Chapter 3
Estimation of a Normal Mean Vector II

As we saw in Chap. 2, the frequentist paradigm is well suited for risk evaluations, but is less useful for estimator construction. It turns out that the Bayesian approach is complementary, as it is well suited for the construction of possibly optimal estimators. In this chapter we take a Bayesian view of minimax shrinkage estimation. In Sect. 3.1 we derive a general sufficient condition for minimaxity of Bayes and generalized Bayes estimators in the known variance case, we also illustrate the theory with numerous examples. In Sect. 3.2 we extend these results to the case when the variance is unknown. Section 3.3 considers the case of a known covariance matrix under a general quadratic loss. The admissibility of Bayes estimators in discussed in Sect. 3.4. Interesting connections to MAP estimation, penalized likelihood methods, and shrinkage estimation are developed in Sect. 3.5. The fascinating connections between Stein estimation and estimation of a predictive density under Kullback-Leibler divergence are outlined in Sect. 3.6.

3.1 Bayes Minimax Estimators

In this section, we derive a general sufficient condition for minimaxity of Bayes and generalized Bayes estimators when $X \sim \mathscr{N}_p(\theta, \sigma^2 I_p)$, 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 is superharmonic. We establish these results in Sect. 3.1.1 and apply them in Sect. 3.1.2 to obtain classes of prior distributions which lead to minimax (generalized and proper) Bayes estimators. Section 3.1.3 will be devoted to minimax multiple shrinkage estimators.

D. Fourdrinier et al., *Shrinkage Estimation*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-02185-6_3

Throughout this section, let $X \sim \mathscr{N}_p(\theta, \sigma^2 I_p)$ (with σ^2 known) and the 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

$$m(x) = K \int_{\mathbb{R}^p} e^{-\frac{\|x-\theta\|^2}{2\sigma^2}} \, d\pi(\theta). \tag{3.1}$$

Recall from Sect. 1.4 that the Bayes estimator corresponding to $\pi(\theta)$ is given by

$$\delta_\pi(X) = X + \sigma^2 \frac{\nabla m(X)}{m(X)}. \tag{3.2}$$

Since the constant K in (3.1) plays no role in (3.2) we will typically take it to be equal to 1 for simplicity. It may happen that an estimator will have the form (3.2) 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.

3.1.1 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. This condition relies on the superharmonicity of the square root of the marginal. Recall from Corollary A.2 in Appendix A.8.3 that a function f from $\mathbb{R}^p$ into $\mathbb{R}$ which is twice weakly differentiable and lower semicontinuous is superharmonic if and only if, for almost every $x \in \mathbb{R}^p$, we have $\Delta f(x) \leq 0$, where Δf is the weak Laplacian of f. Note that, if the function f is analytic, the last inequality holds for any $x \in \mathbb{R}^p$.

Theorem 3.1 *Under the model of this section, an estimator of the form* (3.2) *has finite risk if* $E_\theta\big[\|\nabla m(X)/m(X)\|^2\big] < \infty$ *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 First, note that, as noticed in Example 1.1, the marginal m is a positive analytic function, and so is $\sqrt{m}$.

Using Corollary 2.1 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 $E_\theta\big[\|\nabla m(X)/m(X)\|^2\big] < \infty$. Also, it is minimax provided, for almost any $x \in \mathbb{R}^p$,

$$\mathscr{D}(x) = \frac{\|\nabla m(x)\|^2}{m^2(x)} + 2\,\mathrm{div}\frac{\nabla m(x)}{m(x)} \leq 0\,.$$

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,

$$\begin{aligned} \mathscr{D}(x) &= \frac{2\,m(x)\,\Delta m(x) - \|\nabla m(x)\|^2}{m^2(x)} \qquad (3.3)\\ &= 4\,\frac{\Delta\sqrt{m(x)}}{\sqrt{m(x)}}\,. \end{aligned}$$

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 3.1 *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 (2.19) *with*

$$a\,r(t) = -2\,\frac{m'(t)}{m(t)}\,t \qquad \forall t \ge 0\,;$$

(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 3.1 minimaxity of $\delta_\pi(X)$ follows from the superharmonicity of $\sqrt{m(X)}$, it is worth noting that, in the setting of Corollary 3.1, 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 3.2

(1) *A finite risk (generalized, proper, or pseudo) Bayes estimator of the form* (3.2) *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* (3.2) *is minimax.*

Proof Part (1) follows from the first equality in (3.3), which shows that superharmonicity of m implies superharmonicity of $\sqrt{m}$. Indeed, the superharmonicity of m implies the 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 Green's formula (see also Sect. A.9). More generally, any mixture of superharmonic functions is superharmonic (Sect. A.8). □

Note that the condition of finiteness of risk is superfluous for proper Bayes estimators 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 (1985a) 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 consequently is not analytic. Thus, $m(X)$ cannot be a true marginal (for any generalized prior). However,

$$\nabla m(X) = \frac{-2\,b}{(\|X\|^2)^{b+1}}\,X$$

and

$$\frac{\nabla m(X)}{m(X)} = \frac{-2\,b}{\|X\|^2}\,X,$$

which is weakly differentiable if $p \geq 3$ (see Sect. 2.3). Hence, for $p \geq 3$, the James-Stein estimator

$$\delta_{2b}^{JS}(X) = \left(1 - \frac{2\,b\,\sigma^2}{\|X\|^2}\right) X$$

is a pseudo-Bayes estimator. Also, a simple calculation gives

$$\Delta m(X) = \frac{(-2\,b)[p - 2\,(b + 1)]}{(\|X\|^2)^{b+1}}.$$

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 3.1 gives minimaxity for $0 \leq b \leq p - 2$ which agrees with Theorem 2.2 (with $a = 2b$), while an application of Corollary 3.2 establishes minimaxity for only half of the interval, i.e. $0 \leq b \leq (p - 2)/2$. Thus, while useful, the corollary 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

$$\int_{\|x\| \geq R} m(x)\,dx = \int_{\|x\| \geq R} \frac{1}{(\|X\|^2)^b}\,dX \propto \int_R^\infty \frac{r^{p-1}}{r^{2\,b}}\,dr = \int_R^\infty r^{p-2\,b-1}\,dr$$

and that this integral is finite if and only if $p - 2\,b < 0$. 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 3.2).

3.1.2 Construction of (Proper and Generalized) Minimax Bayes Estimators

Corollary 3.1 provides 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 et al. (1998). Although Corollary 3.1 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 consequently a superharmonic prior density) cannot lead to a proper Bayes estimator.

Theorem 3.2 *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 from Example 1.1 of Sect. 1.4 that the posterior risk is given, for any $x \in \mathbb{R}^p$, by

$$p\,\sigma^2 + \sigma^4\,\frac{m(x)\,\Delta m(x) - \|\nabla m(x)\|^2}{m^2(x)}.$$

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

$$\begin{aligned} r(\pi) &= E^\pi\Big[E_\theta\Big[p\,\sigma^2 + \sigma^4\,\tfrac{2\,m(X)\Delta m(X) - \|\nabla m(X)\|^2}{m^2(X)}\Big]\Big] \\ &= E^m\Big[p\,\sigma^2 + \sigma^4\,\tfrac{2m(X)\Delta m(X) - \|\nabla m(X)\|^2}{m^2(X)}\Big], \end{aligned}$$

since the unbiased estimate of risk does not depend on θ, by definition. Hence, by taking the difference,

$$E^m\left[\frac{\Delta m(X)}{m(X)}\right] = 0\,.$$

Now, since the marginal m is superharmonic ($\Delta m(x) \le 0$ 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

$$\pi(\theta) = k \int_0^\infty \exp\left(-\frac{\|\theta\|^2}{2\,\sigma^2\, v}\right) v^{-p/2}\, h(v)\, dv \tag{3.4}$$

for some constant k and some nonnegative function 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

$$\nabla m(x) = \int_0^\infty \nabla m_v(x)\, h(v)\, dv \tag{3.5}$$

and

$$\Delta m(x) = \int_0^\infty \Delta m_v(x)\, h(v)\, dv \tag{3.6}$$

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

$$I_j(y) = \int_0^\infty \exp(-y/(1+v))\, (1+v)^{-j}\, h(v)\, dv$$

exists for $j \geq p/2$. Hence, with $y = \|x\|^2/2\sigma^2$, we have

$$\begin{aligned} m(x) &= k\, I_{p/2}(y) \\ \nabla m(x) &= -\frac{k}{\sigma^2}\, I_{p/2+1}(y)\, x \\ \Delta m(x) &= -\frac{k}{\sigma^2}\left[p\, I_{p/2+1}(y) - 2\, y\, I_{p/2+2}(y)\right] \\ \|\nabla m(x)\|^2 &= 2\,\frac{k^2}{\sigma^2}\, y\, I^2_{\frac{p}{2}+1}(y). \end{aligned} \tag{3.7}$$

Note that

$$\frac{\|\nabla m(x)\|^2}{m^2(x)} = \frac{2}{\sigma^2}\,\frac{I^2_{p/2+1}(y)}{I^2_{\frac{p}{2}}(y)}\, y \leq \frac{2\, y}{\sigma^2} = \frac{\|x\|^2}{\sigma^4}$$

since $I_{j+p}(y) \leq I_j(y)$. Hence,

$$E_0\left[\frac{\|\nabla m(x)\|^2}{m^2(x)}\right] \leq E_0\left[\frac{\|x\|^2}{\sigma^4}\right] < \infty\,,$$

which, according to Theorem 3.1, guarantees the finiteness of the risk of the Bayes estimator $\delta_\pi(X)$ in (3.2). Furthermore, the unbiased estimator of risk difference (3.3) can be expressed as

$$\begin{aligned} \mathscr{D}(X) = &-\tfrac{2}{\sigma^2}\left[p\, I_{p/2+1}(y) - 2\, y\, I_{p/2+2}(y)\right]/I_{p/2}(y) \\ &-\tfrac{2}{\sigma^2}\left[y\, I^2_{p/2+1}(y)/I^2_{p/2}(y)\right] \\ = &\ \tfrac{2\, I_{p/2+1}(y)}{\sigma^2\, I_{p/2}(y)}\left[\tfrac{2\, y\, I_{p/2+2}(y)}{I_{p/2+1}(y)} - p - \tfrac{y\, I_{p/2+1}(y)}{I_{p/2}(y)}\right]. \end{aligned} \tag{3.8}$$

Then the following intermediate result follows immediately from (3.2) and Theorem 3.1 since finiteness of risk has been guaranteed above.

Lemma 3.1 *The generalized Bayes estimator corresponding to the prior density* (3.4) *is minimax provided*

$$\frac{2\, I_{p/2+2}(y)}{I_{p/2+1}(y)} - \frac{I_{p/2+1}(y)}{I_{p/2}(y)} \leq \frac{p}{y}\,. \tag{3.9}$$

The next theorem gives sufficient conditions on the mixing density $h(\cdot)$ so that the resulting generalized Bayes estimator is minimax.

Theorem 3.3 *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*

$0 \le l_2 \le B$ with $A + 2B \le (p-2)/2$. 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 (3.2) for the prior density (3.4) 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

$$I_k(y) = \int_0^\infty \exp(-y/(1+v))\,(1+v)^{-k}\,h(v)\,dv.$$

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 $k \ge p/2+1$,

$$\begin{aligned}
I_k(y) &= \frac{(1+v)^{-k+2}\exp(-y/(1+v))\,h(v)}{y}\Big|_0^\infty \\
&\quad + \frac{k-2}{y}\int_0^\infty \exp\Big(-\frac{y}{1+v}\Big)\,(1+v)^{-k+1}\,h(v)\,dv \\
&\quad - \frac{1}{y}\int_0^\infty \exp\Big(-\frac{y}{1+v}\Big)\,(1+v)^{-k+2}\,h'(v)\,dv \\
&= -\frac{e^{-y}\,h(0)}{y} + \frac{k-2}{y}\,I_{k-1}(y) \\
&\quad - \frac{1}{y}\int_0^\infty \exp\Big(-\frac{y}{1+v}\Big)\,(1+v)^{-k+2}\,h'(v)\,dv\,. \qquad (3.10)
\end{aligned}$$

Applying (3.10) to both numerators in the left-hand side of (3.9) we have

$$\begin{aligned}
&\frac{2}{I_{p/2+1}(y)}\left[\frac{-e^{-y}\,h(0)}{y} + \frac{p}{2y}\,I_{p/2+1}(y) - \frac{1}{y}\int_0^\infty \exp\Big(-\frac{y}{1+v}\Big)\,(1+v)^{-p/2}\,h'(v)\,dv\right] \\
&-\frac{1}{I_{p/2}(y)}\left[\frac{-e^{-y}\,h(0)}{y} + \frac{p-2}{2y}\,I_{p/2}(y) - \frac{1}{y}\int_0^\infty \exp\Big(-\frac{y}{1+v}\Big)\,(1+v)^{-p/2+1}\,h'(v)\,dv\right] \\
&\le \frac{p+2}{2y} - \frac{2\int_0^\infty \exp\Big(-\frac{y}{1+v}\Big)\,(1+v)^{-p/2+2}\,h'(v)\,dv}{y\,I_{p/2+1}(y)} \\
&+ \frac{\int_0^\infty \exp\Big(-\frac{y}{1+v}\Big)\,(1+v)^{-p/2+1}\,h'(v)\,dv}{y\,I_{p/2}(y)}
\end{aligned}$$

since $I_{p/2+1}(y) < I_{p/2}(y)$. Then it follows from Lemma 3.1 that $\delta_\pi(X)$ is minimax provided, for any $y \ge 0$,

$$J_p^y \leq p - \frac{p+2}{2} = \frac{p-2}{2},$$

where

$$J_p^y = -2\, E_{p/2+1}^y\left[(V+1)\frac{h'(V)}{h(V)}\right] + E_{p/2}^y\left[(V+1)\frac{h'(V)}{h(V)}\right]$$

and where $E_k^y[f(V)]$ is the expectation of $f(V)$ with respect to the random variable V with density $g_k^y(v) = \exp(-y/(1+v))\,(1+v)^{-k}\,h(v)/I_k(y)$. Now upon 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

$$J_p^y = 2\,E_{p/2+1}^y\,[l_1(V) + l_2(V)] - E_{p/2}^y\,[l_1(V) + l_2(V)]$$

$$\leq 2\,E_{p/2+1}^y\,[l_1(V)] - E_{p/2}^y\,[l_1(V)] + 2\,E_{p/2+1}^y\,[l_2(V)]$$

since $l_2 \geq 0$. Also

$$E_{p/2+1}^y\,[l_1(V)] \leq E_{p/2}^y\,[l_1(V)]$$

since l_1 is nondecreasing. Then

$$J_p^y \leq E_{p/2}^y\,[l_1(V)] + 2\,E_{p/2+1}^y\,[l_2(V)] \leq A + 2\,B \leq \frac{p-2}{2}.$$

since $l_1 \leq A$ and $l_2 \leq B$ and by the assumptions on A and B. The result follows. □

The following corollary allows the construction of mixing distributions so that the conditions of the theorem are met and the resulting (generalized or proper) Bayes estimators are minimax.

Corollary 3.3 *Let $\psi = \psi_1 + \psi_2$ be a continuous function such that $\psi_1 \leq C$ and is nondecreasing, while $0 \leq \psi_2 \leq D$, and where $C \leq -2D$. Define, for $v > 0$, $h(v) = \exp\left[-\frac{1}{2}\int_{v_0}^{v}\frac{2\,\psi(u)+p-2}{u+1}\,du\right]$ where $v_0 \geq 0$. Assume also that $\lim_{v\to\infty} h(v)/(1+v)^{p/2-1} = 0$ and that $I_{p/2}(y) = \int_0^\infty \exp(-y/(1+v))\,(1+v)^{-p/2}\,h(v)\,dv < \infty$.*

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

$$-(v+1)\,\frac{h'(v)}{h(v)} = \psi_1(v) + \psi_2(v) + \frac{p-2}{2}.$$

Setting $l_1(v) = \psi_1(v) + (p-2)/2$ and $l_2(v) = \psi_2(v)$, the result follows from Theorem 3.1 with $A = (p-2)/2 + C$ and $B = D$. □

Note that finiteness of $I_{p/2}(y)$ in Corollary 3.2 is assured if we strengthen the limit condition to $\lim_{v\to\infty} h(v)/(1+v)^{p/2-1-\epsilon} = 0$ for some $\epsilon > 0$, since this implies that, for $h(v)/(1+v)^{p/2} \leq M/(1+v)^{1+\epsilon}$ for some $M > 0$ and any $v > 0$. Thus

$$\begin{aligned} I_{p/2}(y) = \int_0^\infty \exp(-y/(1+v))\,(1+v)^{-p/2}\,h(v)\,dv &\leq \int_0^\infty (1+v)^{-p/2}\,h(v)\,dv \\ &\leq \int_0^\infty \frac{M}{(1+v)^{1+\epsilon}}\,dv \\ &< \infty\,. \end{aligned}$$

3.1.3 Examples

An interesting and useful class of examples results from the choice

$$\psi(v) = \alpha + \beta/v + \gamma/v^2 \tag{3.11}$$

for some $(\alpha, \beta, \gamma) \in \mathbb{R}^3$. A simple calculation shows

$$\begin{aligned} h(v) &= \exp\left[-\int_{v_0}^{v} \frac{\alpha + \beta/u + \gamma/u^2 + (p-2)/2}{u+1}\,du\right] \\ &\propto (v+1)^{\beta-\alpha-\gamma-\frac{p-2}{2}}\,v^{\gamma-\beta}\exp\left(\frac{\gamma}{v}\right). \end{aligned} \tag{3.12}$$

Example 3.1 (The Strawderman 1971 prior) Suppose $\alpha \leq 0$ and $\beta = \gamma = 0$ so that $h(v) \propto (v+1)^{-\alpha-(p-2)/2}$. 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 3.1 require $\lim_{v\to\infty} h(v)/(1+v)^{p/2-1} = \lim_{v\to\infty}(v+1)^{-\alpha-(p-2)} = 0$ and this is satisfied if $\alpha > 2-p$. Also

$$\begin{aligned} I_{p/2}(y) &= \int_0^\infty \exp(-y/(1+v))\,(1+v)^{-p/2}\,h(v)\,dv \\ &\propto \int_0^\infty \exp(-y/(1+v))\,(1+v)^{-\alpha-p+1}\,h(v)\,dv \\ &\leq \int_0^\infty (1+v)^{-\alpha-p+1}\,h(v)\,dv \\ &< \infty \end{aligned}$$

if $\alpha > 2-p$ as above. Hence in this case the corresponding generalized Bayes estimator is minimax if $2-p < \alpha \leq 0$ (which requires $p \geq 3$).

Furthermore it is proper Bayes minimax if $\int_0^\infty (1+v)^{-\alpha-(p-2)/2}\,dv < \infty$ which is equivalent to $2-p/2 < \alpha \leq 0$. This latter condition requires $p \geq 5$ and demonstrates the existence of proper Bayes minimax estimators for $p \geq 5$. We will see below that this is the class of priors studied in Strawderman (1971) under the alternative parametrization $\lambda = 1/(1+v)$.

Example 3.2 Consider $\psi(v)$ given by (3.11) with $\alpha \leq 0$, $\beta \leq 0$ and $\gamma \leq 0$. Here we take $\psi_1(v) = \psi(v)$, $\psi_2(v) = 0$, and $C = D = 0$. The minimaxity conditions of Corollary 3.2 require

$$\lim_{v \to \infty} h(v)/(1+v)^{p/2-1} = \lim_{v \to \infty} (v+1)^{\beta-\alpha-\gamma-p+2} v^{\gamma-\beta} \exp(\gamma/v) = 0.$$

This implies $2 - p < \alpha \leq 0$. The finiteness condition on

$$\begin{aligned} I_{p/2}(y) &= \int_0^\infty \exp(-y/(1+v))\,(1+v)^{-p/2}\,h(v)\,dv \\ &\propto \int_0^\infty e^{-\frac{y}{1+v}}(v+1)^{\beta-\alpha-\gamma-p+1} v^{\gamma-\beta} \exp(\gamma/v)\,dv \end{aligned}$$

also requires $2 - p < \alpha \leq 0$. Therefore, minimaxity is ensured as soon as $2 - p < \alpha \leq 0$.

Furthermore, the minimax estimator will be proper Bayes if

$$\int_0^\infty h(v)\,dv \propto \int_0^\infty (1+v)^{\beta-\alpha-\gamma-(p-2)/2}\, v^{\gamma-\beta}\, \exp(\gamma/v)\,dv < \infty.$$

This holds if $2 - \frac{p}{2} < \alpha \leq 0$ as in Example 3.1.

Example 3.3 Suppose $\alpha \leq 0$, $\beta > 0$, and $\gamma < 0$ and take

$$\psi_1(v) = \alpha + (\gamma/v)(1/ + \beta/\gamma) I_{[0,-2\gamma/\beta]}(v)\,,$$

$$\psi_2(v) = (\gamma/v)(1/v + \beta/\gamma)\,\mathbb{1}_{[-2\gamma/\beta,\infty]}(v)\,,$$

for $C = \alpha$ and $D = -\beta^2/4\gamma$.

Note first that $\psi_1(v)$ is monotone nondecreasing and bounded above by α; also, $0 \leq \psi_2(v) \leq -\beta^2/4\gamma$. Therefore, we require $C = \alpha < -2D = \beta^2/2\gamma$. The conditions $\lim_{v\to\infty} h(v)/(1+v)^{p/2-1} = 0$ and $\int_0^\infty \exp(-y/(1+v))\,(1+v)^{-p/2}\,h(v)\,dv < \infty$ are, as in Example 3.2, $2 - p < \alpha \leq 0$.

Thus, $\delta_\pi(X)$ is minimax for $2 - p < \alpha \leq \beta^2/2\gamma < 0$. The condition for integrability of h is also, as in Example 3.2, i.e. $2 - \frac{p}{2} < \alpha \leq \beta^2/2\gamma < 0$.

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 representation of $\delta_\pi(X)$.

For simplicity, in the following examples, we assume $\sigma^2 = 1$.

Example 3.4 (Student-t priors) In this example we take $\psi(v)$ as in Examples 3.2 and 3.3 with the specific choices $\alpha = (m-p+4)/2 \leq 0$, $\beta = (m\,(1-\varphi)+2)/2$, and $\gamma = -m\,\varphi/2 \leq 0$, where $m \geq 1$. In this case $h(v) = C\,v^{-(m+2)/2}\exp(-m\,\varphi/2\,v)$, an inverse gamma density. 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 1982, p.33 or Robert 1994, p.174). If $\sigma^2 \neq 1$, the scale of the t-distribution is $\varphi\,\sigma$.

For various different values of m and φ, either the conditions of Example 3.2 or the conditions of Example 3.3 apply. Both examples require $\alpha = (m-p+4)/2 \leq 0$, or equivalently $1 \leq m \leq p-4$ (so that $p \geq 5$), and $\gamma = -m\,\varphi/2 \leq 0$.

Example 3.2 requires $\beta = \big(m\,(1-\varphi)+2\big)/2 < 0$, or equivalently, $\varphi \geq (m+2)/m$. The condition for minimaxity $2-p < \alpha \leq 0$ is satisfied since it is equivalent to $m > -p$. Furthermore the condition for proper Bayes minimaxity, $2-\frac{p}{2} < \alpha \leq 0$, is satisfied as well since it reduces to $m > 0$. Hence, if $\varphi \geq (m+2)/m$, the scaled p-variate t prior distribution leads to a proper Bayes minimax estimator for $p \geq 5$ and $m \leq p-4$.

On the other hand, when $\varphi < (m+2)/m$, or equivalently, $\beta > 0$, the conditions of Example 3.3 are applicable. Considering the proper Bayes case only, the condition for minimaxity of the Bayes estimator is

$$2-\frac{p}{2} < \alpha = \frac{m-p+4}{2} \leq \frac{\beta^2}{2\gamma} \leq \frac{\beta^2}{2\gamma} = -\frac{1}{4}\,\frac{\big(m\,(1-\varphi)+2\big)^2}{m\,\varphi}.$$

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

$$\varphi \geq \frac{p-2}{m}\left[1-\sqrt{1-\left(\frac{m+2}{p-2}\right)^2}\,\right] > 0\,.$$

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 $m + 2/m \leq p-4$, 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 Sect. 2.4.2 applied to Baranchik-type estimators of the form $\big(1 - a\,r(\|X\|^2)/\|X\|^2\big)\,X$. 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 $r(\cdot)$ function is closely connected to the function $v \mapsto -(v+1)h'(v)/h(v)$. To see this, note that

$$
\begin{aligned}
\delta_\pi(X) &= X + \sigma^2 \frac{\nabla m(X)}{m(X)} \\
&= \left(1 - \frac{I_{p/2+1}(y)}{I_{p/2}(y)}\right) X \qquad \text{from (3.2) with } y = \|X\|^2/2\sigma^2
\end{aligned}
$$

$$
= \left(1 - \frac{1}{y}\left(\frac{p-2}{2} - \frac{\int_0^\infty e^{-\frac{y}{1+v}} (1+v)^{-p/2} [(v+1)h'(v)/h(v)]\, dv - e^{-y} h(0)}{I_{p/2}(y)}\right)\right) X
$$

$$
= \left(1 - \frac{2\sigma^2}{\|X\|^2}\left(\frac{p-2}{2} + E^y_{p/2}\left[-\frac{(V+1)h'(V)}{h(V)}\right] - \frac{e^{-\frac{\|X\|^2}{2\sigma^2}} h(0)}{I_{p/2}(\frac{\|X\|^2}{2\sigma^2})}\right)\right) X\,,
$$

where $E^y_k(f)$ is as in the proof of Theorem 3.1, the second to last equality following from (3.4).

Hence, the Bayes estimator is of Baranchik form with

$$
ar(\|X\|^2) = 2\left(\frac{p-2}{2} + E^{\frac{\|X\|^2}{2\sigma^2}}_{p/2}\left[-\frac{(V+1)h'(V)}{h(V)}\right] - \frac{e^{-\frac{\|X\|^2}{2\sigma^2}} h(0)}{I_{p/2}(\frac{\|X\|^2}{2\sigma^2})}\right).
$$

□

Recall, as in the proof of Theorem 3.1, that the density $g^y_k(V)$ has a monotone decreasing likelihood ratio in k, but notice also that it has a monotone increasing likelihood ratio (actually as an exponential family) in y.

Hence, if $-\frac{(v+1)h'(v)}{h(v)}$ is nondecreasing, it follows that r is nondecreasing since $e^{-y}/I_{p/2}(y)$ is also nondecreasing. Then the following corollary is immediate from Theorem 3.3.

Corollary 3.4 *Suppose the prior is of the form* (3.4) *where* $-(v+1)\, h'(v)/h(v)$ *is nondecreasing and bounded above by* $A > 0$. *Then, the generalized Bayes estimator is minimax provided* $A \leq \frac{p-2}{2}$.

Proof As noted, $r(\cdot)$ is nondecreasing and is bounded above by $p - 2 + 2A \leq 2(p-2)$. □

Corollary 3.3 yields an alternative proof for the minimaxity of the generalized Bayes estimator in Example 3.1.

Finally, as indicated earlier in this section, an alternative parametrization has often been used in minimaxity proofs for the 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 Sect. 1.7. Here the distribution of $\theta \mid v \sim \mathcal{N}_p(0, v\sigma^2 X)$ and the unconditional density of v is the mixing density $h(v)$. The conditional distribution of θ given X and v is $\mathcal{N}_p(\frac{v}{1+v}X, \frac{V}{1+v}\sigma^2 I_p)$. 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

$$\theta \mid \lambda \sim \mathscr{N}(0, \frac{1-\lambda}{\lambda}\sigma^2 I) \tag{3.13}$$

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.

In particular, for the prior distribution of Example 3.1 where the mixing density on v is $h(v) = C\,(1+v)^{-\alpha-(p-2)/2}$, the corresponding mixture density on λ is given by $g(\lambda) = C\lambda^{\alpha+\frac{p-2}{2}-2} = C\lambda^{\beta}$ and ($\beta = \alpha + p/2 - 3$). The resulting prior is proper Bayes minimax if $2 - p/2 < \alpha \le 0$ or equivalently, $-1 < \beta \le /2 - 3$ (and $p \ge 5$). Note that, if $p \ge 6$, $\beta = 0$ satisfies the conditions and consequently the mixing prior $g(\lambda) \equiv 1$ on $0 \le \lambda \le 1$, i.e. the uniform prior on the shrinkage factor λ gives a proper Bayes minimax estimator. This class of priors is often referred to as the Strawderman priors.

To formalize the above discussion further we present a version of Theorem 3.3 in terms of the mixing distribution on λ. The proof follows from Theorem 3.3 and the change of variable $\lambda = 1/(1+v)$.

Corollary 3.5 *Let θ have the hierarchical prior $\theta \mid \lambda \sim \mathscr{N}_p(0, (\{1-\lambda\}/\lambda)\,\sigma^2\, I_p)$ where $\lambda \sim g(\lambda)$ for $0 \le \lambda \le 1$. Assume that $\lim_{\lambda\to 0} g(\lambda)\lambda^{p/2+1} = 0$ 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 $l_1^*(\lambda) \le A^*$, $0 \le l_2^*(\lambda) \le B^*$ with $A^* + 2B^* \le p/2 - 3$.*

Then the generalized Bayes estimator is minimax. Furthermore, if $\int_0^1 g(\lambda) d\lambda < \infty$, the estimator is also proper Bayes.

Example 3.5 (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 $0 < a \le p/2 - 2$ and $b \ge 1$.

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 $-p/2 < a \leq p/2 - 2$ and $b \geq (p + 2a + 2)(3p/2 + a)^{-1}$. This bound in b is in (0, 1) for $a < p/2 - 2$. 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.

3.1.4 Multiple Shrinkage Estimators

In this subsection, we consider a class of estimators that adaptively choose a point (or subspace) toward which to shrink. George (1986a,b) originated work in this area and the results in this section are largely due to him. The basic fact 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 the following result from Corollary 3.1.

Theorem 3.4 *Let $m_\alpha(x)$ be a family of twice weakly differentiable nonnegative superharmonic functions and $G(x)$ a positive measure such that $m(x) = \int m_\alpha(x)\, dG(x) < \infty$, for all $x \in \mathbb{R}^p$.*

Then the (generalized, proper, or pseudo) Bayes estimator

$$\delta(X) = X + \sigma^2 \frac{\nabla m(X)}{m(X)}$$

is minimax provided $E[\|\nabla m\|^2/m^2] < \infty$.

The following corollary for finite mixtures is useful.

Corollary 3.6 *Suppose that $m_i(x)$ is superharmonic and $E[\|\nabla m_i(X)\|^2/m_i^2(X)] < \infty$ 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 $W_i(X) = m_i(X)/\sum_{i=1}^n m_i(X)$ for $0 < W_i(X) < 1, \sum_{i=1}^n W_i(X) = 1$ is minimax. (Note that $E_\theta[\|\nabla m(X)\|^2/m^2(X)] < \sum_{i=1}^n E_\theta[\|\nabla m_i(X)\|^2/m^2(X_i)] < \infty$.)

Example 3.6

(1) Multiple shrinkage James-Stein estimator. Suppose we have several possible points $X_1, X_2, \ldots, X_n$ toward which to shrink. Recall that $m_i(x) = (1/\|x - X_i\|^2)^{(p-2)/2}$ is superharmonic if $p \geq 3$ and the corresponding pseudo-Bayes estimator is $\delta_i(X) = X_i + \left(1 - (p-2)\,\sigma^2/\|X - X_i\|^2\right)(X - X_i)$. Hence, if $m(x) = \sum_{i=1}^{n} m_i(x)$, the resulting minimax pseudo Bayes estimator is given by

$$\delta(X) = \sum_{i=1}^{n} \left[X_i + (1 - \frac{(p-2)\sigma^2}{\|X - X_i\|^2})(X - X_i) \right] W_i(X)$$

where $W_i(X) \propto \left(1/\|X - X_i\|^2\right)^{(p-2)/2}$ 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

$$m_i(x) = \begin{cases} C \, \exp\left(\frac{\|x - X_i\|^2}{2\,\sigma^2}\right) & \text{if } \|x - X_i\|^2 < (p-2)\,\sigma^2 \\ \left(\frac{1}{\|x - X_i\|^2}\right) & \text{if } \|x - X_i\|^2 \geq (p-2)\,\sigma^2 \end{cases}$$

where $C = \left(1/(p-2)\,\sigma^2\right)^{(p-2)/2} e^{(p-2)/2}$ 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

$$\frac{\nabla m_i(X)}{m_i(X)} = \begin{cases} -\frac{X - X_i}{\sigma^2} & \text{if } \|X - X_i\|^2 < (p-2)\sigma^2, \\ -\frac{(p-2)}{\|X - X_i\|^2} & \text{otherwise.} \end{cases}$$

The adaptive combination is again minimax by the corollary and inherits the usual advantages of the positive-part estimator over the James-Stein estimator.

Note that 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 $(p-2)/2$ 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 3.3 for $p-2 < a \le 2\,(p-2)$ 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 then $\pi(\theta) = \sum_{i=1}^n \pi_i(\theta)$ is also superharmonic as is $m(x) = \sum_{i=1}^n m_i(x)$.

For example, $\pi_i(\theta) = \left(1/b + \|\theta - X_i\|^2\right)^{a/2}$, for $b \ge 0$ and $0 \le a \le p-2$, is a suitable prior. Interestingly, according to a heuristic of Brown (1971), $m(x)$ in this case should behave for large $\|x\|^2$ as $\sum_{i=1}^n 1/\left(b + \|x - X_i\|^2\right)^{a/2}$, 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 that shrink adaptively toward subspaces. Further examples can be found in George (1986a,b), Ki and Tsui (1990) and Wither (1991).

3.2 Bayes Estimators in the Unknown Variance Case

3.2.1 A Class of Proper Bayes Minimax Admissible Estimators

In this subsection, we give a class of hierarchical Bayes minimax estimators for the model

$$X \sim \mathcal{N}_p(\theta, \sigma^2 I_p) \quad S \sim \sigma^2 \chi_k^2\,, \tag{3.14}$$

where S is independent of X, under scale invariant squared error loss

$$L(\theta, \delta(X,S)) = \frac{\|\delta(X,S) - \theta\|^2}{\sigma^2}\,. \tag{3.15}$$

We reparameterize σ^2 as $1/\eta$ and consider the following hierarchically, on the unknown parameters, structured prior(θ, η), which is reminiscent of the hierarchical version of the Strawderman prior in (3.13),

$$\begin{aligned} \theta|\lambda,\eta &\sim \mathcal{N}_p\left(0, \frac{1}{\eta}\,\frac{1-\lambda}{\lambda}\, I_p\right) \\ \eta &\sim Gamma\left(\frac{b}{2}, \frac{c}{2}\right) \\ \lambda &\sim (1+a)\,\lambda^a, \quad 0 < \lambda < 1\,. \end{aligned} \tag{3.16}$$

Lemma 3.2 *For the model* (3.14) *and loss* (3.15)*, the (generalized or proper) Bayes estimator of* θ *is given by*

$$\delta(X, S) = \left(1 - \frac{S}{\|X\|^2} r(\|X\|^2, S)\right) X \tag{3.17}$$

where

$$r(\|X\|^2, S) = \frac{\|X\|^2}{\|X\|^2 + c} \frac{\int_0^{(\|X\|^2+c)/S} u^{A+1} \left(\frac{1}{u+1}\right)^{B+1} du}{\int_0^{(\|X\|^2+c)/S} u^{A} \left(\frac{1}{u+1}\right)^{B+1} du} \tag{3.18}$$

where

$$A = \frac{p+a+b}{2} \quad \text{and} \quad B = \frac{p+k+b-2}{2} \tag{3.19}$$

provided $A > -1$, $A - B < 0$, *and* $c > 0$.

Proof Under the loss in (3.15) the Bayes estimator for the model in (3.16) is given by

$$\delta(X, S) = \frac{E[\theta\,\eta|X, S]}{E[\eta|X, S]}\,. \tag{3.20}$$

Expressing the expectation in the numerator of (3.20) gives

$$\begin{aligned} E[\theta\,\eta|X, S] &= \int_0^\infty \int_0^1 \int_{\mathbb{R}^p} \theta\,\eta^{p/2+1} \left(\frac{\lambda\,\eta}{1-\lambda}\right)^{p/2} \\ &\qquad \times \exp\!\left(-\frac{\eta}{2}\left[\|x-\theta\|^2 + \frac{\lambda}{1-\lambda}\,\|\theta\|^2\right]\right) \eta^{(k+b-2)/2} \\ &\qquad \times \lambda^{(b+a)/2} \exp\left(-\frac{\eta}{2}\,(S + \lambda\,c)\right) d\theta\,d\eta\,d\lambda \\ &= \int_0^\infty \int_0^1 (1-\lambda)\lambda^A \eta^B \exp\left(-\frac{\eta}{2}\,(S + \lambda(\|x\|^2 + c))\right) d\eta\,d\lambda \end{aligned} \tag{3.21}$$

upon integrating with respect to θ and evaluating with the constants in (3.19). Similarly, for the denominator in (3.20)

$$\begin{aligned} E[\eta|X, S] &= \int_0^\infty \int_0^1 \int_{\mathbb{R}^p} \eta^{p/2+1} \left(\frac{\lambda\,\eta}{1-\lambda}\right)^{p/2} \\ &\qquad \times \exp\!\left(-\frac{\eta}{2}\left[\|x-\theta\|^2 + \frac{\lambda}{1-\lambda}\,\|\theta\|^2\right]\right) \eta^{(k+b-2)/2} \end{aligned}$$

$$\times \lambda^{(b+a)/2} \exp\left(-\frac{\eta}{2}\,(S+\lambda\, c)\right) d\theta\, d\eta\, d\lambda$$

$$= \int_0^\infty \int_0^1 \eta^B \lambda^A \exp\left(-\frac{\eta}{2}\,(S+\lambda(\|x\|^2+c))\right) d\eta d\lambda. \tag{3.22}$$

Therefore from (3.21) and (3.22) the Bayes estimator in (3.20) has the form

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

where

$$r(\|X\|^2,S) = \frac{\|X\|^2}{S}\, \frac{\int_0^\infty \int_0^1 \eta^B\, \lambda^{A+1}\, \exp\left(-\frac{\eta\, S}{2}\, \left(1+\lambda\, \frac{\|x\|^2+c}{S}\right)\right) d\eta\, d\lambda}{\int_0^\infty \int_0^1 \eta^B\, \lambda^{A}\, \exp\left(-\frac{\eta\, S}{2}\, \left(1+\lambda\, \frac{\|x\|^2+c}{S}\right)\right) d\eta\, d\lambda}$$

$$= \frac{\|X\|^2/S}{(\|X\|^2+c)\, S}\, \frac{\int_0^{(\|X\|^2+c)/S} \int_0^\infty \eta^B\, u^{A+1}\, \exp\left(-\frac{\eta\, S}{2}\,(1+u)\right) d\eta\, du}{\int_0^{(\|X\|^2+c)/S} \int_0^\infty \eta^B\, u^{A}\, \exp\left(-\frac{\eta\, S}{2}\,(1+u)\right) d\eta\, du}$$

$$= \frac{\|X\|^2}{\|X\|^2+c}\, \frac{\int_0^{(\|X\|^2+c)/S} u^{A+1} \left(\frac{1}{u+1}\right)^{B+1} du}{\int_0^{(\|X\|^2+c)/S} u^{A} \left(\frac{1}{u+1}\right)^{B+1} du},$$

with the change of variable $u = \lambda\,(\|X\|^2 + c)/S$ is made in the next to last step. □

The properties of $r(\|X\|^2, S)$ in Lemma 3.2 are given in the following result.

Lemma 3.3 *The function $r(\|X\|^2, S)$ given in (3.18) satisfies the following properties:*

(i) *$r(\|X\|^2, S)$ is nondecreasing in $\|X\|^2$ for fixed S;*
(ii) *$r(\|X\|^2, S)$ is nonincreasing in S for fixed $\|X\|^2$; and*
(iii) $0 \le r(\|X\|^2, S) \le (A+1)/(B-A-1) = (p+a+b+2)/(k-a-4)$

provided the conditions of Lemma 3.2 hold.

Proof Note first that $\int_0^t u\, f(u)\, du / \int_0^t f(u)\, du$ is nondecreasing in t for any integrable nonnegative function $f(\cdot)$. Hence Part (i) follows since $r(\|X\|^2, S)$ is the product of two nonnegative nondecreasing functions $\|X\|^2/\|X\|^2 + c$ and $\int_0^{(\|X\|^2+c)/S} u\, f(u)\, du / \int_0^{(\|X\|^2+c)/S} f(u)\, du$ for $f(u) = u^A\,(1+u)^{-(B+1)}$.

Part (ii) follows from a similar reasoning since the first term is constant in S and $(\|X\|^2 + c)/S$ is decreasing in S.

To show Part (iii) note that, by Parts (i) and (ii),

$$\begin{aligned}0 \le r(\|X\|^2, S) &\le \lim_{\substack{\|X\|^2\to\infty\\ S\to 0}} r(\|X\|^2, S)\\ &\le \frac{\int_0^\infty u^{A+1}\left(\frac{1}{u+1}\right)^{B+1} du}{\int_0^\infty u^{A}\left(\frac{1}{u+1}\right)^{B+1} du}\\ &= \frac{\int_0^1 \lambda^{B-A-2}\,(1-\lambda)^{A+1}}{\int_0^1 \lambda^{B-A-1}\,(1-\lambda)^{A}}\\ &= \frac{A+1}{B-A-1}\\ &= \frac{p+a+b+2}{k-a-4}\,,\end{aligned}$$

expressing the beta functions and according to the values of A and B. □

We also need the following straightforward generalization of Corollary 2.6. The proof is left to the reader.

Corollary 3.7 *Under model* (3.14) *and loss* (3.15) *an estimator of the form*

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

is minimax provided

(i) $r(\|X\|^2, S)$ *is nondecreasing in* $\|X\|^2$ *for fixed* S*;*
(ii) $r(\|X\|^2, S)$ *is nonincreasing in* S *for fixed* $\|X\|^2$*; and*
(iii) $0 \le r(\|X\|^2, S) \le 2\,(p-2)/(k+2)$.

Combining Lemmas 3.2 and 3.3 and Corollary 3.7 gives the following result.

Theorem 3.5 *For the model* (3.14)*, loss* (3.15) *and hierarchical prior* (3.16)*, the generalized or proper Bayes estimator in Lemma* 3.2 *is minimax provided*

$$\frac{p+a+b+2}{k-a-4} \le \frac{2\,(p-2)}{k+2}\,. \tag{3.23}$$

Furthermore, if $p \ge 5$*, there exist values of* $a > -2$ *and* $b > 0$ *which satisfy* (3.23)*, i.e. such that the estimator is proper Bayes, minimax and admissible.*

Proof The first part is immediate. To see the second part, note that it suffices, if $a = -2 + \epsilon\; b = \delta$, for $\epsilon, \delta > 0$, that

$$\frac{p}{k-2} < \frac{p+\epsilon+\delta}{k-2-\epsilon} \le \frac{2\,(p-2)}{k+2}$$

equivalently $p > 4\,\frac{k-2}{k-6}$. Hence, for $p \geq 5$ and k sufficiently large, $k > 2\,(3\,p-4)/(p-4)$, there are values of a and b such that the priors are proper. □

Note that there exist values of a and b satisfying (3.23) and the assumptions of Lemma 3.2 whenever $p \geq 3$.

Strawderman (1973) gave the first example of a generalized and proper Bayes minimax estimators in the unknown variance setting. Zinodiny et al. (2011) also give classes of generalized and proper Bayes minimax estimators along somewhat similar lines as the above. The major difference is that the prior distribution on η $(= 1/\sigma^2)$ in the above development is also hierarchical, as it also depends on λ.

3.2.2 The Construction of a Class of Generalized Bayes Minimax Estimators

In this subsection we extend the generalized Bayes results of Sect. 3.1.2, using the ideas in Maruyama and Strawderman (2005) and Wells and Zhou (2008), to consider point estimation of the mean of a multivariate normal when the variance is unknown. Specifically, we assume the following model in (3.14) and the scaled squared loss function in (3.15).

In order to derive the (formal) Bayes estimator we reparameterize the model in (3.14) by replacing σ by η^{-1}. The model then becomes

$$\begin{aligned} &X \sim \mathcal{N}_p(\theta, \eta^{-2} I_p), \quad S \sim s^{k/2-1}\,\eta^k \exp(s\,\eta^2/2),\\ &\theta \sim \mathcal{N}_p(0, \nu\,\eta^{-2} I_p), \quad \nu \sim h(\nu), \quad \eta \sim \eta^d\,, \eta > 0\,, \end{aligned} \tag{3.24}$$

for some constant d. Under this model, the prior for θ is a scale mixture of normal distributions. Note that the above class of priors cannot be proper due to the impropriety of the distribution of η. However, as a consequence of the form of this model, the resulting generalized Bayes estimator is of the Baranchik form (3.17), with $r(\|X\|^2, S) = r(F)$, where $F = ||X||^2/S$.

We develop sufficient conditions on k, p, and $h(\nu)$ such that the generalized Bayes estimators with respect to the class of priors in (3.24) are minimax under the invariant loss function in (3.15). Maruyama and Strawderman (2005) and Wells and Zhou (2008) were able to obtain such sufficient conditions by applying the bounds and monotonicity results of Baranchik (1970), Efron and Morris (1976), and Fourdrinier et al. (1998).

Before we derive the formula for the generalized Bayes estimator under the model (3.24), we impose three regularity conditions on the parameters of priors. These conditions are easily satisfied by many hierarchical priors. These three conditions are assumed throughout this section.

C1: $A > 1$ where $A = \frac{d+k+p+3}{2}$;

C2: $\int_0^1 \lambda^{\frac{p}{2}-2} h\left(\frac{1-\lambda}{\lambda}\right) d\lambda < \infty$; and

C3: $\lim_{\nu \to \infty} \frac{h(\nu)}{(1+\nu)^{p/2-1}} = 0$.

Now, as in Sect. 3.1, we will first find the form of the Bayes estimator and then show that it satisfies some sufficient conditions for minimaxity. We start with the following lemma that corresponds to (3.2) in the known variance case and (3.18) in the previous subsection.

Lemma 3.4 *Under the model in* (3.24)*, the generalized Bayes estimator can be written as*

$$\delta(X, S) = X - R(F)\, X = X - \frac{r(F)}{F}\, X, \tag{3.25}$$

where $F = ||X||^2/S$*,*

$$R(F) = \frac{\int_0^1 \lambda^{p/2-1}\,(1+\lambda\, F)^{-A}\, h\left(\frac{1-\lambda}{\lambda}\right) d\lambda}{\int_0^1 \lambda^{p/2-2}\,(1+\lambda\, F)^{-A}\, h\left(\frac{1-\lambda}{\lambda}\right) d\lambda}, \tag{3.26}$$

and

$$r(F) = F\, R(F)\,. \tag{3.27}$$

Proof Under the loss function (3.15), the generalized Bayes estimator for the model (3.24) is

$$\begin{aligned}
\delta(X, S) &= \frac{E(\frac{\theta}{\sigma^2}|X, S)}{E(\frac{1}{\sigma^2}|X, S)}\\
&= \frac{\int_0^\infty h(\nu) \int_0^\infty [(\eta^2)^{A-\frac{1}{2}} e^{-\frac{1}{2}\eta^2 S} \int_{\mathbb{R}^p} (\frac{1}{2\pi\nu\eta^{-2}})^{\frac{p}{2}} \theta e^{-\frac{1}{2}\eta^2(\frac{||\theta||^2}{\nu}+||X-\theta||^2)} d\theta] d\eta\, d\nu}{\int_0^\infty h(\nu) \int_0^\infty [(\eta^2)^{A-\frac{1}{2}} e^{-\frac{1}{2}\eta^2 S} \int_{\mathbb{R}^p} (\frac{1}{2\pi\nu\eta^{-2}})^{\frac{p}{2}} e^{-\frac{1}{2}\eta^2(\frac{||\theta||^2}{\nu}+||X-\theta||^2)} d\theta] d\eta d\nu}\\
&= \left(1 - \frac{\int_0^\infty [(\frac{1}{1+\nu}) h(\nu) (\frac{1}{1+\nu})^{\frac{p}{2}} \int_0^\infty (\eta^2)^{A-\frac{1}{2}} e^{-\frac{1}{2}\eta^2(S+\frac{||X||^2}{1+\nu})}\, d\eta]\, d\nu}{\int_0^\infty [h(\nu)(\frac{1}{1+\nu})^{\frac{p}{2}} \int_0^\infty (\eta^2)^{A-\frac{1}{2}} e^{-\frac{1}{2}\eta^2(S+\frac{||X||^2}{1+\nu})}\, d\eta]\, d\nu}\right) X\\
&= \left(1 - \frac{\int_0^\infty (\frac{1}{1+\nu}) h(\nu) (\frac{1}{1+\nu})^{\frac{p}{2}} (1+\frac{F}{1+\nu})^{-A}\, d\nu}{\int_0^\infty h(\nu) (\frac{1}{1+\nu})^{\frac{p}{2}} (1+\frac{F}{1+\nu})^{-A}\, d\nu}\right) X.
\end{aligned} \tag{3.28}$$

Letting $\lambda = (1+\nu)^{-1}$, $\delta(X, S) = (1 - R(F))X$, 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 (3.2), that is, $\delta^{\pi}(X) = X + \sigma^2 \frac{\nabla m(X)}{m(X)}$.

The form of the Bayes estimator given in (3.25) gives an analogous form with the unknown variance replaced by a multiple of the usual unbiased estimator. In particular, define the "quasi-marginal"

$$\mathbf{M}(x, s) = \int\int f_X(x)\, f_S(s)\, \pi(\theta, \sigma^2)\, d\theta\, d\sigma^2$$

where

$$f_X(x) = \left(\frac{1}{2\pi\sigma^2}\right)^{p/2} e^{-\frac{1}{2\sigma^2}||x-\theta||^2}$$

and

$$f_S(s) = \frac{1}{2^{k/2}\Gamma(k/2)} s^{k/2-1} (\sigma^2)^{-k/2} e^{-\frac{s}{2\sigma^2}}.$$

A straightforward calculation shows $\mathbf{M}(x, s)$ is proportional to

$$\int_0^\infty h(\nu) \int_0^\infty [(\eta^2)^{A-\frac{3}{2}} e^{-\frac{1}{2}\eta^2 s} \int_{\mathbb{R}^p} (\frac{1}{2\pi\nu\eta^{-2}})^{\frac{p}{2}} e^{-\frac{1}{2}\eta^2(\frac{||\theta||^2}{\nu}+||x-\theta||^2)} d\theta] d\eta d\nu.$$

It is interesting to note the unknown variance analog of (3.2) is

$$\delta(X, S) = X - \frac{1}{2}\frac{\nabla_X \mathbf{M}(X, S)}{\nabla_S \mathbf{M}(X, S)}.$$

Lastly, note that the exponential term in the penultimate expression in the representation of $\delta(X, S)$ in (3.28) (that comes from the normal sampling distribution assumption) cancels. Hence there is a sort of robustness with respect to the sampling distribution. We will develop this theme in greater detail in Chap. 6 in the setting of spherically symmetric distributions.

3.2.2.1 Preliminary Results

The minimax property of the generalized Bayes estimator is closely related to the behavior of the $r(F)$ and $R(F)$ functions, which is in turn closely related to the behavior of

$$g(\nu) = -(\nu + 1)\frac{h'(\nu)}{h(\nu)}. \tag{3.29}$$

Fourdrinier et al. (1998) gave a detailed analysis of the type of function in (3.29). However, their argument was deduced from the superharmonicity of the square root of a marginal condition. Baranchik (1970) and Efron and Morris (1976) gave certain regularity conditions on the shrinkage function $r(\cdot)$ such that an estimator

$$\widehat{\theta}(X, S) = X - \frac{r(F)}{F} X \tag{3.30}$$

is minimax under the loss function (3.15) for the model (3.14). Both results require an upper bound on $r(F)$ and a condition on how fast $R(F) = r(F)/F$ decreases with F. Both theorems follow from a general result for spherically symmetric distributions given in Chap. 6 (Proposition 6.1), or by applying Theorem 2.5 in a manner similar to that in Corollary 2.3. The proofs are left to the reader.

Theorem 3.6 (Baranchik 1970) *Assume that $r(F)$ is increasing in F and $0 \leq r(F) \leq 2\,(p-2)/(k+2)$. Then any point estimator of the form (3.30) is minimax.*

Theorem 3.7 (Efron and Morris 1976) *Define $c_k = \frac{p-2}{k+2}$. Assume that $0 \leq r(F) \leq 2\,c_k$, that for all F with $r(F) < 2c_k$,*

$$\frac{F^{p/2-1}\, r(F)}{(2 - r(F)/c_k)^{1+2\,c_k}} \text{ is increasing in } F, \tag{3.31}$$

and that, if an F_0 exists such that $r(F_0) = 2c_k$, then $r(F) = 2\,c_k$ for all $F \geq F_0$. With the above assumptions, the estimator $\widehat{\theta}(X, S) = X - r(F)/F\ X$ is minimax.

Consequently, to apply these results one has to establish an upper bound for $r(F)$ in (3.27) and the monotonicity property for some variant of $r(F)$. The candidate we use is $\widetilde{r}(F) = F^c r(F)$ with a constant c. Note that the upper bound $2\,c_k$ is exactly the same upper bound needed in Corollary 3.7(iii). We develop the needed results below.

First note that if $h(\nu)$ is a continuously differentiable function on $[0, \infty)$, and regularity conditions C1, C2 and C3 hold, then the integrations by parts used in Lemmas 3.5 and 3.6 are valid.

Lemma 3.5 *Assume the regularity conditions C1, C2 and C3, and that $g(\nu) \leq M$, where M is a positive constant and $g(\nu)$ is defined as in (3.29). Then, for the $r(F)$ function (3.27), we have*

$$0 \leq r(F) \leq \frac{\frac{p}{2} - 1 + M}{A - \frac{p}{2} - M},$$

where A is defined in condition C1.

Proof By the definition in (3.26), $R(F) \geq 0$. Then $r(F) = F R(F) \geq 0$. 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

$$\begin{aligned} FI_{\frac{p}{2}-1,A,h}(F) &= \int_0^1 \lambda^{p/2-1} h\left(\frac{1-\lambda}{\lambda}\right) d\left[\frac{(1+\lambda F)^{1-A}}{1-A}\right] \\ &= \lambda^{\frac{p}{2}-1} h\left(\frac{1-\lambda}{\lambda}\right) \frac{(1+\lambda F)^{1-A}}{1-A}|_0^1 + \frac{1}{A-1}\int_0^1 (1+\lambda F)^{-A}(1+\lambda F) \\ &\quad \left[\left(\frac{p}{2}-1\right)\lambda^{\frac{p}{2}-2} h\left(\frac{1-\lambda}{\lambda}\right) - \frac{1}{\lambda^2}\lambda^{\frac{p}{2}-1} h'\left(\frac{1-\lambda}{\lambda}\right)\right] d\lambda. \end{aligned}$$

By C1 and C3, we know that the first term of the right hand side is nonpositive. The second term of the right hand side can be written as $N_1 + N_2 + N_3 + N_4$ where

$$N_1 = \frac{1}{A-1}\int_0^1 (1+\lambda F)^{-A}\left(\frac{p}{2}-1\right)\lambda^{\frac{p}{2}-2} h\left(\frac{1-\lambda}{\lambda}\right) d\lambda = \frac{\frac{p}{2}-1}{A-1} I_{\frac{p}{2}-2,A,h}(F),$$

$$\begin{aligned} N_2 &= \frac{1}{A-1}\int_0^1 (1+\lambda F)^{-A}\lambda^{\frac{p}{2}-2} h'\left(\frac{1-\lambda}{\lambda}\right)\left(\frac{-\lambda}{\lambda^2}\right) d\lambda \\ &= \frac{I_{\frac{p}{2}-2,A,h}(F)}{A-1} \frac{\int_0^1 \lambda^{\frac{p}{2}-2}(1+\lambda F)^{-A} g(\frac{1-\lambda}{\lambda}) h(\frac{1-\lambda}{\lambda})\, d\lambda}{\int_0^1 \lambda^{\frac{p}{2}-2}(1+\lambda F)^{-A} h(\frac{1-\lambda}{\lambda})\, d\lambda} \\ &\le \frac{M}{A-1} I_{\frac{p}{2}-2,A,h}(F), \end{aligned}$$

$$N_3 = \frac{\frac{p}{2}-1}{A-1} FI_{\frac{p}{2}-1,A,h}(F) = \frac{(\frac{p}{2}-1)r(F)}{A-1} I_{\frac{p}{2}-2,A,h}(F),$$

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

$$(A-1)r(F) \leq \left(\frac{p}{2}-1\right)+M+\left(\frac{p}{2}-1\right)r(F)+Mr(F) \Rightarrow r(F) \leq \frac{\frac{p}{2}-1+M}{A-\frac{p}{2}-M}.$$

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 establish the minimaxity of the generalized Bayes estimator. The following lemma is analogous to Theorem 3.3 in the known variance case.

Lemma 3.6 *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* $0 \leq l_2(\nu) \leq c$*, then* $\widetilde{r}(F) = F^c r(F)$ *is nondecreasing.*

Proof By taking the derivative, we only need to show (since $r(F) = FR(F)$)

$$0 \leq FR'(F) + (1+c)R(F), \tag{3.32}$$

which is equivalent to

$$0 \leq F\frac{I'_{\frac{p}{2}-1,A,h}(F)I_{\frac{p}{2}-2,A,h}(F) - I'_{\frac{p}{2}-2,A,h}(F)I_{\frac{p}{2}-1,A,h}(F)}{I^2_{\frac{p}{2}-2,A,h}(F)} + (1+c)\frac{I_{\frac{p}{2}-1,A,h}(F)}{I_{\frac{p}{2}-2,A,h}(F)}.$$

This in turn equivalent to

$$\begin{aligned} &-FI'_{\frac{p}{2}-1,A,h}(F)I_{\frac{p}{2}-2,A,h}(F) \\ &\quad\leq -FI'_{\frac{p}{2}-2,A,h}(F)I_{\frac{p}{2}-1,A,h}(F) + (1+c)I_{\frac{p}{2}-2,A,h}(F)I_{\frac{p}{2}-1,A,h}(F). \end{aligned} \tag{3.33}$$

Now note that

$$-FI'_{a,A,h}(F) = \int_0^1 \lambda^a (1+\lambda F)^{-A} h\left(\frac{1-\lambda}{\lambda}\right)\frac{A\lambda F}{1+\lambda F}d\lambda.$$

Define the intergral operator

$$J_a\left(f(u)\right) = \int_0^F u^a(1+u)^{-A} f(u)\, du.$$

Therefore,

$$J_a\left(h\left(\frac{F-u}{u}\right)\right) = \int_0^F u^a(1+u)^{-A} h\left(\frac{F-u}{u}\right) du$$

and

$$J_a\left(\frac{Au}{1+u}h\left(\frac{F-u}{u}\right)\right)=\int_0^F u^a(1+u)^{-A}\frac{Au}{1+u}h\left(\frac{F-u}{u}\right)du.$$

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

$$J_a\left(h\left(\frac{F-u}{u}\right)\right)=F^{a+1}I_{a,A,h}(F).$$

Now, with this new notation, it follows that (3.33) is equivalent to

$$\frac{J_{\frac{p}{2}-1}(\frac{Au}{1+u}h(\frac{F-u}{u}))}{J_{\frac{p}{2}-1}(h(\frac{F-u}{u}))}\le\frac{J_{\frac{p}{2}-2}(\frac{Au}{1+u}h(\frac{F-u}{u}))}{J_{\frac{p}{2}-2}(h(\frac{F-u}{u}))}+(1+c).\tag{3.34}$$

Using integration by parts, we have

$$J_a\left(\frac{Au}{1+u}h\left(\frac{F-u}{u}\right)\right)=\int_0^F u^a(1+u)^{-A}h\left(\frac{F-u}{u}\right)\frac{Au}{1+u}du$$

$$=-u^{a+1}h\left(\frac{F-u}{u}\right)(1+u)^{-A}|_0^F$$

$$+\int_0^F(1+u)^{-A}\left[(a+1)u^ah\left(\frac{F-u}{u}\right)+u^{a+1}h'\left(\frac{F-u}{u}\right)\left(\frac{-F}{u^2}\right)\right]du.$$

Hence, (3.34) is equivalent to

$$\frac{-F^{\frac{p}{2}}h(0)(1+F)^{-A}}{J_{\frac{p}{2}-1}(h(\frac{F-u}{u}))}+\left(\frac{p}{2}\right)$$

$$+\frac{\int_0^F u^{\frac{p}{2}-1}(1+u)^{-A}h(\frac{F-u}{u})\left[\frac{h'(\frac{F-u}{u})}{h(\frac{F-u}{u})}(\frac{-F}{u})\right]du}{\int_0^F u^{\frac{p}{2}-1}(1+u)^{-A}h(\frac{F-u}{u})\,du}$$

$$\le\frac{-F^{\frac{p}{2}-1}h(0)(1+F)^{-A}}{J_{\frac{p}{2}-2}(h(\frac{F-u}{u}))}+\left(\frac{p}{2}-1\right)$$

$$+\frac{\int_0^F u^{\frac{p}{2}-2}(1+u)^{-A}h(\frac{F-u}{u})\left[\frac{h'(\frac{F-u}{u})}{h(\frac{F-u}{u})}(\frac{-F}{u})\right]du}{\int_0^F u^{\frac{p}{2}-2}(1+u)^{-A}h(\frac{F-u}{u})\,du}+(1+c). \tag{3.35}$$

Since $-(v+1)h'(v)/h(v)=l_1(v)+l_2(v)$ (3.35) is equivalent to

$$\frac{-h(0)(1+F)^{-A}}{I_{\frac{p}{2}-1,A,h}(F)}+\frac{J_{\frac{p}{2}-1}(h(\frac{F-u}{u})l_1(\frac{F-u}{u}))}{J_{\frac{p}{2}-1}(h(\frac{F-u}{u}))}+\frac{J_{\frac{p}{2}-1}(h(\frac{F-u}{u})l_2(\frac{F-u}{u}))}{J_{\frac{p}{2}-1}(h(\frac{F-u}{u}))}$$
$$\le\frac{-h(0)(1+F)^{-A}}{I_{\frac{p}{2}-2,A,h}(F)}+\frac{J_{\frac{p}{2}-2}(h(\frac{F-u}{u})l_1(\frac{F-u}{u}))}{J_{\frac{p}{2}-2}(h(\frac{F-u}{u}))}+\frac{J_{\frac{p}{2}-2}(h(\frac{F-u}{u})l_2(\frac{F-u}{u}))}{J_{\frac{p}{2}-2}(h(\frac{F-u}{u}))}+c. \tag{3.36}$$

It is clear that $I_{\frac{p}{2}-1,A,h}(F)\le I_{\frac{p}{2}-2,A,h}(F)$, so we then have

$$\frac{-h(0)(1+F)^{-A}}{I_{\frac{p}{2}-1,A,h}(F)}\le\frac{-h(0)(1+F)^{-A}}{I_{\frac{p}{2}-2,A,h}(F)}$$

which accounts for the first terms on the left and right hand sides of (3.36). As for the second term on each side of (3.36) note that the hypothesis $l_1(\nu)$ is increasing in ν implies that for all fixed F, $l_1(\frac{F-u}{u})$ is decreasing in u. When $t<u$, we have

$$\frac{(1+u)^{-A}u^{\frac{p}{2}-2}h(\frac{F-u}{u})\,1\!\!1\{u\le F\}}{(1+t)^{-A}t^{\frac{p}{2}-2}h(\frac{F-t}{t})\,1\!\!1\{t\le F\}}\le\frac{(1+u)^{-A}u^{\frac{p}{2}-1}h(\frac{F-u}{u})\,1\!\!1\{u\le F\}}{(1+t)^{-A}t^{\frac{p}{2}-1}h(\frac{F-t}{t})\,1\!\!1\{t\le F\}}.$$

By a monotone likelihood ratio argument, we have

$$\frac{J_{\frac{p}{2}-1}(h(\frac{F-u}{u})l_1(\frac{F-u}{u}))}{J_{\frac{p}{2}-1}(h(\frac{F-u}{u}))}=\frac{\int_0^F u^{\frac{p}{2}-1}(1+u)^{-A}h(\frac{F-u}{u})l_1(\frac{F-u}{u})}{\int_0^F u^{\frac{p}{2}-1}(1+u)^{-A}h(\frac{F-u}{u})\,du}$$
$$\le\frac{\int_0^F u^{\frac{p}{2}-2}(1+u)^{-A}h(\frac{F-u}{u})l_1(\frac{F-u}{u})\,du}{\int_0^F u^{\frac{p}{2}-2}(1+u)^{-A}h(\frac{F-u}{u})\,du}=\frac{J_{\frac{p}{2}-2}(h(\frac{F-u}{u})l_1(\frac{F-u}{u}))}{J_{\frac{p}{2}-2}(h(\frac{F-u}{u}))}.$$

Finally, note that since $0\le l_2(v)\le c$ for the third term on each side of (3.36) we have

$$0\le\frac{J_{\frac{p}{2}-i}(l_2(\frac{F-u}{u})h(\frac{F-u}{u}))}{J_{\frac{p}{2}-i}(h(\frac{F-u}{u}))}\le c\quad\text{for } i=1,2.$$

Therefore we established the inequality (3.36) and the proof is complete. □

3.2.2.2 Minimaxity of the Generalized Bayes Estimators

In this subsection we apply Lemmas 3.4, 3.5, 3.6 and Theorems 3.6 and 3.7 to show minimaxity of the generalized Bayes estimator (3.25).

Theorem 3.8 *Assume that $g(\nu) = -(\nu + 1)\, h'(\nu)/h(\nu)$ is increasing in ν, $g(\nu) \leq M$, where M is a positive constant, and*

$$\frac{p-2+2M}{k+3+d-2M} \leq 2\,\frac{p-2}{k+2}\,.$$

Then $\delta(X, S)$ in (3.25) is minimax.

Proof Let $l_2(\nu) = 0$ and $l_1(\nu) = g(\nu)$. By applying Lemma 3.6 to the case $c = 0$, we have $r(F)$ increasing in F. Applying the bound in Lemma 3.5, we can get $0 \leq r(F) \leq 2\frac{p-2}{m+2}$. Therefore, by Lemma 3.4, $\delta(X, S)$ is minimax. □

It is interesting to make connections to the result in Faith (1978). Faith (1978) considered generalized Bayes estimator for $\mathcal{N}_p(\theta, I_p)$ and showed that when $g(\nu)$ is increasing in ν and $M \leq \frac{p-2}{2}$, the generalized Bayes estimator would be minimax. By taking $k \to \infty$, we deduce the same conditions as Faith (1978). The next lemma is a variant of Alam (1973) for the known variance case.

Theorem 3.9 *Define $c_k = \frac{p-2}{k+2}$. If there exists $b \in (0, 1]$ and $c = \frac{b(p-2)}{4+4(2-b)c_k}$, such that $0 \leq r(F) \leq (2-b)c_k$, and $F^c r(F)$ is increasing in F, then the generalized Bayes estimator $\delta(X, S)$ in (3.25) is minimax.*

Proof By taking the derivative of the Efron and Morris' condition, (3.31) can be satisfied by requiring

$$0 \leq 2\left(\frac{p}{2} - 1\right) R(F)\left(2 - \frac{r(F)}{c_m}\right) + 4r'(F)(1 + r(F)). \tag{3.37}$$

Since $r(F) \leq (2-b)c_k$, then (3.37) is satisfied at the point where $r'(F) \geq 0$. Since $r(F) \leq (2-b)c_k$ with $\beta = (2-b)c_k$

$$4r'(F)(1+\beta) \leq 4r'(F)(1 + r(F)), \tag{3.38}$$

at the point where $r'(F) < 0$. We now have

$$\begin{aligned}
0 &\leq (4 + 4\beta)(cR(F) + R(F) + FR'(F)) \\
&= 2b\left(\frac{p}{2} - 1\right) R(F) + 4r'(F)(1+\beta) \\
&\leq 2\left(\frac{p}{2} - 1\right) R(F)\left(2 - \frac{r(F)}{c_k}\right) + 4r'(F)(1 + r(F))
\end{aligned}$$

since $F^c r(F)$ is increasing in F. Thus, for all values of F, we have proven (3.37), and combining with the bound on the $r(F)$ function, we have proven 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 $r(F)$. 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 3.4, 3.5, 3.6, and 3.9 gives the following theorem.

Theorem 3.10 *If there exists $b \in (0,1]$ such that $g(\nu) = l_1(\nu) + l_2(\nu) \le M$, and $l_1(\nu)$ is increasing in ν, $0 \le l_2(\nu) \le c = \frac{b(p-2)}{4+4(2-b)\frac{p-2}{k+2}}$, and $\frac{p-2+2M}{k+3+d-2M} \le \frac{(2-b)(p-2)}{k+2}$, then the generalized Bayes estimator $\delta(X, S)$ in (3.25) is minimax.*

3.2.2.3 Examples of the Priors in (3.24)

In this subsection, we will give several examples to which our results can be applied and make some connection to the existing literature found in Maruyama and Strawderman (2005) and Fourdrinier et al. (1998).

Example 3.7 Maruyama and Strawderman (2005) considered the priors with $h(\nu) \propto \nu^b(1+\nu)^{-a-b-2}$ for $b > 0$ and show that $r(F) \le \frac{\frac{p}{2}+a+1}{\frac{k}{2}+\frac{d}{2}-a-\frac{1}{2}}$ (in terms of the Maruyama and Strawderman (2005) notation $d = 2e + 1$). Condition C1 is equivalent to the condition that $d + k + p > -1$. C2 and C3 are equivalent here, and both are equivalent to the condition that $a + \frac{p}{2} + 1 > 0$. Then, using Theorem 3.8, we have $g(\nu) = a + 2 - b\nu^{-1}$. The condition that $g(\nu)$ is increasing in ν is equivalent to the condition that $b \ge 0$. Clearly, we can let $M = a + 2$. Then the condition of Theorem 3.8 is that

$$\frac{k}{2} + \frac{d}{2} - \frac{1}{2} > a \quad \text{and} \quad \frac{\frac{p}{2} + a + 1}{\frac{k}{2} + \frac{d}{2} - a - \frac{1}{2}} \le 2c_k.$$

A close examination of the Maruyama and Strawderman (2005) proof shows that their upper bound on $r(F)$ is sharp. This implies that our bound in Lemma 3.5 cannot be relaxed.

Example 3.8 Generalized Student-t priors correspond to a mixing distribution of the form

$$h(\nu) = c(\nu+1)^{\beta-\alpha-\gamma-\frac{p-2}{2}}\nu^{\gamma-\beta}e^{\frac{\gamma}{\nu}}.$$

Consider the following two cases. 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 both cases,

$$\ln h(\nu) = (\beta - \alpha - \gamma - \frac{p-2}{2}) \ln(1+\nu) + (\gamma - \beta) \ln \nu + \frac{\gamma}{\nu}$$

and

$$g(\nu) = \left(\frac{p-2}{2} + \alpha + \gamma - \beta\right) + \frac{(1+\nu)(\beta-\gamma)}{\nu} + \frac{\gamma(1+\nu)}{\nu^2} = \frac{p-2}{2} + \alpha + \frac{\beta}{\nu} + \frac{\gamma}{\nu^2}.$$

Clearly, $g(\nu)$ is monotonic in the first case, and minimaxity of the generalized Bayes estimator follows when

$$0 \le \frac{p-2+\alpha}{\frac{k}{2} + \frac{1}{2} + \frac{d}{2} - \frac{p}{2} - \alpha} \le \frac{p-2}{\frac{k}{2}+1}$$

in addition to the 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, the conditions reduce to those in Example 3.4 for the case of known variance.

Next we consider spherical multivariate Student-t priors with f degrees of freedom and a scale parameter τ and with $\alpha = \frac{f-p+4}{2}$, $\beta = \frac{f(1-\tau)+2}{2}$, and $\gamma = -\frac{f\tau}{2}$. The case of $\tau = 1$ is of particular interest but does not necessarily give a monotonic $r(\cdot)$ function. However, we can use the result in Theorem 3.10 to show that the generalized Bayes estimator is minimax under the following conditions for $f \le p-4$, suppose there exists a constant $b \in (0, 1]$ such that

$$\frac{p + f + \frac{1}{f}}{k + 1 + d - f - \frac{1}{f}} \le (2-b)\frac{p-2}{k+2},$$
$$\frac{1}{2f} \le c = \frac{b(p-2)}{4 + 4(2-b)\frac{p-2}{k+2}}. \tag{3.39}$$

Condition (3.39) can be established by observing that for this case,

$$g(\nu) = \frac{p-2}{2} + \alpha + \frac{\beta}{\nu} + \frac{\gamma}{\nu^2} = \frac{f}{2} + 1 + \frac{1}{\nu} - \frac{f}{2\nu^2}$$

is clearly nonmonotonic. We then let $M = \frac{f}{2} + 1 + \frac{1}{2f}$ and apply Lemma 3.5 to get the upper bound on $r(\cdot)$. We define $l_1(\nu) = g(\nu) - \frac{1}{2f}$ when $\nu \le f$ and $l_1(\nu) = \frac{f}{2} + 1$ otherwise. We also define $l_2(\nu) = \frac{1}{2f}$ when $\nu \le f$ and $l_2(\nu) = \frac{1}{\nu} - \frac{f}{2\nu^2}$ otherwise. By applying Lemma 3.6, we get condition (3.39).

The spherical multivariate Cauchy prior corresponds to the case $f = 1$. If $k = O(p)$ and $d = 3$, then condition (3.39) reduces to $p \ge 5$, $\frac{p+2}{k+2} \le (2-b)\frac{p-2}{k+2}$, and $\frac{1}{2} \le \frac{b(p-2)}{4+8-4b}$.

3.3 Results for Known Σ and General Quadratic Loss

3.3.1 Results for the Diagonal Case

Much of this section is based on the review in Strawderman (2003). We begin with a discussion of the multivariate normal case where $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$ is diagonal, which we assume throughout this subsection. Let

$$X \sim \mathcal{N}_p(\theta, \Sigma) \tag{3.40}$$

and the loss be equal to a weighted sum of squared errors loss

$$L(\theta, \delta) = (\delta - \theta)^{\mathrm{T}} D(\delta - \theta) = \sum_{i=1}^{p} (\delta_i - \theta_i)^2 d_i \,. \tag{3.41}$$

The results in Sects. 2.3, 2.4 and 3.1 extend by the use of Stein's lemma in a straightforward way to give the following basic theorem.

Theorem 3.11 *Let X have the distribution* (3.40) *and let the loss be given by* (3.41).

(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*

$$\begin{aligned} R(\delta, \boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}((\delta - \boldsymbol{\theta})^{\mathrm{T}} D(\delta - \boldsymbol{\theta})) \\ &= tr(\Sigma D) + E_\theta \left[\sum_{i=1}^{p} \sigma_i^4 d_i \left(g_i^2(X) + 2 \frac{\partial g_i(X)}{\partial X_i} \right) \right]. \end{aligned}$$

(2) *If $\boldsymbol{\theta} \sim \pi(\theta)$, 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 $\boldsymbol{\theta} \sim \pi(\theta)$, 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*

$$\begin{aligned} R(\delta_m, \theta) &= \text{tr}(\Sigma D) \\ &\quad + E_\theta \left[\frac{2m(X) \sum_{i=1}^{p} \sigma_i^4 d_i \partial m^2(X)/\partial^2 X_i}{m^2(X)} - \frac{\sum_{i=1}^{p} \sigma_i^4 d_i \left(\partial m(X)/\partial X_i\right)^2}{m^2(X)} \right] \\ &= \text{tr}(\Sigma D) + 4\, E_\theta \left[\frac{\sum_{i=1}^{p} \sigma_i^4 d_i \partial^2 \sqrt{m(X)}/\partial^2 X_i}{\sqrt{m(X)}} \right]. \end{aligned}$$

(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 nonpositive, the proper (generalized, pseudo) Bayes* $\delta_m(X)$ *is minimax.*

The proof follows closely to that of corresponding results in Sects. 2.3, 2.4 and 3.1. The result is essentially 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 straightforward result from Strawderman (2003).

Lemma 3.7 *Suppose* $\eta(X)$ *is such that* $\Delta\eta(X) = \sum_{i=1}^{p} \partial^2\eta(X)/\partial^2 X_i^2 \le 0$ *(i.e.* $\eta(X)$ *is superharmonic). Then* $\eta^*(X) = \eta(\Sigma^{-1}D^{-1/2}X)$ *is such that* $\sum_{i=1}^{p} \sigma_i^4 d_i \partial^2 \eta^*(X)/\partial^2 X_i \le 0$.

Note, that for any scalar a, if $\eta(X)$ is superharmonic, then so is $\eta(aX)$. This leads to the following result.

Theorem 3.12 *Suppose* X *has the distribution* (3.40) *and the loss is given by* (3.41).

(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*

$$\delta_m(X) = X + \frac{2m'(X^{\mathrm{T}}\Sigma^{-1}D^{-1}\Sigma^{-1}X)D^{-1}\Sigma^{-1}X}{m(X^{\mathrm{T}}\Sigma^{-1}D^{-1}\Sigma^{-1}X)}$$

is minimax.

(3) *Suppose the prior distribution* $\pi(\theta)$ *has the hierarchical structure* $\theta|\lambda \sim \mathcal{N}_p(0, A_\lambda)$ *for* $\lambda \sim h(\lambda)$, $0 < \lambda < 1$, *where* $A_\lambda = (c/\lambda)\Sigma D\Sigma - \Sigma$, *c is such that* A_1 *is positive definite, and* $h(\lambda)$ *satisfies the conditions of Theorem* 3.12. *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 Part (1) follows directly from Parts (3) and (4) of Theorem 3.11 and Lemma 3.7. Part (2) follows from Part (1) and Part (2) of Theorem 3.11 with a straightforward calculation.

For Part (3), first note that $\theta|\lambda \sim \mathcal{N}_p(0, A_\lambda)$ and $X - \theta|\lambda \sim \mathcal{N}_p(0, \Sigma)$. Thus, $X - \theta$ and θ are conditionally independent given λ. Hence we have $X|\lambda \sim \mathcal{N}_p(0, A_\lambda + \Sigma)$. It follows that

$$m(X) \propto \int_0^1 \lambda^{p/2} \exp\left[-\frac{\lambda}{c} \left(X^{\mathrm{T}} \Sigma^{-1} D^{-1} \Sigma^{-1} X \right) \right] h(\lambda)\, d\lambda$$

but $m(X) = \eta\left(X^{\mathrm{T}} \Sigma^{-1} D^{-1} \Sigma^{-1} X/c\right)$, where $\sqrt{\eta\,(X^{\mathrm{T}} X)}$ is superharmonic by Theorem 3.11. Hence, by Part (2), $\delta_\pi(X)$ is minimax (and proper or generalized Bayes depending on whether $h(\lambda)$ is integrable or not).

Since superharmonicity of $\eta(X)$ implies the superharmonicity of $\sqrt{\eta\,(X)}$, Part (4) follows from Part (1) and the superharmonicity of mixtures of superharmonic functions. □

Example 3.9 (Pseudo-Bayes minimax estimators) When $\Sigma = D = \sigma^2 I$, we saw in Sect. 3.3 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 3.12 that $m(X^{\mathrm{T}} \Sigma^{-1} D^{-1} \Sigma^{-1} X) = (1/X^{\mathrm{T}} \Sigma^{-1} D^{-1} \Sigma^{-1} X)^b$ has associated with it the pseudo-Bayes estimator $\delta_m(X) = (1 - \frac{2bD^{-1}\Sigma^{-1}}{(X^{\mathrm{T}} \Sigma^{-1} D^{-1} \Sigma^{-1} X)})X$. This estimator is minimax for $0 < b \leq 2(p-2)$.

Example 3.10 (Hierarchical proper Bayes minimax estimator) As suggested by Berger (1976) suppose the prior distribution has the hierarchical structure $\theta|\lambda \sim \mathcal{N}_p(0, A_\lambda)$ where $A_\lambda = c\Sigma D \Sigma - \Sigma$, $c > 1/\min(\sigma_i^2 d_i)$ and $h(\lambda) = (1+b)\lambda^b$ for $0 < \lambda < 1$ and $-1 < b \leq \frac{(p-6)}{2}$. The resulting proper Bayes estimator will be minimax for $p \geq 5$ by part (3) of Theorem 3.12 and Example 3.9. For $p \geq 3$, the estimator $\delta_\pi(X)$ given in part (3) of Theorem 3.12 is a generalized Bayes minimax estimator 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). Also see the development in Berger and Strawderman (1996) and Kubokawa and Strawderman (2007).

Example 3.11 (Multiple shrinkage minimax estimators) It follows from Example 3.9 and Theorem 3.12 that $m(X) = \sum_{i=1}^{k} \left[\frac{1}{(X-\nu_i)^{\mathrm{T}}\Sigma^{-1}D^{-1}\Sigma^{-1}(X-\nu_i)}\right]^b$ satisfies the conditions of Theorem 3.12 (4) for $0 < b \leq (p-2)/2$. and hence

$$\delta_m(X) = X - \frac{2b\sum_{i=1}^{k}\left[D^{-1}\Sigma^{-1}(X-\nu_i)\right]\Big/\left[(X-\nu_i)^{\mathrm{T}}\Sigma^{-1}D^{-1}\Sigma^{-1}(X-\nu_i)\right]^{b+1}}{\sum_{i=1}^{k} 1\Big/\left[(X-\nu_i)^{\mathrm{T}}\Sigma^{-1}D^{-1}\Sigma^{-1}(X-\nu_i)\right]^{b}} \tag{3.42}$$

is a minimax multiple shrinkage (pseudo-Bayes) estimator.

If, as in Example 3.11 we used the generalized prior

$$\pi(\theta) = \sum_{i=1}^{k}\left[\frac{1}{(\theta-\nu_i)^{\mathrm{T}}\Sigma^{-1}D^{-1}\Sigma^{-1}(\theta-\nu_i)}\right]^b,$$

the resulting generalized Bayes (as opposed to pseudo-Bayes) estimators is minimax for $0 < b \leq (p-2)/2$.

3.3.2 General Σ and General Quadratic Loss

In this section, we generalize the above results to the case of

$$X \sim \mathcal{N}_p(\theta, \Sigma), \tag{3.43}$$

where Σ is a general positive definite covariance matrix and the loss is given by

$$L(\theta, \delta) = (\delta - \theta)^{\mathrm{T}} Q (\delta - \theta), \tag{3.44}$$

where Q is a general positive definite matrix. We will see that this case can be reduced to the canonical form $\Sigma = I$ and $Q = \mathrm{diag}(d_1, d_2, \ldots, d_p) = D$. We continue to follow the development in Strawderman (2003).

The following well known fact will be used repeatedly to obtain the desired generalization.

Lemma 3.8 *For any pair of positive definite matrices, Σ and Q, there exits a non-singular matrix A such that $A\Sigma A^{\mathrm{T}} = I$ and $(A^{\mathrm{T}})^{-1}QA^{-1} = D$ where D is diagonal.*

Using this fact we can now present the canonical form of the estimation problem.

Theorem 3.13 *Let $X \sim \mathcal{N}_p(\theta, \Sigma)$ and suppose that the loss is $L_1(\delta, \theta) = (\delta - \theta)^{\mathrm{T}} Q(\delta - \theta)$. Let A and D be as in Lemma 3.8 and let $Y = AX \sim \mathcal{N}_p(\upsilon, I_p)$, where $\upsilon = A\theta$ and $L_2(\delta, \upsilon) = (\delta - \upsilon)^{\mathrm{T}} D(\delta - \upsilon)$.*

(1) *If $\delta_1(X)$ is an estimator with risk function $R_1(\delta_1, \boldsymbol{\theta}) = E_\theta L_1(\delta_1(X), \theta)$, then the estimator $\delta_2(Y) = A\delta_1(A^{-1}Y)$ has risk function $R_2(\delta_2, \upsilon) = R_1(\delta_1, \theta) = E_\theta L_2(\delta_2(Y), \upsilon)$.*

(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 $\pi_2(\upsilon) = \pi_1(A^{-1}\upsilon)$ (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 To establish Part (1) note that the risk function

$$\begin{aligned} R_2(\delta_2, \upsilon) &= E_\theta L_2[\delta_2(Y), \upsilon] \\ &= E_\theta[(\delta_2(Y) - \upsilon)^{\mathrm{T}} D(\delta_2(Y) - \upsilon)] \\ &= E_\theta[(A\delta_1(A^{-1}(AX)) - A\theta)^{\mathrm{T}} D(A\delta_1(A^{-1}(AX)) - A\theta)] \\ &= E_\theta[(\delta_1((X) - \theta)^{\mathrm{T}} A^{\mathrm{T}} DA(\delta_1(X) - \theta)] \\ &= E_\theta[(\delta_1((X) - \theta)^{\mathrm{T}} Q(\delta_1(X) - \theta)] \\ &= R_1(\delta_1, \theta). \end{aligned}$$

Since the Bayes estimator for any quadratic loss is the posterior mean and $\theta \sim \pi_1(\theta)$ and $\upsilon = A\theta \sim \pi_2(\upsilon) = \pi_1(A^{-1}\upsilon)$ (ignoring constants), then Part (2) follows by noting that

$$\delta_2(Y) = E[\upsilon|Y] = E[A\theta|Y] = E[A\theta|AX] = A\ E[\theta|X] = A\ \delta_1(X) = A\delta_1(A^{-1}Y).$$

Lastly, Part (3) follows directly from Part (1). □

Note that 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 $D = P\Sigma^{1/2} Q \Sigma^{1/2} P^{\mathrm{T}}$ satisfy the requirements of the theorem.

Example 3.12 Proceeding as we did in Example 3.9 and applying Theorem 3.13, $m(X^{\mathrm{T}} \Sigma^{-1} Q^{-1} \Sigma^{-1} X) = (X^{\mathrm{T}} \Sigma^{-1} Q^{-1} \Sigma^{-1} X)^{-b}$ has associated with it, the pseudo-Bayes minimax James-Stein estimators is

$$\delta_m(X) = \left(1 - \frac{2\,b\,Q^{-1}\Sigma^{-1}}{\left(X^{\mathrm{T}}\,\Sigma^{-1} Q^{-1} \Sigma^{-1} X\right)}\right) X,$$

for $0 < b \leq 2\,(p - 2)$.

Generalizations of Example 3.10 to hierarchical Bayes minimax estimators and generalizations of Example 3.11 to multiple shrinkage estimators are straightforward. We omit the details.

3.4 Admissibility of Bayes Estimators

Recall from Sect. 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 also 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; 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.

In this section, we will study the admissibility of estimators corresponding to priors which are variance mixtures of normal distributions for the case of $X \sim \mathcal{N}_p(\theta, I)$ and quadratic loss $\|\delta - \theta\|^2$ as in Sect. 3.1.2. In particular, we consider prior densities of the form (3.4) and establish a connection between admissibility and the behavior of the mixing (generalized) density $h(v)$ at infinity. The analysis will be based on Brown (1971), Theorem 1.2. An Abelian Theorem (see, e.g., Widder (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 $\lim_{x \to a} f(x)/g(x) = 1$. Here is an adaptation of the Abelian theorem in Widder that meets our needs.

Theorem 3.14 *Assume $g : \mathbb{R}^+ \to \mathbb{R}$ has a Laplace transform $f(s) = \int_0^\infty g(t) e^{-st}\, dt$ that 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 (1946) 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 3.9 *Suppose* $X \sim \mathcal{N}_p(\theta, I_p)$*,* $L(\theta, \delta) = \|\delta - \theta\|^2$ *and* $\pi(\theta)$ *is given by* (3.4) *where* $h(v) \sim K\, v^a$ *as* $v \to \infty$ *with* $a < (p-2)/2$ *and where* $v^{-p/2}\, h(v)$ *is integrable in a neighborhood of* 0*. Then*

(1) $\pi(\theta) \sim K\, (\|\theta\|^2)^{a-(p-2)/2}\, \Gamma((p-2)/2 - a)$ *as* $\|\theta\|^2 \to \infty$,
$m(x) \sim K(\|x\|^2)^{a-(p-2)/2}\, \Gamma((p-2)/2 - a)$ *as* $\|x\|^2 \to \infty$,
and therefore $\pi(\|x\|^2) \sim m(\|x\|^2)$ *as* $\|x\|^2 \to \infty$,

(2) $\|\delta(x) - x\|$ *is uniformly bounded, where* δ *is the generalized Bayes estimator corresponding to* π.

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 $g(t) = t^{\frac{p}{2}-2} h(1/t) \sim K t^{\frac{p-4}{2}-a}$ as $t \to 0_+$. Therefore, by Theorem 3.14, $\pi(\theta) \sim K(\|\theta\|^2)^{a-\frac{p-2}{2}}\, \Gamma\left(\frac{p-2}{2} - a\right)$ as $\|\theta\|^2 \to \infty$. Similarly

$$\begin{aligned} m(x) &= \int_0^\infty e^{-\frac{\|\theta\|^2}{2(1+v)}} (1+v)^{-\frac{p}{2}}\, h(v)\, dv \quad \left(\text{for } t = \frac{1}{1+v}\right) \\ &= \int_1^\infty e^{-\frac{\|\theta\|^2}{2} t} t^{\frac{p}{2}-2} h\left(\frac{1-t}{t}\right) dt. \end{aligned}$$

We note that 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 3.14,

$$m(x) \sim K(\|x\|^2)^{a-\frac{p-2}{2}}\, \Gamma\left(\frac{p-2}{2} - a\right) \text{ as } \|x\|^2 \to \infty,$$

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

$$\|\delta(x)-x\|^2 \sim \left[\frac{(\|x\|^2)^{a-\frac{p}{2}}\ \Gamma(\frac{p}{2}-a)}{\|x\|^2)^{a-\frac{p-2}{2}}\ \Gamma(\frac{p-2}{2}-a)}\right]^2 \|x\|^2$$

$$\sim \left(\frac{p-2}{2}-a\right)^2 \frac{1}{\|x\|^2} \text{ as } \|x\|^2 \to \infty.$$

Since $\delta(x)-x$ is in $\mathscr{C}^\infty$ 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 3.15 *For priors $\pi(\theta)$ of the form (3.4) with mixing density $h(v) \sim v^a$ as $v \to \infty$, the corresponding generalized Bayes estimator δ is admissible if and only if $a \le 0$.*

Proof (Admissibility if $a \le 0$) By Lemma 3.9, we have $\bar{m}(r) = m^*(r^2) \sim K^*(r^2)^{a-(p-2)/2}$, with $m(x) = m^*(\|x\|^2)$. Thus, for any $\epsilon > 0$, there is an $r_0 > 0$ such that, for $r > r_0$, $\bar{m}(r) \le (1+\epsilon)K^* r^{2a-(p-2)}$. Since $\|\delta(x)-x\|$ is uniformly bounded,

$$\int_{r_0}^{\infty} (r^{p-1}\bar{m}(r))^{-1}\, dr \ge (K^*(1+\epsilon))^{-1} \int_{r_0}^{\infty} r^{-(2a+1)}\, dr = \infty$$

if $a \ge 0$. Hence, $\delta(x)$ is admissible if $a \le 0$, by Theorem 1.2.

(Inadmissibility if $a > 0$) Similarly, we have, for $r \ge r_0$,

$$\underline{m}(r) = \frac{1}{m^*(r^2)} \sim \frac{1}{K^*}(r^2)^{\frac{p-2}{2}-a},$$

$$\underline{m}(r) \le \frac{1}{(1-\epsilon)K^*} r^{p-2-2a},$$

and

$$\int_0^{\infty} r^{1-p}\underline{m}(r)\, dr \le \frac{1}{K^*}\int_{r_0}^{\infty} r^{-(1+2a)}\, dr < \infty$$

if $a > 0$. Thus $\delta(x)$ is inadmissible if $a > 0$. □

Example 3.13 (Continued) Recall for the Strawderman prior that $h(v) = C(1+v)^{-\alpha-(\frac{p-2}{2})} \sim v^a$ as $v \to \infty$ for $a = -(\alpha + \frac{p-2}{2})$.

The above theorem implies that the generalized Bayes estimator is admissible if and only if $\alpha + \frac{p-2}{2} \ge 0$ or $1 - \frac{p}{2} \le \alpha$. We previously established minimaxity when $2-p < \alpha \le 0$ for $p \ge 3$ and propriety of the prior when $2-\frac{p}{2} < \alpha \le 0$ for $p \ge 5$.

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 3.15. Hence the bound for admissibility, $a \leq 0$, differs from the bound for propriety, $a < -1$, by 1.

3.5 Connections to Maximum a Posteriori Estimation

3.5.1 Hierarchical Priors

As we have seen in previous sections of this chapter, the classical Stein estimate and its positive-part modification can be motivated in a number of ways, perhaps most commonly as empirical Bayes estimates (i.e., posterior means) under a normal hierarchical model in which $\theta \sim \mathcal{N}_p(0, \psi I_p)$ where ψ, viewed as a hyperparameter, is estimated. In this section we look at shrinkage estimation through the lens of maximum a posteriori (MAP) estimation. The development of this section follows Strawderman and Wells (2012).

The class of proper Bayes minimax estimators constructed in Sect. 3.1 relies on the use of a hierarchically specified class of proper prior distributions $\pi_S(\theta, \kappa)$. In particular, for the prior in Strawderman (1971), $\pi_S(\theta, \kappa)$ is specified according to

$$\theta|\kappa \sim \mathcal{N}_p(0, g(\kappa)I_p), \;\; \pi_S(\kappa) = \kappa^{-a}(1-a)^{-1}\, 1\!\!1_{[0<\kappa<1]}, \tag{3.45}$$

where $g(\kappa) = (1-\kappa)/\kappa$ and the constant a satisfies $0 \leq a < 1$, i.e., $\pi_S(\kappa)$ is a Beta$(1-a, 1)$ probability distribution. Suppose $a = 1/2$; then, utilizing the transformation $\psi = g(\kappa) > 0$ in (3.45), we obtain the equivalent specification

$$\theta|\psi \sim \mathcal{N}_p(0, \psi I_p), \;\; \pi_S(\psi) \propto \left(\frac{1}{1+\psi}\right)^{\frac{3}{2}} 1\!\!1_{[\psi>0]}. \tag{3.46}$$

Two interesting alternative formulations of (3.46) are given below for the case $p = 1$ and generalized later for arbitrary p. In what follows, we let Gamma(τ, ξ) denote a random variable with probability density function

$$g(x|\tau, \xi) = \frac{\xi^\tau}{\Gamma(\tau)} x^{\tau-1} e^{-x\xi}\, 1\!\!1_{[x>0]} \quad \text{for} \;\; \tau > 0 \quad \text{and} \quad \xi > 0$$

and Exp(ξ) corresponds to the choice $\tau = 1$ (i.e., an exponential random variable in its rate parametrization).

For $p = 1$, the marginal prior distribution on θ induced by (3.46) is equivalent to that obtained under the specification

$$\theta|\psi, \lambda \sim \mathscr{N}(0, \psi), \quad \psi|\lambda \sim \text{Exp}\left(\frac{\lambda^2}{2}\right), \quad \lambda|\alpha \sim \text{HN}(\alpha^{-1}), \tag{3.47}$$

where $\alpha = 1$ and $\text{HN}(\zeta)$ denotes the half-normal density

$$f(x|\zeta) = \sqrt{\frac{2}{\pi\,\zeta}} \exp\left\{-\frac{x^2}{2\,\zeta}\right\} 1\!\!1_{[x>0]} \quad \text{for} \quad \zeta > 0.$$

The marginal prior distribution on θ induced by (3.46) is also equivalent to that obtained under the alternative specification

$$\theta|\lambda \sim \text{Laplace}(\lambda), \quad \lambda|\alpha \sim \text{HN}(\alpha^{-1}), \tag{3.48}$$

where $\alpha = 1$ and $\text{Laplace}(\lambda)$ denotes a random variable with the Laplace (double exponential) probability density function

$$f(y|\lambda) = \frac{\lambda}{2} e^{-\lambda|y|}\, 1\!\!1_{[y\in\mathbb{R}]}.$$

This result follows from Griffin and Brown (2010). Define

$$\theta|\psi, \omega \sim \mathscr{N}(0, \psi), \quad \psi|\omega \sim \text{Exp}(\omega), \quad \omega|\delta, \alpha \sim \text{Gamma}(1/2, \alpha) \tag{3.49}$$

as a hierarchically specified prior distribution for θ, ψ and ω. The resulting marginal prior distribution for θ, obtained by integrating out ψ and ω, is exactly the quasi-Cauchy distribution of Johnstone and Silverman (2004); see Griffin and Brown (2010) for details. Carvalho et al. (2010) showed that this distribution also coincides with the marginal prior distribution for θ induced by taking $a = 1/2$ in (3.45). The transformation $\lambda = \sqrt{2\omega}$ in (3.49) leads directly to (3.47) upon setting $\alpha = 1$; (3.48) is then obtained by integrating out ψ in (3.47).

3.5.2 *The Positive-Part Estimator and Extensions as MAP Estimators*

Takada (1979) showed that a positive-part type minimax estimator

$$\delta^c_{JS+}(X) = \left(1 - \frac{c}{\|X\|_2^2}\right)_+ X, \tag{3.50}$$

where $(t)_+ = \max(t, 0)$, is also the MAP estimator under a certain class of hierarchically specified generalized prior distributions, say $\pi_T(\theta, \kappa) = \pi(\theta|\kappa)\pi_T(\kappa)$. For the specific choice $c = p - 2$ in (3.50), Takada's prior reduces to

$$\theta|\kappa \sim \mathcal{N}_p(0, g(\kappa)I_p), \quad \pi_T(\kappa) \propto (1-\kappa)^{p/2}\kappa^{-1}\, 1\!\!1_{[0<\kappa<1]}. \tag{3.51}$$

The improper prior (3.51) evidently behaves similarly to Strawderman's proper prior (3.45) (i.e., for $a = 1/2$). Notably, the numerator $(1-\kappa)^{p/2}$ in $\pi_T(\kappa)$ explicitly offsets the contribution of $(1-\kappa)^{-p/2}$ arising from the determinant of the variance matrix $g(\kappa)\, I_p$ in the conditional prior specification $\theta|\kappa$. Under the monotone decreasing variable transformation $\psi = g(\kappa) > 0$, (3.51) implies an alternative representation that is analogous to (3.46):

$$\theta|\psi \sim \mathcal{N}_p(0, \psi I_p), \quad \pi_T(\psi) \propto \psi^{p/2}\left(\frac{1}{1+\psi}\right)^{p/2+1} 1\!\!1_{[\psi>0]}. \tag{3.52}$$

We observe that the proper prior (3.46) and improper prior (3.52) (almost) coincide when $p = 1$; in particular, multiplying the former by $\psi^{1/2}$ yields the latter. In view of the fact that (3.46) and (3.47) lead to the same marginal prior on $\boldsymbol{\theta}$ when $p = 1$, one is led to question whether a deeper connection between these two prior specifications might exist. Supposing $p \geq 1$, consider the following straightforward generalization of (3.47):

$$\theta|\psi, \lambda \sim \mathcal{N}_p(0, \psi I_p), \;\; \psi|\lambda \sim \text{Gamma}\left(\frac{p+1}{2}, \frac{\lambda^2}{2}\right), \;\; \lambda|\alpha \sim \text{HN}(\alpha^{-1}). \tag{3.53}$$

Integrating λ out of the higher level prior specification the resulting marginal (proper) prior for ψ reduces to

$$\pi(\psi|\alpha) \propto \psi^{-1/2}\psi^{p/2}\left(\frac{1}{1+\frac{\psi}{\alpha}}\right)^{\frac{p}{2}+1} 1\!\!1_{[\psi>0]}. \tag{3.54}$$

For $\alpha = 1$ and any $p \geq 1$, we now observe that the proper prior (3.54) is simply the improper prior $\pi_T(\psi)$ in (3.52) multiplied by $\psi^{-1/2}$ and it reduces to Strawderman's prior (3.46) for $p = 1$.

3.5.3 Penalized Likelihood and Hierarchical Priors

Expressed in modern terms of penalization, Takada (1979) proved that the positive-part estimator (3.50) is the solution to a certain penalized likelihood estimation problem in which the penalty (or regularization) term is determined by the prior (3.51). Penalized likelihood estimation, and more generally problems of regularized estimation, have become a very important conceptual paradigm in both statistics and machine learning. Such methods suggest principled estimation and model selection procedures for a variety of high-dimensional problems. The statistical literature on penalized likelihood estimators has exploded, in part due to success in constructing procedures for regression problems in which one can simultaneously select variables and estimate their effects. The class of penalty functions leading to procedures with good asymptotic frequentist properties have singularities at the origin; important examples of separable penalties include the least absolute shrinkage and selection operator (LASSO) , Tibshirani (1996), smoothly clipped absolute deviation (SCAD), Fan and Li (2001), and minimax concave penalties (MCP) Zhang (2010). In fact, most such penalties utilized in the literature behave similarly to the LASSO penalty near the origin, differing more in their respective behaviors away from the origin, where control of estimation bias for those parameters not estimated to be zero becomes the driving concern. Generalizations of the LASSO penalty have been proposed to deal with correlated groupings of parameters, such as those that might arise in problems with features that can be sensibly ordered, as in the fused LASSO in Tibshirani et al. (2005), or separated into distinct subgroups as in the group LASSO in Yuan and Lin (2006). In such problems, the use of these penalties serves a related purpose.

The LASSO was initially formulated as a least squares estimation problem subject to a ℓ_1 constraint on the parameter vector. The more well-known penalized likelihood version arises from a Lagrange multiplier formulation of a convex relaxation of a ℓ_0 non-convex optimization problem. Since the underlying objective function is separable in the parameters, the underlying estimation problem is evidently directly related to the now-classical problem of estimating a bounded normal mean. From a decision theoretic point of view, if $X \sim \mathcal{N}(\theta, 1)$ for $|\theta| \leq \lambda$, then the projection of the usual estimator dominates the unrestricted MLE, but cannot be minimax for quadratic loss because it is not a Bayes estimator. Casella and Strawderman (1981) showed that the unique minimax estimator of θ is the Bayes estimator corresponding to a two-point prior on $\{-\lambda, \lambda\}$ for λ sufficiently small. Casella and Strawderman (1981) further showed that the uniform boundary Bayes estimator, $\lambda \tanh(\lambda x)$, is the unique minimax estimator if $\lambda < \lambda_0 \approx 1.0567$. They also considered three-point priors supported on $\{-\lambda, 0, \lambda\}$ and obtained sufficient conditions for such a prior to be least favorable. Marchand and Perron (2001) considered the multivariate extension, $X \sim \mathcal{N}_p(\theta, I_p)$ with $\|\theta\|_2 \leq \lambda$ and showed that the Bayes estimator with respect to a boundary uniform prior dominates the MLE whenever $\lambda \leq \sqrt{p}$ under squared error loss.

It has long been recognized that the class of penalized likelihood estimators also has a Bayesian interpretation. For example, in the canonical version of the LASSO problem, minimizing

$$\frac{1}{2}\|X - \theta\|_2^2 + \lambda\|\theta\|_1, \quad ||\theta||_1 = \sum_{i=1}^{p} |\theta_i| \tag{3.55}$$

with respect to θ is easily seen to be equivalent to computing the MAP estimator of θ under a model specification in which $X \sim \mathcal{N}_p(\theta, I_p)$ and θ has a prior distribution satisfying $\theta_i \overset{iid}{\sim}$ Laplace(λ). It is easily shown that the solution to (3.55) is $\widehat{\theta_i}(X) = \text{sign}(X_i)(|X_i| - \lambda)_+$, $i = 1, \ldots, p$. The critical hyperparameter λ, though regarded as fixed for the purposes of estimating θ, is typically estimated in some ad hoc manner (e.g., cross validation), resulting in an estimator with an empirical Bayes flavor.

The Laplace prior inherent in the LASSO minimization problem (3.55) has broad connections to estimation under hierarchical prior specifications that lead to scale mixtures of normal distributions. As pointed out above, the conditional prior distribution of $\theta|\lambda$ obtained by integrating out ψ in (3.47) is exactly Laplace(λ). More generally, the conditional distribution for $\theta|\lambda$ under the hierarchical prior specification (3.53) is a special case of the class of multivariate exponential power distributions in Gomez-Sanchez-Manzano et al. (2008); in particular, we obtain

$$\pi(\theta|\lambda) \propto \lambda^p \exp\{-\lambda\|\theta\|_2\}, \tag{3.56}$$

a direct generalization of the Laplace distribution that arises when $p = 1$. Treating λ as fixed hyperparameter, computation of the resulting MAP estimator under the previous model specification $X \sim \mathcal{N}_p(\theta, I_p)$ reduces to determining the value of θ that minimizes

$$\frac{1}{2}\|X - \theta\|_2^2 + \lambda\|\theta\|_2. \tag{3.57}$$

The resulting estimator is easily shown to be

$$\delta_{GL}(X) = \left(1 - \frac{\lambda}{\|X\|_2}\right)_+ X, \tag{3.58}$$

an estimator that coincides with the solution to the canonical version of the grouped LASSO problem involving a single group of parameters (see Yuan and Lin 2006) and equals $\widehat{\theta}(X) = \mathrm{sign}(X)(|X| - \lambda)_+$ for the case where $p = 1$.

Consider the problem of estimating θ in the canonical setting $X \sim \mathcal{N}_p(\theta, I_p)$. In view of the fact that (3.53) leads to (3.56) upon integrating out ψ, our starting point is the (possibly improper) generalized class of joint prior distributions $\pi(\theta, \lambda|\alpha, \beta)$, which we define in the following hierarchical fashion

$$\begin{aligned} \pi(\theta|\lambda, \alpha, \beta) &\propto \lambda^p \exp\{-\lambda\|\theta\|_2\}, \\ \pi(\lambda|\alpha, \beta) &\propto \lambda^{-p} \exp\{-\alpha(\lambda - \beta)^2\}, \end{aligned} \tag{3.59}$$

where $\alpha, \beta > 0$ are hyperparameters. Equivalently,

$$\pi(\theta, \lambda|\alpha, \beta) \propto \exp\{-\lambda\|\theta\|_2\} \exp\{-\alpha(\lambda - \beta)^2\}. \tag{3.60}$$

The prior on λ is an improper modification of that given in (3.53), in which a location parameter β is introduced and the factor λ^{-p} is introduced to offset the contribution λ^p in (3.56). This construction mimics the idea underlying the prior used by Takada (1979) to motivate (3.50) as a MAP estimator.

Considering (3.60) as motivation for defining a new class of hierarchical penalty functions, Strawderman and Wells (2012) propose deriving the MAP estimator for (θ, λ) through minimizing the objective function

$$G(\theta, \lambda) = \frac{1}{2}\|X - \theta\|_2^2 + \lambda\|\theta\|_2 + \alpha(\lambda - \beta)^2 \tag{3.61}$$

jointly in $\theta \in \mathbb{R}^p$ and $\lambda > 0$, where $\alpha > 1/2$ and $\beta > 0$ are fixed. The resulting estimator for θ takes the closed form

$$\delta^{(\alpha,\beta)}(X) = w_{\alpha,\beta}(\|X\|_2)X, \tag{3.62}$$

where

$$w_{\alpha,\beta}(s) = \begin{cases} 0 & s \le \beta \\ \nu_\alpha\left(1 - \frac{\beta}{s}\right) & \beta < s \le 2\alpha\beta \\ 1 & s > 2\alpha\beta \end{cases}$$

for $\nu_\alpha = 2\alpha/(2\alpha - 1)$. Equivalently, we may write

$$w_{\alpha,\beta}(s) = \begin{cases} \nu_\alpha\left(1 - \frac{\beta}{s}\right)_+ & s \le 2\alpha\beta \\ 1 & s > 2\alpha\beta \end{cases}$$

demonstrating that (3.62) has the flavor of a range-modified positive-part estimator. A detailed derivation of this estimator is in Strawderman and Wells (2012).

Some interesting special cases of the estimator (3.62) arise when considering specific values of α, β and p. For example, letting $\alpha \to \infty$, we obtain (for $\beta > 0$)

$$\delta^{(\beta)}(X) = \left(1 - \frac{\beta}{\|X\|_2}\right)_+ X; \tag{3.63}$$

upon setting $\beta = \lambda$, we evidently recover (3.58); subsequently, setting $\lambda = \sqrt{p-2}$, one then obtains an obvious modification of (3.50) for the case where $c = p - 2$:

$$\delta^*_{PP}(X) = \left(1 - \frac{\sqrt{p-2}}{\|X\|_2}\right)_+ X \tag{3.64}$$

In the special case $p = 1$, the estimator (3.62) reduces to

$$\delta^M(X) = \begin{cases} 0 & \text{if } |X| \le \beta \\ \frac{2\alpha}{2\alpha-1}(X - \text{sign}(X)\beta) & \text{if } \beta < |X| \le 2\alpha\beta \\ X & \text{if } |X| > 2\alpha\beta \end{cases}. \tag{3.65}$$

As shown in Strawderman et al. (2013), (3.65) is also the solution to the penalized minimization problem

$$\frac{1}{2}(X-\theta)^2+\rho(\theta;\alpha,\beta),$$

where $\beta>0, \alpha>1/2$ and

$$\rho(t;\alpha,\beta)=\beta\int_0^{|t|}(1-\frac{z}{2\alpha\beta})_+\,dz,\quad t\in\mathbb{R}.$$

This optimization problem is the univariate equivalent of the penalized likelihood estimation problem considered in Zhang (2010), who referred to $\rho(t;\alpha,\beta)$ as MCP. It follows that (3.65) is equivalent to the univariate MCP thresholding operator; consequently, (3.62) may be regarded as a generalization of this operator for thresholding a vector of parameters. Zhang (2010) showed that the LASSO, SCAD, and MCP belong to a family of quadratic spline penalties with certain sparsity and continuity properties. MCP turns out to be the simplest penalty that results in an estimator that is nearly unbiased, sparse and continuous. As demonstrated above, MCP also has an interesting Bayesian motivation under a hierarchical modeling strategy. Strawderman et al. (2013) undertook a more detailed study of the connections between MCP, the hierarchically penalized estimator, and proximal operators for the case of $p=1$. They also compared this estimator to several others through consideration of frequentist and Bayes risks.

3.6 Estimation of a Predictive Density

Consider a parametric model $\{\mathscr{Y},(\mathscr{P}'{}_\mu)_{\mu\in\Omega}\}$ where $\mathscr{Y}$ is the sample space, Ω is the parameter space and $\mathscr{P}'=\{p(y|\mu):\mu\in\Omega\}$ is a class of densities of $\mathscr{P}'{}_\mu$ with respect to a σ -finite measure. In addition, suppose an observed value x of the random variable X follows a model $\{\mathscr{X},(\mathscr{P}_\mu)_{\mu\in\Omega}\}$ indexed by the same parameter. In this section, we examine the problem of estimating the true density $p'(.|\mu)\in\mathscr{P}'$ of a random variable Y. In this context $p'(\cdot|\mu)$ is referred to as the predictive density of Y.

Let the density $\hat{q}(y|x)$ (belonging to some class of models $\mathscr{C}\supset\mathscr{P}'$) be an estimate, based on the observed data x, of the true density $p(y|\mu)$. Aitchison (1975) proposed using the Kullback and Leibler (1951) divergence, defined in (3.66) below, as a loss function for estimating $p(y|\mu)$.

The class of estimates $\mathscr{C}$ can be identical to the class $\mathscr{P}'$, that is, for any $y\in\mathscr{Y}$

$$\hat{q}(y|x)=p(y|\mu=\hat{\mu}(x))$$

where $\hat{\mu}$ is some estimate of μ. This type of density estimator is called the "plug-in density estimate" associated with the estimate $\hat{\mu}$. Alternatively, one may choose

$$\hat{q}(y|x) = \int_{\Omega} p(y|\mu)\, d\pi(\mu|x)$$

where $d\pi(\mu|x)$ may be a weight function (measure) or an *a posteriori* density associated with a priori measure $\pi(\mu)$. In this case, the class $\mathscr{C}$ will be broader than the class of the models $\mathscr{P}'$. Aitchison (1975) showed that this latter method is preferable to the plug-in approach for several families of probability distributions by comparing their risks induced by the Kullback-Leibler divergence.

3.6.1 The Kullback-Leibler Divergence

First, recall the definition of the Kullback-Leibler divergence and some of its properties.

Lemma 3.10 *The Kullback-Leibler divergence (relative entropy) $D_{KL}(p, q)$ between two densities p and q is defined by*

$$D_{KL}(p, q) = E_p\left[\log \frac{p}{q}\right] = \int \log\left[\frac{p(x)}{q(x)}\right] p(x)\, dx \geq 0 \qquad (3.66)$$

and equality is achieved if and only if $p = q$, $p-$almost surely.

Note that the divergence can be finite only if the support of the density p is contained in the support of the density q. By convention, we define $0 \log \frac{0}{0} = 0$.

Proof By definition of 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 equality, using Jensen's inequality, if and only if $p = q$, p -almost surely. Note that the lemma is true if q is assumed only to be a subdensity (mass less than or equal to 1). □

The Kullback-Leibler divergence is not a true distance since it is not symmetric and it does not satisfy the triangle inequality. But it appears as the natural discrepancy measure in information theory. An important property, given in the following lemma, is that it is strictly convex.

Lemma 3.11 *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* $0 \leq \lambda \leq 1$,

$$D_{KL}(\lambda\, p_1 + (1-\lambda)\, p_2, \lambda\, q_1 + (1-\lambda)\, q_2) \leq \lambda D_{KL}(p_1, q_1) + (1-\lambda) D_{KL}(p_2, q_2)\,, \tag{3.67}$$

with strict inequality unless $(p_1, p_2) = (q_1, q_2)$ *a.e. with respect to* $p_1 + p_2$.

Proof Note that $f(t) = t\ \log(t)$ 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 above values of α_1, α_2, t_1 and t_2 gives

$$(\lambda p_1 + (1-\lambda)p_2) \log \frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2} \leq \lambda p_1 \log \frac{p_1}{q_1} + (1-\lambda)p_2 \log \frac{p_2}{q_2}.$$

Finally, by integrating the latter term, (3.67) and the strict convexity follow from the strict convexity of the function f. □

3.6.2 The Bayesian Predictive Density

Assume in the rest of this subsection that $p(x|\mu)$ and $p'(y|\mu)$ are densities with respect to the Lesbegue measure. For any estimator $\hat{p}(\cdot|x)$ of the density $p'(y|\mu)$, define the Kullback-Leibler loss by

$$\mathrm{KL}(\mu, \hat{p}(\cdot|x)) = \int p'(y|\mu) \log\left[\frac{p'(y|\mu)}{\hat{p}(y|x)}\right] dy \tag{3.68}$$

and its corresponding risk as

$$\mathscr{R}_{\mathrm{KL}}(\mu,\hat{p}) = \int p(x|\mu)\left[\int p'(y|\mu)\log\left[\frac{p'(y|\mu)}{\hat{p}(y|x)}\right]dy\right]dx. \tag{3.69}$$

We say that the density estimate $\hat{p}_2$ dominates the density estimate $\hat{p}_1$ if, for any $\mu \in \Omega$, $\mathscr{R}_{\mathrm{KL}}(\mu,\hat{p}_1) - \mathscr{R}_{\mathrm{KL}}(\mu,\hat{p}_2) \le 0$, with strict inequality for at least some value of μ.

In the Bayesian framework we will compare estimates using Bayes risk. We will consider the class, more general than Aitchison (1975), of all subdensities,

$$\mathscr{D} = \left\{q(\cdot|x)|\int q(y|x)\,dy \le 1 \quad \text{for all } x\right\}.$$

Lemma 3.12 (Aitchison 1975) *The Bayes risk*

$$r_\pi(\hat{p}) = \int \mathscr{R}_{KL}(\mu,\hat{p})\,\pi(\mu)\,d\mu$$

is minimized by

$$\hat{p}_\pi(y|x) = \int p'(y|\mu)\,p(\mu|x)\,d\mu = \frac{\int p'(y|\mu)\,p(x|\mu)\pi(\mu)\,d\mu}{\int p(x|\mu)\,\pi(\mu)\,d\mu}. \tag{3.70}$$

We call $\hat{p}_\pi$ the Bayesian predictive density.

Proof The difference between the Bayes risks of $\hat{p}_\pi$ and another competing subdensity estimator $\hat{q}$ is

$$\begin{aligned}
r_\pi(\hat{q}) - r_\pi(\hat{p}_\pi) &= \int_\Omega\left[\int_{\mathscr{X}}\left\{\int_{\mathscr{Y}} p'(y|\mu)\log\frac{\hat{p}_\pi(y|x)}{\hat{q}(y|x)}\,dy\right\}p(x|\mu)\,dx\right]\pi(\mu)\,d\mu \\
&= \int_\Omega\left[\int_{\mathscr{X}}\left\{\int_{\mathscr{Y}} p'(y|\mu)\log\frac{\hat{p}_\pi(y|x)}{\hat{q}(y|x)}\,dy\right\}p(x|\mu)\,\pi(\mu)\,dx\right]d\mu \\
&= \int_\Omega\left[\int_{\mathscr{X}}\left\{\int_{\mathscr{Y}} p'(y|\mu)\log\frac{\hat{p}_\pi(y|x)}{\hat{q}(y|x)}\,dy\right\}p(\mu|x)\,m(x)\,dx\right]d\mu\,.
\end{aligned}$$

Rearranging the order of integration thanks to Fubini'sTheorem gives

$$\begin{aligned}
r_\pi(\hat{q}) - r_\pi(\hat{r}) &= \int_{\mathscr{X}}\left[\int_{\mathscr{Y}}\left\{\int_\Omega p(\mu|x)\,p'(y|\mu)\,d\mu\right\}\log\frac{\hat{p}_\pi(y|x)}{\hat{q}(y|x)}\,dy\right]m(x)\,dx \\
&= \int_{\mathscr{X}}\left[\int_{\mathscr{Y}}\hat{p}_\pi(y|x)\log\frac{\hat{p}_\pi(y|x)}{\hat{q}(y|x)}\,dy\right]m(x)\,dx \\
&= \int_{\mathscr{X}} D_{\mathrm{KL}}(\hat{p}_\pi(.|x),\hat{q}(.|x))\,m(x)\,dx \ge 0.
\end{aligned}$$

□

3.6.3 Sufficiency Reduction in the Normal Case

Let $X_{(n)} = (X_1, \dots, X_n)$ and $Y_{(m)} = (Y_1, \dots, Y_m)$ be independent *iid* samples from p-dimensional normal distributions $\mathcal{N}_p(\mu, \Sigma_1)$ and $\mathcal{N}_p(\mu, \Sigma_2)$ with unknown common mean μ and known positive definite covariance matrices Σ_1 and Σ_2. On the basis of an observation $x_{(n)} = (x_1, \dots, x_n)$ from $X_{(n)}$, consider the problem of estimating the true predictive density $p'(y_{(m)}|\mu)$ of $y_{(m)} = (y_1, \dots, y_m)$, under the Kullback-Leibler loss. For a prior density $\pi(\mu)$, the Bayesian predictive density is given by

$$\hat{p}_\pi(y_{(m)}|x_{(n)}) = \frac{\int_\Omega p'(y_{(m)}|\mu)\, p(x_{(n)}|\mu)\, \pi(\mu)\, d\mu}{\int_\Omega p(x_{(n)}|\mu)\, \pi(\mu)\, d\mu}. \tag{3.71}$$

For simplicity, we consider the case where $\Sigma_1 = \Sigma_2 = I_p$. According to Komaki (2001) the Bayesian predictive densities satisfy

$$\int_{\mathbb{R}^{pm}} p'(y_{(m)}|\mu) \log \frac{p'(y_{(m)}|\mu)}{\hat{p}_\pi(y_{(m)}|x_{(n)})}\, dy_{(m)} = \int_{\mathbb{R}^p} p'(\bar{y}_m|\mu) \log \frac{p'(\bar{y}_m|\mu)}{\hat{p}_\pi(\bar{y}_m|\bar{x}_n)}\, d\bar{y}_m \tag{3.72}$$

where, denoting by $\phi_p(\cdot, |\mu, \Sigma)$ the density of $\mathcal{N}_p(\mu, \Sigma)$, in the left-hand side of (3.72),

$$p'(y_{(m)}|\mu) = \prod_{i=1}^{m} \phi_p(y_i, |\mu, I_p)$$

while, in the right-hand side of (3.72),

$$p'(\bar{y}_m|\mu) = \phi_p(\bar{y}_m|\mu, I_p/m)$$

with $\bar{y}_m = \sum_{j=1}^m y_j/m$. Similarly, $\hat{p}_\pi(y_{(m)}|x_{(n)})$ corresponds to the conditional density of the $p \times m$ matrix $y_{(m)}$ given the $p \times m$ matrix $x_{(n)}$ while $\hat{p}_\pi(\bar{y}_m|\bar{x}_m)$ corresponds to the conditional density of the $p \times 1$ vector $\bar{y}_m$ given the $p \times 1$ vector $\bar{x}_n = \sum_{i=1}^n x_i/n$.

To see this sufficiency reduction, use the fact that

$$\sum_{i=1}^{m} \|y_i - \mu\|^2 = \sum_{i=1}^{m} \|y_i - \bar{y}_m\|^2 + m\,(\|\bar{y}_m - \mu\|)^2.$$

Then we can express $p'(y_{(m)}|\mu)$ as

$$p'(y_{(m)}|\mu) = \frac{1}{(2\,\pi)^{mp/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{m} \|y_i - \bar{y}_m\|^2\right) \exp\left(-\frac{m}{2}\,(\|\bar{y}_m - \mu\|)^2\right)$$
$$= \frac{m^{p/2}}{(2\pi)^{(m-1)p/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{m} \|y_i - \bar{y}_m\|^2\right) p(\bar{y}_m|\mu). \tag{3.73}$$

Similarly, it follows that

$$p(x_{(n)}|\mu) = \frac{n^{p/2}}{(2\pi)^{(n-1)p/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \|x_i - \bar{x}_m\|^2\right) p(\bar{x}_m|\mu)\,.$$

By replacing these expressions in the form of the predictive density in (3.71), we get

$$\hat{p}_\pi(y_{(m)}|x_{(n)})$$
$$= \left\{\frac{m^{p/2}}{(2\pi)^{(m-1)p/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{m} \|y_i - \bar{y}_m\|^2\right)\right\} \frac{\int p'(\bar{y}_m|\mu)\, p(\bar{x}_m|\mu)\, \pi(\mu)\, d\mu}{\int p(\bar{x}_m|\mu)\, \pi(\mu)\, d\mu}$$
$$= \left\{\frac{m^{p/2}}{(2\pi)^{(m-1)p/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{m} \|y_i - \bar{y}_m\|^2\right)\right\} \hat{p}_\pi(\bar{y}_m|\bar{x}_m). \tag{3.74}$$

Finally, for (3.73) and (3.74), 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 predictive density estimator is equal to the risk of the Bayesian predictive density associated to π in the reduced model $X \sim \mathcal{N}_p(\mu, \frac{1}{n} I_p)$ and $Y \sim \mathcal{N}_p(\mu, \frac{1}{m} I_p)$. 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 different estimators of μ, δ_1 and δ_2. That is, for $i = 1, 2$, define

$$\hat{p}_i(y_{(m)}|x_{(n)}) = p'(y_{(m)}|\mu = \delta_i(x_{(n)})). \tag{3.75}$$

The difference in risk between $\hat{p}_2$ and $\hat{p}_1$ is given by

$$\begin{aligned}
\Delta\mathscr{R}_{\mathrm{KL}}(\hat{p}_2, \hat{p}_1) &= \mathscr{R}_{\mathrm{KL}}(\mu, \hat{p}_2) - \mathscr{R}_{\mathrm{KL}}(\mu, \hat{p}_1) \\
&= \int p(x_{(n)}|\mu) \int p(y_{(m)}|\mu) \log \frac{\hat{p}_1(y_{(m)}|x_{(n)})}{\hat{p}_2(y_{(m)}|x_{(n)})}\, dy_{(m)}\, dx_{(n)} \\
&= \int p(x_{(n)}|\mu) \int p(y_{(m)}|\mu) \Big(\frac{1}{2} \sum_{i=1}^{m} \|\delta_2(x_{(n)}) - y_i\|^2 \\
&\qquad - \frac{1}{2} \sum_{i=1}^{m} \|\delta_1(x_{(n)}) - y_i\|^2 \Big)\, dy_{(m)}\, dx_{(n)} \,.
\end{aligned}$$

By the independence of $X_{(n)}$ and $Y_{(m)}$ this can be reexpressed in terms of expectations as

$$\begin{aligned}
&\Delta\mathscr{R}_{\mathrm{KL}}(\hat{p}_2, \hat{p}_1) \\
&= \frac{1}{2} \sum_{i=1}^{m} E_{X_{(n)},Y_{(m)}} \Big(\|\delta_2(X_{(n)}) - \mu + \mu - Y_i\|^2 - \|\delta_1(X_{(n)}) - \mu + \mu - Y_i\|^2 \Big) \\
&= \frac{m}{2} E_{X_{(n)},Y_{(m)}} \Big[\|\delta_2(X_{(n)}) - \mu\|^2 - \|\delta_1(X_{(n)}) - \mu\|^2 \Big] \\
&\quad + \sum_{i=1}^{m} E_{X_{(n)},Y_{(m)}} \Big(\big[(\delta_2(X_{(n)}) - \mu)(\mu - Y_i)\big] - \big[(\delta_1(X_{(n)}) - \mu)(\mu - Y_i)\big] \Big) \\
&= \frac{m}{2} \Big(E_{X_{(n)}} \Big[\|\delta_2(X_{(n)}) - \mu\|^2 \Big] - E_{X_{(n)}} \Big[\|\delta_1(X_{(n)}) - \mu\|^2 \Big] \Big) \\
&= \frac{m}{2} \Big[\mathscr{R}_Q(\delta_2, \mu) - \mathscr{R}_Q(\delta_1, \mu) \Big],
\end{aligned}$$

which shows that the risk difference between $\hat{p}_2$ and $\hat{p}_1$ is proportional to the risk difference between δ_2 and δ_1.

Note that, by completeness of the statistics $\bar{X}_n$, it suffices to consider only estimates of μ that depend only on $\bar{X}_n$.

3.6.4 Properties of the Best Invariant Density

In this subsection, we restrict our attention to location models. We assume $X \sim p(x|\mu) = p(x - \mu)$ and $Y \sim p'(y|\mu) = p'(y - \mu)$, where p and p' are two known possibly different densities. A density $\hat{q}$ is called invariant (equivariant) with 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 $q(y+a|x+a) = q(y|x)$. The following result shows that the risk of an invariant predictive density is constant.

Lemma 3.13 *The invariant predictive densities with respect to the location group of translations have constant risk.*

Proof By the property of invariance, the risk of an invariant density $\hat{q}$ is equal to

$$\begin{aligned}
\mathscr{R}(\mu, \hat{q}) &= \int \log \frac{p'(y-\mu)}{\hat{q}(y|x)}\, p(x-\mu)\, p'(y-\mu)\, dy\, dx \\
&= \int \log \frac{p'(y-\mu)}{\hat{q}(y-\mu|x-\mu)}\, p(x-\mu)\, p'(y-\mu)\, dy\, dx \\
&= \int \log \frac{p(z')}{\hat{q}(z'|z)}\, p(z)\, p'(z')\, dz'\, dz, \qquad (3.76)
\end{aligned}$$

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 invariant predictive density which minimizes this risk is known as the best invariant predictive density.

Lemma 3.14 *The best invariant predictive density is the Bayesian predictive density $\hat{p}_U$ associated with the Lebesgue measure on $\mathbb{R}^p$, $\pi(\mu) = 1$, is given by*

$$\hat{p}_U(y|x) = \frac{\int_{\mathbb{R}^p} p'(y|\mu)\, p(x|\mu)\, d\mu}{\int_{\mathbb{R}^p} p(x|\mu)\, d\mu}. \qquad (3.77)$$

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. As noted in the previous section, if $\hat{q}$ is an invariant predictive density, $\hat{q}(y|x) = \hat{q}(y-x|0) = \hat{q}(y-x)$, by an abuse of notation. Hence,

$$\begin{aligned}
\mathscr{R}(\mu, \hat{q}) - \mathscr{R}(\mu, \hat{p}) &= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} \left[\log \frac{\hat{p}(y-x)}{\hat{q}(y-x)}\right] p(x-\mu)p'(y-\mu)\, dx\, dy \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} \left[\log \frac{\hat{p}(z'-z)}{\hat{q}(z'-z)}\right] p(z)p'(z')\, dz\, dz' \\
&= \int_{\mathbb{R}^p} \left[\log \frac{\hat{p}(t)}{\hat{q}(t)}\right] \hat{p}(t)\, dt, \qquad (3.78)
\end{aligned}$$

which is always positive by the inequality in (3.66). The result of the equality in (3.78), and hence the lemma, follows from the fact that $\hat{p}(t) = \hat{p}(y-x) = \hat{p}_U(y|x)$, that is,

$$\begin{aligned}
\hat{p}(t) &= \int_{\mathbb{R}^p} p(z)\, p'(z+t)\, dz \\
&= \int_{\mathbb{R}^p} p(z)\, p'(z+y-x)\, dz \\
&= \int_{\mathbb{R}^p} p(x-\mu)\, p'(y-\mu)\, d\mu \\
&= \frac{\int_{\mathbb{R}^p} p'(y|\mu)\, p(x|\mu)\, d\mu}{\int_{\mathbb{R}^p} p(x|\mu)\; d\mu} \qquad (3.79)
\end{aligned}$$

which is the expression of $\hat{p}_U$ given in (3.70) 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. 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.

We will now show that $\hat{p}_U$ is minimax in location problems.

Lemma 3.15 *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 We show minimaxity using Lemma 1.8. Consider a sequence $\{\pi_k\}$ of normal $\mathcal{N}_p(0, k\, I_p)$ priors. The difference of Bayes risk between $\hat{p}_U$ and $\hat{p}_{\pi_k}$, is given by

$$\begin{aligned}
r(\hat{p}_U, \pi_k) - r(\hat{p}_{\pi_k}, \pi_k) &= \int_{\mathbb{R}^p} \left[\mathscr{R}(\mu, \hat{p}_U) - \mathscr{R}(\mu, \hat{p}_{\pi_k})\right] \pi_k(\mu)\; d\mu \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} \log \frac{\hat{p}_{\pi_k}(y|x)}{\hat{p}_U(y|x)} p(y|\mu)\, p(x|\mu)\pi_k(\mu)\; dy\; dx\; d\mu \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} \log \frac{\hat{p}_{\pi_k}(y|x)}{\hat{p}_U(y|x)} \left[\int_{\mathbb{R}^p} p(y|\mu)\, p(x|\mu)\, \pi_k(\mu)\; d\mu\right] dy\; dx \\
&= E^{X,Y}_{\pi_k} \log \frac{\hat{p}_{\pi_k}(Y|X)}{\hat{p}_U(Y|X)} \qquad (3.80)
\end{aligned}$$

where $E^{x,y}_{\pi_k}$ denotes the expectation with respect to the joint marginal of (X, Y),

$$m_{\pi_k}(x, y) \;=\; \int_{\mathbb{R}^p} p(y|\mu)\, p(x|\mu)\, \pi_k(\mu)\; d\mu.$$

Since $r(\hat{p}_U, \pi_k) = \mathscr{R}(\mu, \hat{p}_U)$ ($\hat{p}_U$ has constant risk) it suffices to show (3.80) tends to 0 as k tends to infinity. By simplifying one gets

$$
\begin{aligned}
& r(\hat{p}_U, \pi_k) - r(\hat{p}_{\pi_k}, \pi_k) \\
&= E_{\pi_k}^{X,Y}\left[\log\left(\frac{\int p(x,y|\mu)\,\pi_k(\mu)\,d\mu}{\int p(x|\mu)\,\pi_k(\mu)\,d\mu}\frac{1}{\int p(x,y|\mu)\,d\mu}\right)\right] \\
&= E_{\pi_k}^{X,Y}\left[-\log\frac{\int p(x,y|\mu)\,\pi_k(\mu)\frac{1}{\pi_k(\mu)}\,d\mu}{\int p(x,y|\mu)\,\pi_k(\mu)\,d\mu} - \log\left(\int p(x|\mu)\,\pi_k(\mu)\,d\mu\right)\right] \\
&= E_{\pi_k}^{X,Y}\left[-\log E_{\mu|X,Y}\frac{1}{\pi_k(\mu)} - \log\left(\int p(x|\mu)\,\pi_k(\mu)\,d\mu\right)\right]
\end{aligned}
$$

where $E_{\mu|X,Y}$ denotes the expectation with respect to the posterior of μ given (X, Y). An application of Jensen's inequality gives

$$
\begin{aligned}
& r(\hat{p}_U, \pi_k) - r(\hat{p}_{\pi_k}, \pi_k) \\
&\leq E_{\pi_k}^{X,Y} E_{\mu|X,Y}\log\pi_k(\mu) - E_{\pi_k}^{X,Y}\left[\int p(X|\mu)\,\log\pi_k(\mu)\,d\mu\right]. \qquad (3.81)
\end{aligned}
$$

By developing the expectations, it follows that

$$
\begin{aligned}
E_{\pi_k}^{X,Y} E_{\mu|X,Y}\log\pi_k(\mu) &= \iint m_{\pi_k}(x,y)\frac{\int p(x,y|\mu)\pi_k(\mu)\log(\pi_k(\mu))d\mu}{m_{\pi_k}(x,y)}dxdy \\
&= \iiint \pi_k(\mu)\log(\pi_k(\mu))\,d\mu\,dxdy \\
&= \int \pi_k(\mu)\log(\pi_k(\mu))d\mu. \qquad (3.82)
\end{aligned}
$$

Similarly, by integrating with respect to y and by interchanging between μ and μ' we have

$$
\begin{aligned}
& E_{\pi_k}^{X,Y}\left[\int p(X|\mu)\,\log\pi_k(\mu)\,d\mu\right] \\
&= \iiiint p(x|\mu')p(y|\mu')\pi_k(\mu')p(x|\mu)\log\pi_k(\mu)\,d\mu' d\mu dxdy \\
&= \iiint \pi_k(\mu')p(x|\mu)p(x|\mu')\log\pi_k(\mu)d\mu'\,dx\,d\mu \\
&= \iiint \pi_k(\mu)p(x|\mu)p(x|\mu')\log\pi_k(\mu')d\mu\,dxd\mu'. \qquad (3.83)
\end{aligned}
$$

By grouping the expressions (3.81), (3.83) and (3.84) and making the changes of variables $z = x - \mu$ and $z' = x - \mu'$ it follows that

$$
\begin{aligned}
&r(\hat{p}_U, \pi_k) - r(\hat{p}_{\pi_k}, \pi_k) \\
&\leq \iiint p(x|\mu)p(x|\mu')\pi_k(\mu)\left[\log(\pi_k(\mu)) - \log(\pi_k(\mu'))\right] d\mu d\mu'\, dx \\
&= \iiint \pi_k(\mu)p(x-\mu)p(x-\mu')\log\left(\frac{\pi_k(\mu)}{\pi_k(\mu')}\right) d\mu\, dz\, dz' \\
&= \iiint \pi_k(\mu)p(z)p(z')\log\left(\frac{\pi_k(\mu)}{\pi_k(\mu+z-z')}\right) d\mu\, dz\, dz'. \qquad (3.84)
\end{aligned}
$$

In view of the form $\pi_k(\mu)$, the term on the right in (3.84) can be written as

$$
\begin{aligned}
&E_{\pi_k}E_{Z,Z'}\log\left(\frac{\pi_k(\mu)}{\pi_k(\mu+Z-Z')}\right) \\
&= E_{\pi_k}E_{Z,Z'}\frac{1}{2k}\left(\|\mu+Z-Z'\| - \|\mu^2\|\right) \\
&= E_{\pi_k}E_{Z,Z'}\left[\frac{1}{2k}\left(\|Z\|^2 + \|Z'\|^2 + 2\langle\mu, Z-Z'\langle\right)\right] \\
&= E_{Z,Z'}\left[\frac{1}{2k}\left(\|Z\|^2 + \|Z'\|^2\right)\right],
\end{aligned}
$$

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 by Lemma 1.8. □

This result is in Liang and Barron (2004), a more direct proof for the Gaussian case can be found in George et al. (2006) and is given in the next section.

3.6.5 An Explicit Expression for $\hat{p}_U$ and Its Risk in the Normal Case

We now give an explicit expression of $\hat{p}_U$, described the previous subsections, in the Gaussian setting. Let $X \sim \mathcal{N}_p(\mu, v_x I_p)$ and $Y \sim \mathcal{N}_p(\mu, v_y I_p)$.

Lemma 3.16 *The Bayesian predictive density associated with the uniform prior on* $\mathbb{R}^p$, $\pi(\mu) \equiv 1$, *is given by the following expression*

$$
\hat{p}_U(y|x) = \frac{1}{\left((2\pi)(v_y+v_x)\right)^{p/2}}\exp\left(-\frac{\|y-x\|^2}{2(v_x+v_y)}\right). \qquad (3.85)
$$

Proof For $W = (v_y X + v_x Y)/(v_x + v_y)$ and $v_w = (v_x v_y)/(v_x + v_y)$ it is clear that $W \sim \mathcal{N}_p(\mu, v_w I_p)$, by the independence of X and Y. Further, note that

$$\frac{\|x-\mu\|^2}{2v_x}+\frac{\|y-\mu\|^2}{2v_y}=\frac{\|\mu-w\|^2}{2v_w}+\frac{\|y-x\|^2}{2(v_x+v_y)}\,. \tag{3.86}$$

By definition, and through the previous representation, it follows that

$$\begin{aligned}
\hat{p}_U(y|x) &= \frac{\displaystyle\int_{\mathbb{R}^p} p(y|\mu,v_y)\,p(x|\mu,v_x)\,d\mu}{\displaystyle\int_{\mathbb{R}^p} p(x|\mu,v_x)\,d\mu} \\
&= \int_{\mathbb{R}^p} \frac{1}{(2\pi)^p(v_y\,v_x)^{p/2}}\exp\left(-\frac{\|x-\mu\|^2}{2\,v_x}-\frac{\|y-\mu\|^2}{2\,v_y}\right)d\mu \\
&= \int_{\mathbb{R}^p} \frac{1}{(2\pi)^p(v_y\,v_x)^{p/2}}\exp\left(-\frac{\|\mu-w\|^2}{2\,v_w}\right)\exp\left(-\frac{\|y-x\|^2}{2\,(v_x+v_y)}\right)d\mu \\
&= \frac{(2\pi\,v_w)^{p/2}}{(2\pi)^p(v_y\,v_x)^{p/2}}\exp\left(-\frac{\|y-x\|^2}{2(v_x+v_y)}\right) \\
&= \frac{1}{\big((2\pi)\,(v_y+v_x)\big)^{p/2}}\exp\left(-\frac{\|y-x\|^2}{2\,(v_x+v_y)}\right).
\end{aligned}$$

□

Note that the risk of $\hat{p}_U$ is constant, as we have previously seen for invariant densities. Given the form of $\hat{p}_U(.|x)$ it follows that the Kullback-Liebler divergence is

$$\begin{aligned}
&\mathrm{KL}(\hat{p}_U(.|x),\mu) \\
&= \int p(y|\mu,v_y)\log\frac{p(y|\mu,v)}{\hat{p}_U(y|x)}\,dy \\
&= E^Y\left[\log\frac{p(Y|\mu,v)}{\hat{p}_U(Y|x)}\right] \\
&= E^Y\left[-\frac{p}{2}\log\frac{v_y}{v_x+v_y}-\frac{1}{2v_y}\|Y-\mu\|^2+\frac{1}{2(v_x+v_y)}\|Y-x\|^2\right] \\
&= -\frac{p}{2}\log\frac{v_y}{v_x+v_y}-\frac{p}{2}+E^Y\left[\frac{1}{2(v_x+v_y)}\left(\|Y-\mu\|^2+\|\mu-x\|^2\right)\right] \\
&= \left[-\frac{p}{2}\log\frac{v_y}{v_x+v_y}-\frac{p}{2}+\frac{pv_y}{2(v_x+v_y)}\right]+\frac{1}{2(v_x+v_y)}\|\mu-x\|^2. \qquad (3.87)
\end{aligned}$$

Hence, we can conclude that the risk of $\hat{p}_U$ is

$$\begin{aligned}\mathscr{R}_{\text{KL}}(\hat{p}_U, \mu) &= E^X \left[\text{KL}(\hat{p}_U, \mu, X)\right] \\ &= \left[-\frac{p}{2}\log\frac{v_y}{v_x+v_y} - \frac{p}{2} + \frac{p v_y}{2(v_x+v_y)}\right] + \frac{p v_x}{2(v_x+v_y)} \\ &= -\frac{p}{2}\log\left(\frac{v_y}{v_x+v_y}\right) = \frac{p}{2}\log\left(1+\frac{v_x}{v_y}\right). \qquad (3.88)\end{aligned}$$

In the framework of the *iid* sampling model presented in Sect. 3.6.3 with $\varSigma_1 = \varSigma_2 = I_p$, we can express the risk as

$$\mathscr{R}_{\text{KL}}(\hat{p}_U, \mu) = \frac{p}{2}\log\left(1+\frac{m}{n}\right).$$

A predictive density is called the plug-in relative to an estimator δ if it has the form

$$\hat{p}_\delta(y|x) = \frac{1}{(2\pi v_y)^{p/2}}\exp\left(-\frac{1}{2}\frac{\|y-\delta(x)\|^2}{v_y}\right).$$

The predictive plug-in density, which corresponds to the standard estimator of the mean, μ, $\delta_0(X) = X$, is

$$\hat{p}_\delta(y|x) = \frac{1}{\left(2\pi v_y\right)^{p/2}}\exp\left[-\frac{1}{2}\frac{\|y-x\|^2}{v_y}\right].$$

We can directly verify that the predictive density $\hat{p}_U$ dominates the plug-in density $\hat{p}_{\delta_0}$ for any $\mu \in \mathbb{R}^p$. In fact, their difference in risk is

$$\begin{aligned}\triangle\mathscr{R}_{\text{KL}}(\hat{p}_U, \hat{p}_{\delta_0}) &= E^{X,Y}\left(\log\frac{\hat{p}_U(Y|X)}{\hat{p}_{\delta_0}(Y|X)}\right) \\ &= -\frac{p}{2}\log\left(\frac{v_x+v_y}{v_y}\right) - \frac{1}{2}\left[\frac{1}{v_x+v_y} - \frac{1}{v_y}\right]E^{X,Y}\left(\|Y-X\|^2\right).\end{aligned}$$

Since $E^{X,Y}\left(\|Y-X\|^2\right)$ equals

$$\begin{aligned}E^{X,Y}\left(\|Y-\mu\|^2\right) + E^{X,Y}\left(\|X-\mu\|^2\right) - 2\left\langle E^{X,Y}(Y-\mu), E^{X,Y}(X-\mu)\right\rangle \\ = p(v_x+v_y),\end{aligned}$$

we have

$$\triangle\mathscr{R}_{\mathrm{KL}}(\hat{p}_U, \hat{p}_{\delta_0}) = -\frac{p}{2}\left[\log\left(1+\frac{v_x}{v_y}\right) - \frac{v_x}{v_y}\right] > 0\,.$$

Surprisingly, the predictive density $\hat{p}_U$ has similar properties to 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 predictive density using the harmonic prior, $\pi(\mu) = \|\mu\|^{2-p}$. George et al. (2006) extended the analogy with point estimation. We give some of this development next.

Lemma 3.17 (George et al. 2006, Lemma 2) *For $W = (v_y\, X + v_x\, Y)/(v_x + v_y)$ and $v_w = (v_x\, v_y)/(v_x + v_y)$, let $m_\pi(W; v_w)$ and $m_\pi(X; v_x)$ be the marginals of W and X, respectively, relative to the a prior π. Then*

$$\hat{p}_\pi(y|X) = \frac{m_\pi(W; v_w)}{m_\pi(X; v_x)}\;\hat{p}_U(y|X) \tag{3.89}$$

where $\hat{p}_U(\cdot|X)$ is the Bayes estimator associated with the uniform prior on $\mathbb{R}^p$ given by (3.85). In addition, for any prior measure π, the Kullback-Leibler risk difference between $\hat{p}_U(\cdot|x)$ and the Bayesian predictive density $\hat{p}_\pi(\cdot|x)$ is given by

$$\mathscr{R}_{KL}(\mu, \hat{p}_U) - \mathscr{R}_{KL}(\mu, \hat{p}_\pi) = E_{\mu,v_w}\left[\log\, m_\pi(W; v_w)\right] - E_{\mu,v_x}\left[\log\, m_\pi(X; v_x)\right] \tag{3.90}$$

where $E_{\mu,v}$ denotes the expectation with respect to the normal $\mathscr{N}_p(\mu, vI_p)$ distribution.

Proof The marginal density of (X, Y) associated with π is equal to

$$\begin{aligned}\hat{p}_\pi(x, y) =& \int_{\mathbb{R}^p} p(x|\mu, v_x)\, p(y|\mu, v_y)\, \pi(\mu)\; d\mu\\ =& \int_{\mathbb{R}^p} \frac{1}{(2\pi v_x)^{p/2}} \exp\left(-\frac{\|x-\mu\|^2}{2v_x}\right) \frac{1}{(2\pi v_y)^{p/2}} \exp\left(-\frac{\|y-\mu\|^2}{2v_y}\right) \pi(\mu)\; d\mu.\end{aligned}$$

Applying (3.85) and (3.86) it follows that

$$\begin{aligned}\hat{p}_\pi(x, y) &= \frac{1}{(2\pi)^p\,(v_x\, v_y)^{p/2}} \int_{\mathbb{R}^p} \exp\left(-\frac{\|y-x\|^2}{2(v_x+v_y)}\right) \exp\left(-\frac{\|\mu-w\|^2}{2v_w}\right) \pi(\mu)\, d\mu\\ &= \frac{(2\pi v_w)^{p/2}}{(2\pi)^p\,(v_x\, v_y)^{p/2}}\, \exp\left(-\frac{\|y-x\|^2}{2(v_x+v_y)}\right) m_\pi(w; v_w)\\ &= \hat{p}_U(y|x)\, m_\pi(w; v_w).\end{aligned}$$

Since $\hat{p}_\pi(y|x) = \hat{p}_\pi(x, y)/m_\pi(x)$, (3.89) follows.

Hence, we can write the risk difference as

$$
\begin{aligned}
&\mathscr{R}_{\mathrm{KL}}(\mu, \hat{p}_U) - \mathscr{R}_{\mathrm{KL}}(\mu, \hat{p}_\pi) \\
&= \int\int p(x|\mu, v_x)\, p(y|\mu, v_y) \log \frac{\hat{p}_\pi(y|x)}{\hat{p}_U(y|x)}\, dy\, dx \\
&= \int\int p(x|\mu, v_x)\, p(y|\mu, v_y) \log \frac{m_\pi(W(x, y); v_w)}{m_\pi(x; v_x)}\, dy\, dx \\
&= E^{X,Y} \log m_\pi(W(X, Y); v_w) - E^{X,Y} \log m_\pi(X; v_x) \\
&= E^W \log m_\pi(W|v_w) - E^X \log m_\pi(X|v_x).
\end{aligned}
$$

□

Using this lemma, George et al. (2006) gave a simple proof of the result of Liang and Barron (2004) for the Gaussian setting. By taking the same sequence of priors $\{\pi_k\} = \mathscr{N}_p(0, kI_p)$, the difference of the Bayes risk equals (using constancy of the risk of $\hat{p}_U$)

$$
\begin{aligned}
\mathscr{R}_{\mathrm{KL}}(\mu, \hat{p}_U) - r(\pi_k, \hat{p}_{\pi_k}) &= \int \pi_k(\mu) \left[E_{\mu,v_w} \log m_{\pi_k}(W, v_w) - E_{\mu,v_x} \log m_{\pi_k}(X, v_x)\right] d\mu \\
&= \int \pi_k(\mu) \Bigg[E_{\mu,v_w} \log \left\{(2\pi(v_w + k))^{-p/2} \exp\left(-\frac{\|W\|^2}{2(v_w + k)}\right)\right\} \\
&\quad - E_{\mu,v_x} \log \left\{(2\pi(v_x + k))^{-p/2} \exp\left(-\frac{\|X\|^2}{2(v_x + k)}\right)\right\} \Bigg] d\mu \\
&= \int \pi_k(\mu) \Bigg[- p/2 \log(2\pi(v_w + k)) - \frac{pv_w}{2(v_w + k)} \\
&\quad + p/2 \log(2\pi(v_x + k)) + \frac{pv_x}{2(v_x + k)} \Bigg] d\mu \\
&= -\frac{p}{2} \log \frac{v_w + k}{v_x + k} - \frac{pv_w}{2(v_w + k)} + \frac{pv_x}{2(v_x + k)}.
\end{aligned}
$$

Hence, we see that $\lim_{k\to\infty} r(\pi_k, \hat{p}_U) - r(\pi_k, \hat{p}_{\pi_k}) = 0$ and so, $\hat{p}_U$ is minimax by Lemma 1.8. George et al. (2006) also show that the best predictive invariant density is dominated by any Bayesian predictive density relative to a superharmonic prior. This result parallels the result of Stein for the estimation of the mean under quadratic loss and the use differential operators discussed in Sect. 2.6. The following lemma from George et al. (2006) allows us to give sufficient conditions for domination. We use Stein's identity in the proof.

Lemma 3.18 *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,*

$$\begin{aligned}\frac{\partial}{\partial v} E \log m_\pi(z; v) &= E_{\mu,v}\left[\frac{\Delta m_\pi(Z; v)}{m_\pi(Z; v)} - \frac{1}{2}\|\nabla \log m_\pi(Z; v)\|^2\right] \\ &= E_{\mu,v}\left[2\,\frac{\Delta\sqrt{m_\pi(Z; v)}}{\sqrt{m_\pi(Z; v)}}\right]. \end{aligned} \tag{3.91}$$

Proof For any $v_w \le v \le v_x$,

$$\begin{aligned} m_\pi(z; v) &= \int_{\mathbb{R}^p} \frac{1}{(2\pi\, v)^{p/2}} \exp\left(-\frac{\|z-\mu\|^2}{2v}\right)\pi(\mu)\; d\mu \\ &= \left(\frac{v_x}{v}\right)^{p/2} \int_{\mathbb{R}^p} \frac{1}{(2\pi\, v_x)^{p/2}} \exp\left(-\frac{v_x}{v}\,\frac{\|z-\mu\|^2}{2v_x}\right)\pi(\mu)\; d\mu \\ &\le \left(\frac{v_x}{v}\right)^{p/2} m_\pi(z; v_x) < \infty. \end{aligned}$$

Hence, the marginal m_π is finite. Setting $Z' = (Z-\mu)/\sqrt{v} \sim \mathcal{N}(0, I)$,

$$\begin{aligned} \frac{\partial}{\partial v} E_{\mu,v} \log m_\pi(Z; v) &= \frac{\partial}{\partial v}\int p(z|\mu, v) \log\left(m_\pi(z; v)\, dz\right) \\ &= \frac{\partial}{\partial v}\int p(z'|0, 1) \log\left(m_\pi(\sqrt{v}z' + \mu; v)\right)\, dz' \\ &= E_{Z'}\frac{(\partial/\partial v) m_\pi(\sqrt{v}Z' + \mu; v)}{m_\pi(\sqrt{v}Z' + \mu; v)} \end{aligned} \tag{3.92}$$

where

$$\begin{aligned} \frac{\partial}{\partial v} \quad m_\pi(\sqrt{v}z' + \mu; v) &= \frac{\partial}{\partial v}\int \frac{1}{(2\pi v)^{p/2}} \exp\left\{-\frac{\|\sqrt{v}z' + \mu - \mu'\|^2}{2v}\right\}\pi(\mu')\, d\mu' \\ = \frac{1}{(2\pi\, v)^{p/2}} &\int \left(-\frac{p}{2\,v} + \frac{\|z-\mu'\|^2}{2\,v^2} - \frac{\|z'\|^2}{2\,v} - \frac{2\langle z', \mu-\mu'\rangle}{2\,v^{3/2}}\right) p(z|\mu')\,\pi(\mu')\, d\mu' \\ = \frac{\partial}{\partial v} m_\pi(z; v) &- \int \frac{\langle z-\mu, z-\mu'\rangle}{2\,v^2}\, p(z|\mu')\,\pi(\mu')\, d\mu'. \end{aligned} \tag{3.93}$$

Note that

$$\nabla_z m_\pi(z, v) = \int \frac{-(z-\mu)}{v} p(z|\mu)\pi(\mu) d\mu \tag{3.94}$$

and

$$\begin{aligned}\Delta_z m_\pi(z, v) &= \int \left[\frac{-p}{v} + \frac{\|z-\mu\|^2}{v^2}\right] p(z|\mu)\pi(\mu)d\mu \\ &= 2\,\frac{\partial}{\partial v} m_\pi(z; v). \end{aligned} \tag{3.95}$$

It follows that

$$E_{Z'} \frac{(\partial/\partial v) m_\pi(\sqrt{v}Z' + \mu; v)}{m_\pi(\sqrt{v}Z' + \mu; v)} = E_{\mu,v}\left(\frac{1}{2}\frac{\Delta m_\pi(Z; v)}{m_\pi(Z; v)} + \frac{\langle Z-\mu, \nabla \log m_\pi(Z; v)\rangle}{2\,v}\right).$$

Hence, using Stein's identity,

$$\begin{aligned} E_{\mu,v}\left[\frac{(Z-\mu)^{\mathrm{T}}\nabla \log m_\pi(Z; v)}{2\,v}\right] &= E_{\mu,v}\left[\frac{1}{2}\Delta \log m_\pi(Z; v)\right] \\ &= E_{\mu,v}\left[\frac{1}{2}\left(\frac{\Delta m_\pi(Z; v)}{m_\pi(Z; v)} - \|\nabla \log m_\pi(Z; v)\|^2\right)\right], \end{aligned}$$

which is the desired result. □

Lemmas 3.17 and 3.18 gives a result regarding minimaxity and domination from George et al. (2006). This result reveals parallels to those on minimax estimation of mean under quadratic loss in Sect. 3.1.1. Its proof is contained in the proof of Theorem 3.17.

Theorem 3.16 *Assume that $m_\pi(z; v_x)$ is finite for any z in $\mathbb{R}^p$. If $\Delta m_\pi \le 0$ for all $v_w \le v \le v_x$, then the Bayesian predictive density $\hat{p}_\pi(y|x)$ is minimax and dominates $\hat{p}_U$ (when π is not the uniform itself). If $\Delta \pi \le 0$, then the Bayesian predictive density $\hat{p}_\pi(y|x)$ is minimax and dominates $\hat{p}_U$ (when π is uniform).*

The next result from Brown et al. (2008) illuminates the link between the two problems of estimating the predictive density under the Kullback-Leibler loss and estimating the mean under quadratic loss. The result expresses this link in terms of risk differences.

Theorem 3.17 *Suppose the prior $\pi(\mu)$ is such that the marginal $m_\pi(z; v)$ is finite for any $z \in \mathbb{R}^p$. Then,*

$$\mathscr{R}_{KL}(\mu, \hat{p}_U) - \mathscr{R}_{KL}(\mu, \hat{p}_\pi) = \frac{1}{2}\int_{v_w}^{v_x} \frac{1}{v^2}\left(\mathscr{R}_Q^v(\mu, X) - \mathscr{R}_Q^v(\mu, \hat{\mu}_{\pi,v})\right) dv. \tag{3.96}$$

Proof From (3.90) and (3.91) it follows

$$\begin{aligned}\mathscr{R}_{\mathrm{KL}}(\mu, \hat{p}_U) - \mathscr{R}_{\mathrm{KL}}(\mu, \hat{p}_\pi) &= \int_{v_w}^{v_x} -\frac{\partial}{\partial v} E_{\mu,v}[\log m_\pi(Z; v)]\; dv \\ &= \int_{v_w}^{v_x} E_{\mu,v}\left[2\,\frac{\Delta\sqrt{m_\pi(Z; v)}}{\sqrt{m_\pi(Z; v)}}\right] dv. \qquad (3.97)\end{aligned}$$

On the other hand, Stein (1981) showed that

$$\mathscr{R}_Q^v(\mu, X) - \mathscr{R}_Q^v(\mu, \hat{\mu}_{\pi,v}) = -4v^2 E_{\mu,v} \frac{\Delta\sqrt{m_\pi(Z; v)}}{\sqrt{m_\pi(Z; v)}}. \qquad (3.98)$$

Hence substituting (3.98) in the integral (3.97) gives (3.96). □

It is worth noting that using (3.88) and (3.96) leads to the following expression for the Kullback-Liebler risk of $\hat{p}_U$:

$$\begin{aligned}\frac{1}{2}\int_{v_w}^{v_x} \frac{1}{v^2}\left(\mathscr{R}_Q^v(\mu, X)\right)\; dv &= \frac{1}{2}\int_{v_w}^{v_x} \frac{p}{v}\; dv \\ &= \frac{p}{2}\log\frac{v_x}{v_w} \\ &= \frac{p}{2}\log\left(1 + \frac{v_x}{v_y}\right). \\ &= \mathscr{R}_{\mathrm{KL}}(\mu, \hat{p}_U)\,. \qquad (3.99)\end{aligned}$$

The area of predictive density estimation continues to develop. Recent research covers the case of restricted parameter (Fourdrinier et al. 2011), general α-divergence losses (Maruyama and Strawderman 2012; Boisbunon and Maruyama 2014), integrated $L1$ and $L2$ losses (Kubokawa et al. 2015, 2017). For a general review, see George and Xu (2010).