Chapter 2
Estimation of a Normal Mean Vector I

2.1 Introduction

This chapter is concerned with estimating the p-dimensional mean vector of a multivariate normal distribution under quadratic loss. Most of the chapter will be concerned with the case of a known covariance matrix of the form $\Sigma = \sigma^2 I_p$ and "usual quadratic loss," $L(\theta, \delta) = \|\delta - \theta\|^2 = (\delta - \theta)^{\mathrm{T}}(\delta - \theta)$. Generalizations to known general covariance matrix Σ, and to general quadratic loss, $L(\theta, \delta) = (\delta - \theta)^t Q(\delta - \theta)$, where Q is a $p \times p$ symmetric non-negative definite matrix will also be considered. Let $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ where σ^2 is assumed known and it is desired to estimate the unknown vector $\theta \in \mathbb{R}^p$. The "usual" estimator of θ is $\delta_0(X) = X$, in the sense that it is the maximum likelihood estimator (MLE), the uniformly minimum variance unbiased estimator (UMVUE), the least squares estimator (LSE), and under a wide variety of loss functions it is the minimum risk equivariant estimator (MRE), and is minimax. The estimator $\delta_0(X)$ is also admissible under a wide class of invariant loss functions if $p = 1$ or 2. However, Stein (1956) showed that X is inadmissible if $p \geq 3$ for the loss $L(\theta, \delta) = \|\delta - \theta\|^2$. This result was surprising at the time and has led to a large number of developments in multi-parameter estimation. One important aspect of this "Stein phenomenon" (also known as the Stein paradox at one time, see Efron and Morris 1977) is that it illustrates the difference between estimating one component at a time and simultaneously estimating the whole mean vector. Indeed, if we wish to estimate any particular component, θ_i, of the vector θ, then the estimator $\delta_{0i}(X) = X_i$ remains admissible whatever the value of p (see for example Lehmann and Casella (1998), Lemma 5.2.12). James and Stein (1961) showed that the estimator $\delta_a^{JS}(X) = (1 - a\,\sigma^2 / \|X\|^2)X$ dominates $\delta_0(X)$ for $p \geq 3$ provided $0 < a < 2\,(p-2)$. They also showed that the risk of $\delta_{p-2}^{JS}(X) = \left(1 - (p-2)\,\sigma^2 / \|X\|^2\right) X$ at $\theta = 0$ is equal to $2\,\sigma^2$ for all $p \geq 3$ indicating that substantial gain in risk over the usual estimator is possible for large p, since the risk of $\delta_0(X)$ is equal to the constant $p\,\sigma^2$.

D. Fourdrinier et al., *Shrinkage Estimation*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-02185-6_2

In Sect. 2.2, we will give some intuition into why improvement over $\delta_0(X)$ should be possible in higher dimensions and how much improvement might be expected. Section 2.3 is devoted to Stein's unbiased estimation of risk technique which provides the technical basis of many results in the area of multi-parameter estimation. Section 2.4 is devoted to establishing improved procedures such as the James-Stein estimator. In Sect. 2.5, we will provide a link between Stein's lemma and Stokes' theorem while, in Sect. 2.6, we will give some insight into Stein's phenomenon in terms of nonlinear partial differential operators.

2.2 Some Intuition into Stein Estimation

2.2.1 Best Linear Estimators

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

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

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

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

and solving for a gives

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

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

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

we may estimate a_{opt} as

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

and hence approximate the best linear "estimator"

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

by

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

This is in fact a James-Stein type estimator

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

which is close to the optimal James-Stein estimator (as we will see in Sect. 2.4 $\delta_{p-2}^{JS}(X)$ is optimal if X is normal). Hence, the James-Stein estimator can be viewed as an approximation to the best linear "estimator" that adapts to the value of $\|\theta\|^2$.

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

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

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

It is also interesting to note that normality plays no role in the above discussion indicating that we can expect James-Stein type estimators to improve on $\delta_0(X)$ in a fairly general location vector setting. This will be discussed further in Chaps. 5 and 6 for spherically symmetric distributions.

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

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

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

We will see later in Sect. 2.4 that this is also true for James-Stein-type estimators in the sense that there is very little savings in risk for large $\|\theta\|^2$ but substantial savings for small $\|\theta\|^2$ and especially so for large p.

2.2.2 *Some Geometrical Insight*

The argument here closely follows the discussion presented by Brandwein and Strawderman (1991a). We again suppose $E[X] = \theta \in \mathbb{R}^p$ and $Cov(X) = \sigma^2 I_p$ with σ^2 known. Since $E[\|X\|^2] = \|\theta\|^2 + p\,\sigma^2$, it seems that X is" too long" as an estimator of θ and that perhaps the projection of θ onto X or something close to it would be a better estimator than X. Again, the projection of θ onto X will depend on θ and so will not be a valid estimator, but perhaps we can find a reasonable approximation. Since the projection of θ on X has the form $(1-a)\,X$ we are trying to approximate the constant a. Note $E(\theta - X)^{\mathrm{T}}\theta = 0$, and hence we expect θ and $X - \theta$ to be nearly orthogonal which implies that we expect $0 < a < 1$.

In what follows, we assume that θ and $X-\theta$ are exactly orthogonal. The situation is shown in Fig. 2.1.

From the two right triangles in Fig. 2.1 we note

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

Since

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

reasonable approximations are

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

Hence we have as approximations

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

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

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

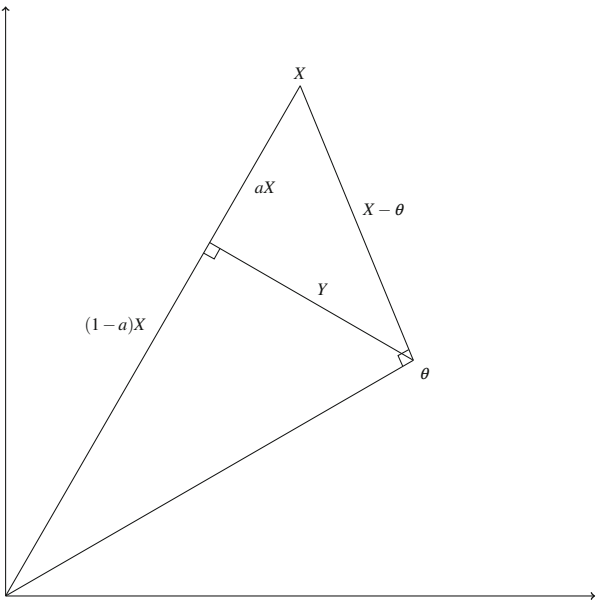

Fig. 2.1 Observation vector X in p dimensions with mean θ orthogonal to $X - \theta$

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

$$(1-a)\,X \cong \left(1 - \frac{p\,\sigma^2}{\|X\|^2}\right) X = \delta_p^{JS}(X)\,, \tag{2.1}$$

remarkably the same James-Stein estimator suggested in Sect. 2.2.1. Once again, note that normality plays no role in the discussion. Stein (1962) gave a similar geometric arguement to contruct confidence sets for θ, centred at (2.1), as the orthogonal projection of θ on X. For more on the geometrical explanation of the inadmissibility of X as a point estimator see Brown and Zhao (2012).

2.2.3 The James-Stein Estimator as an Empirical Bayes Estimator

Assume in this subsection that $X \sim \mathcal{N}_p(\theta, \sigma^2\, I_p)$ with (σ^2 known) and that the prior distribution on θ is $\mathcal{N}_p(0, \tau^2\, I_p)$. As indicated in Sect. 1.4, the Bayes estimator of θ for quadratic loss is the posterior mean of θ given by $\delta(X) = E[\theta \mid X] = (1 - \sigma^2/(\tau^2 + \sigma^2))X$.

If we now assume that τ^2 is unknown we can derive an empirical Bayes estimator as follows; the marginal distribution of X is $\mathcal{N}_p(0, (\sigma^2 + \tau^2)\, I_p)$ and hence $\|X\|^2$,

which is distributed as $(\sigma^2+\tau^2)$ times a chi-square with p degrees of freedom, is a complete sufficient statistic for $\sigma^2+\tau^2$. It follows that $(p-2)/\|X\|^2$ is the UMVUE of $1/(\sigma^2+\tau^2)$ and that $\delta^{JS}_{p-2}(X) = (1-(p-2)\,\sigma^2/\|X\|^2)\,X$ can be viewed as an empirical Bayes estimator of θ.

Here we have explicitly used the assumption of normality but a somewhat analogous argument will be given in Sect. 5.1 for a general multivariate location family.

2.3 Improved Estimators via Stein's Lemma

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

Through

$$\begin{aligned} L(\theta,\delta) &= \|X+\sigma^2 g(X)-\theta\|^2 \\ &= \|X-\theta\|^2 + \sigma^4\,\|g(X)\|^2 + 2\,\sigma^2\,(X-\theta)^{\mathrm{T}} g(X)\,, \end{aligned} \tag{2.2}$$

we will see that the risk of δ is finite if and only if $E_\theta[\|g(X)\|^2] < \infty$. Indeed, considering the expectation of the cross product term in (2.2), we have

$$E_\theta[|(X-\theta)^{\mathrm{T}} g(X)|] \le \big(E_\theta[\|(X-\theta)\|^2]\big)^{1/2}\big(E_\theta[\|g(X)\|^2]\big)^{1/2}\,,$$

by the Cauchy-Schwarz inequality. Therefore, as $E_\theta[\|(X-\theta)\|^2] < \infty$, it suffices that $E_\theta[\|g(X)\|^2] < \infty$ to have $E_\theta[\|X+g(X)-\theta\|^2] < \infty$, that is, $R(\theta, X+g(X)) < \infty$.

Conversely, assume that $R(\theta, X+g(X)) < \infty$. As

$$\begin{aligned} \|g(X)\|^2 &= \|X+g(X)-\theta-(X-\theta)\|^2 \\ &= \|X+g(X)-\theta\|^2 + \|X-\theta\|^2 - 2\,(X-\theta)^{\mathrm{T}}(X+g(X)-\theta) \end{aligned}$$

then applying the above argument gives $E_\theta[\|g(X)\|^2] < \infty$ since, by assumption, $E_\theta[\|X+g(X)-\theta\|^2] < \infty$, $E_\theta[\|(X-\theta)\|^2] < \infty$ and hence using again the Cauchy-Schwarz inequality

$$E_\theta[|(X-\theta)^{\mathrm{T}}(X+g(X)-\theta)|] \le \big(E_\theta[\|(X-\theta)\|^2]\big)^{1/2}\big(E_\theta[\|X+g(X)-\theta\|^2]\big)^{1/2}\,.$$

Under this fineteness condition the risk function of δ is given by

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

Stein's lemma in (2.7) below allows an alternative expression for the last expectation, that is, $E_\theta[(X-\theta)^{\mathrm{T}} g(X)] = \sigma^2 E_\theta[\mathrm{div} g(X)]$ where $\mathrm{div} g(X) = \sum_{i=1}^p \frac{\partial}{\partial X_i} g_i(X)$ under suitable conditions on g. The great advantage that Stein's lemma gives is that the risk function can be expressed as the expected value of a function of X only (and not θ), that is,

$$R(\theta, \delta) = E_\theta[p\,\sigma^2 + \sigma^4 \|g(X)\|^2 + 2\,\sigma^4\,\mathrm{div} g(X)], \tag{2.3}$$

and hence the expression

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

can be interpreted as an unbiased estimate of the risk of δ (see Corollary 2.1 (3)). Actually, as X is a complete sufficient statistic, this unbiased estimator is the uniformly minimum variance unbiased estimator of the risk. To see that $E_\theta[(X-\theta)^{\mathrm{T}} g(X)] = \sigma^2 E_\theta[\mathrm{div} g(X)]$ is quite easy if g is sufficiently smooth. Suppose first that $p = 1$ and g is absolutely continuous. We show in Sect. A.5 in the Appendix that $\lim_{x\to\pm\infty} g(x)\exp\{-(x-\theta)^2/2\sigma^2\} = 0$ as soon as $E_\theta[|g'(X)|] < \infty$ (see also Hoffmann 1992 where g is assumed to be continuously differentiable). Then a simple integration by parts gives

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

In higher dimensions, let $g = (g_1, \ldots, g_p)$ be a function from $\mathbb{R}^p$ into $\mathbb{R}^p$. Also, for any $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ and for fixed $i = 1, \ldots, p$, set $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$ and, with a slight abuse of notation, $x = (x_i, x_{-i})$. Then, using the independence of X_i and X_{-i}, we have

$$\begin{aligned}
E_\theta[(X_i-\theta_i)\, g_i(X)] &= E_\theta\Big[E_\theta[(X_i-\theta_i)\, g_i(X_i, X_{-i}) |\, X_{-i}]\Big] \\
&= E_\theta\Big[E_\theta[\sigma^2\, \partial_i g_i(X_i, X_{-i}) |\, X_{-i}]\Big] \\
&= \sigma^2 E_\theta[\partial_i g_i(X)]\,.
\end{aligned}$$

Now, summing on i gives $E_\theta[(X-\theta)^{\mathrm{T}} g(X)] = \sigma^2\, E_\theta[\mathrm{div} g(X)]$.

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

$$\delta_a^{JS}(X) = \left(1 - \frac{a\,\sigma^2}{\|X\|^2}\right) X \tag{2.4}$$

where the coordinate functions of $g(X) = (a\,\sigma^2/\|X\|^2)\,X$ are not smooth, since g explodes at 0. For this reason, Stein considered a weaker regularity condition for his identity to hold, that he called almost differentiability. In his proof, he essentially required that $g(x) = (g_1(x), g_2(x), \ldots, g_p(x))$ be such that, for each $i = 1, \ldots, p$, the coordinate $g_i(x)$ is absolutely continuous in x_i for almost every x_{-i}. Formally, he stated: "A function h from $\mathbb{R}^p$ into $\mathbb{R}^p$ is said to be almost differentiable if there exists a function $\nabla h = (\nabla_1 h, \ldots, \nabla_p h)$ from $\mathbb{R}^p$ into $\mathbb{R}^p$ such that, for all $z \in \mathbb{R}^p$,

$$h(x+z) - h(x) = \int_0^1 z^{\mathrm{T}}\,\nabla h(x + t\,z)\,dt\,, \tag{2.5}$$

for almost all $x \in \mathbb{R}^p$. A function $g = (g_1, \ldots, g_p)$ from $\mathbb{R}^p$ into $\mathbb{R}^p$ is said to be almost differentiable if all its coordinate functions g_i's are" (see Sect. A.1 in the Appendix for a detailed discussion).

We will establish Stein's identity under the weaker notion of weak differentiability which is of more common use in analysis and also in statistics (see e.g. Johnstone 1988). To this end, recall that the space of functions h from $\mathbb{R}^p$ into $\mathbb{R}$ such that h is locally integrable is defined by

$$L^1_{loc}(\mathbb{R}^p) = \left\{ h : \mathbb{R}^p \to \mathbb{R} \mid \int_K |h(x)|\,dx < \infty \quad \forall K \subset \mathbb{R}^p \text{ with } K \text{ compact} \right\}.$$

Definition 2.1 A locally integrable function h from $\mathbb{R}^p$ into $\mathbb{R}$ is said to be weakly differentiable if there exist p locally integrable functions $\partial_1 h, \ldots \partial_p h$ such that, for any $i = 1, \ldots, p$,

$$\int_{\mathbb{R}^p} h(x)\,\frac{\partial \varphi}{\partial x_i}(x)\,dx = -\int_{\mathbb{R}^p} \partial_i h(x)\,\varphi(x)\,dx \tag{2.6}$$

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

Note that weak differentiability is a global, not local, property. The functions $\partial_i h$ in Definition 2.1 are denoted, as the usual derivatives, by $\partial/\partial x_i$. The vector $\partial h = (\partial_1 h, \ldots, \partial_p h) = (\partial h/\partial x_1, \ldots, \partial h/\partial x_p)$ denotes the weak gradient of h and the scalar $\mathrm{div} g = \sum_{i=1}^p \partial_i g_i$ denotes the weak divergence of g. The following proposition establishes a link between weak differentiability and those aspects of almost differentiability that Stein used (and we will use) in the proof of Stein's lemma.

Proposition 2.1 (Ziemer 1989) *Let h be a locally integrable function from $\mathbb{R}^p$ into $\mathbb{R}$. Then h is weakly differentiable if and only if there exists a representative*

h_0 which is equal to h almost everywhere such that, for any $i = 1, \ldots, p$, the function $h_0(x_i, x_{-i})$ is absolutely continuous in x_i for almost all values of x_{-i} and whose (classical) partial derivatives belong to $L^1_{loc}(\mathbb{R}^p)$. Also the classical partial derivatives of h_0 coincide with the weak partial derivatives of h almost everywhere.

Proposition 2.1 is essentially Theorem 2.1.4. of Ziemer (1989) who deals with functions h in $L^1(\Omega)$ where Ω is an open set of $\mathbb{R}^p$ (and, more generally, in $L^q(\Omega)$ with $q \geq 1$). However, his proof relies only on local integrability of h and its partial derivatives. So, the apparently stronger statement of Proposition 2.1 follows from his arguments. See also Theorem 8.27 of Bressan (2012).

As indicated in Proposition 2.1, the key feature of weak differentiability is the local integrability of the function and of all its partial derivatives. For the functions h of interest to us, the representative h_0 is the function itself so that the weak differentiability follows from the local integrability of h and its derivatives and its absolute continuity along almost all lines parallel to the axes. In particular, as the weak partial derivative is unique up to pointwise almost everywhere equivalence, the weak partial derivative of a continuously differentiable function coincides with the usual derivative (see e. g. Hunter 2014, Chap. 3).

As an example, consider the shrinkage factor $h(x) = x/\|x\|^2$ of the James-Stein estimator in (2.4). In Sect. A.2 in the Appendix, we show that h is weakly differentiable if and only if $p \geq 3$ and that $\operatorname{div} h(x) = (p-2)/\|x\|^2$. We also show that h is not almost differentiable in the sense of Stein given above. This last fact is due to the requirement that h be absolutely continuous in all directions while weak differentiability, in contrast, only requires absolute continuity in directions parallel to the axes. Again we note that Stein only used absolute continuity in the coordinates directions.

We give now a precise statement of Stein's lemma for weakly differentiable functions along the lines of Stein (1981). Note that we will see, in Sect. 2.5, that it is closely related to Stokes' theorem, which will provide an alternative proof.

Theorem 2.1 (Stein's lemma) *Let $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ and let g be a weakly differentiable function from $\mathbb{R}^p$ into $\mathbb{R}^p$. Then*

$$E_\theta[(X-\theta)^{\mathrm{T}} g(X)] = \sigma^2 \, E_\theta[\operatorname{div} g(X)]\,, \tag{2.7}$$

provided, for any $i = 1, \ldots, p$, either

$$E_\theta[|(X_i - \theta_i) g_i(X)|] < \infty \quad \text{or} \quad E_\theta[|\partial_i g_i(X)|] < \infty\,. \tag{2.8}$$

Formula (2.7) is often referred to as Stein's identity.

Proof Let $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ and set

$$\varphi(x) = \frac{\|x-\theta\|^2}{2\sigma^2} \quad \text{and} \quad \phi(x) = \frac{1}{(2\pi\sigma^2)^{p/2}} \exp(-\varphi(x))\,.$$

Equality (2.7) is equivalent to

$$E_\theta[\nabla\varphi(X)^{\mathrm{T}} g(X)] = \sigma^2\, E_\theta[\mathrm{div} g(X)]\,. \tag{2.9}$$

For fixed $i = 1, \ldots, p$, set $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$ and, with a slight abuse of notation, set $x = (x_i, x_{-i})$. Note that

$$\frac{\partial\phi(x)}{\partial x_i} = -\frac{\partial\varphi(x)}{\partial x_i}\,\phi(x)$$

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

$$\phi(x) = \int_{-\infty}^{x_i} -\frac{\partial\varphi(\tilde{x}_i, x_{-i})}{\partial\tilde{x}_i}\,\phi(\tilde{x}_i, x_{-i})\,d\tilde{x}_i = \int_{x_i}^{\infty} \frac{\partial\varphi(\tilde{x}_i, x_{-i})}{\partial\tilde{x}_i}\,\phi(\tilde{x}_i, x_{-i})\,d\tilde{x}_i\,, \tag{2.10}$$

noticing that, by assumption, $\lim_{|x_i|\to\infty}\varphi(x_1, \ldots, x_p) = \infty$ implies

$$\lim_{|x_i|\to\infty}\phi(x_i, x_{-i}) = \frac{1}{(2\,\pi\,\sigma^2)^{p/2}}\,\lim_{|x_i|\to\infty}\exp\big(-\varphi(x_i, x_{-i})\big) = 0\,. \tag{2.11}$$

Fixing $i \in \{1, \ldots, p\}$ and assuming first $E_\theta[|\partial_i g_i(X)|] < \infty$, we can write using (2.10), for almost every x_{-i},

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{\partial g_i(x_i, x_{-i})}{\partial x_i}\,\phi(x_i, x_{-i})\,dx_i \\
&= \int_{-\infty}^{0} \frac{\partial g_i(x_i, x_{-i})}{\partial x_i} \int_{-\infty}^{x_i} -\frac{\partial\varphi(\tilde{x}_i, x_{-i})}{\partial\tilde{x}_i}\,\phi(\tilde{x}_i, x_{-i})\,d\tilde{x}_i\,dx_i \\
&\quad + \int_{0}^{\infty} \frac{\partial g_i(x_i, x_{-i})}{\partial x_i} \int_{x_i}^{\infty} \frac{\partial\varphi(\tilde{x}_i, x_{-i})}{\partial\tilde{x}_i}\,\phi(\tilde{x}_i, x_{-i})\,d\tilde{x}_i\,dx_i \\
&= \int_{-\infty}^{0} -\frac{\partial\varphi(\tilde{x}_i, x_{-i})}{\partial\tilde{x}_i}\,\phi(\tilde{x}_i, x_{-i}) \int_{\tilde{x}_i}^{0} \frac{\partial g_i(x_i, x_{-i})}{\partial x_i}\,dx_i\,d\tilde{x}_i \\
&\quad + \int_{0}^{\infty} \frac{\partial\varphi(\tilde{x}_i, x_{-i})}{\partial\tilde{x}_i}\,\phi(\tilde{x}_i, x_{-i}) \int_{0}^{\tilde{x}_i} \frac{\partial g_i(x_i, x_{-i})}{\partial x_i}\,dx_i\,d\tilde{x}_i\,.
\end{aligned} \tag{2.12}$$

Now, according to Proposition 2.1, as g is weakly differentiable, we may assume without loss of generality that, for each $i = 1, \ldots, p$, the function $g_i(x_i, x_{-i})$ is absolutely continuous in x_i for almost all values of x_{-i} so that

$$-\int_{\tilde{x}_i}^{0} \frac{\partial g_i(x_i, x_{-i})}{\partial x_i}\, dx_i = \int_{0}^{\tilde{x}_i} \frac{\partial g_i(x_i, x_{-i})}{\partial x_i}\, dx_i = g_i(\tilde{x}_i, x_{-i}) - g_i(0, x_{-i})\,.$$

Then (2.12) becomes

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{\partial g_i(x_i, x_{-i})}{\partial x_i}\, \phi(x_i, x_{-i})\, dx_i \\
=&\int_{-\infty}^{\infty} \frac{\partial \varphi(x_i, x_{-i})}{\partial x_i}\, \phi(x_i, x_{-i})\, [g_i(x_i, x_{-i}) - g_i(0, x_{-i})]\, dx_i \\
=&\int_{-\infty}^{\infty} \frac{\partial \varphi(x_i, x_{-i})}{\partial x_i}\, \phi(x_i, x_{-i})\, g_i(x_i, x_{-i})\, dx_i\,,
\end{aligned}$$

since, using again (2.11),

$$-\int_{-\infty}^{\infty} \frac{\partial \varphi(x_i, x_{-i})}{\partial x_i}\, \phi(x_i, x_{-i})\, dx_i = \int_{-\infty}^{\infty} \frac{\partial \phi(x_i, x_{-i})}{\partial x_i}\, dx_i = 0\,.$$

Finally, integrating with respect to x_{-i} gives

$$\begin{aligned}
E_\theta\left[\frac{\partial g_i(X)}{\partial x_i}\right] &= \int_{\mathbb{R}^p} \frac{\partial g_i(x_i, x_{-i})}{\partial x_i}\, \phi(x_i, x_{-i})\, dx_i\, dx_{-i} \\
&= \int_{\mathbb{R}^p} \frac{\partial \varphi(x_i, x_{-i})}{\partial x_i}\, \phi(x_i, x_{-i})\, g_i(x_i, x_{-i})\, dx_i\, dx_{-i} \\
&= E_\theta\left[\frac{\partial \varphi(X)}{\partial x_i}\, g_i(X)\right]
\end{aligned}$$

and hence, summing on i gives (2.9), which is the desired result.

To show (2.7) assuming $E_\theta[|(X_i - \theta_i)^{\mathrm{T}} g_i(X)|] < \infty$ for $i \in \{1, \ldots, p\}$, it suffices to essentially reverse the steps in the above argument. □

The following corollary is immediate from Stein's lemma and the above discussion. Recall that $L(\theta, d) = \|d - \theta\|^2$, $R(\theta, \delta) = E_\theta[L(\theta, \delta(X))] = E_\theta[\|\delta(X) - \theta\|^2]$, and $E_\theta[\|g(X)\|^2] < \infty$ implies that for any $i = 1, \ldots, p$, $E_\theta[|(X_i - \theta_i)\, g_i(X)|] < \infty$.

Corollary 2.1 *Let $g(X)$ be a weakly differentiable function from $\mathbb{R}^p$ into $\mathbb{R}^p$ such that $E_\theta[\|g(X)\|^2] < \infty$. Then*

(1) $R(\theta, X + \sigma^2 g(X)) = E_\theta[p\,\sigma^2 + \sigma^4 (\|g(X)\|^2 + 2\,\mathrm{div} g(X))]$;
(2) $\delta(X) = X + \sigma^2 g(X)$ *is minimax as soon as* $\|g(X)\|^2 + 2\,div\ g(X) \le 0$ *a.e. and dominates X provided there is strict inequality on a set of positive measure; and*
(3) $p\,\sigma^2 + \sigma^4 (\|g(X)\|^2 + 2\,\mathrm{div} g(X))$ *is an unbiased estimator (in fact the UMVUE) of* $R(\theta, X + \sigma^2 g(X))$.

We note once again that $\delta(X)$ is minimax since it dominates (or ties) the minimax estimator X. In the next few sections we apply the above corollary to show domination of the James-Stein estimators and several others over the usual estimator in three and higher dimensions.

2.4 James-Stein Estimators and Other Improved Estimators

In this section, we apply the integration by parts results of Sect. 2.3 to obtain several classes of estimators that dominate the classical minimax estimator $\delta_0(X)$ in dimension 3 and higher. The estimators of James and Stein, Baranchik, and certain estimators shrinking toward subspaces are the main application of this section. Bayes (generalized, proper, and pseudo) are considered in Chap. 3. Throughout this section, except for Theorem 2.4, let $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ and loss be $L(\theta, \delta) = \|\delta - \theta\|^2$. According to Corollary 2.1 it suffices to find weakly differentiable functions g from $\mathbb{R}^p$ into $\mathbb{R}^p$ such that $E_\theta[\|g(X)\|^2] < \infty$ and $\|g(X)\|^2 + 2 \text{ div } g(X) \leq 0$ (with strict inequality on a set of positive measure) in order to show that $\delta(X) = X + \sigma^2 g(X)$ dominates X.

2.4.1 James-Stein Estimators

The class of James-Stein estimators is given by

$$\delta_a^{JS}(X) = \left(1 - \frac{a\,\sigma^2}{\|X\|^2}\right) X. \tag{2.13}$$

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

Theorem 2.2 *Under the above model*

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

$$R(\theta, \delta_a^{JS}) = p\,\sigma^2 + \sigma^4\,(a^2 - 2\,a\,(p-2))E_\theta\left[\frac{1}{\|X\|^2}\right] \tag{2.14}$$

for $p \geq 3$.

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

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

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

Proof Observe that $\delta_a^{JS}(X) = X + \sigma^2 g(X)$ where $g(X) = -a/\|X\|^2\,X$. As noted in Sect. 2.3, $g(X)$ is weakly differentiable if $p \geq 3$. Also $E_\theta[\|g(X)\|^2] = a^2\,E_\theta[1/\|X\|^2]$ is finite if $p \geq 3$ since $\|X\|^2/\sigma^2$ has a non-central χ^2 distribution with p degrees of freedom and non-centrality parameter $\lambda = \|\theta\|^2/2\sigma^2$. Indeed

by the usual Poisson representation of a non-central χ^2, we have $\|X\|^2/\sigma^2 \mid K \sim \chi^2_{p+2K}$ where $K \sim$ Poisson $(\lambda = \|\theta\|^2/2\sigma^2)$ and hence,

$$E_\theta\left[\frac{\sigma^2}{\|X\|^2}\right] = E_\lambda\left[E\left[\frac{1}{\chi^2_{p+2K}}\middle| K\right]\right] = E_\lambda\left[\frac{1}{p+2K-2}\right] \leq \frac{1}{p-2} < \infty \tag{2.15}$$

if $p > 2$.

Also, according to (A.18), for any $x \neq 0$,

$$\operatorname{div}\left(\frac{x}{\|x\|^2}\right) = \frac{p-2}{\|x\|^2}. \tag{2.16}$$

Hence,

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

and by Corollary 2.1, for $p \geq 3$,

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

This proves (1).

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

$$R(\theta, \delta_a^{JS}) < p\sigma^2 = R(\theta, \delta_0). \tag{2.17}$$

The minimaxity claim for $0 \leq a \leq 2(p-2)$ follows by replacing $<$ by $\leq$ in (2.17). It is interesting to note that $R\left(\theta, \delta^{JS}_{2(p-2)}\right) \equiv R(\theta, \delta_0) \equiv p\sigma^2$ and, more generally, $R(\theta, \delta_{2(p-2)-a}) \equiv R\left(\theta, \delta_a^{JS}\right)$.

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

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

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

□

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

In Theorem 2.2, the fact that $p \geq 3$ is crucial (which is coherent with the admissibility of X for $p = 1$ and $p = 2$). Actually, a crucial part of the proof uses Stein's identity, which fails to hold if $p = 1, 2$ with $h(x) = x/\|x\|^2$. Indeed, when $p = 1$, $h(x) = 1/x$ and $\mathrm{div}(x) = -1/x^2$ so that $E_0[X^{\mathrm{T}}h(X)] = 1$ and $E_0[\mathrm{div}h(X)] = -\infty$. When $p = 2$, we also have $E_0[X^{\mathrm{T}}h(X)] = 1$ while $E_0[\mathrm{div}h(X)] = 0$ since, for any $x \neq 0$, $\mathrm{div}h(x) = 0$. It is interesting to note that, while the divergence of h exists and is 0 almost everywhere, h is not weakly differentiable since its partial derivatives are not locally integrable as shown in Sect. A.1 in the Appendix.

We may use (2.15) to give upper and lower bounds for the risk of δ_a^{JS} based on the following lemma.

Lemma 2.1 *Let $K \sim$ Poisson(λ). Then, for $b \geq 1$, we have*

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

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

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

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

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

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

Hence the third inequality follows. □

The following bounds on the risk of δ_a^{JS} follow directly from (2.14), (2.15) and Lemma 2.1.

Corollary 2.2 (Hwang and Casella 1982) *For $p \geq 4$ and $0 \leq a \leq 2\,(p-2)$, we have*

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

We note in passing that the upper bound may be improved at the cost of added complexity by using the second inequality in Lemma 2.1. The improved upper bound has the advantage that it is exact at $\theta = 0$. The lower bound is also valid for $p = 3$ and is also exact at $\theta = 0$.

2.4.2 *Positive-Part and Baranchik-Type Estimators*

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

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

$$\delta_a^{JS+}(X) = \left(1 - \frac{a\,\sigma^2}{\|X\|^2}\right)_+ X \tag{2.18}$$

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

$$\delta_{a,r}^{B}(X) = \left(1 - \frac{a\,\sigma^2\,r(\|X\|^2)}{\|X\|^2}\right) X \tag{2.19}$$

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

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

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

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

Theorem 2.3 *The estimator given by* (2.19) *with* $r(\cdot)$ *absolutely continuous, is minimax for* $p \geq 3$ *provided*

(1) $0 < a \leq 2\,(p-2)$*;*
(2) $0 \leq r(\cdot) \leq 1$*; and*
(3) $r(\cdot)$ *is nondecreasing.*

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

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

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

Hence,

$$\begin{aligned} &\|g(X)\|^2 + 2\,\text{div} g(X) \qquad (2.20)\\ &= \frac{a^2\, r^2(\|X\|^2)}{\|X\|^2} - \frac{2\,a\,(p-2) r(\|X\|^2)}{\|X\|^2} - 4\,a\,r'(\|X\|^2) \\ &\leq \frac{r(\|X\|^2)}{\|X\|^2}(a^2 - 2\,a\,(p-2)) - 4\,a\,r'(\|X\|^2) \\ &\leq 0\,. \end{aligned}$$

The first inequality being satisfied by Conditions (2) while the last inequality uses all of Conditions (1), (2), and (3). Hence, minimaxity follows from Corollary 2.1. Under the additional conditions, it is easy to see that the above inequalities become strict on a set of positive measure so that domination over X is guaranteed. □

As an example of a dominating Baranchik-type estimator consider

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

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

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

Theorem 2.4 *Suppose X has a density $f(x-\theta)$ in $\mathbb{R}^p$ such that the function f is symmetric and unimodal in each coordinate separately for each fixed value of the other coordinates. Then, for any finite risk estimator of θ of the form*

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

the positive-part estimator

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

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

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

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

Therefore it suffices to show that, for any nonnegative function $H(X_1^2, \ldots, X_p^2)$, $\theta_i E_\theta[X_i H(X_1^2, \ldots, X_p^2)] \geq 0$. This follows by symmetry if whenever $\theta_i \geq 0$, then $E_\theta[X_i \mid X_i^2 = t_i^2, X_j = t_j\ j \neq i] \geq 0$ for all i ($1 \leq i \leq p$) and all $(t_1, \ldots, t_p)$. However this expression is proportional to

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

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

For the remainder of this current section we return to the assumption that $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$.

The positive-part James-Stein estimators are inadmissible because of a lack of smoothness which precludes them from being generalized Bayes. The Baranchik class however contains "smooth" estimators which are generalized (and even

proper) Bayes and admissible. Baranchik-type estimators will play an important role in Chap. 3.

We close this subsection with a generalization of the Baranchik result in Theorem 2.3. It is apparent from the proof of the theorem that it is only necessary that the second expression in (2.20) be nonpositive (and negative on a set of positive measure) in order for $\delta(X)$ to dominate X. In particular it is not necessary that $r(\cdot)$ be nondecreasing. The following result (see Efron and Morris 1976 and Fourdrinier and Ouassou 2000) gives a necessary and sufficient condition for the unbiased estimator of risk difference, $R(\theta, \delta) - R(\theta, X)$, for $\delta(X) = \left(1 - a\, r(\|X\|^2)/\|X\|^2\right) X$, to be nonpositive. The proof is by direct calculation.

Lemma 2.2 *Let $g(X) = -a\left(r(\|X\|^2)/\|X\|^2\right) X$ where $r(y)$ is an absolutely continuous function from $\mathbb{R}^+$ into $\mathbb{R}$. Then on the set where $r(y) \neq 0$,*

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

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

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

Corollary 2.3 *Suppose $\delta(X) = \left(1 - a\, r(\|X\|^2)/\|X\|^2\right) X$ with*

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

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

Proof The result follows from Corollaries 2.1 and 2.2 by noting that

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

□

An application of Corollary 2.3 gives a useful class of dominating estimators due to Alam (1973).

Corollary 2.4 *Let $\delta(X) = \left(1 - a\, f(\|X\|^2)/(\|X\|^2)^{\tau+1}\right) X$ where $f(y)$ is nondecreasing and absolutely continuous and where $0 \leq a\, f(y)/y^\tau < 2\,(p - 2 - 2\tau)$ for some $\tau \geq 0$. Then $\delta(X)$ is minimax and dominates X if $0 < a\, f(y)/y^\tau$ on a set of positive measure.*

Proof The proof follows from Corollary 2.3 by letting

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

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

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

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

A simple example of a minimax Baranchik-type estimator with a nonmonotone $r(\cdot)$ is given by $r(y) = y^{1-\tau}/(1+y)$ for $0 < \tau < 1$ and $0 < a < 2(p-2-2\,\tau)$. To see this, apply Corollary 2.4 with $f(y) = y/(1+y)$ and note that $f(y)$ is increasing and $0 \leq f(y)/y^{\tau} = r(y) \leq 1$. Note also that $r'(y) = y^{-\tau}[(1-\tau) - \tau\ y]/(1+y)^2$, hence $r(y)$ is increasing for $0 < y < (1-\tau)/\tau^{-1}$ and decreasing for $y > (1-\tau)/\tau^{-1}$.

We will use the above corollaries in Chap. 3 to establish minimaxity of certain Bayes and generalized Bayes estimators.

2.4.3 Unknown Variance

In the development above, it was tacitly assumed that the covariance matrix was known and equal to a multiple of the identity matrix $\sigma^2\, I_p$. Typically, this covariance is unknown and should be estimated. The next result extends Stein's identity (2.7) to the case where it is of the form $\sigma^2\, I_p$ with σ^2 unknown.

Lemma 2.3 *Let $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ and let S be a nonnegative random variable independent of X such that $S \sim \sigma^2 \chi^2_k$. Denoting by E_{θ,σ^2} the expectation with respect to the joint distribution of (X, S), we have the following two results, provided the corresponding expectations exist:*

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

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

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

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

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

Proof Part (1) is Stein's lemma, from Theorem 2.1. Part (2) can be seen as a particular case of Lemma $1(ii)$ (established for elliptically symmetric distributions) of Fourdrinier et al. (2003), although we will present a direct proof. Part (2) also follows from well known identities for chi-square distributions.

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

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

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

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

since

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

□

The following theorem provides an estimate of risk in the setting of an unknown variance when the loss is given by

$$\frac{\|\delta - \theta\|^2}{\sigma^2}\,. \tag{2.21}$$

Theorem 2.5 *Let* $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ *where* θ *and* σ^2 *are unknown and* $p \geq 3$ *and let* S *be a nonnegative random variable independent of* X *such that* $S \sim \sigma^2 \chi^2_k$. *Consider an estimator of* θ *of the form* $\varphi(X, S) = X + S\, g(X, S)$ *with* $E_{\theta,\sigma^2}[S^2\, ||g(X, S)||^2] < \infty$, *where* E_{θ,σ^2} *denotes the expectation with respect to the joint distribution of* (X, S).

Then an unbiased estimator of the risk under loss (2.21) *is*

$$\delta_0(X, S) = p + S\left\{(k+2)\,||g(X,S)||^2 + 2\,\mathrm{div}_X g(X,S) + 2\,S\frac{\partial}{\partial S}||g(X,S)||^2\right\}. \tag{2.22}$$

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

$$\frac{1}{\sigma^2}||X-\theta||^2 + 2\,\frac{S}{\sigma^2}(X-\theta)^{\mathrm{T}} g(X,S) + \frac{S^2}{\sigma^2}||g(X,S)||^2. \tag{2.23}$$

Clearly,

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

and Lemma 2.3 (1) and (2) express, respectively, that

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

With $h(x, s) = s^2\,||g(x, s)||^2$ we have

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

Therefore $R(\theta, \varphi) = E_{\theta,\sigma^2}[\delta_0(X, S)]$ with $\delta_0(X, S)$ given in (2.22), which means that $\delta_0(X, S)$ is an unbiased estimator of the risk $||\varphi(X, S) - \theta||^2/\sigma^2$. □

Corollary 2.5 *Under condition of Theorem* 2.5*, if, for any* $(x, s) \in \mathbb{R}^p \times \mathbb{R}_+$,

(i) $\partial/\partial s\,\|g(x,s)\|^2 \le 0$ *and*
(ii) $(k+2)\,||g(x,s)||^2 + 2\,\mathrm{div}_x g(x,s) + 2 \le 0$,

then $\varphi(X, S)$ *is minimax. It dominates* X *if either inequality is strict on a set of positive measure.*

In the following corollary, we consider an extension of the Baranchik form in Theorem 2.3.

Corollary 2.6 *Let*

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

If r *is nondecreasing and if* $0 < a\,r(\|X\|^2/S) < 2\,(p-2)/(k+2)$*, then* $\delta(X, S)$ *dominates* X *and is minimax.*

Proof Straightforward calculations show that the term in curly brackets in (2.22) equals

$$a\,\frac{r(\|X\|^2/S)}{\|X\|^2}\,((k+2)\,a\,r(\|X\|^2/S)-2\,(p-2))$$
$$-4\,a\,\frac{r'(\|X\|^2/S)}{S}(1+a\,r(\|X\|^2/S)). \tag{2.24}$$

Therefore, if $0 < a\,r(\|X\|^2/S) < 2\,(p-2)/(k+2)$, then $\delta(X, S)$ dominates X and is minimax. □

Note that, in the case $r \equiv 1$, the bound on the constant a is $2\,(p-2)/(k+2)$. This is the estimator developed by James and Stein (1961) using direct methods.

2.4.4 Estimators That Shrink Toward a Subspace

We saw in Sect. 2.4.1, when σ^2 is known, that the James-Stein estimator shrinks toward $\theta = 0$ and that substantial risk savings are possible if θ is in a neighborhood of 0. If we feel that θ is close to some other value, say θ_0, a simple adaptation of the James-Stein estimator that shrinks toward θ_0 may be desirable. Such an estimator is given by

$$\delta^{JS}_{a,\theta_0}(X) = \theta_0 + \left(1 - \frac{a\,\sigma^2}{\|X-\theta_0\|^2}\right)(X-\theta_0). \tag{2.25}$$

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

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

Hence, for $p \geq 3$, δ^{JS}_{a,θ_0} dominates X and is minimax for $0 < a < 2\,(p-2)$, and $a = p-2$ is the optimal choice of a. Furthermore the risk of $\delta^{JS}_{a,\theta_0}(X)$ at $\theta = \theta_0$ is $2\,\sigma^2$ and so large gains in risk are possible in a neighborhood of θ_0. The same argument establishes the fact that, for any estimator, $\delta(X)$, we have $R(\theta, \theta_0 + \delta(X-\theta_0)) = R(\theta-\theta_0, \delta(X))$. Hence any of the minimax estimators of Sects. 2.4.1 and 2.4.2 may be modified in this way and minimaxity will be preserved.

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

$$P_V X + \left(1 - \frac{a\,\sigma^2 r(\|X - P_V X\|^2)}{\|X - P_V X\|^2}\right)(X - P_V X) \tag{2.26}$$

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

A standard canonical representation is helpful. Suppose V is an s-dimensional linear subspace of $\mathbb{R}^p$ and $V^\perp$ is the $p - s$ dimensional orthogonal complement of V. Let $P = (P_1 \;\; P_2)$ be an orthogonal matrix such that the s columns of the $p \times s$ matrix P_1 span V and the $p - s$ columns of the $p \times (p - s)$ matrix P_2 span $V^\perp$.

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

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

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

$$P^{\mathrm{T}} X = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}_p\left(\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}, \sigma^2 \begin{pmatrix} I_s & 0 \\ 0 & I_{p-s} \end{pmatrix}\right)$$

where $P_1 \nu_1 = P_V \theta$ and $P_2 \nu_2 = P_{V^\perp} \theta$ so that

$$\|P_V X\|^2 = \|Y_1\|^2, \quad \|P_{V^\perp} X\|^2 = \|Y_2\|^2$$

and

$$\|P_V (X - \theta)\|^2 = \|Y_1 - \nu_1\|^2, \quad \|P_{V^\perp}(X - \theta)\|^2 = \|Y_2 - \nu_2\|^2.$$

The following result gives risk properties of the estimator (2.26).

Theorem 2.6 *Let V be a subspace of dimension $s \geq 0$. Then, for the estimator (2.26), we have*

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

where Y_2 and ν_2 are as above. Further, if $p - s \geq 3$ and a and $r(y)$ satisfy the assumptions of Theorem 2.3 (or Corollary 2.3 or Corollary 2.4) with $p - s$ in place of p, then $\delta(X)$ is minimax and dominates X if the additional conditions are satisfied.

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

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

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

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

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

is

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

This estimator is minimax if $0 \leq a \leq 2\,(p-s-2)$ and dominates X if $0 < a < 2\,(p-s-2)$ provided $p-s \geq 3$. The uniformly best choice of a is $p-s-2$. If in fact $\theta \in V$, the risk of the corresponding optimal estimator is $(s+2)\,\sigma^2$, since in this case $\nu_2 = P_{V^\perp}\theta = 0$ and $E_\theta\left[\sigma^2\|X-P_V X\|^{-2}\right] = E_0\left[\sigma^2\|Y_2\|^{-2}\right] E\left[1/\chi^2_{p-s}\right] = (p-s-2)^{-1}$. If $\theta \notin V$, then $\nu_2 \neq 0$ and $\|Y_2\|^2$ has a non-central chi-square distribution with $p-s$ degrees of freedom and non-centrality parameter $\|\nu_2\|^2/2\,\sigma^2$.

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

generated by the vector $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}}$. The resulting optimal James-Stein type estimator is

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

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

In the case where σ^2 is unknown, it follows from the results of Sect. 2.4.3 that replacing σ^2 in (2.26) by $S/(k+2)$ results in an estimator that dominates X under squared error loss and is minimax under scaled squared error loss (provided $r(\cdot)$ satisfies the conditions of Theorem 2.6).

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

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

The estimators discussed in this section shrink toward some "vague" prior information that θ is in or near the specified set. Consequently it shrinks toward the set but does not restrict the estimator to lie in the set. In Chap. 7 we will consider estimators that are restricted to lie in a particular set. We will see in Chap. 7 that, although vague and restricted constraints seem conceptually similar, it turns out that the analyses of risk functions in these two settings are quite distinct.

2.5 A Link Between Stein's Lemma and Stokes' Theorem

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

Typically, Stokes' theorem is considered for a more general open set Ω in $\mathbb{R}^p$ with boundary $\partial\Omega$ which could be less smooth than a sphere, and where the function g is often smooth. For example, Stroock (1990) considers a bounded open set Ω in $\mathbb{R}^p$ for which there exists a function φ from $\mathbb{R}^p$ into $\mathbb{R}$ having continuous third order partial derivatives with the properties that $\Omega = \{x \in \mathbb{R}^p \mid \varphi(x) < 0\}$ and the

gradient $\partial\varphi$ of φ vanishes at no point where φ itself vanishes. Further he requires that g has continuous first order partial derivatives in a neighborhood of the closure $\bar{\Omega}$ of Ω. For such an open set, its boundary is $\partial\Omega = \{x \in \mathbb{R}^p \mid \varphi(x) = 0\}$. Then, Stroock states that

$$\int_{\partial\Omega} n^{\mathrm{T}}(x)\, g(x) d\sigma(x) = \int_{\Omega} \mathrm{div} g(x)\, dx \tag{2.27}$$

where $n(x)$ is the outer normal (the unit outward vector) to $\partial\Omega$ at $x \in \partial\Omega$ and σ is the surface measure (the uniform measure) on $\partial\Omega$. He provides an elegant proof of Stokes' theorem in (2.27) through a rigorous construction of the outer normal and the surface measure. It is beyond the scope of this book to reproduce Stroock's proof, especially as the link we wish to make with Stein's identity only needs to deal with Ω being a ball and with $\partial\Omega$ being a sphere. Note that Stroock's conditions are satisfied for a ball of radius $r \geq 0$ centered at $\theta \in \mathbb{R}^p$ with the function $\varphi(x) = \|x - \theta\| - r$. In that context, Stokes' theorem expresses that

$$\int_{S_{r,\theta}} \left(\frac{x-\theta}{\|x-\theta\|} \right)^{\mathrm{T}} g(x)\, d\sigma_{r,\theta}(x) = \int_{B_{r,\theta}} \mathrm{div} g(x)\, dx \tag{2.28}$$

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

In the following, we will show that Stein's identity for continuously differentiable functions can be derived in a straightforward way from this ball-sphere version of Stokes' theorem. Furthermore, and perhaps more interestingly, we will see that the converse is also true: Stein's identity (for which we have an independent proof in Sect. 2.3) implies directly the classical ball-sphere version of Stokes' theorem.

Proposition 2.2 *Let $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ and let g be a continuously differentiable function from $\mathbb{R}^p$ into $\mathbb{R}^p$ such that either*

$$E_\theta[|(X-\theta)^{\mathrm{T}} g(X)|] < \infty \quad \text{or} \quad E_\theta[|\mathrm{div} g(X)|] < \infty\,. \tag{2.29}$$

Then Stein's identity in (2.7) holds, that is,

$$E_\theta[(X-\theta)^{\mathrm{T}} g(X)] = \sigma^2\, E_\theta[\mathrm{div} g(X)]\,. \tag{2.30}$$

Proof Integrating through uniform measures on spheres (see Lemma 1.4), we have

$$\begin{aligned} E_{\theta,\sigma^2}[(X-\theta)^{\mathrm{T}} g(X)] &= \int_{\mathbb{R}^p} (x-\theta)^{\mathrm{T}} g(x)\, \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left(-\frac{\|x-\theta\|^2}{2\sigma^2} \right) dx \\ &= \int_0^\infty \int_{S_{r,\theta}} \left(\frac{x-\theta}{\|x-\theta\|} \right)^{\mathrm{T}} g(x)\, d\sigma_{r,\theta}(x)\, \psi_{\sigma^2}(r)\, dr \end{aligned} \tag{2.31}$$

where

$$\psi_{\sigma^2}(r) = \frac{1}{(2\,\pi\,\sigma^2)^{p/2}}\, r\, \exp\left(-\frac{r^2}{2\,\sigma^2}\right) \tag{2.32}$$

and $\sigma_{r,\theta}$ is the uniform measure on $S_{r,\theta}$. Then applying Stokes' theorem in (2.28) to the inner most integral in (2.31) gives

$$E_{\theta,\sigma^2}[(X-\theta)^{\mathrm{T}} g(X)] = \int_0^\infty \int_{B_{r,\theta}} \mathrm{div} g(x)\, dx\, \psi_{\sigma^2}(r)\, dr\,. \tag{2.33}$$

Now, applying Fubini's theorem to the right-hand side of (2.33), we have

$$\begin{aligned}
\int_0^\infty \int_{B_{r,\theta}} \mathrm{div} g(x) dx\, \psi_{\sigma^2}(r)\, dr &= \int_{\mathbb{R}^p} \mathrm{div} g(x) \int_{\|x-\theta\|}^\infty \psi_{\sigma^2}(r)\, dr\, dx \\
&= \int_{\mathbb{R}^p} \mathrm{div} g(x)\, \frac{1}{(2\,\pi\,\sigma^2)^{p/2}} \left[-\sigma^2 \exp\left(-\frac{r^2}{2\,\sigma^2}\right)\right]_{\|x-\theta\|}^\infty dx \\
&= \sigma^2 \int_{\mathbb{R}^p} \mathrm{div} g(x)\, \frac{1}{(2\,\pi\,\sigma^2)^{p/2}} \exp\left(-\frac{\|x-\theta\|^2}{2\,\sigma^2}\right) dx \\
&= \sigma^2\, E_{\theta,\sigma^2}[\mathrm{div} g(X)]
\end{aligned} \tag{2.34}$$

since, according to (2.32),

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

Therefore combining (2.33) and (2.34) we have that

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

which is Stein's identity in (2.39).

To show Stein's identity in (2.7) assuming $E_\theta[|\mathrm{div} g(X)|] < \infty$, it suffices to essentially reverse the steps in the above development. □

Note that using Stokes' theorem in the proof of Proposition 2.2 allows the weaker condition (2.29) instead of Condition (2.8) used in Theorem 2.1.

Kavian (1993) showed that (2.27) and (2.28) continue to hold for weakly differentiable functions g, provided that g behaves properly in a neighborhood of the boundary. See also Lepelletier (2004). However, Stokes' theorem may fail if g is not sufficiently smooth in a neighborhood of the boundary. For example, it is clear that a weakly differentiable function may be redefined on the boundary of the ball $B_{r,\theta}$ without affecting either its weak differentiability or the integral

of the right-hand side of (2.28). But, by properly defining g on $S_{r,\theta}$, the integral over $S_{r,\theta}$ on the left-hand side of (2.28) may take on any value. For this reason, we develop the following version of Stokes' theorem (for balls and spheres) which will hold simultaneously for almost all r as long as the function g is weakly differentiable. It will be extensively used in extending Stein's lemma to general spherically symmetric distributions in Chaps. 5 and 6. Interestingly, the proof is based on Stein's lemma and completeness of a certain exponential family. We provide an extension to general smooth open sets in Sect. A.5 of the Appendix.

Theorem 2.7 (Fourdrinier and Strawderman 2016) *Let g be a weakly differentiable function from $\mathbb{R}^p$ into $\mathbb{R}^p$. Then (2.28) holds for almost every r.*

Proof Since g is weakly differentiable, the functions $(X-\theta)^{\mathrm{T}}g$ and $\mathrm{div}g$ are locally integrable. The same is true for the functions $g_n = g\,h_n$ where, for $n \in \mathbb{N}$, h_n is a smooth cutoff function such that $h_n(x) = 1$ if $\|x\| < n$, $h_n(x) = 0$ if $\|x\| > n+1$, $h_n \in \mathscr{C}^\infty$, and $h_n(x) \le 1$ for all x. Thus g_n is weakly differentiable and we have $E_\theta[|(X-\theta)^{\mathrm{T}}g_n(X)|] < \infty$ or $E_\theta[|\mathrm{div}g_n(X)|] < \infty$. Hence, Stein's lemma applies to g_n, so that (2.39) holds for g_n, that is,

$$E_\theta[(X-\theta)^{\mathrm{T}}g_n(X)] = \sigma^2\, E_\theta[\mathrm{div}g_n(X)]\,. \tag{2.35}$$

Then, as in (2.31), with ψ_{σ^2} given in (2.32),

$$E_{\theta,\sigma^2}[(X-\theta)^{\mathrm{T}}g_n(X)] = \int_0^\infty \int_{S_{r,\theta}} \left(\frac{x-\theta}{\|x-\theta\|}\right)^{\mathrm{T}} g_n(x)\, d\sigma_{r,\theta}(x)\, \psi_{\sigma^2}(r)\, dr \tag{2.36}$$

and, as in (2.33), we also have

$$\sigma^2\, E_{\theta,\sigma^2}[\mathrm{div}g_n(X)] = \int_0^\infty \int_{B_{r,\theta}} \mathrm{div}g_n(x)\, dx\, \psi_{\sigma^2}(r)\, dr\,. \tag{2.37}$$

Hence, it follows from (2.35), (2.36), and (2.37) that, for all σ^2,

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

Therefore, since the family $\{\psi_{\sigma^2}(r)\}_{\sigma^2>0}$ defined in (2.32) is proportional to a family of densities that is complete as an exponential family, we have

$$\int_{S_{r,\theta}} \left(\frac{x-\theta}{\|x-\theta\|}\right)^{\mathrm{T}} g_n(x)\, d\sigma_{r,\theta}(x) = \int_{B_{r,\theta}} \mathrm{div}g_n(x)\, dx\,, \tag{2.38}$$

for almost every $0 < r < n$. Now, since $g_n(x) = g(x)$ for $\|x\| < n$, it follows that (2.38) holds for g for almost every $r > 0$. □

As a first corollary, it follows that the classical (ball-sphere) version of Stokes' theorem holds for every r when g is continuously differentiable.

Corollary 2.7 *Let g be a continuously differentiable function from $\mathbb{R}^p$ into $\mathbb{R}^p$. Then (2.28) holds for every $r > 0$.*

Proof Because g is continuously differentiable, both sides of (2.38) are continuous. Then, since the equality holds almost everywhere, it must hold for all $r > 0$. □

Note that the proof of Proposition 2.2 remains valid when (2.28) holds for almost every $r > 0$. Hence the following corollary follows from Theorem 2.7 and Proposition 2.2.

Corollary 2.8 (Stein's lemma) *Let $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ and let g be a weakly differentiable function from $\mathbb{R}^p$ into $\mathbb{R}^p$ such that either $E_\theta[|(X-\theta)^{\mathrm{T}} g(X)|] < \infty$ or $E_\theta[|\mathrm{div} g(X)|] < \infty$. Then Stein's identity in (2.7) holds, that is,*

$$E_\theta[(X-\theta)^{\mathrm{T}} g(X)] = \sigma^2 \, E_\theta[\mathrm{div} g(X)] \,. \tag{2.39}$$

Note that, as in Proposition 2.2, Corollary 2.8 uses the weaker condition (2.29) instead of Condition (2.8) which was used in Theorem 2.1.

We have seen for balls and spheres that Stokes' theorem can be directly derived from Stein's identity, for weakly differentiable functions. This result will be particularly important for proving Stein type identities for spherically symmetric distributions in Chaps. 5 and 6. Note that we have in fact obtained a stronger result. It is actually shown that, any time Stein's identity is valid, then the version of Stokes' theorem given in Theorem 2.7 holds as well. This result is particularly interesting when the weak differentiability assumption is not met. For example, Fourdrinier et al. (2006) noticed that this may be the case when dealing with a location parameter restricted to a cone; Stein's identity (2.7) holds but the weak differentiability of the functions at hand is not guaranteed (see also Sect. 7.3).

2.6 Differential Operators and Dimension Cut-Off When Estimating a Mean

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

Indeed, when estimating θ under quadratic loss, improvements on X through unbiased estimation techniques often involve a nonlinear partial differential operator of the form

$$\mathscr{R}g(x) = k\,\mathrm{div}g(x) + \|g(x)\|^2 \tag{2.40}$$

for a certain constant k. A sufficient condition for improvement is typically

$$\mathscr{R}g(x) \leq 0 \tag{2.41}$$

for all $x \in \mathbb{R}^p$ (with strict inequality on a set of positive Lebesgue measure). We will see that (2.41) does not have a nontrivial solution g (i.e. g is not equal to 0 almost everywhere) when the dimension $p \leq 2$, even if we look for solutions with smoothness conditions as weak as possible. Consequently, a necessary dimension condition for (2.41) to have solutions $g \not\equiv 0$ is $p \geq 3$.

Here is a precise statement of this fact.

Theorem 2.8 *Let $k \in \mathbb{R}$ be fixed. When $p \leq 2$, the only weakly differentiable solution g with $\|g\|^2 \in L^1_{loc}(\mathbb{R}^p)$ of*

$$\mathscr{R}g(x) = k\,\mathrm{div}g(x) + \|g(x)\|^2 \leq 0\,, \tag{2.42}$$

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

Note that, in Theorem 2.8, the search for solutions of (2.42) is addressed in the general setting of weakly differentiable functions. The proof will follow the development in Blanchard and Fourdrinier (1999). However, in that paper, the g's are sought in the much larger space of distributions $\mathscr{D}'(\mathbb{R}^p)$ introduced by Schwartz (see Schwartz 1973 for a full account). Note also that the condition $\|g\|^2 \in L^1_{loc}(\mathbb{R}^p)$ is not restrictive. Any estimator of the form $X + g(X)$ with finite risk must satisfy $E_\theta[\|g(X)\|^2] < \infty$ and hence $\|g\|^2$ must be in $L^1_{loc}(\mathbb{R}^p)$.

The proof of Theorem 2.8 is based on the use of the following sequence of so-called test functions. Let φ be a nonnegative infinitely differentiable function on $\mathbb{R}_+$ bounded by 1, identically equal to 1 on $[0, 1]$, and with support on the interval $[0, 2]$ ($\mathrm{supp}(\varphi) = [0, 2]$), which implies that its derivative is bounded. Associate to φ the sequence $\{\varphi_n\}_{n\geq 1}$ of infinitely differentiable functions from $\mathbb{R}^p$ into $[0, 1]$ defined through

$$\forall n \geq 1 \quad \forall x \in \mathbb{R}^p \quad \varphi_n(x) = \varphi\left(\frac{||x||}{n}\right). \tag{2.43}$$

Clearly, for any $n \geq 1$, the function φ_n has compact support B_{2n}, the closed ball of radius $2n$ and centered at zero in $\mathbb{R}^p$. Also, an interesting property that follows from the uniform boundedness of φ', is that, for any $\beta \geq 1$ and for any $j = 1, \ldots, p$, there exists a constant $K > 0$ such that

$$\left|\frac{\partial \varphi_n^\beta}{\partial x_j}(x)\right| \leq \frac{K}{n}\,\varphi_n^{\beta-1}(x)\,. \tag{2.44}$$

Note that, as all the derivatives of φ vanish outside of the compact interval $[1, 2]$ and φ is bounded by 1, (2.44) implies

$$\left| \frac{\partial \varphi_n^\beta}{\partial x_j}(x) \right| \leq \frac{K}{n} \, \mathbb{1}_{C_n}(x) \, . \tag{2.45}$$

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

Proof of Theorem 2.8 Let g be a weakly differentiable function g, with $\|g\|^2 \in L^1_{loc}(\mathbb{R}^p)$, satisfying (2.42). Then, using the defining property (2.6) of weak differentiability (see also Sect. A.1), we have, for any $n \in \mathbb{N}^*$ and any $\beta > 1$,

$$\begin{aligned}
\int_{\mathbb{R}^p} \|g(x)\|^2 \, \varphi_n^\beta(x) \, dx &\leq -k \int_{\mathbb{R}^p} \mathrm{div} g(x) \, \varphi_n^\beta(x) \, dx \\
&= -k \sum_{i=1}^p \int_{\mathbb{R}^p} \frac{\partial}{\partial x_i} g_i(x) \, \varphi_n^\beta(x) \, dx \\
&= k \sum_{i=1}^p \int_{\mathbb{R}^p} g_i(x) \, \frac{\partial}{\partial x_i} \varphi_n^\beta(x) \, dx \\
&= k \int_{\mathbb{R}^p} g^{\mathrm{T}}(x) \, \partial \varphi_n^\beta(x) \, dx \\
&\leq k \int_{\mathbb{R}^p} \|g(x)\| \, \|\partial \varphi_n^\beta(x)\| \, dx \, .
\end{aligned} \tag{2.46}$$

Then, using (2.44), it follows from (2.46) that there exists a constant $C > 0$ such that

$$\begin{aligned}
\int_{\mathbb{R}^p} \|g(x)\|^2 \, \varphi_n^\beta(x) \, dx &\leq \frac{C}{n} \int_{\mathbb{R}^p} \|g(x)\| \, \varphi_n^{\beta-1}(x) \, dx \\
&\leq \frac{C}{n} \left(\int_{\mathbb{R}^p} \varphi_n^{\beta-2}(x) \, dx \right)^{1/2} \left(\int_{\mathbb{R}^p} \|g(x)\|^2 \, \varphi_n^\beta(x) \, dx \right)^{1/2} ,
\end{aligned} \tag{2.47}$$

when applying Schwarz's inequality with $\beta > 2$ and using

$$\|g(x)\| \, \varphi_n^{\beta-1}(x) = \varphi_n^{\beta/2-1}(x) \, \|g(x)\| \, \varphi_n^{\beta/2}(x) \, .$$

Clearly (2.47) is equivalent to

$$\int_{\mathbb{R}^p} \|g(x)\|^2 \, \varphi_n^\beta(x) \, dx \leq \frac{C^2}{n^2} \int_{\mathbb{R}^p} \varphi_n^{\beta-2}(x) \, dx \, . \tag{2.48}$$

Thus, since $\varphi_n = 1$ on B_n and $\varphi_n \geq 0$,

$$\int_{B_n} \|g(x)\|^2 \, dx = \int_{B_n} \|g(x)\|^2 \, \varphi_n^\beta(x) \, dx \leq \int_{\mathbb{R}^p} \|g(x)\|^2 \, \varphi_n^\beta(x) \, dx \,. \quad (2.49)$$

Then, since $\text{supp}\,(\varphi_n) = B_{2n}$ and $0 \leq \varphi_n \leq 1$, using (2.48) gives

$$\int_{B_n} \|g(x)\|^2 \, dx \leq \frac{C^2}{n^2} \int_{\mathbb{R}^p} \varphi_n^{\beta-2}(x) \, dx \leq \frac{C^2}{n^2} \int_{B_{2n}} dx = A \, n^{p-2} \quad (2.50)$$

for some constant $A > 0$. Letting n go to infinity in (2.50) shows that, when $p < 2$, $g = 0$ almost everywhere, which proves the theorem in that case. It also implies that $\|g\|^2 \in L^1(\mathbb{R}^p)$ when $p = 2$.

In the case $p = 2$, the result will follow by applying (2.45). Indeed, it follows from (2.45), (2.49) and the first inequality in (2.47) that, for some constant $C > 0$,

$$\begin{aligned} \int_{B_n} \|g(x)\|^2 \, dx &\leq \frac{C}{n} \int_{C_n} \|g(x)\| \, dx \\ &\leq \frac{C}{n} \left(\int_{C_n} dx \right)^{1/2} \left(\int_{C_n} \|g(x)\|^2 \, dx \right)^{1/2} \end{aligned} \quad (2.51)$$

by Schwarz's inequality. Now, since $p = 2$,

$$\int_{C_n} dx \leq \int_{B_{2n}} dx \propto n^2 \,. \quad (2.52)$$

Hence (2.51) and (2.52) imply that

$$\int_{B_n} \|g(x)\|^2 \, dx \leq A \left(\int_{C_n} \|g(x)\|^2 \, dx \right)^{1/2} , \quad (2.53)$$

for some constant $A > 0$. Since as noted above, $\|g\|^2 \in L^1(\mathbb{R}^p)$, we have

$$\lim_{n \to \infty} \int_{C_n} \|g(x)\|^2 \, dx = 0$$

and consequently (2.53) gives rise to

$$0 = \lim_{n \to \infty} \int_{C_n} \|g(x)\|^2 \, dx = \int_{\mathbb{R}^p} \|g(x)\|^2 \, dx \,.$$

Thus $g = 0$ almost everywhere and gives the desired result for $p = 2$ is obtained.

□

Such a dimension cut-off result implies that the usual Stein inequality $2\,\mathrm{div} g(x) + ||g(x)||^2 \leq 0$, for any $x \in \mathbb{R}^p$, has no nontrivial solution g, with $\|g\|^2 \in L^1_{loc}(\mathbb{R}^p)$ when $p \leq 2$. This reinforces the fact that the MLE X is admissible in dimension $p \leq 2$ when estimating a normal mean. Blanchard and Fourdrinier (1999) (to which we refer for a full account of the dimension cut-off phenomenon) also considered more general nonlinear partial differential inequalities. We will again use their technique in Chap. 8 (for loss estimation) to prove that, for an inequality of the form $k\,\Delta\gamma(x) + \gamma^2(x) \leq 0$, the same dimension cut-off phenomenon occurs for $p \leq 4$ (there is no nontrivial solution γ, with $\gamma^2 \in L^1_{loc}(\mathbb{R}^p)$, when $p \leq 4$).