - \alpha + d/||x||^2$ for $0 \leq d \leq k\,(p-4)$.

Note that the correction to δ_0 is downward (upward) by $d/||x||^2$ for the first (second) problem.

The use of the property that the generating function f is bounded by M gives rise to a constant k which may be small in absolute value and hence may reduce the scope of the possible corrections γ leading to improved estimators δ. In [53], it is given an additional condition, relying on the monotonicity of the ratio K/f, which avoids the use of M. Here is their result.

Theorem 8.8. *Assume that Conditions ($\mathscr{C}$) are satisfied and that the function $\delta_0 - c$ has only one sign change. In the case where $\delta_0 - c$ is first negative and then positive (first positive and then negative), assume that the Laplacian of γ is subharmonic (superharmonic). Finally assume that the functions K and K/f have the same monotonicity (both nonincreasing or both nondecreasing).*

Then a sufficient condition for δ to dominate δ_0 is that γ satisfies the partial differential inequality

$$\forall x \in \mathbb{R}^p \quad \kappa \Delta \gamma(x) + \gamma^2(x) \le 0 \tag{8.65}$$

with

$$\kappa = E_0 \left[\frac{K(\|X\|^2)}{f(\|X\|^2)} \right] .$$

Proof. We consider the case where $\gamma_0 - c$ is first negative and then positive (case where the function $\Delta\gamma$ is assumed subharmonic). The main point is to treat the left hand side of Inequality (8.62); it equals

$$
\begin{aligned}
&E_\theta[K(\|X-\theta\|^2)\,\Delta\gamma(X)] \\
&= \int_{\mathbb{R}^p} K(\|x-\theta\|^2)\,\Delta\gamma(x)\,f(\|x-\theta\|^2)\,dx \\
&= \int_0^\infty \int_{S_{r,\theta}} \Delta\gamma(x)\,dU_{r,\theta}(x)\,K(r^2)\,\frac{2\pi^{p/2}}{\Gamma(p/2)}\,r^{p-1}\,f(r^2)\,dr \qquad (8.66)
\end{aligned}
$$

where $U_{r,\theta}$ is the uniform distribution on the sphere $S_{r,\theta} = \{x \in \mathbb{R}^p \,|\, \|x-\theta\| = r\}$ of radius r and centered at θ. Note that the function $r \mapsto \frac{2\pi^{p/2}}{\Gamma(p/2)}\,r^{p-1}\,f(r^2)$ is the radial density, that is, the density of the radius $R = \|X-\theta\|$. Now the right hand side of (8.66) can be bounded above by

$$
\begin{aligned}
&\int_0^\infty \int_{S_{r,\theta}} \Delta\gamma(x)\,dU_{r,\theta}(x)\,\frac{2\pi^{p/2}}{\Gamma(p/2)} \\
&\qquad r^{p-1}\,f(r^2)\,dr \times \int_0^\infty \frac{K(r^2)}{f(r^2)}\,\frac{2\pi^{p/2}}{\Gamma(p/2)}\,r^{p-1}\,f(r^2)\,dr \qquad (8.67)
\end{aligned}
$$

by covariance inequality, since K/f is nonincreasing (K is nonincreasing according to Lemma) and $r \mapsto \int_{S_{r,\theta}} \Delta\gamma(x)\,dU_{r,\theta}(x)$ is non decreasing by subharmonicity of $\Delta\gamma$ (see e.g. Doob [43]). Therefore we have obtained

$$
E_\theta\left[\frac{K(\|X-\theta\|^2)}{f(\|X-\theta\|^2)}\,\Delta\gamma(X)\right] \le E_0\left[\frac{K(\|X\|^2)}{f(\|X\|^2)}\right] E_\theta[\Delta\gamma(X)] = \kappa E_\theta[\Delta\gamma(X)]
$$

which, through (8.62), implies that the risk difference in (8.59) satisfies

$$
\mathscr{D}_c(\theta,\delta) \le E_\theta[\kappa\,\Delta\gamma(X) + \gamma^2(X)]
$$

and, finally, proves the theorem. □

8.6 Differential operators and dimension cut-off when estimating a loss

In the previous sections, we have seen that, in various distribution settings, unbiased estimators of loss can be improved when the dimension p is greater than or equal to 5. In the normal case, Johnstone [80] formally proved that, when $p \leq 4$, the unbiased loss estimator $\mathit{\Delta}_0(X) \equiv p$ based on the MLE is admissible so that no (global) improvement over it cannot be expected. That situation parallels the dimension phenomenon which occurs when estimating the mean θ: the MLE X is admissible when $p \leq 2$, but inadmissible when $p \geq 3$.

Johnstone's proof uses a Blyth's method. Although it is specific to the normal case, it can be extended to other distributional setting (such as exponential families) so that this dimension cut-off should reflect a more fundamental mathematical phenomenon. Below, we give an insight into the reason of such phenomenon in terms of non linear partial differential operators.

Indeed, when estimating a quadratic loss, improvements on standard loss estimators through unbiased estimation techniques often involve nonlinear partial differential operators whose the necessary nonpositivity will imply higher dimensions. Usually these operators are of the form

$$\mathscr{R}\gamma(x) = k\,\mathit{\Delta}\,\gamma(x) + \gamma^2(x) \tag{8.68}$$

for a certain constant k, the sufficient improvement condition being typically

$$\mathscr{R}\gamma(x) \leq 0 \tag{8.69}$$

for all $x \in \mathbb{R}^p$ (with strict inequality on a set of positive Lebesgue measure). We will see that Inequality (8.69) has no nontrivial solution γ (i.e. γ is not equal to 0 almost everywhere) when the space dimension p is less than or equal to 4, even if we look for solutions with smoothness conditions as weak as possible. Consequently, a necessary dimension condition for (8.69) to have solutions $\gamma \not\equiv 0$ is $p \geq 5$.

Here follows the precise statement of this fact.

Theorem 8.9. *Let $k \in \mathbb{R}$ fixed. When $p \leq 4$, the only solution γ in $L^2_{loc}(\mathbb{R}^p)$ of*

$$\mathscr{R}\gamma(x) = k\,\Delta\,\gamma(x) + \gamma^2(x) \leq 0\,, \tag{8.70}$$

for any $x \in \mathbb{R}$, is $\gamma = 0$ (a.e.).

Note that, in Theorem 8.9, the search of solutions of Inequation (8.70) is addressed in a very general setting. Indeed the γ's are first sought in the space of distributions $\mathscr{D}'(\mathbb{R}^p)$ (see Schwartz [127] for a full account) so that the nature of the result is not due to whichever regularity conditions on the solutions we may choose. Nevertheless it is worth noticing that defining the non linear term γ^2 in (8.70) as a distribution prompts to seek γ in $L^2_{loc}(\mathbb{R}^p)$ (in which case the linear part of $\mathscr{R}\gamma$ is well defined). This is still a wide space for possible solutions γ.

The proof of Theorem 8.9 is based on the use of the following sequence of the so-called test functions. Let φ be a positive infinitely differentiable function on $\mathbb{R}_+$ bounded by 1, identically equal to 1 on $[0,1]$ and with support the interval $[0,2]$ ($\text{supp}(\varphi) = [0,2]$). Associate to φ the sequence $(\varphi_n)_{n\geq 1}$ of infinitely differentiable functions from $\mathbb{R}^p$ into $[0,1]$ defined through

$$\forall n \geq 1 \quad \forall x \in \mathbb{R}^p \quad \varphi_n(x) = \varphi\left(\frac{||x||}{n}\right). \tag{8.71}$$

Clearly, for any $n \geq 1$, the function φ_n has compact support B_{2n}, the closed ball of radius $2n$ and centered at 0 in $\mathbb{R}^p$. Also an interesting property **(see Appendix for a proof)** is that, for any $\beta \geq 2$ and for any $j = 1, \ldots, p$,

$$\left| \frac{\partial^2 \varphi_n^\beta}{\partial x_j^2}(x) \right| \leq \frac{K}{n^2} \varphi_n^{\beta-2}(x). \tag{8.72}$$

Note that, as all the derivatives of φ vanish out of the compact [1, 2] and φ is bounded by 1, Inequality (8.72) can be refined in

$$\left| \frac{\partial^2 \varphi_n^\beta}{\partial x_j^2}(x) \right| \leq \frac{K}{n^2} \, 1\!\!1_{C_n}(x). \tag{8.73}$$

where $1\!\!1_{C_n}$ is the indicator function of the annulus $C_n = \{x \in \mathbb{R}^p | n \leq ||x|| \leq 2n\}$.

(Proof of Theorem 8.9) Let $\gamma \in L^2_{loc}(\mathbb{R}^p)$ satisfying (8.70) Then, through the duality brackets between the space of distributions $\mathscr{D}'(\mathbb{R}^p)$ and the space $\mathrm{C}_0^\infty(\mathbb{R}^p)$ of infinitely differentiable functions on $\mathbb{R}^p$ with compact support **(see Appendix)**, we have, for any $n \in \mathbb{N}^*$ and any $\beta > 0$,

$$\begin{aligned} \int_{\mathbb{R}^p} \gamma^2(x)\, \varphi_n^\beta(x)\, dx &\leq -k \left\langle \Delta\gamma, \varphi_n^\beta \right\rangle \\ &= -k \left\langle \gamma, \Delta\varphi_n^\beta \right\rangle \end{aligned} \tag{8.74}$$

according to **(A1)**. Now, since the distribution γ lies in $L^2_{loc}(\mathbb{R}^p)$, we can express (8.74) as

$$\begin{aligned} \int_{\mathbb{R}^p} \gamma^2(x)\, \varphi_n^\beta(x)\, dx &\leq -k \int_{\mathbb{R}^p} \gamma(x)\, \Delta\varphi_n^\beta\, dx \\ &\leq k \int_{\mathbb{R}^p} |\gamma(x)|\, |\Delta\varphi_n^\beta|\, dx. \end{aligned} \tag{8.75}$$

Then, using (8.72), it follows from (8.74) that there exists a constant $C > 0$ such that

$$\begin{aligned}
&\int_{\mathbb{R}^p} \gamma^2(x)\, \varphi_n^\beta(x)\, dx \\
&\leq \frac{C}{n^2} \int_{\mathbb{R}^p} |\gamma(x)|\, \varphi_n^{\beta-2}(x)\, dx \\
&\leq \frac{C}{n^2} \left(\int_{\mathbb{R}^p} \varphi_n^{\beta-4}(x)\, dx \right)^{1/2} \left(\int_{\mathbb{R}^p} \gamma^2(x)\, \varphi_n^\beta(x)\, dx \right)^{1/2} \qquad (8.76)
\end{aligned}$$

applying Schwarz's inequality with $\beta > 4$ and

$$\gamma(x)\, \varphi_n^{\beta-2}(x) = \varphi_n^{\beta/2-2}(x)\, \gamma(x)\, \varphi_n^{\beta/2}(x)\,.$$

Clearly (8.76) is equivalent to

$$\int_{\mathbb{R}^p} \gamma^2(x)\, \varphi_n^\beta(x)\, dx \leq \frac{C^2}{n^4} \int_{\mathbb{R}^p} \varphi_n^{\beta-4}(x)\, dx\,. \qquad (8.77)$$

Thus, since $\mathbb{1}_{B_n} \leq \varphi_n \leq 1$ with $\text{supp}\varphi_n \subset B_n$, restricting the first integral of (8.77) over B_n leads to

$$\int_{B_n} \gamma^2(x)\, dx \leq \frac{C^2}{n^4} \int_{B_n} dx = A\, n^{p-4} \qquad (8.78)$$

for some constant $A > 0$. Letting n go to infinity in (8.78) shows that, when $p < 4$, $\gamma = 0$ almost everywhere, which proves the theorem in that case.

Consider now the case $p = 4$. Note that the above reasoning for $p = 4$ guarantees that $\gamma \in L^2(\mathbb{R}^p)$. The result will follow in applying Inequality (8.73). Indeed it follows from (8.73) and the first inequality in (8.76) that, for some constant $C > 0$,

$$\begin{aligned}
\int_{B_n} \gamma^2(x)\, dx &\leq \frac{C}{n^2} \int_{C_n} |\gamma(x)|\, dx \\
&\leq \frac{C}{n^2} \left(\int_{C_n} dx \right)^{1/2} \left(\int_{C_n} \gamma^2(x)\, dx \right)^{1/2} \qquad (8.79)
\end{aligned}$$

by Schwarz's inequality. Now

$$\int_{C_n} dx \leq \int_{B_{2n}} dx \propto n^4 \tag{8.80}$$

since $p = 4$. Hence (8.79) and (8.80) imply that, for some constant $A > 0$,

$$\int_{B_n} \gamma^2(x)\, dx \leq A \left(\int_{C_n} \gamma^2(x)\, dx \right)^{1/2} . \tag{8.81}$$

As $\gamma \in L^2(\mathbb{R}^p)$, we have

$$\lim_{n\to\infty} \int_{C_n} \gamma^2(x)\, dx = 0$$

and hence (8.81) gives rise to

$$0 = \lim_{n\to\infty} \int_{C_n} \gamma^2(x)\, dx = \int_{\mathbb{R}^p} \gamma^2(x)\, dx ,$$

which implies that $\gamma = 0$ almost everywhere and gives the desired result for $p = 4$. □

The above proof of Theorem 8.9 follows the approach of [19] (to which we refer to for a full account on that dimension cut-off phenomenon) where more general non linear partial differential inequalities are considered. Thus it can be shown that the usual Stein's inequality $2\,\text{div} g(x) + ||g(x)||^2 \leq 0$, for any $x \in \mathbb{R}^p$, has no nontrivial solution g in $\left(L^2_{loc}(\mathbb{R}^p)\right)^p$ when $p \leq 2$, which reinforces the fact that the MLE X is admissible in dimension $p \leq 2$ when estimating a normal mean.

8.7 Discussion

There are several areas of the theory of loss estimation that we have not discussed. Our primary focus has been on location parameters for the multivariate normal and spherical distributions. Loss estimation for exponential families is addressed in [97] [98] and [123]. In [97] and [98] Lele develops improved loss estimators for point estimators in the general setup of Hudson's [77] subclass of continuous exponential family. Hudson's family essentially includes distributions for which the Stein-like identities hold; explicit calculations and loss estimators are given for the gamma distribution, as well as for improved scaled quadratic loss estimators in the Poisson setting for the Clevenson-Zidak [40] estimator. Rukhin [123] studies the posterior loss estimator for a Bayes estimate (under quadratic loss) for the canonical parameter of a linear exponential family.

As point out in the introduction, in the known variance normal setting Johnstone [80] used a version of Blyth's lemma to show that the constant loss estimate p is admissible if $p \leq 4$. Lele [98] give some additional sufficient conditions for admissibility in the general exponential family and works out the precise details for the Poisson model. Rukhin [123] considers loss functions for the simultaneous estimate of θ and $L(\theta, \varphi(X))$ and deduced some interesting admissibility results.

A number of researchers have investigated improved estimators of a covariance matrix, Σ, under the Stein loss, $L_S(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma}\Sigma^{-1}) - \log|\hat{\Sigma}\Sigma^{-1}| - \ p$, using an unbiased estimation of risk technique. In the normal case, [**?**], [71], [133], [**?**], and [**?**] propose improved estimators that dominate the sample covariance under

$L_S(\hat{\Sigma},\Sigma)$. In [91], it is shown that the domination of these improved estimators over the sample covariance estimator are robust with respect to the family of elliptical distributions. To date, there has not been any work on improving the unbiased estimate of $L_S(\hat{\Sigma},\Sigma)$.

In addition to the theoretical ideas discussed in the previous sections there are very practical applications of loss estimation. The primary application of loss estimation ideas is to model selection. It is shown in [60] that improved loss estimators gives more accurate model selection procedures. [8] study model selection strategies based on penalized empirical loss minimization and point out the equivalence between loss estimation and data-based complexity penalization. It is shown that any good loss estimate may be converted into a data-based penalty function and the performance of the estimate is governed by the quality of the loss estimate. Furthermore, a selected model that minimizes the penalized empirical loss achieves an almost optimal trade-off between the approximation error and the expected complexity, provided that the loss estimate on which the complexity is based is an approximate upper bound on the true loss. The key point to stress is that there is a fundamental dependence on the notions of good complexity regularization and good loss estimation. The ideas in this review lay the theoretical foundation for the construction such loss estimators and model selection rules as well as give a decision theoretic analysis of their statistical properties.

In linear models the notion of degrees of freedom plays the important role as a model complexity measure in various model selection criteria, such as Akaike information criterion (AIC) [2] , Mallow's C_p [104], and Bayesian information criterion

(BIC) [126], and generalized cross-validation (GCV) [41]. In regression the degrees of freedom are the trace of the so-called "hat" matrix. Efron ([46]) pointed out that the theory of Stein's unbiased risk estimation is central to the ideas underlying the calculation of the degrees of freedom of certain regression estimators.

Specifically, let Y be a random vector having a n-variate normal distribution $\mathcal{N}(\theta,\sigma^2 I_n)$ with unknown p-dimensional mean θ and identity covariance matrix $\sigma^2 I_n$. Let $\hat{\theta} = \varphi(Y)$ be an estimate of θ. In regression one focuses is how accurate φ can be in predicting using a new response vector y^{new}. Under the quadratic loss, the prediction risk is $E\{||Y^{new} - \theta||^2\}/n$. Efron [46] notes that

$$E\{||\varphi - \theta||^2\} = E\{||Y - \varphi(Y)||^2 - n\sigma^2\} + 2\sum_{i=1}^{n} \operatorname{Cov}(\varphi_i, Y_i). \tag{8.82}$$

This expression suggests a natural definition of the degrees of freedom for an estimator φ as $df(\varphi) = \sum_{i=1}^{n} \operatorname{Cov}(\varphi_i, Y_i)/\sigma^2 = E_\theta[(Y-\theta)^t \varphi(Y)]/\sigma^2$. Thus one can define a C_p-type quantity

$$C_p(\varphi) = \frac{||Y - \varphi||^2}{n} + \frac{2df(\varphi)}{n}\sigma^2 \tag{8.83}$$

which has the same expectations the true prediction error but may not an estimate if $df(\varphi)$ and σ^2 are unknown. However if φ is weak differentiable and $\hat{\sigma}^2$ is an unbiased estimate of σ^2, the integration by parts formula in Lemma 3.1 implies that $df(\varphi)\ \sigma^2 = E_\theta[\operatorname{div}\varphi(Y)\ \hat{\sigma}^2]$, hence $\operatorname{div}\varphi\ \hat{\sigma}^2$ is unbiased estimate for the complexity parameter term, $df(\varphi)\ \sigma^2$, in (8.83). Therefore an unbiased estimate for the prediction error is

$$C_p^*(\varphi) = \frac{||Y - \varphi||^2}{n} + \frac{2\operatorname{div}\varphi}{n}\hat{\sigma}^2. \tag{8.84}$$

Note that if φ is a linear estimator ($\varphi = \mathbf{S}y$ for some matrix $\mathbf{S}$ independent of Y) then it is easy to show that this definition coincides with the definition of generalized degrees of freedom given by Hastie and Tibshirani [73] since $\text{div}\varphi = \text{tr}(\mathbf{S})$. Note that if φ also depends on $\hat{\sigma}^2$ then (8.83) needs to be augmented by additional derivative terms with respect to $\hat{\sigma}^2$ as in the proof of Theorem 3.5.

Other approaches for estimating the complexity term penalty involve the use of resampling methods ([46], [148]) to directly estimate the prediction error. A K-fold cross-validation randomly divides the original sample into K part, and rotates through each part as a test sample and uses the remainder as a training sample. Cross-validation provides an approximately unbiased estimate of the prediction error, although the its variance can be large. Other commonly used resampling techniques are the nonparametric and parametric bootstrap methods.

A number of new regularized regression methods have been recently been developed, starting with Ridge regression [74], followed by the Lasso [142], the Elastic Net [149], and LARS [**?**]. Each of these estimates are weakly differentiable and have the form of a general shrinkage estimate, thus the prediction error estimate in ((8.84) may be applied to construct a model selection procedure). Zou, Hastie and Tibshirani [150] use this idea to develop a model selection method for the Lasso. In some situations verifying the weak differentiability of φ may be complicated.

Loss estimates have been used to derive nonparametric penalized empirical loss estimates in the context of function estimation, which adapt to the unknown smoothness of the function of interest. See [7] and Donoho and [42] for more details.

Chapter A

Appendix

A.1 Weakly differentiable functions

For $\Omega \subset \mathbb{R}^n$ an open set and for $p \in \mathbb{R}$ such that $1 \leq p \leq \infty$, recall that the space of functions f from Ω into $\mathbb{R}$ such that f^p is locally integrable is denoted and defined by

$$L^p_{loc}(\Omega) = \left\{ f : \Omega \to \mathbb{R} / \int_K |f(x)|^p \, dx < \infty \quad \forall K \subset \Omega \text{ with } K \text{ compact} \right\} .$$

Then a function $f \in L^p_{loc}(\Omega)$ is said weakly differentiable if there exist n functions $g_1, \dots g_n$ in $L^p_{loc}(\Omega)$ such that, for any $i = 1, \dots, n$,

$$\int_\Omega f(x) \, \frac{\partial \varphi}{\partial x_i}(x) \, dx = - \int_\Omega g_i(x) \, \varphi(x) \, dx \tag{A.1}$$

for any $\varphi \in \mathscr{C}^\infty_c(\Omega)$.

The space of the functions f in $L^p_{loc}(\Omega)$ satisfying (A.1) is the Sobolev space $W^{1,p}_{loc}(\Omega)$. The functions g_i are the i-th partial weak derivatives of f and are denoted, as the usual derivatives, by $g_i = \partial f / \partial x_i$. They are unique in the sense that any function $\tilde{g}_i$ which satisfies (A.1) is equal almost everywhere to g_i [1]. Naturally, the vector $\nabla f = (\partial f/\partial x_1, \ldots, \partial f/\partial x_n)$ is referred to as the weak gradient of f.

Note that, in (A.1), it is just required that the function f^p is locally integrable but not necessarily in $L^p(\Omega)$ (which will not be the case for many functions of interest, see examples below). The subspace of $W^{1,p}_{loc}(\Omega)$ of the functions f in $L^p(\Omega)$ satisfying (A.1) is the Sobolev space $W^{1,p}(\Omega)$. In terms of the distributions theory of Schwartz (see e.g. Schwartz (1973)), the derivatives $\partial f/\partial x_i$ are seen as distributions and one can view $W^{1,p}_{loc}(\Omega)$ (respectively $W^{1,p}(\Omega)$) as the space of functions $f \in L^p_{loc}(\Omega)$ (respectively $f \in L^p(\Omega)$) such that the distributions $\partial f/\partial x_i \in L^p_{loc}(\Omega)$ (respectively $\partial f/\partial x_i \in L^p(\Omega)$).

It is clear that, if f is continuously differentiable, then f is weakly differentiable, the usual derivative and the weak derivative of f coinciding. Thus (A.1) appears to be the usual integration by part formula, where the usual term in brackets vanishes since the function φ has compact support. In that sense, this notion extends the usual notion of differentiability, which remains a basis to determine the expressions of the weak derivatives as illustrated by the following examples.

Example A.1. Let f be the function defined, for any $(x_1, \ldots, x_n) \in \mathbb{R}^n$, by

$$f(x_1, \ldots, x_n) = \frac{1}{2} \sum_{j=1}^{n} (|x_j| + x_j)$$

Even if f is locally integrable, obviously it has no partial derivatives at 0 in the usual sense since the absolute value function has no derivative f at 0. However, for any

[1] This can be derived from the fact that, if $h \in L^1_{loc}(\Omega)$ is such that $\int_\Omega h(x)\,\varphi(x)\,dx = 0$ for any $\varphi \in \mathscr{C}^\infty_c(\Omega)$, then $h = 0$ a.e on Ω.

$i = 1, \ldots, n$, a i-th partial weak derivative $\partial f / \partial x_i$ can be defined through

$$\frac{\partial f}{\partial x_i}(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if} \quad x_i > 0 \\ \\ 0 & \text{if} \quad x_i < 0 \end{cases} . \tag{A.2}$$

Indeed, as clearly a constant function has all its weak derivatives equal to 0 and as the weak differentiability is linear, it suffices to consider the case where $n = 1$. Then, for $\varphi \in \mathscr{C}_c^\infty(\mathbb{R})$, we have

$$\int_{\mathbb{R}} f(x)\,\varphi'(x)\,dx = \int_0^\infty x\,\varphi'(x)\,dx = -\int_0^\infty \varphi(x)\,dx = -\int_{\mathbb{R}} f'(x)\varphi(x)\,dx$$

with

$$f'(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ \\ 0 & \text{if} \quad x < 0 \end{cases} , \tag{A.3}$$

where the second equality holds by the usual integration by part formula (which has a null bracketed term since φ has compact support) and where the third equality follows from the definition of $f'(x)$ in (A.3). As f' is clearly locally integrable, we have the stated result.

Note that the function f' in (A.3) is not itself weakly differentiable (and hence the function f in (A.2)). Indeed for $\varphi \in \mathscr{C}_c^\infty(\mathbb{R})$, we have

$$\int_{\mathbb{R}} f(x)\,\varphi'(x)\,dx = \int_0^\infty \varphi'(x)\,dx = -\varphi(0)$$

since φ has compact support in $\mathbb{R}$. On the other hand, since $f(x)$ is constant and equal to 1 for $x \geq 0$ and null for $x < 0$, the natural candidate for a weak derivative is the constant function identically equal to 0, which implies

$$-\int_{\mathbb{R}} f'(x)\,\varphi(x)\,dx = 0\,.$$

Hence these two last integrals cannot be equal for any choice of φ such that $\varphi(0) \neq 0$.

Along the same lines, another example of a non weakly differentiable function is the sign function f defined, for $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ such that $x_i \neq 0$ for all $i = 1, \ldots, p$, by $f(x) = \text{sgn}(x) = (\text{sgn}(x_1), \ldots, \text{sgn}(x_n)) = (x_1/|x_1|, \ldots, x_n/|x_n|)$. Note that it is not necessary to define f everywhere.

However functions which have singularities, and hence are not differentiable in the usual sense, may be weakly differentiable. The shrinkage function of the James-Stein estimator is of particular interest. In the following example, we specify the conditions under which it is a weakly differentiable function.

Example A.2. Let f be the function on $\mathbb{R}^n$ defined for any $x \neq 0$ by $f(x) = 1/||x||^2$, its value at 0 being any real constant. Note that $f(x)$ explodes at 0 so that its lack of smoothness implies that f cannot be differentiable at 0. However we will see that f has partial weak derivatives in $L^1_{loc}(\mathbb{R}^n)$ as soon as $n \geq 4$). More precisely, for $i = 1, \ldots, n$ fixed, the i-th weak derivative corresponds to the usual one at $x = (x_1, \ldots, x_n) \neq 0$, that is,

$$\frac{\partial f}{\partial x_i}(x) = -\frac{2x_i}{||x||^4}\,. \tag{A.4}$$

First the local integrability of f can be seen through its integral on any fixed ball B_R of radius R and centered at 0, that is,

$$\int_{B_R} f(x)\,dx \propto \int_0^R \frac{1}{r^2}\,r^{n-1}\,dr = \int_0^R r^{n-3}\,dr$$

which is finite when $n-3 > -1$, that is, for $n \geq 3$. Now, for the function in (A.4), we have

$$\int_{B_R} \frac{|x_i|}{||x||^4}\,dx \leq \int_{B_R} \frac{1}{||x||^3}\,dx \propto \int_0^R r^{n-4}\,dr$$

which is finite when $n-4 > -1$, that is, for $n \geq 4$.

To see that (A.4) is the expected weak derivative note that, for $\varphi \in \mathscr{C}_c^\infty(\mathbb{R}^n)$, we can write

$$\int_{\mathbb{R}^n} \frac{1}{||x||^2}\,\frac{\partial\varphi}{\partial x_i}(x)\,dx = \int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} \frac{1}{x_i^2 + \sum_{j\neq i} x_j^2}\,\frac{\partial\varphi}{\partial x_i}(x_1,\ldots,x_n)\,dx_i\,dx_{-i} \qquad \text{(A.5)}$$

where $dx_{-i} = dx_1 \ldots dx_{i-1}\,dx_{i+1} \ldots dx_n$. In (A.5), we can assume $\sum_{j\neq i} x_j^2 > 0$ since the only case where $\sum_{j\neq i} x_j^2 = 0$ is when all the x_j's are null (in fact, the most exterior integral is on $\mathbb{R}^n - \{0\}$). Hence, in the most inner integral, the function $x_i \to 1/\left(x_i^2 + \sum_{j\neq i} x_j^2\right)$ has the necessary regularity to apply the usual integration by part formula, with a vanishing bracketed term since φ has compact support. Thus we have

$$\int_{\mathbb{R}} \frac{1}{x_i^2 + \sum_{j\neq i} x_j^2}\,\frac{\partial\varphi}{\partial x_i}(x_1,\ldots,x_n)\,dx_i = -\int_{\mathbb{R}} \frac{2x_i}{\left(x_i^2 + \sum_{j\neq i} x_j^2\right)^2}\,\varphi(x_1,\ldots,x_n)\,dx_i.$$

Replacing in (A.5) gives the expected result with the derivative given in (A.4). Thus $f \in W^{1,1}_{loc}(\mathbb{R}^n)$ for $n \geq 4$.

Following the same way it can be shown that, for $q > 0$, the function f given by $f(x) = 1/||x||^q$ for $x \neq 0$ is in $W^{1,p}_{loc}(\mathbb{R}^n)$ as soon as $n > p\,(q+1)$, its i-th derivative being equal to $-q x_i/||x||^{q+2}$.

We already posed the question of the possible weak differentiability of a weak derivative (and we replied in the negative for f' in (A.3)). A positive answer involves twice weak differentiability. For a function f from an open set Ω of $\mathbb{R}^n$ into $\mathbb{R}$, this is of course defined through the weak differentiability of all the weak partial derivatives $\partial f/\partial x_i$; these second weak partial derivatives are denoted by $\partial^2 f/\partial x_j \partial x_i$. Then $\Delta\gamma = \sum_{i=1}^{p} \partial^2 f/\partial x_i^2$ is referred to the weak Laplacian of f. The space of the functions f in $L^p_{loc}(\Omega)$ having second weak partial derivatives in $L^p_{loc}(\Omega)$ is the Sobolev space $W^{2,p}_{loc}(\Omega)$.

As an example, it can be seen, following Example A.2, that the preceding function $x \mapsto 1/||x||^q$ is in $W^{2,p}_{loc}(\mathbb{R}^n)$ for $n > p(q+2)$, which is the integrability condition of the second derivatives. Thus, taking $p = 1$ and $q = 2$, the function $x \mapsto 1/||x||^2$ is in $W^{2,1}_{loc}(\mathbb{R}^n)$ for $n \geq 5$; under that condition, its weak Laplacian equals $\Delta(1/||x||^2) = -2(n-4)/||x||^4$. Also, taking $q = n-2$ gives rise to the fundamental harmonic function $x \mapsto 1/||x||^{n-2}$. Note that, although it is an infinitely differentiable function in $\mathbb{R}^n\backslash\{0\}$ (and, in fact, it is an analytic function), it is not a twice weakly differentiable function on the entire space $\mathbb{R}^n$ (it does not belong to $W^{2,1}_{loc}(\mathbb{R}^n)$) since the above integrability condition is violated (with $p = 1$, $p(q+2) = n$).

As the functions in $W^{1,p}_{loc}(\Omega)$ possess a weakened property of differentiability, it is natural to consider the possible link with the notion of absolute continuity for functions from $\mathbb{R}$ into $\mathbb{R}$. Recall that a function f of one real variable (having possibly complex values) is said to be absolutely continuous if, for any $\varepsilon > 0$, there exists $\eta > 0$ such that

$$\sum_{i=1}^{N}(b_i - a_i) < \eta \quad \Rightarrow \quad \sum_{i=1}^{N}|f(b_i) - f(a_i)| < \varepsilon\,,$$

whenever $(a_1,b_1),\dots,(a_N,b_N)$ are disjoint segments. Rudin [122] observes that the equation

$$f(x) - f(a) = \int_a^x f'(t)\,dt \tag{A.6}$$

holds for all x in some interval $[a,b]$ if and only if f is absolutely continuous on $[a,b]$ (f is differentiable almost everywhere and $f'(t)$ is its derivative at any t where it exists). The following proposition shows that, for a fixed open interval I of $\mathbb{R}$, any function f in $W^{1,p}_{loc}(I)$ satisfies (A.6) almost everywhere. More precisely, among the functions equivalent to f, a continuous representent on $\bar{I}$ can be choosen. Furthermore a converse is given.

Proposition A.1. *Let I be an open interval in $\mathbb{R}$. For any $f \in W^{1,p}_{loc}(I)$ with weak derivative f', there exists a function $\tilde{f}$, continuous on $\bar{I}$ and almost everywhere equal to f on I, such that, for any $x \in \bar{I}$ and for any $a \in \bar{I}$,*

$$\tilde{f}(x) - \tilde{f}(a) = \int_a^x f'(t)\,dt\,, \tag{A.7}$$

which implies that $\tilde{f}$ is actually absolutely continuous on I. Conversely, if a function f is in $L^p_{loc}(I)$ and is absolutely continuous on I, and if its derivative (which consequently exists almost everywhere and is measurable) is in $L^p_{loc}(I)$, then f belongs to $W^{1,p}_{loc}(I)$ and its derivative and weak derivative coincide.

Before giving a proof, note that a consequence of Proposition A.1 (recalled by Johnstone [80]) is that a function locally integrable in an open set of $\mathbb{R}^n$ is weakly differentiable if and only if it is (equivalent to) a function which is absolutely continuous on almost all line segments parallel to the co-ordinate axes, and has partial derivatives (existing almost everywhere) which are locally integrable. The following proposition gives a precise statement.

Proposition A.2. *For an open set $\Omega \subset \mathbb{R}^n$ and for $p \in \mathbb{R}$ such that $1 \leq p \leq \infty$, let $f : \Omega \to \mathbb{R}$ be a function in $L^p_{loc}(\Omega)$. Then $f \in W^{1,p}_{loc}(\Omega)$ if and only if f is absolutely continuous in each variable (on segments in Ω) for almost all values of the other variables and its first partial derivatives (which consequently exist almost everywhere and are measurable) are in $L^p_{loc}(\Omega)$. Furthermore its partial and weak derivatives coincide almost everywhere.*

Proof of Proposition A.1

Let $f \in W^{1,p}_{loc}(I)$ with weak derivative f'. For $a \in I$ fixed, set, for any $x \in I$,

$$\bar{f}(x) = \int_a^x f'(t)\,dt\,.$$

Note that, since $f' \in L^p_{loc}(I)$, the function $\bar{f}$ is well defined and is continuous (and hence is locally integrable).

Now, with α and β denoting the bounds of I, for any $\varphi \in \mathscr{C}^\infty_c(I)$, we have

$$\begin{aligned}\int_I \bar{f}(x)\,\varphi'(x)\,dx &= -\int_\alpha^a \int_x^a f'(t)\,dt\,\varphi'(x)\,dx + \int_a^\beta \int_a^x f'(t)\,dt\,\varphi'(x)\,dx \\ &= -\int_\alpha^a \int_\alpha^t \varphi'(x)\,dx\,f'(t)\,dt + \int_a^\beta \int_t^\beta \varphi'(x)\,dx\,f'(t)\,dt \quad \text{(A.8)}\end{aligned}$$

by applying Fubini's theorem. As φ' is continuous, it follows that the two inner most integrals in (A.8) may be expressed respectively as $\varphi(t) - \varphi(\alpha)$ and $\varphi(\beta) - \varphi(t)$ and hence

$$\int_I \bar{f}(x)\,\varphi'(x)\,dx = -\int_\alpha^\beta \varphi(t)\,f'(t)dt \tag{A.9}$$

since, φ has compact support, we have $\varphi(\alpha) = \varphi(\beta) = 0$.

Equality (A.9) implies that $\bar{f}$ has f' as a weak derivative. In addition it states that, for any $\varphi \in \mathscr{C}_c^\infty(I)$,

$$\int_I (f(x) - \bar{f}(x))\,\varphi'(x)\,dx = 0\,.$$

It follows that there exists a contant c such that we have $f - \bar{f} = c$ almost everywhere[2]. Then setting $\tilde{f}(x) = \bar{f}(x) + c$, for any $x \in I$, and $\tilde{f}(\alpha) = \lim_{x\to\alpha} f(x)$ and $\tilde{f}(\beta) = \lim_{x\to\beta} f(x)$ yields a function, almost everywhere equal to f on I, continuous on $\bar{I}$ and satisfying (A.7). The fact that $\tilde{f}$ is absolutely continuous on I follows from the comment on equation (A.6) given before the statement of Proposition A.1.

Conversely assume that $f \in L^p_{loc}(I)$ is absolutely continuous with $f' \in L^p_{loc}(I)$ as a derivative defined almost everywhere. For a fixed $\varphi \in \mathscr{C}_c^\infty(I)$, the product $\psi = f\varphi$ is absolutely continuous with derivative ψ' satisfying

$$\psi'(x) = f'(x)\,\varphi(x) + f(x)\,\varphi'(x) \quad \text{a.e.} \tag{A.10}$$

As ψ inherits a compact support from φ, we have

$$\int_I \psi'(x)\,dx = 0$$

and hence it follows from (A.10) that

$$\int_I f'(x)\,\varphi(x)\,dx = -\int_I f(x)\,\varphi'(x)\,dx\,. \tag{A.11}$$

[2] This can be shown as follows.

Equality (A.11) and the fact that f' was assumed to lie in $L^p_{loc}(I)$ give the desired result. □

Formula (A.1) illustrates integration by parts with vanishing bracketed term. To establish his identity, Stein (1981) highlights the necessity of such a regularity condition through the notion of "almost differentiability": he considers g as almost differentiable if there exists a function $\nabla g = (\partial_1 g, \ldots, \partial_p g)'$ such that, for all $z \in \mathbb{R}^p$,

$$g(x+z) - g(x) = \int_0^1 z' \nabla g(x+tz)\, dt. \tag{A.12}$$

Note that, from the above propositions, the notions of weak differentiability and of almost differentiability through (A.1) and (A.12) respectively are equivalent. This has been underlined by Johnstone [80]; he gave, refering to [113], a statement similar to Proposition A.2.

A.2 More on Stein's identity

We noticed above that the sign function is not weakly differentiable so that no unbiased estimator of $(X-\theta)^t \operatorname{sgn}(X)$ can be derived from Stein's identity since its basic validity condition is not filled. Actually, when $X \sim \mathscr{N}(\theta, I_p)$, such an unbiased estimator does not exist.

First, it is easy see that it suffices to consider $p = 1$ and it is straightforward to calculate the corresponding expectation $A_\theta = E_\theta[(X-\theta)\operatorname{sgn}(X)]$ since we have

$$\begin{aligned}
A_\theta &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \operatorname{sgn}(x) \, \frac{d}{dx} \left[-\exp\left(-\frac{(x-\theta)^2}{2} \right) \right] dx \\
&= \frac{1}{\sqrt{2\pi}} \left\{ - \left[-\exp\left(-\frac{(x-\theta)^2}{2} \right) \right]_{-\infty}^{0} + \left[-\exp\left(-\frac{(x-\theta)^2}{2} \right) \right]_{0}^{\infty} \right\} \\
&= \frac{1}{\sqrt{2\pi}} \left\{ \exp\left(-\frac{\theta^2}{2} \right) + \exp\left(-\frac{\theta^2}{2} \right) \right\} \\
&= \sqrt{\frac{2}{\pi}} \left\{ \exp\left(-\frac{\theta^2}{2} \right) \right\} .
\end{aligned}$$

Now assume that there exists a function ϕ such that

$$E_\theta[\phi] = \sqrt{\frac{2}{\pi}} \left\{ \exp\left(-\frac{\theta^2}{2} \right) \right\} . \tag{A.13}$$

Then, deriving with respect to θ, it follows that

$$\frac{d}{d\theta} E_\theta[\phi(X)] = -\sqrt{\frac{2}{\pi}} \, \theta \exp\left(-\frac{\theta^2}{2} \right) = -\theta \, E_\theta[\phi(X)]$$

and hence, since we deal with an exponential family, deriving under the integral sign gives obtain

$$\begin{aligned}
\int_{-\infty}^{\infty} \phi(x) \frac{\partial}{\partial x} \left\{ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2} \right) \right\} dx &= \int_{-\infty}^{\infty} \phi(x)(x-\theta) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2} \right) dx \\
&= -\theta \int_{-\infty}^{\infty} \phi(x) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2} \right) dx
\end{aligned}$$

that is, after simplifying,

$$E_\theta[X \, \phi(X)] = 0 \, . \tag{A.14}$$

Therefore, as (A.14) is satisfied for all θ, by completeness of the normal family, we have $x \, \phi(x) = 0$ almost everywhere, and consequently, $\phi(x) = 0$ almost everywhere. This contradicts (A.13) proving that ϕ cannot be an unbiased estimator of $(X - \theta) \operatorname{sgn}(X)$.

A.3 Differentiation of marginal densities

When considering a density prior $\pi(\theta)$ for the normal density

$$\frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{||x-\theta||^2}{2}\right)$$

differentiation of the marginal density

$$m(x)=\int_{\mathbb{R}^p}\frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{||x-\theta||^2}{2}\right)\pi(\theta)\,d\theta$$

is easily tractable since we are dealing with an exponential family. Thus, deriving under the integral sign, we have that the gradient of m is expressed as

$$\begin{aligned}\nabla m(x) &= \int_{\mathbb{R}^p}\frac{1}{(2\pi)^{p/2}}\nabla\exp\left(-\frac{||x-\theta||^2}{2}\right)\pi(\theta)\,d\theta\\ &= \int_{\mathbb{R}^p}\frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{||x-\theta||^2}{2}\right)(\theta-x)\,\pi(\theta)\,d\theta\,.\end{aligned}$$

The following lemma shows that, for a general spherically symmetric density $f(||x-\theta||^2)$ a similar formula is valid.

Lemma A.1. *Let $f(||x-\theta||^2)$ be a spherically symmetric density and $\pi(\theta)$ a prior density (possibly improper) such that, for any $x\in\mathbb{R}^p$, the marginal density*

$$m(x)=\int_{\mathbb{R}^p}f(||x-\theta||^2)\,\pi(\theta)\,d\theta$$

exists. If the generating function f is absolutely continuous then, for almost any $x\in\mathbb{R}^p$,

$$\begin{aligned}\nabla m(x) &= \int_{\mathbb{R}^p}\nabla f(||x-\theta||^2)\,\pi(\theta)\,d\theta\\ &= \int_{\mathbb{R}^p}2\,f'(||x-\theta||^2)\,(x-\theta)\,\pi(\theta)\,d\theta\,.\end{aligned}$$

Proof. Let $j \in \{1,\ldots,p\}$ and $z \in \mathbb{R}$ fixed. For any $x = (x_1,\ldots,x_{j-1},x_j,x_{j+1},\ldots,x_p)$ in $\mathbb{R}^p$, denote by $x_{(z)} = (x_1,\ldots,x_{j-1},z,x_{j+1},\ldots,x_p)$ the vector obtained by replacing the j-th component x_j of x by z.

As f is absolutely continuous, the function $z \to f(||x_{(z)} - \theta||^2)$ is also absolutely continuous so that, for $a \in \mathbb{R}$, we have

$$f(||x_{(z)} - \theta||^2) - f(||x_{(a)} - \theta||^2) = \int_a^z \frac{\partial}{\partial x_j} f(||x-\theta||^2)\, dx_j\,.$$

Therefore

$$\begin{aligned} m(x_{(z)}) &= \int_{\mathbb{R}^p} \int_a^z \frac{\partial}{\partial x_j} f(||x-\theta||^2)\, dx_j\, \pi(\theta)\, d\theta + m(x_{(a)}) \\ &= \int_a^z \int_{\mathbb{R}^p} \frac{\partial}{\partial x_j} f(||x-\theta||^2)\, \pi(\theta)\, d\theta\, dx_j + m(x_{(a)}) \end{aligned} \tag{A.15}$$

by Fubini's theorem. Equation (A.15) means that the function $z \to m(x_{(z)})$ is absolutely continuous, and hence differentiable almost everywhere. More precisely, it entails that the partial derivative of $m(x)$ with respect to x_j equals

$$\frac{\partial}{\partial x_j} m(x) = \int_{\mathbb{R}^p} \frac{\partial}{\partial x_j} f(||x-\theta||^2)\, \pi(\theta)\, d\theta$$

which is the desired result. □

A.4 Monotonicity of expectations

In the following, when X has a uniform distribution on a sphere of radius r, we consider expectations of $R^{2q}\, \|X\|^{-2q}$. We mention in our notations the dimension of the spaces in which spheres and balls lie. Thus $U^{p+k}_{R,\theta}$ stands for the uniform distribution on the sphere $S^{p+k}_{R,\theta} = \{(x,u) \in \mathbb{R}^{p+k} / \|(x,u) - (\theta,0)\| = R\}$, in $\mathbb{R}^{p+k}$, of radius R and centered at $(\theta,0) \in \mathbb{R}^{p+k}$, while $V^p_{R,\theta}$ holds for the uniform distribution on the ball $B^p_{R,\theta} = \{x \in \mathbb{R}^p / \|x-\theta\| \le R\}$, in $\mathbb{R}^{p+k}$, of radius R and centered at $\theta \in \mathbb{R}^p$. We essentially extend a result given by Fourdrinier and Strawderman [58] who considered the case where $q = 1$.

Lemma A.2. *Let $q > 0$. Then, for any fixed $\theta \in \mathbb{R}^p$, the function*

$$f_\theta : R \longmapsto R^{2q} \int_{S^{p+k}_{R,\theta}} \frac{1}{\|x\|^{2q}} \, dU^{p+k}_{R,\theta}(x,u) \qquad \text{(A.16)}$$

is nondecreasing for $p \ge 2\,(q+1)$ and $k \ge 0$. Also, for any fixed R, this monotonicity is reversed in $\|\theta\|$.

Proof. Note that, by invariance, f_θ depends on θ only through $\|\theta\|$. With the change of variable

$$(y,v) = \left(\frac{x-\theta}{R}, \frac{u}{R}\right),$$

we have

$$f_\theta(R) = \int_{S^{p+k}_{1,0}} \frac{1}{\|y + \frac{\theta}{R}\|^{2q}} \, d\mathscr{U}^{p+k}_{1,0}(y,v) = f^*_{\|\theta\|}(R)\,.$$

Hence, integrating with respect to the uniform distribution on $\{\theta \in \mathbb{R}^p / \|\theta\| = R_0\}$,

$$\begin{aligned} f_\theta(R) &= \int_{S^p_{R_0,0}} \int_{S^{p+k}_{1,0}} \frac{1}{\|z + \frac{\theta}{R}\|^{2q}} \, dU^{p+k}_{1,0}(y,v) \, dU^p_{R_0,0}(\theta) \\ &= \int_{S^{p+k}_{1,0}} \int_{S^p_{R_0,0}} \frac{1}{\|z + \frac{\theta}{R}\|^{2q}} \, dU^p_{R_0,0}(\theta) \, dU^{p+k}_{1,0}(y,v) \end{aligned}$$

by Fubini's theorem. In the inner integral, the change of variable $z = \theta/R + y$ leads to

$$f_\theta(R) = \int_{S^{p+k}_{1,0}} \int_{S^p_{R_0/R,y}} \frac{1}{\|z\|^{2q}} \, dU^p_{R_0/R,y}(z) \, dU^{p+k}_{1,0}(y,v)$$

As the function $1/\|z\|^{2q}$ is superharmonic for $p \geq 2\,(q+1)$, the inner integral is nonincreasing in R_0/R for each y, and hence, nondecreasing in R and, for any fixed R, nonincreasing in R_0. □

Lemma A.3. *Let $q > 0$ and let $r(t)$ be a nonnegative and nondecreasing function on $[0,\infty)$ such that $r(t)/t^q$ is nonincreasing. Then, for any fixed $\theta \in \mathbb{R}^p$, the function*

$$f_\theta : R \longmapsto R^{2q} \int_{S^{p+k}_{R,\theta}} \frac{r(\|x\|^2)}{\|x\|^{2q}}\, dU^{p+k}_{R,\theta}(x,u) \tag{A.17}$$

is nondecreasing for $p \geq 1$ and $k \geq 2$.

Proof. Under $U^{p+k}_{R,\theta}$, it is well known that the marginal distribution of $(x,u) \mapsto x$ is absolutely continuous with unimodal density $\frac{1}{R^p}\,\psi\left(\frac{\|x-\theta\|^2}{R^2}\right)$ for all $k \geq 2$ where $\psi(t) = (1-t)^{k/2-1}\mathbf{1}_{[0,1]}(t)$. Then f_θ can be written as

$$f_\theta(R) = \int_{B^p_{1,0}} \frac{r(R^2\,\|z+\frac{\theta}{R}\|^2)}{\|z+\frac{\theta}{R}\|^{2q}}\,\psi(\|z\|^2)\,dz\,.$$

For any $R_1 \leq R_2$, we have, by nondecreasing monotonicity of $r(t)$,

$$f_\theta(R_1) \leq \int_{B^p_{1,0}} \frac{r(R_2^2\|z+\frac{\theta}{R_1}\|^2)}{\|z+\frac{\theta}{R_1}\|^{2q}}\,\psi(\|z\|^2)\,dz\,.$$

Furthermore nonincreasing monotonicity of $r(t)/t$ in t implies that the function $r(R_2^2\,\|z+\theta/R_1\|^2)/\|z+\theta/R_1\|^{2q}$ is symmetric and unimodal in z about $-\frac{\theta}{R_1}$. Hence, by Anderson's theorem (see [4]),

$$\begin{aligned}\int_{B^p_{1,0}} \frac{r(R_2^2\|z+\frac{\theta}{R_1}\|^2)}{\|z+\frac{\theta}{R_1}\|^{2q}}\,\psi(\|z\|^2)\,dz &\leq \int_{B^p_{1,0}} \frac{r(R_2^2\|z+\frac{\theta}{R_2}\|^2)}{\|z+\frac{\theta}{R_2}\|^{2q}}\,\psi(\|z\|^2)\,dz \\ &= f_\theta(R_2)\,.\end{aligned}$$

□

As, if $(X,U) \sim U_{R,\theta}^{p+2}$, then $X \sim V_{R,\theta}^{p}$, the following corollary of Lemma A.3 is immediate.

Corollary A.1. *Under the conditions of Lemma A.3, the function*

$$R \longmapsto R^{2q} \int_{B_{R,\theta}^{p}} \frac{r(\|x\|^2)}{\|x\|^{2q}} \, dV_{R,\theta}^{p}(x) \tag{A.18}$$

is nondecreasing for $p \geq 1$.

References

[1] Aitchison J (1975) Goodness of prediction fit. Biometrika 62:547–554

[2] Akaike H (1973) Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory 25:267–281

[3] Alam K (1973) A family of admissible minimax estimators of the mean of a multivariate normal distribution. Annals of Statistics 1:517–525

[4] Anderson TW (1955) The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. Proc Amer Math Sot 6:170–176

[5] Baranchik A (1970) A family of minimax estimators of the mean of a multivariate normal distribution. Annals of Mathematical Statistics 41:642–645

[6] Barlow RE, Proschan F (1981) Statistical Theory of Reliability and Life Testing: Probability Models. TO BEGIN WITH

[7] Barron AR, Birgé L, Massart P (1999) Risk bound for model selection via penalization. Probability Theory and Related Fields 113:301–413

[8] Bartlett P, Boucheron S, Lugosi G (2002) Model selection and error estimation. Machine Learning 48:85–113

[9] Berger JO (1975) Minimax estimation of location vectors for a wide class of densities. The Annals of Statistics 3(6):1318–1328

[10] Berger JO (1976) title. ? p ?

[11] Berger JO (1985) Decision Theory and Bayesian Analysis. Springer, New York

[12] Berger JO (1985) The frequentist viewpoint and conditioning. In: Cam LL, Olshen R (eds) Proceedings of the Berkeley Conference in Honor of Jersy Neyman and Jack Kiefer, vol 1, Wadsworth, Monterey, California, pp 15–44

[13] Berger JO (1985) In the defense of likelihood principle: Axiomatics and coherency. In: J M Bernardo DVL M H DeGroot, Smith AFM (eds) Bayesian Statistics II, Norh Holland, Amsterdam, pp 33–65

[14] Berger JO (1985) Statistical Decision Theory and Bayesian Analysis. Springer-Verlag

[15] Berger JO, Srinivasan C (1978) Generalized bayes estimators in multivariate problems. The Annals of Statistics 6(4):783–801

[16] Berger JO, Strawderman WE (1996) Choice of hierarchical priors. admissibility in estimation of normal means. Ann Statist (24):931–951

[17] Berger JO, Strawderman WE, Tang D (2005) Posterior propriety and admissibility of hyperpriors in normal hierarchical models. Ann Statist (33):606–646

[18] Bickel PJ, Doksum K (2006) Mathematical Statistics. Printice Hall

[19] Blanchard D, Fourdrinier D (1999) Non trivial solutions of non-linear partial differential inequations and order cut-off. Rendiconti di Matematica 19:137–154

[20] Blyth CR (1951) On minimax statistical procedures and their admissibility. Annals of Mathematical Statistics 22:22–42

[21] Bock ME (1988) Shrinkage estimator: Pseudo-bayes estimators for normal mean vectors. In: Gupta SS, Berger J (eds) Statistical Decision Theory and Related Topics 4, vol 1, Springer-Verlag, New York, pp 281–298

[22] Bondar JV, Milnes P (1981) Amenability: A survey for statistical applications of Hunt-Stein theorem and related conditions on groups. Zeitschr Wahrsch Verw Geb 57:103–128

[23] Brandwein AC, Strawderman WE (1978) title. ? p ?

[24] Brandwein AC, Strawderman WE (1980) Minimax estimation of location parameters for spherically symmetric distributions with concave loss. Annals of Statistics 8:279–284

[25] Brandwein AC, Strawderman WE (1991) Generalizations of james-stein estimators under spherical symmetry distributions with concave loss. Annals of Statistics 19:1639–1650

[26] Brandwein AC, Strawderman WE (1991) Improved estimates of location in the presence of unknown scale. Journal of Multivariate Analysis 39:305–314

[27] Brown L (1968) Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters. Annals of Mathematical Statistics 39:24–48

[28] Brown LD (1966) On the admissibility of invariant estimators in one or more location parameter. Annals of Mathematical Statistics 37:1087–1136

[29] Brown LD (1971) Admissible estimators, recurrent diffusions, and insoluble boundary value problems. Annals of Mathematical Statistics 42:855–903

[30] Brown LD (1979) A heuristic method for determining admissibility of estimators–with applications. The Annals of Statistics 7(5):960–994, DOI 10.1214/aos/1176344782, URL http://dx.doi.org/10.1214/aos/1176344782

[31] Brown LD (1986) Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory, Lecture Notes-Monograph Series, vol 9. Institute of Mathematical Statistics, Hayward, California

[32] Brown LD, Hwang JT (1982) A unified admissibility proof. In: Gupta SS, Berger J (eds) Statistical Decision Theory and Related Topics III, vol 1, Academic Press, New York, pp 205–230

[33] Brown LD, Hwang JT (1989) Admissibility of confidence estimators. Tech. rep., Statistical Center Cornell University

[34] Brown LD, George I, Xu X (2008) Admissible predictive density estimation. Annals of Statistics 36:1156–1170

[35] Casella G, Berger RL (2001) Statistical Inference, 2nd edn. Duxbury Press, Belmont, California

[36] Cellier D, Fourdrinier D (1990) Sur les lois à symétrie elliptique. In: Séminaire de probabilités (Strasbourg), vol 24, Springer-Verlag, Berlin, Heidelberg, New York, pp 320–328

[37] Cellier D, Fourdrinier D (1995) Shrinkage estimators under spherical symmetry for the general linear model. Journal of Multivariate Analysis 52:338–351

[38] Cellier D, Fourdrinier D, Robert C (1989) Robust shrinkage estimators of the location parameter for elliptically symmetric distributions. Journal of Multivariate Analysis 29:39–52

[39] Chou JP, Strawderman WE (1990) Minimax estimation of means of multivariate normal mixtures. Journal of Multivariate Analysis 35(2):141–150

[40] Clevenson M, Zidek J (1975) Simultaneous estimation of the mean of independent poisson laws. Journal of the American Statistical Association 70:698–705

[41] Craven P, Wahba G (1979) Smoothing noisy data with spline functions. Numerische Mathematik 31:377–403

[42] Donoho DL, Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association 90:1200–1244

[43] Doob JL (1984) Classical potential theory and its probabilistics counterpart. Springer, Berlin, Heidelberg, New York

[44] Eaton ML (1989) Group Invariance Applications in Statistics. Institute of Mathematical Statistics, Hayward, CA

[45] Efron B (1976) title. ? p ?

[46] Efron B (2004) The estimation of prediction error: covariance penalties and cross-validation. Journal of the American Statistical Association 81:461–470

[47] Efron B, Morris C (1976) Families of minimax estimators of the mean of a multivariate normal distribution. Annals of Statistics 4(1):11–21

[48] Efron B, Morris C (1977) Stein's paradox in statistics. Scientific American 236(5):119–127

[49] Faith RE (1978) Minimax bayes point estimators of a multivariate normal mean. Journal of Multivariate Analysis 8:372–379

[50] Fang KT (1990) title. ?, ?

[51] Fang KT, Zhang YT (1990) Generalized Multivariate Analysis. Springer, Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong

[52] Fang KT, Kotz KS, Ng KW (1990) Symmetric Multivariate and Related Distributions. Chapman and Hall, New York

[53] Fourdrinier D, Lepelletier P (2008) Estimating a general function of a quadratic function. Annals of the Institute of Statistical Mathematics 60:85–119

[54] Fourdrinier D, Ouassou I (2000) Estimation of the mean of a spherically symmetric distribution with constraints on the norm. Canadian Journal of Statistics 28:399–415

[55] Fourdrinier D, Strawderman WE (1996) A paradox concerning shrinkage estimators: Should a known scale parameter be replaced by an estimated value in the shrinkage factor? Journal of Multivariate Analysis 59:109–140

[56] Fourdrinier D, Strawderman WE (2003) On Bayes and unbiased estimators of loss. Annals of the Institute of Statistical Mathematics 55:803–816

[57] Fourdrinier D, Strawderman WE (2008) Generalized bayes minimax estimators of location vector for spherically symmetric distributions. Journal of Multivariate Analysis 99(4):735–750

[58] Fourdrinier D, Strawderman WE (2008) A unified and generalized set of shrinkage bounds on minimax stein estimates. Journal of Multivariate Analysis 99(10):2221–2233

[59] Fourdrinier D, Strawderman WE (2010) Robust generalized Bayes minimax estimators of location vectors for spherically symmetric distribution with unknown scale. IMS Collections, Institute of Mathematical Statistics, A Festschrift for Lawrence D Brown 6:249–262

[60] Fourdrinier D, Wells MT (1994) Comparaisons de procédures de sélection d'un modèle de régression: Une approche décisionnelle. CR Acad Sci Paris Serie I 319:865–870

[61] Fourdrinier D, Wells MT (1995) Loss estimation for spherically symmetric distributions. Journal of Multivariate Analysis 53:311–331

[62] Fourdrinier D, Strawderman, Wells MT (1998) On the construction of bayes minimax estimators. Annals of Statistics 26:660–671

[63] Fourdrinier D, Strawderman WE, Wells MT (2003) Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix. Journal of Multivariate Analysis 85:24–39

[64] Fourdrinier D, Marchand E, Strawderman WE (2004) On the inevitability of a paradox in shrinkage estimation for scale mixtures of normals. Journal of Statistical Planning and Inference 121:37–51

[65] Fourdrinier D, Strawderman, Wells MT (2006) Estimation of a location parameter with restrictions or "vague information" for spherically symmetric distributions. The Annals of the Institute of Statistical Mathematics 58:73–92

[66] Fourdrinier D, Kortbi O, Strawderman W (2008) Bayes minimax estimators of the mean of a scale mixture of multivariate normal distributions. Journal

of Multivariate Analysis 99(1):74–93

[67] George EI (1986) A formal bayes multiple shrinkage estimator. Communications in Statistics-Theory and Methods 15(7):2099–2114

[68] George EI (1986) Minimax multiple shrinkage estimation. The Annals of Statistics 14(1):188–205

[69] George EI, Feng L, Xu X (2006) Improved minimax predictive densities under Kullback-Leibler loss. Annals of Statistics 34:78–91

[70] Green EJ, Strawderman WE (1991) A James-Stein type estimator for combining unbiased and possibly biased estimators. Journal of the American Statistical Association 86:1001–1006

[71] Haff LR (1979) An identity for the wishart distribution with applications. Journal of Multivariate Analysis 9:531–544

[72] Hartigan JA (2004) Uniform priors on convex sets improve risk. Statistics and Probability Letters 67:285–288

[73] Hastie T, Tibshirani R (1990) Generalized Additive Models. Chapman and Hall, London

[74] Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12:55–67

[75] Hoffmann K (1992) Improved estimation of distribution parameters: Stein-type estimators. Teubner-Verlag, Stuttgart, Leipzig

[76] Hsieg F, Hwang JTG (1993) Admissibility under the frequentist's validity constraint in estimating the loss of the least-squares estimator. Journal of Multivariate Analysis 44:279–285

[77] Hudson HM (1978) A natural identity for exponential families with applications in multiparameter estimation. Annals of Statistics 6:473–484

[78] James W, Stein C (1961) Estimation with quadratic loss. In: Proc. Fourth Berkeley Symp. Math. Statist. Prob., University of California Press, pp 361–379

[79] Johnson, Kotz (1972) Distributions in statistics: continuous multivariate distributions. Wiley, New York

[80] Johnstone I (1988) On inadmissibility of some unbiased estimates of loss. In: Gupta SS, Berger J (eds) Statistical Decision Theory and Related Topics IV, vol 1, Springer-Verlag, New York, pp 361–379

[81] Kagan AM, Linnik UV, Rao CR (1973) Characterization problems in mathematical statistics. Wiley, New York

[82] Katz MW (1961) Admissible and minimax estimates of parameters in truncated spaces. Annals of Mathematical Statistics 32(1):136–142

[83] Kavian O (1993) Introduction à la théorie des points critiques et applications aux problèmes elliptiques. Springer-Verlag, Paris, France

[84] Ki F, Tsui KW (1990) Multiple-shrinkage estimators of means in exponential families. The Canadian Journal of Statistics / La Revue Canadienne de Statistique 18:pp. 31–46

[85] Kiefer J (1957) Invariance, minimax sequential estimation, and continuous time processes. Annals of Mathematical Statistics 28:573–601

[86] Kiefer J (1975) Conditional confidence approach in multi-decision problems. In: Krishnaiah PR (ed) Multivariate Analysis 4, Academic Press, New York

[87] Kiefer J (1976) Admissibility of conditional confidence procedures. Annals of Statistics 4:836–865

[88] Kiefer J (1977) Conditional confidence statements and confidence estimators. Journal of the American Statistical Association 72:789–827

[89] Komaki F (2001) A shrinkage predictive distribution for multivariate normal observables. Biometrika 88:859–864

[90] Kubokawa T (1991) An approach to improving the james-stein estimator. Journal of Multivariate Analysis 36:121–126

[91] Kubokawa T, Srivastava MS (1999) Robust improvement in estimation of a covariance matrix in an elliptically contoured distribution. The Annals of Statistics 27:600–609

[92] Kubokawa T, Srivastava MS (2001) Robust improvement in estimation of a mean matrix in an elliptically contoured distribution. Journal of Multivariate Analysis 76:138–152

[93] Kubokawa T, Strawderman W (2005) title. ? p ?

[94] Kullback S, Leibler RA (1951) On information and sufficiency. Annals of Mathematical Statistics 22:79–86

[95] Lehmann EL, Casella G (1998) Theory of point estimation, 2nd edn. Springer-Verlag

[96] Lehmann EL, Sheffé H (1950) Completeness, similar regions and unbiased estimates. Sankhyā 17:305–340

[97] Lele C (1992) Inadmissibility of loss estimators. Statistics and Decision 10:309–322

[98] Lele C (1993) Admissibility results in loss estimation. Annals of Statistics 21:378–390

[99] Lepelletier P (2004) Sur les rgions de confiance : amlioration, estimation d'un degr de confiance conditionnel. PhD thesis, Universit de Rouen, France

[100] Liang F, Barron A (2004) Exact minimax strategies for predictive density estimation, data compression and model selection. IEEE Transactions on Information Theory 50:2708–2726

[101] Liese F, Miescke KJ (2008) Statistical Decision Theory. Springer

[102] Lindley DV (1962) Discussion on professor Stein's paper. Journal of the Royal Statistical Society 24:285–287

[103] Lu KL, Berger JO (1989) Estimation of normal means: frequentist estimation of loss. Annals of Statistics 17:890–906

[104] Mallows C (1973) Some comments on Cp. Technometrics 15:661–675

[105] Marchand E, Strawderman WE (2003) title. ? p ?

[106] Maruyama Y (1998) A unified and broadened class of admissible minimax estimators of a multivariate normal mean. Journal of Multivariate Analysis 64:196–205

[107] Maruyama Y (2003) Admissible minimax estimators of a mean vector of scale mixtures of multivariate normal distributions. Journal of Multivariate Analysis 84:274–283

[108] Maruyama Y (2003) A robust generalized bayes estimator improving on the James-Stein estimator for spherically symmetric distributions. Statistics and Decisions 21:69–77

[109] Maruyama Y, Strawderman WE (2005) A new class of generalized Bayes minimax ridge regression estimators. Annals of Statistics 33:1753–1770

[110] Maruyama Y, Strawderman WE (2009) An extended class of minimax generalized Bayes estimators of regression coefficients. Journal of Multivariate Analysis 100:2155–2166

[111] Maruyama Y, Takemura A (2006) ? ?

[112] Maruyama Y, Takemura A (2008) Admissibility and minimaxity of generalized bayes estimators for spherically symmetric family. Journal of Multivariate Analysis 99:1

[113] Morrey CB (1966) Multiple Integrals in the Calculus of Variations. Springer-Verlag, Berlin, Heidelberg, New York

[114] Muirhead RJ (1982) Aspects of multivariate statistics. John Wiley and Sons, New York

[115] Murray GD (1977) A note on the estimation of probability density functions. Biometrika 64:150–152

[116] Nachbin L (1965) The Haar Integral. D. Van Nostrand Company, Toronto, New York, London

[117] Ng VM (1980) On the estimation of parametric density functions. Biometrika 67:505–(506

[118] Philoche JL (1977) Une condition de validité pour le test F. Statistique et analyse des données pp 37–59

[119] du Plessis N (1970) An Introduction to Potential Theory. Hafner, Darien, CT

[120] Robert C, Casella G (1994) Improved confidence estimators for the usual multivariate normal confidence set. In: Gupta SS, Berger J (eds) Statistical Decision Theory and Related Topics 5, Springer-Verlag, New York

[121] Robert CP (1994) The Bayesian choice: A Decision Theoretic Motivation. Springer-Verlag, New York

[122] Rudin W (1966) Real and complex analysis. McGraw-Hill, New York, St. Louis, San Francisco, Toronto, London, Sydney

[123] Rukhin AL (1988) Estimated loss and admissible loss estimators. In: Gupta SS, Berger J (eds) Statistical Decision Theory and Related Topics 4, vol 1, Springer-Verlag, New York, pp 409–418

[124] Sacks (1963) title. ? p ?

[125] Sandved E (1968) Ancillary statistics and estimation of the loss in estimation problems. Annals of Mathematical Statistics 39:1756–1758

[126] Schwartz G (1978) Estimating the dimension of a model. Annals of Statistics 6:461–464

[127] Schwartz L (1973) Théorie des distributions. Hermann

[128] Sengupta, Sen (1991) Shrinkage estimation in a restricted parameter space. Sankhyā 53:389–411

[129] Shao J (2003) Mathematical Statistics. Springer

[130] Stein C (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 1, University of California Press, Berkeley, pp 197–206

[131] Stein C (1964) Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. Ann Inst Statist Math 16:155–160

[132] Stein C (1973) Estimation of the mean of a multivariate normal distribution. In: Proc. Prague Symp. Asymptotic Statist., pp 345–381

[133] Stein C (1977) Estimating the covariance matrix, unpublished manuscript

[134] Stein C (1981) Estimation of the mean of multivariate normal distribution. Annals of Statistics 9:1135–1151

[135] Stoer, Witzgall (1970) title. ? p ?

[136] Strawderman WE (1971) Proper Bayes minimax estimators of the multivariate normal mean. The Annals of Mathematical Statistics pp 385–388

[137] Strawderman WE (1974) title. ? p ?

[138] Strawderman WE (1992) title. ? p ?

[139] Strawderman WE (2001) title. ? p ?

[140] Strawderman WE (2003) title. ? p ?

[141] Stroock DW (1990) A concise introduction to the theory of integration. World Scientific, Singapore, New Jersey, London, Hong Kong

[142] Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58:267–288

[143] Wan ATK, Zou G (2004) On unbiased and improved loss estimation for the mean of a multivariate normal distribution with unknown variance. Journal of Statistical Planning and Inference 119:17–22

[144] Wells M (1989) ? ? p ?

[145] Wells M, Zhou (2008) ? ? p ?

[146] Widder DV (1946) The Laplace Transform. Princeton University Press, Princeton

[147] Wither C (1991) A class of multiple shrinkage estimators. Annals of the Institute of Statistical Mathematics 43:147–156

[148] Ye J (1998) On meausuring and correcting the effects of data mining and model selection. Journal of the American Statistical Association 93:120–131

[149] Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67:301–320

[150] Zou H, Hastie T, Tibshirani R (2007) On the degrees of freedom of the lasso. The Annals of Statistics 35:2173–2192