On Improved Loss Estimation for Shrinkage Estimators

Dominique Fourdrinier and Martin T. Wells

Abstract. Let X be a random vector with distribution P_θ where θ is an unknown parameter. When estimating θ by some estimator $\varphi(X)$ under a loss function $L(\theta, \varphi)$, classical decision theory advocates that such a decision rule should be used if it has suitable properties with respect to the frequentist risk $R(\theta, \varphi)$. However, after having observed $X = x$, instances arise in practice in which φ is to be accompanied by an assessment of its loss, $L(\theta, \varphi(x))$, which is unobservable since θ is unknown. A common approach to this assessment is to consider estimation of $L(\theta, \varphi(x))$ by an estimator δ, called a loss estimator. We present an expository development of loss estimation with substantial emphasis on the setting where the distributional context is normal and its extension to the case where the underlying distribution is spherically symmetric. Our overview covers improved loss estimators for least squares but primarily focuses on shrinkage estimators. Bayes estimation is also considered and comparisons are made with unbiased estimation.

Key words and phrases: Conditional inference, linear model, loss estimation, quadratic loss, risk function, robustness, shrinkage estimation, spherical symmetry, SURE, unbiased estimator of loss, uniform distribution on a sphere.

1. INTRODUCTION

Suppose X is an observable from a distribution P_θ parameterized by an unknown parameter θ. In classical decision theory, it is usual, after selecting an estimation procedure $\varphi(X)$ of θ, to evaluate it through a loss criterion, $L(\theta, \varphi(X))$, which represents the cost incurred by the estimate $\varphi(X)$ when the unknown parameter equals θ. In the long run, as it depends on the particular value of X, this loss cannot be appropriate to assess the performance of the estimator φ. Indeed, to be valid (in the frequentist sense), a global evaluation of such a statistical procedure should be based on averages over all

Dominique Fourdrinier is Professor, Université de Rouen, LITIS EA 4108, Avenue de l'Université, BP 12, 76801 Saint-Étienne-du-Rouvray, France, and Visiting Professor, Department of Statistical Science, Cornell University, Comstock Hall, Ithaca, New York 14853, USA (e-mail: dominique.fourdrinier@univ-rouen.fr). Martin T. Wells is Professor, Department of Statistical Science, Cornell University, Comstock Hall, Ithaca, New York 14853, USA (e-mail: mtw1@cornell.edu).

the possible observations. Consequently, it is common to report the risk $R(\theta, \varphi) = E_\theta[L(\theta, \varphi(X)]$ as a measure of the efficiency of φ (E_θ denotes expectation with respect to P_θ). Thus we have at our disposal a long-run performance of $\varphi(X)$ for each value of θ. However, although this notion of risk can effectively be used in comparing $\varphi(X)$ with other estimators, it is inaccessible since θ is unknown. The usual frequentist risk assessment is the maximum risk $\overline{R}_\varphi = \sup_\theta R(\theta, \varphi)$.

By construction, this least favorable report of the estimation procedure is non-data-dependent [as we were guided by a global notion of accuracy of $\varphi(X)$]. However, there exist situations where the fact that the observation X has such or such value may influence the judgment on a statistical procedure. A particularly edifying example is given by the following simple confidence interval estimation (which can be viewed as a loss estimation problem). Assume that the observable is a couple (X_1, X_2) of independent copies of a random variable X satisfying, for $\theta \in \mathbb{R}$,

$$P[X = \theta - 1] = P[X = \theta + 1] = \tfrac{1}{2}.$$

Then it is clear that the confidence interval for θ defined by

$$I(X_1, X_2) = \left\{\theta \in \mathbb{R} \,\middle|\, \left|\frac{X_1 + X_2}{2} - \theta\right| < \frac{1}{2}\right\}$$

satisfies

$$\mathbb{1}_{[\theta \in I(X_1, X_2)]} = \begin{cases} 1, & \text{if } X_1 \neq X_2, \\ 0, & \text{if } X_1 = X_2, \end{cases}$$

so that it suffices to observe (X_1, X_2) in order to know exactly whether $I(X_1, X_2)$ contains θ or not.

The previous (ad hoc) example indicates that data-dependent reports are relevant. When $X = x$ the loss, $L(\theta, \varphi(x))$, itself could serve as a perfect measure of the accuracy of φ if it were available (which it is not since θ is unknown). It is natural to estimate $L(\theta, \varphi(x))$ by a data-dependent estimator $\delta(X)$, a new estimator called a loss estimator. Such an estimator can serve as a data-dependent assessment (instead of $\overline{R}_\varphi$). This is a conditional approach in the sense that the accuracy assessment is made on a data-dependent quantity, the loss, instead of the risk.

To evaluate the extent to which $\delta(X)$ successfully estimates $L(\theta, \varphi(X))$, another loss is required and it has become standard, for simplicity, to use the squared error

$$(1.1) \quad L^*(\theta, \varphi(X), \delta(X)) = \big(\delta(X) - L(\theta, \varphi(X))\big)^2.$$

Insofar as we are thinking in terms of long-run frequencies, we adopt a frequentist approach to evaluating the performance of L^* by averaging over the sampling distribution of X given θ, that is, by using a new notion of risk

$$(1.2) \quad \begin{aligned} \mathcal{R}(\theta, \varphi, \delta) &= E_\theta[L^*(\theta, \varphi(X), \delta(X))] \\ &= E_\theta\big[(\delta(X) - L(\theta, \varphi(X)))^2\big]. \end{aligned}$$

As $\overline{R}_\varphi$ reports on the worst possible situation (the maximum risk), we may expect that a competitive data-dependent report $\delta(X)$ should improve on $\overline{R}_\varphi$ under the risk (1.2), that is, for all θ, $\delta(X)$ satisfies

$$(1.3) \quad \mathcal{R}(\theta, \varphi, \delta) \leq \mathcal{R}(\theta, \varphi, \overline{R}_\varphi).$$

More generally, a reference loss estimator δ_0 will be dominated by a competitive estimator δ if, for all θ,

$$(1.4) \quad \mathcal{R}(\theta, \varphi, \delta) \leq \mathcal{R}(\theta, \varphi, \delta_0),$$

with strict inequality for some θ.

Unlike the usual estimation setting where the quantity of interest is a function of the parameter θ, loss estimation involves a function of both θ and X (the data). This feature may make the statistical analysis more difficult but it is clear that the usual notions of minimaxity, admissibility, etc., and their methods of proof can be directly adapted to that situation. Also, although frequentist interpretability was evoked above, in case we would be interested in a Bayesian approach, it is easily seen that this approach would consist of the usual Bayes estimator φ_B of θ and the posterior loss $\delta_B(X) = E[L(\theta, \varphi_B)|X]$.

The problem of estimating a loss function has been considered by Sandved [43] who developed a notion of unbiased estimator of $L(\theta, \varphi(X))$ in various settings. However, the underlying conditional approach traces back to Lehmann and Sheffé [37] who estimated the power of a statistical test. Kiefer, in a series of papers [33–35], developed conditional and estimated confidence theories. A subjective Bayesian approach was compared by Berger [4–6] with the frequentist paradigm. Jonhstone [32] considered (in)admissibility of unbiased estimators of loss for the maximum likelihood estimator $\varphi_0(X) = X$ and for the James–Stein estimator $\varphi^{\text{JS}}(X) = (1 - (p-2)/\|X\|^2)X$ of a p-variate normal mean θ. For $\varphi_0(X) = X$, the unbiased estimator of the quadratic loss $L(\theta, \varphi_0(X)) = \|\varphi_0(X) - \theta\|^2$, that is, the loss estimator δ_0 which satisfies, for all θ,

$$(1.5) \quad E_\theta[\delta_0] = E_\theta[L(\theta, \varphi_0(X))] = R(\theta, \varphi_0),$$

is $\delta_0 = \overline{R}_\varphi = p$. Johnstone proved that (1.3) is satisfied with the competitive estimator $\delta(X) = p - 2(p-4)/\|X\|^2$ when $p \geq 5$, the risk difference between δ_0 and δ being expressed as $-4(p-4)^2 E_\theta[1/\|X\|^4]$. For the James–Stein estimator φ^{JS}, the unbiased estimator of loss is itself data-dependent and equal to $\delta_0^{\text{JS}}(X) = p - (p-2)^2/\|X\|^2$. Jonhstone showed that improvement on δ_0^{JS} can be obtained with $\delta^{\text{JS}}(X) = p - (p-2)^2/\|X\|^2 + 2p/\|X\|^2$ when $p \geq 5$, with strict inequality in (1.4) for all θ since the difference in risk between δ^{JS} and δ_0^{JS} equals $-4p^2 E_\theta[1/\|X\|^2]$.

In Section 2, we develop the quadratic loss estimation problem for a p-normal mean. After a review of the basic ideas, a new class of loss estimators is constructed in Section 2.1. In Section 2.2, we turn our focus on some interesting and surprising behavior of Bayesian assessments; this paradoxical result is illustrated in a general inadmissibility theorem. Section 3 is devoted to the case where the variance is unknown. Extensions to the spherical case are given in Section 4. In Section 4.1, we consider the general case of a spherically symmetric distribution around a fixed vector $\theta \in \mathbb{R}^p$ and in Section 4.2 these ideas are then generalized

to the case where a residual vector is available. We conclude by mentioning a number of applied and theoretical developments of loss estimation not covered in this overview. The Appendix gives some necessary background material and technical results.

2. ESTIMATING THE QUADRATIC LOSS OF A *p*-NORMAL MEAN WITH KNOWN VARIANCE

2.1 Dominating Unbiased Estimators of Loss

Let X be a p-variate normally distributed $\mathcal{N}(\theta, I_p)$ random vector with unknown mean θ and identity covariance matrix I_p. To estimate θ, the observable X is itself a reference estimator (it is the maximum likelihood estimator (m.l.e.) and it is an unbiased estimator of θ) so that it is convenient to write any estimator of θ through X as $\varphi(X) = X + g(X)$, for a certain function g from $\mathbb{R}^p$ into $\mathbb{R}^p$. Under squared error loss $\|\varphi(X) - \theta\|^2$, the (quadratic) risk of φ is defined by

$$(2.1) \qquad R(\theta, \varphi) = E_\theta[\|\varphi(X) - \theta\|^2],$$

where E_θ denotes the expectation with respect to $\mathcal{N}(\theta, I_p)$.

Clearly, the risk of the m.l.e. X equals p and in general $\varphi(X)$ will be a reasonable estimator only if its risk is finite. It is easy to see (Lemma A.1 in Appendix A.1) through Schwarz's inequality that this is the case as soon as

$$(2.2) \qquad E_\theta[\|g(X)\|^2] < \infty,$$

which we will assume in the following (it can be also seen that this condition is in fact necessary to guarantee the risk finiteness).

To improve on the m.l.e. X when $p \geq 3$ [i.e., to have $R(\theta, \varphi) \leq p$], Stein [48] exhibited (under certain differentiability conditions that we recall below) an unbiased estimator of the risk of $\varphi(X)$, that is, a function $\delta_0(X)$ (depending only on X and not on θ) for which

$$(2.3) \qquad R(\theta, \varphi) = E_\theta[\delta_0(X)].$$

This suggests a natural estimator of the loss $\|\varphi(X) - \theta\|^2$ since (2.3) implies that

$$(2.4) \qquad E_\theta[\|\varphi(X) - \theta\|^2] = E_\theta[\delta_0(X)]$$

and hence is an unbiased estimator of the loss. Stein [48] proved more precisely that $\delta_0(X) = p + 2 \cdot \operatorname{div} g(X) + \|g(X)\|^2$ [where $\operatorname{div} g(X)$ stands for the divergence of $g(X)$, i.e., $\operatorname{div} g(X) = \sum_{i=1}^p \partial_i g_i(X)$]. One can see that δ_0 may change sign so that, as an estimator of loss (which is nonnegative), it cannot be completely satisfactory, and hence, is likely to be improved upon.

Any competitive loss estimator $\delta(X)$ can be written as $\delta(X) = \delta_0(X) - \gamma(X)$ for a certain function $\gamma(X)$ which can be interpreted as a correction to $\delta_0(X)$. Note that, for the m.l.e. [i.e., if $g(X) = 0$], we may expect that an improvement on $\delta_0(X) = p$ would be obtained with a nonnegative function $\gamma(X)$ satisfying the requirement expressed by condition (1.3). Note also that, similarly to the finiteness risk condition (2.2), we will require that

$$(2.5) \qquad E_\theta[\gamma^2(X)] < \infty$$

to assure that the risk of $\delta(X)$ is finite (see Appendix A.1).

Using straightforward algebra, the risk difference $\mathcal{D}(\theta, \varphi, \delta) = \mathcal{R}(\theta, \varphi, \delta) - \mathcal{R}(\theta, \varphi, \delta_0)$ simplifies to

$$(2.6) \quad \begin{aligned} \mathcal{D}(\theta, \varphi, \delta) &= E_\theta[\gamma^2(X) - 2\gamma(X)\delta_0(X)] \\ &\quad + 2E_\theta[\gamma(X)\|\varphi(X) - \theta\|^2]. \end{aligned}$$

Conditions for which $\mathcal{D}(\theta, \varphi, \delta) \leq 0$ will be formulated after finding an unbiased estimate of the term $\gamma(X)\|\varphi(X) - \theta\|^2$ in the last expectation. We briefly review the flow of ideas of those techniques.

For a function g from $\mathbb{R}^p$ into $\mathbb{R}^p$, the Stein's identity (see Stein [48]) states that

$$(2.7) \qquad E_\theta[(X - \theta)^t g(X)] = E_\theta[\operatorname{div} g(X)]$$

provided that these expectations exist. Here Stein specified that g was almost differentiable. Weak differentiability is needed to integrate shrinkage functions $g(X)$, intervening in the James–Stein estimators, of the form $g(X) = -aX/\|X\|^2$ which are not differentiable in the usual sense [such a $g(X)$ explodes at zero]. This notion is equivalent (and it is of more common use in analysis) to the statement that g belongs to the Sobolev space $W_{\mathrm{loc}}^{1,1}(\mathbb{R}^p)$ of weakly differentiable functions. That equivalence was noticed by Johnstone [32].

Recall that a locally integrable function γ from $\mathbb{R}^p$ into $\mathbb{R}$ is said to be weakly differentiable if there exist p functions $h_1, \ldots, h_p$ locally integrable on $\mathbb{R}^p$ such that, for any $i = 1, \ldots, p$,

$$(2.8) \quad \int_{\mathbb{R}^p} \gamma(x) \frac{\partial \varphi}{\partial x_i}(x)\, dx = -\int_{\mathbb{R}^p} h_i(x)\varphi(x)\, dx$$

for any infinitely differentiable function φ on $\mathbb{R}^p$ with compact support. The functions h_i are the ith partial weak derivatives of γ. Their common notation is $\partial \gamma / \partial x_i$ and the vector $\nabla \gamma = (\partial \gamma / \partial x_1, \ldots, \partial \gamma / \partial x_p)^t$ is referred to as the weak gradient of γ.

Note that (2.8) usually holds when γ is continuously differentiable, that is, when $h_i = \partial \gamma / \partial x_i$, the standard

partial derivative, is continuous. Thus, via (2.8), the extension to weak differentiability consists in a propriety of integration by parts with vanishing bracketed term. Naturally a function $g = (g_1, \ldots, g_p)$ from $\mathbb{R}^p$ into $\mathbb{R}^p$ is said to be weakly differentiable if each of its components g_j is weakly differentiable. In that case, the function $\operatorname{div} g = \sum_{i=1}^p \partial g_i / \partial x_i$ is referred to as the weak divergence of g; this is the operator intervening in the Stein's identity (2.7).

When dealing with an unbiased estimator of a quantity of the form $\|X - \theta\|^2 \gamma(X)$, where γ is a function from $\mathbb{R}^p$ into $\mathbb{R}$, writing

$$(2.9) \quad \|X - \theta\|^2 \gamma(X) = (X - \theta)^t (X - \theta) \gamma(X)$$

naturally leads to an iteration of Stein's identity (2.7) and involves twice weak differentiability of γ. This is of course defined through the weak differentiability of all the weak partial derivatives $\partial \gamma / \partial x_i$; these second weak partial derivatives are denoted by $\partial^2 \gamma / \partial x_j \, \partial x_i$. Thus γ belongs to the Sobolev space $W_{\text{loc}}^{2,1}(\mathbb{R}^p)$ and $\Delta \gamma = \sum_{i=1}^p \partial^2 \gamma / \partial x_i^2$ is referred to as the weak Laplacian of γ.

By (2.9) and (2.7), we have

$$
\begin{aligned}
(2.10) \quad & E_\theta[\|X - \theta\|^2 \gamma(X)] \\
& = E_\theta[\operatorname{div}((X - \theta)^t \gamma(X))] \\
& = E_\theta[p\gamma(X) + (X - \theta)^t \nabla \gamma(X)]
\end{aligned}
$$

by the product rule for the divergence operator. Then, applying again (2.7) to the last term in (2.10) gives

$$
\begin{aligned}
(2.11) \quad E_\theta[(X - \theta)^t \nabla \gamma(X)] & = E_\theta[\operatorname{div}(\nabla \gamma(X)) \\
& = E_\theta[\Delta \gamma(X)]
\end{aligned}
$$

by definition of the Laplacian operator. Finally, gathering (2.10) and (2.11), we have that

$$
\begin{aligned}
(2.12) \quad & E_\theta[\|X - \theta\|^2 \gamma(X)] \\
& = E_\theta[p\gamma(X) + \Delta \gamma(X)].
\end{aligned}
$$

We are now in a position to provide an unbiased estimator of the difference in risk $\mathcal{D}(\theta, \varphi, \delta)$ in (2.6). Its nonpositivity will be a sufficient condition for $\mathcal{D}(\theta, \varphi, \delta) \le 0$ and hence for δ to improve on δ_0. Indeed we have

$$
\begin{aligned}
& \|\varphi(X) - \theta\|^2 \\
& = \|X + g(X) - \theta\|^2 \\
& = \|g(X)\|^2 + 2(X - \theta)^t g(X) + \|X - \theta\|^2
\end{aligned}
$$

so that, according to (2.7) and (2.12),

$$
\begin{aligned}
& E_\theta[\|\varphi(X) - \theta\|^2 \gamma(X)] \\
& = E_\theta[\gamma(X)\|g(X)\|^2 + 2\operatorname{div}(\gamma(X)g(X)) \\
& \hspace{3cm} + p\gamma(X) + \Delta \gamma(X)].
\end{aligned}
$$

Therefore, as $\operatorname{div}(\gamma(X)g(X)) = \gamma(X)\operatorname{div} g(X) + \nabla \gamma(X)^t g(X)$ and as $\delta_0(X) = p + 2\operatorname{div} g(X) + \|g(X)\|^2$, the risk difference $\mathcal{D}(\theta, \varphi, \delta)$ in (2.6) reduces to

$$\mathcal{D}(\theta, \varphi, \delta) = E_\theta[\gamma^2(X) + 4\nabla \gamma(X)^t g(X) + 2\Delta \gamma(X)],$$

so that a sufficient condition for $\mathcal{D}(\theta, \varphi, \delta)$ to be nonpositive is

$$(2.13) \quad \gamma^2(x) + 4\nabla \gamma(x)^t g(x) + 2\Delta \gamma(x) \le 0$$

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

The question now arises of determining a "best" correction γ satisfying (2.13). The following theorem provides a way to associate to the function g a suitable correction γ which satisfies (2.13) in the case where $g(x)$ is of the form $g(x) = \nabla m(x)/m(x)$ for a certain nonnegative function m. This is the case when φ is a Bayes estimator of θ related to a prior π, the function m being the corresponding marginal (see Brown [10]). Bock [8] showed that, through the choice of m, such estimators constitute a wide class of estimators of θ (which are called pseudo-Bayes estimators when the function m does not correspond to a true prior π).

THEOREM 2.1. *Let m be a nonnegative function which is also superharmonic* (*respectively subharmonic*) *on $\mathbb{R}^p$ such that $\nabla m/m \in W_{\text{loc}}^{1,1}(\mathbb{R}^p)$. Let ξ be a real-valued function, strictly positive and strictly subharmonic* (*respectively superharmonic*) *on $\mathbb{R}^p$ such that*

$$(2.14) \qquad E_\theta\left[\left(\frac{\Delta \xi(X)}{\xi(X)}\right)^2\right] < \infty.$$

Assume also that there exists a constant $K > 0$ such that, for any $x \in \mathbb{R}^p$,

$$(2.15) \qquad m(x) > K \frac{\xi^2(x)}{|\Delta \xi(x)|}$$

and let $K_0 = \inf_{x \in \mathbb{R}^p} m(x) \frac{|\Delta \xi(x)|}{\xi^2(x)}$.

Then the unbiased loss estimator δ_0 of the estimator φ of θ defined by $\varphi(X) = X + \nabla m(X)/m(X)$ is dominated by the estimator $\delta = \delta_0 - \gamma$, where the correction term γ is given, for any $x \in \mathbb{R}^p$ such that $m(x) \ne 0$, by

$$(2.16) \qquad \gamma(x) = -\alpha \operatorname{sgn}(\Delta \xi(x)) \frac{\xi(x)}{m(x)},$$

as soon as $0 < \alpha < 2K_0$.

PROOF.   The domination condition will be shown by proving that the risk difference is less than zero. We only consider the case where m is superharmonic and ξ is strictly subharmonic, the case where m is subharmonic and ξ is strictly superharmonic being similar.

First note that the finiteness risk condition (2.5) is guaranteed by the condition in (2.14) and the fact that (2.15) implies that, for any $x \in \mathbb{R}^p$,

$$\gamma^2(x) = \alpha^2 \frac{\xi^2(x)}{m^2(x)} \le \frac{\alpha^2}{K_0^2}\left(\frac{\Delta \xi(x)}{\xi(x)}\right)^2.$$

Further note that, for a shrinkage function g of the form $g(x) = \nabla m(x)/m(x)$, the left-hand side of (2.13) can be expressed as

$$\begin{aligned}(2.17) \quad \mathcal{R}\gamma(x) &= \gamma^2(x) \\ &+ 2\left\{2\frac{\Delta(m(x)\gamma(x))}{m(x)} - \gamma(x)\frac{\Delta m(x)}{m(x)}\right\}\end{aligned}$$

and hence, for γ in (2.16), as

$$\begin{aligned}(2.18) \quad \mathcal{R}\gamma(x) &= \alpha^2 \frac{\xi^2(x)}{m^2(x)} \\ &+ 2\alpha\left\{-\frac{\Delta \xi(x)}{m(x)} + \frac{\xi(x)\Delta m(x)}{m^2(x)}\right\}.\end{aligned}$$

Now, since m is superharmonic and ξ is positive, it follows from (2.18) that

$$\mathcal{R}\gamma(x) \le \frac{\alpha}{m(x)}\left\{\frac{\alpha\xi^2(x)}{m(x)} - 2\Delta \xi(x)\right\}$$

and hence, by subharmonicity of ξ, the inequality in (2.15) and the definition of K_0, that

$$(2.19) \quad \mathcal{R}\gamma(x) < \frac{\alpha}{m(x)}\{\alpha - 2K_0\}\frac{\xi^2(x)}{m(x)}.$$

Finally, since $0 < \alpha < 2K_0$, the inequality in (2.19) gives $\mathcal{R}\gamma(x) < 0$, which is the desired result.   □

As an example, consider $m(x) = 1/\|x\|^{p-2}$, that is, the fundamental harmonic function which is superharmonic on the entire space $\mathbb{R}^p$ (see Du Plessis [17]). Then we have $\nabla m(x)/m(x) = -(p-2)/\|x\|^2$ and $\varphi(X)$ is the James–Stein estimator whose unbiased estimator of loss is $\delta_0(X) = p - (p-2)^2/\|X\|^2$. First note that $\nabla m/m \in W_{\text{loc}}^{1,1}(\mathbb{R}^p)$ for $p \ge 3$. Now choosing, for any $x \ne 0$, the function $\xi(x) = 1/\|x\|^p$ gives rise to $\Delta \xi(x) = 2p/\|x\|^{p+2} > 0$ and hence to

$$\frac{\xi^2(x)}{|\Delta \xi(x)|} = \frac{1}{2p}\frac{1}{\|x\|^{p-2}},$$

which means that condition (2.15) is satisfied with $K < 2p$. Also we have

$$\left(\frac{\Delta \xi(x)}{\xi(x)}\right)^2 = \frac{4p^2}{\|x\|^4}$$

which implies that the condition in (2.14) is satisfied for $p \ge 5$. Now it is clear that the constant K_0 is equal to $2p$ and that the correction term γ in (2.16) equals, for any $x \ne 0$, $\gamma(x) = -\alpha/\|x\|^2$. Finally, Theorem 2.1 guarantees that an improved loss estimator over the unbiased estimator of loss $\delta_0(X)$ is $\delta(X) = \delta_0(X) + \alpha/\|x\|^2$ for $0 < \alpha < 4p$, which is Johnstone's result [32] for the James–Stein estimator.

Similarly Johnstone's result for $\varphi(X) = X$ can be constructed with $m(x) = 1$ (which is both subharmonic and superharmonic) and with the choice of the superharmonic function $\xi(x) = 1/\|x\|^2$, for which $K_0 = 2(p-4)$, so that $\delta(x) = p - \alpha/\|x\|^2$ dominates p for $0 < \alpha < 4(p-4)$.

We have shown that the unbiased estimator of loss can be dominated. Often one may wish to add a frequentist-validity constraint to a loss estimation problem. Specifically in our problem, the frequentist-validity constraint for some estimator δ would be $E_\theta[\delta(X)] \ge E_\theta[\delta_0(X)]$ for all θ. Kiefer [35] suggested that conditional and estimated confidence assessments should be conservatively biased, that is, the average reported loss should be greater than or equal to the average actual loss. Under such a frequentist-validity condition Lu and Berger [40] gave improved loss estimators for several of the most important Stein-type estimators. One of their estimators is a generalized Bayes estimator, suggesting that Bayesians and frequentists can potentially agree on a conditional assessment of loss.

A possible problem with the improved estimator defined in (2.16) is that it may be negative, which is undesirable since we are estimating a nonnegative quantity. A simple remedy to this problem is to use a positive-part estimator. If we define the positive-part as $\delta^+ = \max\{\delta, 0\}$, the loss difference between δ^+ and δ is $(\delta - L(\theta, \varphi))^2 - (\delta^+ - L(\theta, \varphi))^2 = (\delta^2 - 2\delta L(\theta, \varphi))\mathbb{1}_{\delta \le 0}$, hence it is always nonnegative. Therefore the risk difference is positive, which implies that δ^+ dominates δ. It would be of interest to find an estimator that dominates δ^+.

In the context of variance estimation, despite warnings on its inappropriate behavior (Stein [46], Brown [9]) the decision-theoretic approach to the normal variance estimation is typically based on the standardized quadratic loss function, where overestimation of the

variance is much more severely penalized than underestimation, thus leading to presumably too small estimates. Similarly in loss estimation under quadratic loss, the overestimation of the loss is also much more severely penalized than underestimation. A possible alternative to quadratic loss would be a Stein-type loss. Suppose $\varphi(X)$ is an estimator of θ under $\|\theta - \varphi(X)\|^2$ and let $\delta(X)$ be an estimator of $\|\theta - \varphi(X)\|^2$ for $\delta(X) > 0$. Then we can define the Stein-type loss for evaluating $\delta(X)$ as

$$
L(\theta, \varphi(X), \delta(X)) = \frac{\|\theta - \varphi(X)\|^2}{\delta(X)}
$$
(2.20)
$$
- \log \frac{\|\theta - \varphi(X)\|^2}{\delta(X)} - 1.
$$

The analysis of the loss estimates under the Stein-type loss is more challenging but can be carried out using the integration-by-parts tools developed in this section.

2.2 Dominating the Posterior Risk

In the previous sections, we have seen that the unbiased estimator of loss should be often dismissed since it can be dominated. When a (generalized) Bayes estimator of θ is available, incorporating the same prior information for estimating the loss of this Bayesian estimator is coherent, and we may expect that the corresponding Bayes estimator is a good candidate to improve on the unbiased estimator of loss. However, somewhat surprisingly, Fourdrinier and Strawderman [22] found that, in the normal setting considered in Section 2.1, the unbiased estimator often dominates the corresponding generalized Bayes estimator of loss for priors which give minimax estimators in the original point estimation problem. They also gave a general inadmissibility result for a generalized Bayes estimator of loss. While much of their focus is on pseudo-Bayes estimators, in this section, we essentially present their results on generalized Bayes estimators.

For a given generalized prior π, we denote the generalized marginal by m and the generalized Bayes estimator of θ by

$$
(2.21) \qquad \varphi_m(X) = X + \frac{\nabla m(X)}{m(X)}.
$$

Then (see Stein [48]) the unbiased estimator of risk of $\varphi_m(X)$ is

$$
(2.22) \quad \delta_0(X) = p + 2 \frac{\Delta m(x)}{m(X)} - \frac{\|\nabla m(X)\|^2}{m^2(X)}
$$

while the posterior risk of $\varphi_m(X)$ is

$$
(2.23) \quad \delta_m(X) = p + \frac{\Delta m(X)}{m(X)} - \frac{\|\nabla m(X)\|^2}{m^2(X)}.
$$

Domination of $\delta_0(X)$ over $\delta_m(X)$ is obtained thanks to the fact that their risk admits $(\Delta m(X)/m(X))^2 - 2\Delta^{(2)}m(X)/m(X)$ as an unbiased estimator of their risk difference, that is,

$$
\mathcal{R}(\theta, \varphi_m, \delta_0) - \mathcal{R}(\theta, \varphi_m, \delta_m)
$$
(2.24)
$$
= E_\theta \left[\left(\frac{\Delta m(X)}{m(X)} \right)^2 - 2 \frac{\Delta^{(2)}m(X)}{m(X)} \right],
$$

where $\Delta^{(2)}m = \Delta(\Delta m)$ is the bi-Laplacian of m (see [22]). Thus the above domination will occur as soon as

$$
(2.25) \quad \left(\frac{\Delta m(X)}{m(X)} \right)^2 - 2 \frac{\Delta^{(2)}m(X)}{m(X)} \leq 0.
$$

Applicability of that last condition is underlined by the remarkable fact that if the prior π satisfies (2.25), that is, if

$$
(2.26) \quad \left(\frac{\Delta \pi(\theta)}{\pi(\theta)} \right)^2 - 2 \frac{\Delta^{(2)}\pi(\theta)}{\pi(\theta)} \leq 0,
$$

then (2.25) is satisfied for the marginal m.

As an example, Fourdrinier and Strawderman [22] considered $\pi(\theta) = (\|\theta\|^2/2 + a)^{-b}$ (where $a \geq 0$ and $b \geq 0$) and showed that, if $p \geq 2(b + 3)$ then (2.26) holds and hence δ_0 dominates δ_m. Since π is integrable if and only if $b > \frac{p}{2}$ (for $a > 0$), the prior π is improper whenever this condition for domination of δ_0 over δ_m holds. Of course, whenever π is proper, the Bayes estimator δ_m is admissible provided its Bayes risk is finite.

Inadmissibility of the generalized Bayes loss estimator is not exceptional. Thus, in [22], the following general inadmissibility result is given; its proof is parallel to the proof of Theorem 2.1.

THEOREM 2.2. *Let m be a nonnegative function such that $\nabla m/m \in W_{\mathrm{loc}}^{1,1}(\mathbb{R}^p)$. Let ξ be a real-valued function satisfying the conditions of Theorem* 2.1. *Then δ_m is inadmissible and a class of dominating estimators is given by*

$$
\delta_m(X) + \alpha \operatorname{sgn}(\Delta \xi(X)) \frac{\xi(X)}{m(X)} \quad \text{for } 0 < \alpha < 2K_0.
$$

Note that, unlike Theorem 2.1, neither the superharmonicity condition nor the subharmonicity condition on m is needed. Note also that Theorem 2.2 gives conditions of improvement on δ_m while Theorem 2.1 looks for improvements on δ_0. As we saw, often δ_0 dominates δ_m. So it is not surprising that the proofs of the

two theorems are parallel; more precisely, it suffices to suppress, in the proof of Theorem 2.1, the superharmonicity (or subharmonicity) condition on m to obtain the proof of Theorem 2.2.

In [22], it is suggested that the inadmissibility of the generalized Bayes (or pseudo-Bayes) estimator is due to the fact that the loss function $(\delta(x) - \|\varphi(x) - \theta\|^2)^2$ may be inappropriate. The possible deficiency of this loss is illustrated by the following simple result concerning estimation of the square of a location parameter in $\mathbb{R}^1$.

Suppose $X \in \mathbb{R} \sim f((X - \theta)^2)$ such that $E_\theta[X^4] < \infty$. Consider estimation of θ^2 under loss $(\delta - \theta^2)^2$. The generalized Bayes estimator δ_π of θ^2 with respect to the uniform prior $\pi(\theta) \equiv 1$ is given by

$$\delta_\pi(X) = \frac{\int \theta^2 f((X - \theta)^2) \, d\theta}{\int f((X - \theta)^2) \, d\theta} = X^2 + E_0[X^2].$$

Since this estimator has constant bias $2E_0[X^2]$, it is dominated by the unbiased estimator $X^2 - E_0[X^2]$ (the risk difference is $4(E_0[X^2])^2$). Hence δ_π is inadmissible for any $f(\cdot)$ such that $E_\theta[X^4] < \infty$.

2.3 Examples of Improved Estimators

In this subsection, we give some examples of Theorems 2.1 and 2.2. The only example up to this point of an improved estimator over the unbiased estimator of loss $\delta_0(X)$ is $\delta(X) = \delta_0(X) + \alpha/\|x\|^2$ for $0 < \alpha < 4p$, which is Johnstone's result [32]. Although the shrinkage factor in Theorems 2.1 and 2.2 is the same, in the examples below we will only focus on improvements of posterior risk.

As an application of Theorem 2.2, let $\xi_b(x) = (\|x\|^2 + a)^{-b}$ (with $a \geq 0$ and $b \geq 0$). It can be shown that we have $\Delta\xi_b(x) < 0$ for $a \geq 0$ and $0 < 2(b + 1) < p$. Also $\Delta\xi_b(x) > 0$ if $a = 0$ and $2(b + 1) > p$. Furthermore

$$\frac{\xi_b^2(x)}{|\Delta\xi_b(x)|} = \frac{1}{2b|p - 2(b + 1)\|x\|^2/(\|x\|^2 + a)|}$$
$$\cdot \frac{1}{(\|x\|^2 + a)^{b-1}}.$$

(a) Suppose that $0 < 2(b + 1) < p$ and $a \geq 0$. Then

$$\frac{\xi_b^2(x)}{|\Delta\xi_b(x)|} \leq \frac{1}{2b(p - 2(b + 1))} \frac{1}{(\|x\|^2 + a)^{b-1}}$$

and $E_\theta[(\Delta\xi_b(X)/\xi_b(X))^2] < \infty$ since it is bounded from above by a quantity proportional to $E_\theta[(\|X\|^2 + a)^{-2}]$, which is finite for $a > 0$ or for $a = 0$ and $p > 4$.

Suppose that $m(x)$ is greater than or equal to some multiple of $(\|x\|^2 + a)^{1-b}$ or equivalently

$$(2.27) \quad m(x) \geq \frac{k}{2b(p - 2(b + 1))} \frac{1}{(\|x\|^2 + a)^{b-1}}$$

for some $k > 0$. Theorem 2.2 implies that $\delta_m(X)$ is inadmissible and is dominated by

$$\delta_m(X) - \frac{\alpha}{m(X)(\|X\|^2 + a)^b}$$

for $0 < \alpha < 4b(p - 2(b + 1)) \inf_{x \in \mathbb{R}^p} (m(x)(\|x\|^2 + a)^{b-1})$. Note that the improved estimators shrink toward 0.

Suppose, for example, that $m(x) \equiv 1$. Then (2.27) is satisfied for $b \geq 1$. Here $\varphi_m(X) = X$ and $\delta_m(X) = p$. Choosing $b = 1$, an improved class of estimators is given by $p - \frac{\alpha}{\|X\|^2 + a}$ for $0 < \alpha < 4(p - 4)$. The case $a = 0$ is equivalent to Johnstone's result for this marginal.

(b) Suppose that $2(b + 1) > p > 4$ and $a = 0$. Then

$$\frac{\xi_b^2(x)}{|\Delta\xi_b(x)|} = \frac{1}{2b(2(b + 1) - p)} \frac{1}{\|x\|^{2(b-1)}}.$$

A development similar to the above implies that, when $m(x)$ is greater than or equal to some multiple of $\|x\|^{2(1-b)}$, an improved estimator is

$$\delta_m(X) + \frac{\alpha}{m(X)\|X\|^{2b}}$$

for $0 < \alpha < 4b(2(b + 1) - p) \inf_{x \in \mathbb{R}^p} (m(x)\|x\|^{2(b-1)})$.

Note that, in this case, the correction term is positive and hence the estimators expand away from 0. Note also that this result only works for $a = 0$ and hence applies to pseudo-marginals which are unbounded in a neighborhood of 0. Since all marginals corresponding to a generalized prior π are bounded, this result can never apply to generalized Bayes procedures but only to pseudo-Bayes procedures.

Suppose, for example, that $m(x) = \|x\|^{2-p}$. Here $\varphi_m(X) = (1 - \frac{p-2}{\|x\|^2})X$ is the James–Stein estimator and $\delta_m(X) = p - \frac{(p-2)^2}{\|X\|^2}$. In particular, the above applies for $b - 1 = \frac{p-2}{2}$, that is, for $b = \frac{p}{2} > \frac{p-2}{2}$. An improved estimator is given by $\delta_m(X) + \frac{\gamma}{\|X\|^2}$ for $0 < \gamma < 4p$. This again agrees with Johnstone's result for James–Stein estimators.

3. ESTIMATING THE QUADRATIC LOSS OF A p-NORMAL MEAN WITH UNKNOWN VARIANCE

In Section 2 it was assumed that the covariance matrix was known and equal to the identity matrix I_p.

Typically, the covariance is unknown and should be estimated. In the case where it is of the form $\sigma^2 I_p$ with σ^2 unknown, Wan and Zou [51] showed that, for the invariant loss $\|\varphi(X) - \theta\|^2/\sigma^2$, Johnstone's result [32] can be extended when estimating the loss of the James–Stein estimator. In fact, the general framework considered in Section 2 can be extended to the case where σ^2 is unknown, and we show that a condition parallel to Condition (2.13) can be found.

Before stating the main result for the unknown variance case, we need an extension of Stein's identity involving the sample variance.

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

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

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

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

(ii) *if $h(x, s)$ is a function from $\mathbb{R}^p \times \mathbb{R}_+$ into $\mathbb{R}$ such that, for any $s \in \mathbb{R}_+$, $h(\cdot, s)$ is weakly differentiable, 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 (i) is just Stein's lemma (cf. [48]). Part (ii) can be seen as a particular case of Lemma 1(ii) (established for elliptically symmetric distributions) of Fourdrinier et al. [23], although we will present a direct proof. The joint distribution of (X, S) can be viewed as resulting, in the setting of the canonical form of the general linear model, from the distribution of $(X, U) \sim \mathcal{N}((\theta, 0), \sigma^2 I_{p+k})$ with $S = \|U\|^2$. Then we can write

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

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

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

since

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

The following theorem provides an extension of results in Section 2 to the setting of an unknown variance. The necessary conditions to insure the finiteness of the risks are given in Appendix A.1.

THEOREM 3.1. *Let $X \sim \mathcal{N}(\theta, \sigma^2 I_p)$ where θ and σ^2 are unknown and $p \geq 5$ and let S be a nonnegative random variable independent of X and such that $S \sim \sigma^2 \chi_k^2$. Consider an estimator of θ of the form $\varphi(X, S) = X + Sg(X, S)$ 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 invariant loss $\|\varphi(X, S) - \theta\|^2/\sigma^2$ is

$$\delta_0(X, S)$$
$$(3.1) = p + S\left\{(k + 2)\|g(X, S)\|^2 + 2\operatorname{div}_X g(X, S)\right.$$
$$\left. + 2S\frac{\partial}{\partial S}\|g(X, S)\|^2\right\}.$$

Its risk $\mathcal{R}(\theta, \sigma^2, \varphi, \delta_0) = E_{\theta,\sigma^2}[(\delta_0(X, S) - \|\varphi(X, S) - \theta\|^2/\sigma^2))^2]$ is finite as soon as $E_{\theta,\sigma^2}[S^2\|g(X, S)\|^4] < \infty$, $E_{\theta,\sigma^2}[(S\operatorname{div}_X g(X, S))^2] < \infty$ and $E_{\theta,\sigma^2}[(S^2\frac{\partial}{\partial S}\|g(X, S)\|)^2] < \infty$.

Furthermore, for any function $\gamma(X)$ such that $E_{\theta,\sigma^2}[\gamma^2(X)] < \infty$, the risk difference $\mathcal{D}(\theta, \sigma^2, \varphi, \delta) = \mathcal{R}(\theta, \sigma^2, \varphi, \delta) - \mathcal{R}(\theta, \sigma^2, \varphi, \delta_0)$ between the estimators $\delta(X, S) = \delta_0(X, S) - S\gamma(X)$ and $\delta_0(X, S)$ is given by

$$E_{\theta,\sigma^2}\left[S^2\left\{\gamma^2(X) + \frac{2}{k + 2}\Delta\gamma(X)\right.\right.$$
$$(3.2)$$
$$\left.\left. + 4g^t(X, S)\nabla\gamma(X) + 4\gamma(X)\|g(X, S)\|^2\right\}\right].$$

Therefore a sufficient condition for $\mathcal{D}(\theta, \sigma^2, \varphi, \delta)$ to be nonpositive, and hence for $\delta(X, S)$ to improve on $\delta_0(X, S)$, is

$$\gamma^2(x) + \frac{2}{k+2}\Delta\gamma(x) + 4g^t(x, s)\nabla\gamma(x)$$
(3.3)
$$+ 4\gamma(x)\|g(x, s)\|^2 \leq 0$$

for any $x \in \mathbb{R}^p$ and any $s \in \mathbb{R}_+$.

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

$$\frac{1}{\sigma^2}\|X - \theta\|^2 + 2\frac{S}{\sigma^2}(X - \theta)^t g(X, S)$$
(3.4)
$$+ \frac{S^2}{\sigma^2}\|g(X, S)\|^2.$$

Clearly $E_{\theta,\sigma^2}[\sigma^{-2}\|X - \theta\|^2] = p$ and Lemma 3.1 implies that

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

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

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

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

$$\left.\left. + 2S\frac{\partial}{\partial S}\|g(X, S)\|^2\right\}\right].$$

Therefore $R(\theta, \varphi) = E_{\theta,\sigma^2}[\delta_0(X, S)]$ with $\delta_0(X, S)$ given in (3.1), which means that $\delta_0(X, S)$ is an unbiased estimator of the invariant loss $\|\varphi(X, S) - \theta\|^2/\sigma^2$. The fact that the risk $\mathcal{R}(\theta, \sigma^2, \varphi, \delta_0)$ of $\delta_0(X)$ is finite is shown in Lemma A.1.

Now consider the finiteness of the risk of the alternative loss estimator $\delta(X, S) = \delta_0(X, S) - S\gamma(X)$. It is easily seen that its difference in loss $d(\theta, \sigma^2, X, S)$ with $\delta_0(X, S)$ can be written as

$$d(\theta, \sigma^2, X, S)$$

$$= \left(\delta_0(X, S) - \frac{1}{\sigma^2}\|\varphi(X) - \theta\|^2 - S\gamma(X)\right)^2$$

(3.5)
$$- \left(\delta_0(X, S) - \frac{1}{\sigma^2}\|\varphi(X) - \theta\|^2\right)^2$$

$$= S^2\gamma^2(X)$$

$$- 2S\gamma(X)\left(\delta_0(X, S) - \frac{1}{\sigma^2}\|\varphi(X) - \theta\|^2\right).$$

Hence, since $E_{\theta,\sigma^2}[\|\varphi(X, S) - \theta\|^2/\sigma^2] < \infty$ as the risk of the estimator $\varphi(X, S)$, the condition $E_{\theta,\sigma^2}[\gamma^2(X)] < \infty$ ensures that the expectation of the loss in (3.5), that is, the risk difference $\mathcal{D}(\theta, \sigma^2, \varphi, \delta)$ is finite. Then $\mathcal{R}(\theta, \sigma^2, \varphi, \delta) < \infty$ since $\mathcal{R}(\theta, \sigma^2, \varphi, \delta_0) < \infty$.

We now express the risk difference $\mathcal{D}(\theta, \sigma^2, \varphi, \delta) = E_{\theta,\sigma^2}[d(\theta, \sigma^2, X, S)]$. Using (3.1) and expanding $\|\varphi(X, S) - \theta\|^2/\sigma^2$ give that $d(\theta, \sigma^2, X, S)$ in (3.5) can be written as $d(\theta, \sigma^2, X, S) = A(X, S) + B(\theta, \sigma^2, X, S)$ where

$$A(X, S) = S^2\gamma^2(X) - 2pS\gamma(X)$$

$$- 2(k + 2)S^2\gamma(X)\|g(X, S)\|^2$$
(3.6)
$$- 4S^2\gamma(X)\,\text{div}_X\,g(X, S)$$

$$- 4S^3\gamma(X)\frac{\partial}{\partial S}\|g(X, S)\|^2$$

and

$$B(\theta, \sigma^2, X, S) = 2\frac{S^3}{\sigma^2}\gamma(X)\|g(X, S)\|^2$$

(3.7)
$$+ 2\frac{S}{\sigma^2}\gamma(X)\|X - \theta\|^2$$

$$+ 4\frac{S^2}{\sigma^2}\gamma(X)(X - \theta)^t g(X, S).$$

Through Lemma 3.1(ii) with $h(x, s) = 2\frac{s^3}{\sigma^2}\gamma(x) \cdot \|g(x, s)\|^2$, the expectation of the first term in the right-hand side of (3.7) equals

$$E_{\theta,\sigma^2}\left[2\frac{S^3}{\sigma^2}\gamma(X)\|g(X, S)\|^2\right]$$

(3.8)
$$= E_{\theta,\sigma^2}\left[2(k + 4)S^2\gamma(X)\|g(X, S)\|^2\right.$$

$$\left. + 4S^3\gamma(X)\frac{\partial}{\partial S}\|g(X, S)\|^2\right].$$

An iterated application of Lemma 3.1(i) to the expectation of the second term in the right-hand side of (3.7) allows to write

$$E_{\theta,\sigma^2}\left[2\frac{S}{\sigma^2}\gamma(X)\|X - \theta\|^2\right]$$

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

$$= E_{\theta,\sigma^2}[2\,\text{div}_X\{S\gamma(X)(X - \theta)\}]$$

$$= E_{\theta,\sigma^2}[2pS\gamma(X) + 2S(X - \theta)^t\nabla\gamma(X)]$$

$$= E_{\theta,\sigma^2}[2pS\gamma(X) + 2\sigma^2 S\Delta\gamma(X)]$$

which, as $S \sim \sigma^2 \chi_k^2$ entails that $E[S^2/(k+2)] = E[\sigma^2 S]$ and as S is independent of X, gives

$$(3.9) \quad \begin{aligned} &E_{\theta,\sigma^2}\left[2\frac{S}{\sigma^2}\gamma(X)\|X-\theta\|^2\right] \\ &= E_{\theta,\sigma^2}\left[2pS\gamma(X) + 2\frac{S^2}{k+2}\Delta\gamma(X)\right]. \end{aligned}$$

As for the third term in the right-hand side of (3.7), its expectation can also be expressed using Lemma 3.1(i) as

$$(3.10) \quad \begin{aligned} &E_{\theta,\sigma^2}\left[4\frac{S^2}{\sigma^2}\gamma(X)(X-\theta)^t g(X,S)\right] \\ &= E_{\theta,\sigma^2}[4S^2 \operatorname{div}_X\{\gamma(X)g(X,S)\}] \\ &= E_{\theta,\sigma^2}[4S^2\gamma(X)\operatorname{div}_X\{g(X,S)\} \\ &\qquad\qquad + 4S^2 g(X,S)^t \nabla\gamma(X)] \end{aligned}$$

by the product rule for the divergence. Finally, gathering (3.8), (3.9) and (3.10) yields an expression of (3.7) which, with (3.6), gives the integrand term of (3.2), which is the desired result. $\square$

As an example, consider the James–Stein estimator with unknown variance

$$\varphi^{\mathrm{JS}}(X,S) = X - \frac{p-2}{k+2}\frac{S}{\|X\|^2}X.$$

Here the shrinkage factor is the product of a function of S with a function of X so that, through routine calculation, the unbiased estimator of loss is

$$\delta_0(X,S) = p - \frac{(p-2)^2}{k+2}\frac{S}{\|X\|^2}.$$

For a correction of the form $\gamma(x) = -d/\|x\|^2$ with $d \geq 0$, it is easy to check that the expression in (3.3) equals

$$\begin{aligned} &d^2 + 4\frac{p-4}{k+2}d - 8\frac{p-2}{k+2}d - 4\left(\frac{p-2}{k+2}\right)^2 d \\ &= d\left(d - \frac{4}{k+2}\left[p + \frac{(p-2)^2}{k+2}\right]\right) \end{aligned}$$

which is negative for $0 < d < \frac{4}{k+2}[p + \frac{(p-2)^2}{k+2}]$ and gives domination of $p - \frac{(p-2)^2}{k+2}\frac{S}{\|X\|^2} + \frac{d}{\|x\|^2}$ over $p - \frac{(p-2)^2}{k+2}\frac{S}{\|X\|^2}$. This condition recovers the result of Wan and Zou [51] who considered the case $d = \frac{2}{k+2}[p + \frac{(p-2)^2}{k+2}]$.

4. EXTENSIONS TO THE SPHERICAL CASE

4.1 Estimating the Quadratic Loss of the Mean of a Spherical Distribution

In the previous sections the loss estimation problem was considered for the normal distribution setting. The normal distribution has been generalized in two important directions, first as a special case of the exponential family and second as a spherically symmetric distribution. In this section we will consider the latter. There are a variety of equivalent definitions and characterizations of the class of spherically symmetric distributions; a comprehensive review is given in [20]. We will use the representation of a random variable from a spherically symmetric distribution, $X = (X_1,\ldots,X_p)^t$, as $X \stackrel{d}{=} RU^{(p)} + \theta$, where $R = \|X - \theta\|$ is a random radius, $U^{(p)}$ is a uniform random variable on the p-dimensional unit sphere, where R and $U^{(p)}$ are independent. In such a situation, the distribution of X is said to be spherically symmetric around θ and we write $X \sim \mathrm{SS}_p(\theta)$. We also extend, in Section 4.2, these results to the case where the distribution of X is spherically symmetric and when a residual vector U is available (which allows an estimation of the variance factor σ^2).

Assume $X \sim \mathrm{SS}_p(\theta)$ and suppose we wish to estimate $\theta \in \mathbb{R}^p$ by a decision rule $\delta(X)$ using quadratic loss. Suppose that we also use quadratic loss to assess the accuracy of loss estimate $\delta(X)$; then the risk of this loss estimate is given by (1.2). In [26], the problem of estimating the loss when $\varphi(X) = X$ is the estimate of the location parameter θ is considered. The estimate φ is the least squares estimator and is minimax among the class of spherically symmetric distributions with bounded second moment. Furthermore, if one assumes the density of X exists and is unimodal, then φ is also the maximum likelihood estimator.

The unbiased constant estimate of the loss $\|X - \theta\|^2$ is $\delta_0 = E_\theta[R^2]$. Note that δ_0 is independent of θ, since $E_\theta[\|X-\theta\|^2] = E_0[\|X\|^2]$. Fourdrinier and Wells [26] showed that the unbiased estimator δ_0 can be dominated by $\delta_0 - \gamma$, where γ is a particular superharmonic function for the case where the sampling distribution is a scale mixture of normals and in more general spherical cases.

The development of the results depends on some interesting extensions of the classical Stein identities in (2.7) and (2.12) to the general spherical setting. Since the distribution of X, say P_θ, is spherically symmetric around θ, for every bounded function f, we have $E_\theta[f] = EE_{R,\theta}[f] = \int_{\mathbb{R}_+} E_{R,\theta}[f]\rho(dR)$, where ρ

is the distribution of the radius, namely the distribution of the norm $\|X - \theta\|$ under P_θ and where E and $E_{R,\theta}$ denote respectively the expectation with respect to the radial distribution and uniform distribution $U_{R,\theta}$ on the sphere $S_{R,\theta} = \{x \in \mathbb{R}^p \mid \|x - \theta\| = R\}$ of radius R and center θ. To deduce the various risk domination results it suffices to work conditionally on the radius, that is to say to replace P_θ by $U_{R,\theta}$ in the risk expressions. Let $\sigma_{R,\theta}$ denote the area measure on $S_{R,\theta}$. Therefore, for every Borel measurable set A, $U_{R,\theta}(A) = \sigma_{R,\theta}(A)/\sigma(S_{R,\theta}) = \Gamma(p/2)\sigma_{R,\theta}(A)/2\pi^{p/2}R^{p-1}$. Define the volume measure $\tau_{R,\theta}$ on the ball $B_{R,\theta} = \{x \in \mathbb{R}^p \mid \|x - \theta\| \leq R\}$ of radius R and center θ and denote the uniform distribution on $B_{R,\theta}$ as $V_{R,\theta}$. Hence, for every Borel measurable set A, $V_{R,\theta}(A) = \tau_{R,\theta}(A)/\tau_{R,\theta}(B_{R,\theta}) = p\Gamma(p/2)\tau_{R,\theta}(A)/2\pi^{p/2}R^p$. Suppose γ is a weakly differentiable vector-valued function; then by applying the Divergence Theorem for weakly differentiable functions to the definition of the expectation we have

$$E_\theta[(X - \theta)^t \gamma(X) \mid \|X - \theta\| = R]$$

$$(4.1) \qquad = \int_{S_{R,\theta}} (x - \theta)^t \gamma(x) U_{R,\theta}(dx)$$

$$= \frac{R}{\sigma_{R,\theta}(S_{R,\theta})} \int_{B_{R,\theta}} \operatorname{div} \gamma(x)\, dx.$$

If γ is a real-valued function, then it follows from (4.1) and the product rule applied to the vector-valued function $(x - \theta)\gamma(x)$ that

$$E_\theta[\|X - \theta\|^2 \gamma(X) \mid \|X - \theta\| = R]$$

$$= \int_{S_{R,\theta}} (x - \theta)^t (x - \theta)\gamma(x) U_{R,\theta}(dx)$$

$$(4.2) \qquad = \frac{R}{\sigma_{R,\theta}(S_{R,\theta})}$$

$$\cdot \int_{B_{R,\theta}} [p\gamma(x) + (x - \theta)^t \nabla\gamma(x)]\, dx.$$

Our first extension of Theorem 2.1 is to the class of spherically symmetric distributions that are scale mixtures of normal distributions. Well-known examples in the class of densities include the double exponential, multivariate t-distribution (hence, the multivariate Cauchy distribution). Let $\phi(x; \theta, I)$ be the probability density function of a random vector X with a normal distribution with mean vector θ and identity covariance matrix. Suppose that there is a probability measure on $\mathbb{R}_+$ such that the probability density function p_θ may be expressed as

$$(4.3) \qquad p_\theta(x|\theta) = \int_0^\infty \phi(x; \theta, I/\varsigma) G(d\varsigma).$$

One can think of Υ being a random variable with distribution G; the conditional distribution of X given $\Upsilon = \varsigma$, $X|\Upsilon = \varsigma$, is $N_p(\theta, I/\varsigma)$. This class contains some heavy-tailed distributions, possibly with no moments. It is well known (see [20]) that, if a spherical distribution has a density p_θ, it is of the form $p_\theta(x) = g(\|x - \theta\|^2)$ for a measurable positive function g (called the generating function).

In the scale mixture of normals setting the unbiased estimate, δ_0, of risk equals

$$E[R^2] = E_\theta[\|X - \theta\|^2] = p \int_0^\infty \varsigma^{-1} G(d\varsigma).$$

It is easy to see that the risk of the unbiased estimator δ_0 is finite if and only if $E_\theta[\|X - \theta\|^4] < \infty$, which holds if

$$(4.4) \qquad \int_0^\infty \varsigma^{-2} G(d\varsigma) < \infty.$$

The main theorem in [26] is the following domination result of an improved estimator of loss over the unbiased loss estimator.

THEOREM 4.1. *Assume the distribution of X is a scale mixture of normal random variables as in (4.3) such that (4.4) is satisfied and such that*

$$(4.5) \qquad \int_{\mathbb{R}_+} \varsigma^{p/2} G(d\varsigma) < \infty.$$

Also, assume that the shrinkage function γ is twice weakly differentiable on $\mathbb{R}^p$ and satisfies $E_\theta[\gamma^2] < \infty$, for every $\theta \in \mathbb{R}^p$. Then a sufficient condition for $\delta_0 - \gamma$ to dominate δ_0 is that γ satisfies the differential inequality

$$(4.6) \quad k\Delta\gamma + \gamma^2 < 0 \quad \text{with } k = 2 \frac{\int_{\mathbb{R}_+} \varsigma^{p/2} G(d\varsigma)}{\int_{\mathbb{R}_+} \varsigma^{p/2-2} G(d\varsigma)}.$$

As an example let $\gamma(x) = c/\|x\|^2$ where c is a positive constant. Note that γ is twice weakly differentiable only when $p > 4$ (thus its Laplacian exists as a locally integrable function). Then it may be shown that $\Delta\gamma(x) = -2c(p-4)/\|x\|^4$. Hence $k\Delta(x) + \gamma^2(x) = -2kc(p-4)/\|x\|^4 + c^2/\|x\|^4 < 0$ if $-2kc(p-4) + c^2 < 0$, that is, $0 < c < 2k(p-4)$. It is easy to see that the optimal value of c for which this inequality is the most negative equals $k(p-4)$, so an interesting estimate in this class of γ's is $\delta = \delta_0 - k(p-4)/\|x\|^2$ ($p > 4$). This is precisely the estimate proposed by [32] in the normal distribution case $N_p(\theta, I)$ where $k = 2$; recall, in that case $\delta_0 = p$.

In this example, we have assumed that the dimension p is greater than 4. In general we can have domination

as long as the assumptions of the theorem are valid. Actually, Blanchard and Fourdrinier [7] showed explicitly that, when $p \leq 4$, the only solution γ in $L^2_{\text{loc}}(\mathbb{R}^p)$ of the inequality $k\Delta\gamma + \gamma^2 \leq 0$ is $\gamma \equiv 0$, almost everywhere with respect to the Lebesgue measure λ. Now, in the normal setting $N_p(\theta, I/\varsigma)$, an unbiased estimator of the risk difference between an estimator $\delta = \delta_0 - \gamma$ and δ_0 is $2\varsigma^{-2}\Delta\gamma + \gamma^2$. Hence, for dimensions 4 or less, it is impossible to find an estimator $\delta = \delta_0 - \gamma$ whose unbiased estimate of risk is always less than that of δ_0. Indeed we cannot have $E_\theta[2\varsigma^{-2}\Delta\gamma + \gamma^2] < 0$, for some θ, without having $\lambda[\varsigma^{-2}\Delta\gamma(x) + \gamma^2(x) < 0] > 0$, which entails that $\lambda[\gamma(x) \neq 0] > 0$.

In the case of scale mixture of normal distributions, the conjecture of admissibility of δ_0 for lower dimensions, although it is probably true, remains open. Indeed, under conditions of Theorem 4.1, $k\Delta\gamma + \gamma^2$ is no longer an unbiased estimator of the risk difference and $E_\theta[k\Delta\gamma + \gamma^2]$ is only its upper bound. The use of Blyth's method would need to specify the distribution of X (i.e., the mixture distribution G). It is worth noting that dimension-cutoff also arises through the finiteness of $E_\theta[\gamma^2]$ when using the classical shrinkage function $c/\|x\|^2$.

In order to prove Theorem 4.1 we need some additional technical results. The first lemma gives some important properties of superharmonic functions and is found in Du Plessis [17] and the second lemma links the integral of the gradient on a ball with the integral of the Laplacian.

LEMMA 4.1. *If γ is a real-valued superharmonic function, then*:

(i) $\int_{S_{R,\theta}} \gamma(x) U_{R,\theta}(dx) \leq \int_{B_{R,\theta}} \gamma(x) V_{R,\theta}(dx)$,
(ii) *both of the integrals in* (i) *are decreasing in R*.

PROOF.   See Sections 1.3 and 2.5 in [17].   □

LEMMA 4.2. *Suppose γ is a twice weakly differentiable function. Then*

$$\int_{B_{R,\theta}} (x - \theta)^t \nabla\gamma(x) V_{R,\theta}(dx)$$
$$= \frac{p\Gamma(p/2)}{2\pi^{p/2}} \frac{1}{R^p} \int_0^R r \int_{B_{r,\theta}} \Delta\gamma(x)\, dx\, dr.$$

PROOF.   Since the density of the distribution of the radius under $V_{R,\theta}$ is $(p/R^p)r^{p-1}\mathbb{1}_{[0,R]}(r)$, we have

$$\int_{B_{R,\theta}} (x - \theta)^t \nabla\gamma(x) V_{R,\theta}(dx)$$
$$= \int_0^R \int_{S_{r,\theta}} (x - \theta)^t \nabla\gamma(x) U_{r,\theta}(dx) \frac{p}{R^p} r^{p-1}\, dr.$$

The result follows from applying (4.1) to the innermost integral of the right-hand side of this equality and by recalling the fact that $\sigma_{r,\theta}(S_{r,\theta}) = (2\pi^{p/2}/\Gamma(p/2)) \cdot r^{p-1}$.   □

PROOF OF THEOREM 4.1.   Denoting by ρ the distribution of the radius $\|X - \theta\|$, the risk difference between δ_0 and $\delta_0 - \gamma$ equals $\alpha(\theta) + \beta(\theta)$ where

(4.7)
$$\alpha(\theta) = \int_{\mathbb{R}_+} \alpha_R(\theta)\rho(dR) \quad \text{and}$$
$$\beta(\theta) = \int_{\mathbb{R}_+} \beta_R(\theta)\rho(dR)$$

with

(4.8)
$$\alpha_R(\theta) = 2R^2 \int_{B_{R,\theta}} \gamma(x) V_{R,\theta}(dx)$$
$$- 2\lambda_0 \int_{S_{R,\theta}} \gamma(x) U_{R,\theta}(dx)$$

and

(4.9)
$$\beta_R(\theta) = 2\frac{R^2}{p} \int_{B_{R,\theta}} (x - \theta)^t \nabla\gamma(x) V_{R,\theta}(dx)$$
$$+ \int_{S_{R,\theta}} \gamma^2(x) U_{R,\theta}(dx).$$

Indeed, the risk difference conditional on the radius R equals

$$\int_{S_{R,\theta}} [2\|x - \theta\|^2 \gamma(x) - 2\lambda_0\gamma(x) + \gamma^2(x)] U_{R,\theta}(dx)$$

and the result follows from (4.2) applied to the first term between brackets.

Let us first deal with $\alpha(\theta)$ considering the first term in (4.8). We have from the definition of $V_{R,\theta}$ and an application of Fubini's theorem

(4.10)
$$\int_{\mathbb{R}_+} R^2 \int_{B_{R,\theta}} \gamma(x) V_{R,\theta}(dx)\rho(dR)$$
$$= p\frac{\Gamma(p/2)}{2\pi^{p/2}} \int_{\mathbb{R}_+} R^{2-p} \int_{B_{R,\theta}} \gamma(x)\, dx\rho(dR)$$
$$= p\frac{\Gamma(p/2)}{2\pi^{p/2}} \int_{\mathbb{R}^p} \gamma(x) \int_{\|x-\theta\|}^{+\infty} R^{2-p}\rho(dR)\, dx.$$

Now, for fixed $\varsigma \geq 0$, in the normal case $N_p(\theta, I/\varsigma)$ the distribution ρ_ς of the radius has the density f_ς of the form $f_\varsigma(R) = \frac{\varsigma^{p/2}}{2^{p/2-1}\Gamma(p/2)} R^{p-1} \exp\{-\frac{\varsigma R^2}{2}\}$ and

$\delta_0 = \frac{p}{\varsigma}$. Thus the expression (4.10) becomes

$$\int_{\mathbb{R}_+} R^2 \int_{B_{R,\theta}} \gamma(x) V_{R,\theta}(dx)\rho(dR)$$

$$= \frac{p\varsigma^{p/2}}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} \gamma(x) \int_{\|x-\theta\|}^{+\infty} R \exp\left\{-\frac{\varsigma R^2}{2}\right\} dR\,dx$$

$$= \frac{p\varsigma^{p/2-1}}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} \gamma(x) \exp\left\{-\frac{\varsigma}{2}\|x-\theta\|^2\right\} dx$$

$$= \frac{p}{\varsigma} \int_{\mathbb{R}_+} \int_{S_{R,\theta}} \gamma(x) U_{R,\theta}(dx)\rho_\varsigma(dR),$$

the last equality holding since $X \overset{D}{=} RU^{(p)}$. Turning back to (4.7) and (4.8) and using the mixture representation with mixing distribution G, the expression of $\alpha(\theta)$ is written as

$$\alpha(\theta) = 2p \int_{\mathbb{R}_+} \left(\frac{1}{\varsigma} - \frac{\delta_0}{p}\right)$$

(4.11)
$$\cdot \int_{\mathbb{R}^p} \gamma(x) \left(\frac{\varsigma}{2\pi}\right)^{p/2}$$

$$\cdot \exp\left(-\frac{\varsigma}{2}\|x-\theta\|^2\right) dx\,G(d\varsigma).$$

It can be easily seen that the innermost integral in (4.11) is proportional to

$$\int_0^\infty \int_{S_{(u/\varsigma)^{1/2},\theta}} \gamma(x)\,dU_{S_{(u/\varsigma)^{1/2},\theta}} u^{p/2-1} \exp\left(-\frac{u}{2}\right) du$$

and hence is nondecreasing in ς by superharmonicity of γ induced by the inequality in (4.6) and by Lemma 4.1(ii). Thus, since $\delta_0 = p/\varsigma$ for fixed ς, the expression for $\alpha(\theta)$ in (4.11) is a nonpositive covariance with respect to G.

We can now treat the integral of the expression $\beta(\theta)$ in the same manner. The function $x \to (x-\theta)^t \nabla\gamma(x)$ and the function $x \to \nabla\gamma(x)$ taking successively the role of the function γ, we obtain

$$\int_{\mathbb{R}_+} \frac{R^2}{p} \int_{B_{R,\theta}} (x-\theta)^t \nabla\gamma(x) V_{R,\theta}(dx)\rho_\varsigma(dR)$$

$$= \frac{1}{\varsigma} \int_{\mathbb{R}_+} \int_{S_{R,\theta}} (x-\theta)^t \nabla\gamma(x) U_{R,\theta}(dx)\rho_\varsigma(dR)$$

$$= \frac{1}{\varsigma} \int_{\mathbb{R}_+} \frac{R^2}{p} \int_{B_{R,\theta}} \nabla\gamma(x)\,dx\rho_\varsigma(dR)$$

$$= \frac{\varsigma^{p/2-2}}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} \nabla\gamma(x) \exp\left\{-\frac{\varsigma}{2}\|x-\theta\|^2\right\} dx$$

by applying (4.1) for the second equality and remembering that $\triangle\gamma = \mathrm{div}(\nabla\gamma)$. Therefore by the Fubini

Theorem $\beta(\theta)$ can be reexpressed as

$$\beta(\theta) = \int_{\mathbb{R}^p} \left(2\triangle\gamma(x)\right.$$

$$\cdot \frac{\int_{\mathbb{R}_+} \varsigma^{p/2-2} \exp(-\varsigma\|x-\theta\|^2/2)G(d\varsigma)}{\int_{\mathbb{R}_+} \varsigma^{p/2} \exp(-\varsigma\|x-\theta\|^2/2)G(d\varsigma)}$$

(4.12)
$$\left. + \gamma^2(x)\right)$$

$$\cdot \int_{\mathbb{R}_+} \left(\frac{\varsigma}{2\pi}\right)^{p/2}$$

$$\cdot \exp\left(-\frac{\varsigma}{2}\|x-\theta\|^2\right) G(d\varsigma)\,dx.$$

Now, through a monotone likelihood ratio argument, the ratio of integrals in (4.12) can be seen to be bounded from below by the constant k in (4.6). Hence the inequality in (4.6) gives

$$\beta(\theta) \le \int_{\mathbb{R}^p} \left(k\triangle\gamma(x) + \gamma^2(x)\right)$$

$$\cdot \int_{\mathbb{R}_+} \left(\frac{\varsigma}{2\pi}\right)^{p/2}$$

$$\cdot \exp\left(-\frac{\varsigma}{2}\|x-\theta\|^2\right) G(d\varsigma)\,dx$$

$$< 0.$$

Finally, remembering that $\alpha(\theta)$ is nonpositive, it follows that the risk difference $\alpha(\theta) + \beta(\theta)$ between δ_0 and $\delta_0 - \gamma$ is negative, which proves the theorem. $\square$

The improved loss estimator result in Theorem 4.1 for scale mixture of normal distributions family was extended to a more general family of spherically symmetric distributions in [26]. In this setting the conditions for improvement rest on the generating function g of the spherical density p_θ. A sufficient condition for domination of δ_0 has the usual form $k\Delta\gamma + \gamma^2 \le 0$.

THEOREM 4.2. *Assume the spherical distribution of X with generating function g has finite fourth moment. Assume the function γ is nonnegative and twice weakly differentiable on $\mathbb{R}^p$ and satisfies $E_\theta[\gamma^2] < \infty$. If, for every $s \ge 0$,*

(4.13)
$$\frac{\int_s^\infty g(z)\,dz}{2g(s)} \le \frac{\delta_0}{p}$$

and if there exists a constant k such that, for any $s \ge 0$,

(4.14)
$$0 < k < \frac{\int_s^\infty zg(z)\,dz - s\int_s^\infty g(z)\,dz}{2g(s)},$$

then a sufficient condition for $\delta_0 - \gamma$ to dominate δ_0 is that γ satisfies the differential inequality

$$k\Delta\gamma + \gamma^2 < 0.$$

We have shown that one can dominate the unbiased constant estimator of loss by a shrinkage-type estimator. As in the normal case one may wish to add a frequentist-validity constraint to the loss estimation problem. It is easy to show that the only frequentist valid estimator of the form δ_0 would be the only frequentist valid loss estimator. The proof of this result follows from a randomization of the origin technique as in Hsieh and Hwang [30].

4.2 Estimating the Quadratic Loss of the Mean of a Spherical Distribution with a Residual Vector

In this section, we extend the ideas of the previous sections to a spherically symmetric distribution with a residual vector. We first develop an unbiased estimator of the loss and then construct a dominating shrinkage-type estimator. An important feature of our results is that the proposed loss estimates dominate the unbiased estimates for the entire class of spherically symmetric distributions. That is, the domination results are robust with respect to spherical symmetry.

Let $(X, U) \sim \mathrm{SS}(\theta, 0)$ where $\dim X = \dim \theta = p$ and $\dim U = \dim 0 = k$ ($p + k = n$). For convenience of notation, here (X, U) and $(\theta, 0)$ represent $n \times 1$ vectors (see Appendix A.2 for more details on this model). Unlike Section 4.1, the dimension of the observable (X, U) is greater than the dimension of the estimand θ. This model arises as the canonical form of the following seemingly more general model, the general linear model. Let V be an $n \times p$ matrix (of full rank p) which is often referred to as the design matrix. Suppose an $n \times 1$ vector Y is observed such that $Y = V\beta + \varepsilon$ where β is a $p \times 1$ vector of (unknown) regression coefficients and ε is an $n \times 1$ vector with a spherically symmetric distribution about 0. A common alternative representation of this model is $Y = \eta + \varepsilon$ where ε is as above and η is in the column space of V.

To understand this representation in terms of the general linear model, let $G = (G_1^t, G_2^t)^t$ be an $n \times n$ orthogonal matrix partitioned such that the first p rows of G (i.e., the rows of G_1 considered as column vectors) span the column space of V. Now let

$$\binom{X}{U} = \binom{G_1}{G_2} Y = \binom{G_1}{G_2} V\beta + G\varepsilon = \binom{\theta}{0} + G\varepsilon$$

with $\theta = G_1 V\beta$ and $G_2 V\beta = 0$ since the rows of G_2 are orthogonal to the columns of V. It follows from

the definition that (X, U) has a spherically symmetric distribution about $(\theta, 0)$. In this sense, the model given above is the canonical form of the general linear model.

The usual estimator of θ is the orthogonal projector X. A class of competing point estimators which are also considered is of the form $\varphi = X - \|U\|^2 g(X)$; g is a measurable function from $\mathbb{R}^p$ into $\mathbb{R}^p$. This class of estimators is closely related to Stein-like estimators (when estimating the mean of a normal distribution, the square of the residual term $\|u\|$ is used as an estimate of the unknown variance). Their domination properties are robust with respect to spherical symmetry (cf. [11] and [12]). We will first consider estimation of the loss of the usual least squares estimator X, then estimation of the loss of the more general shrinkage estimator φ. In order to assure the finiteness of their risk of the usual estimator X and the risk of the shrinkage estimator φ, we need two hypotheses (H1) and (H2) given in [11].

In the spherical case in Section 3, the risk of X was constant with respect to θ. Thus this risk provides an unbiased estimator of the loss, that is, $pE[R^2]/n$, which is subject to the knowledge of $E[R^2]$. Its properties, as the properties of any improved estimator, may depend on the specific underlying distribution. An important feature of the results in this subsection is that we propose an unbiased estimator δ_0 of the loss of X which is available for every spherically symmetric distribution (with finite fourth moment), that is, $\delta_0(X, U) = p\|U\|^2/k$. Thus we do not need to know the specific distribution, and we get robustness with an estimator which is no longer constant. Notice δ_0 makes sense because $p < n$ (i.e., $k \geq 1$).

In this subsection, we consider estimation of θ by X so that, as in the work of Fourdrinier and Wells [25], we deal with estimating the loss $\|X - \theta\|^2$. An unbiased estimator of that loss is given by $\delta_0(X, U) = p\|U\|^2/k$, that we write $\delta_0(U)$ since it depends only on U. The unbiasedness of δ_0 follows from Corollary A.1 by taking $q = 0$ and $\gamma \equiv 1$. The goal of this subsection is to prove the domination of the unbiased estimator δ_0 by a competing estimator δ of the form

$$(4.15) \qquad \delta(X, U) = \delta_0(U) - \|U\|^4 \gamma(X),$$

where γ is a nonnegative function. It is important to notice that the "residual term" $\|U\|$ appears explicitly in the shrinkage function. It has been noted in [11] that the use of this term allows fewer assumptions about the distributions than when it does not appear. Specifically, this including gives a robustness property to the results, since they are valid for the entire class of spherically symmetric distributions.

We require the real-valued function γ to be twice weakly differentiable, in order to include basic examples, which are not twice differentiable. The following domination result is given in [25]. We will see below that it appears as a consequence of a more general result when shrinkage estimators of θ are involved.

THEOREM 4.3. *Assume that $p \geq 5$, the distribution of (X, U) has a finite fourth moment and the function γ is twice weakly differentiable on $\mathbb{R}^p$ and there exists a constant β such that $\gamma(t) \leq \beta/\|t\|^2$. A sufficient condition under which the estimator δ in* (4.15) *dominates the unbiased estimator δ_0 is that γ satisfies the differential inequality*

$$(4.16) \qquad \gamma^2 + \frac{2}{(k+4)(k+6)} \triangle \gamma \leq 0.$$

The standard example where $\gamma(t) = d/\|t\|^2$ for all $t \neq 0$ with $d > 0$ satisfies the conditions of the theorem. More precisely, it is easy to deduce that $\triangle \gamma(t) = -2d(p-4)/\|t\|^4$ and thus the sufficient condition of the theorem is written as $0 < d \leq 4(p-4)/(k+4)(k+6)$, which only occurs when $p \geq 5$. Straightforward calculus shows that the optimal value of d is given by $2(p-4)/(k+4)(k+6)$. The optimal constant in [11] is equal to $2(p-4)$. The extra terms in the denominator compensate for the $\|U\|^4$ term in our estimator.

We now consider the estimation of the loss of a class of shrinkage estimators considered in [11] (with a slight modification of their form in order to have notations coherent with those of the previous sections), that is, location estimators of the form

$$(4.17) \qquad \varphi_g = X + \|U\|^2 g(X),$$

where g is a weakly differentiable function from $\mathbb{R}^p$ into $\mathbb{R}^p$. In [11] it is shown that, if $\|g\|^2 \leq -2 \operatorname{div} g/(k+2)$, then φ_g dominates X, under quadratic loss for all spherically symmetric distributions with a finite second moment. A general example of a member of this class of estimators is with $g(X) = -r(\|X\|^2)\frac{A(X)}{b(X)}$, where r is a positive differentiable and nondecreasing function, A is a positive definite symmetric matrix and b is a positive definite quadratic form of $\mathbb{R}^p$. When r is equal to some constant a, A is the identity on $\mathbb{R}^p$ and the quadratic form b is the usual norm, g reduces to $a/\|X\|^2$. It can be shown that the optimal choice of a equals $(p-2)/(k+2)$. A member of the class is $\varphi_r = X - (p-2)\frac{\|U\|^2}{k+2}\frac{X}{\|X\|^2}$, the James–Stein estimator used when the variance is unknown as in Section 3.

In Proposition 2.3.1 of Section 2.3 of [11], it is shown that an unbiased estimator of the loss of the shrinkage estimator φ_g is given by

$$\delta_0^g(X, U) = \frac{p}{k}\|U\|^2$$
$$(4.18)$$
$$+ \left(\|g(X)\|^2 + \frac{2}{k+2} \operatorname{div} g(X) \right) \|U\|^4.$$

As in Theorem 4.3 above, the unbiased estimator of the loss can be improved by a shrinkage estimator of the loss. Thus the competing estimator we consider is

$$(4.19) \quad \delta_\gamma^g(X, U) = \delta_0^g(X, U) - \|U\|^4 \gamma(X),$$

where γ is a nonnegative function. Note that (4.19) is a true shrinkage estimator, while Johnstone's [32] optimal loss estimate for the normal case is an expanding estimator. This is not contradictory since we are using a different estimator than Johnstone and he was only dealing with the normal case. If $g \equiv 0$, the following result reduces to Theorem 4.3.

THEOREM 4.4. *Assume that $p \geq 5$, the distribution of (X, U) has a finite fourth moment and the function γ is twice weakly differentiable on $\mathbb{R}^p$ and there exists a constant β such that $\gamma(t) \leq \beta/\|t\|^2$. A sufficient condition under which the estimator δ_γ^g given in* (4.19) *dominates the unbiased estimator δ_0^g is that γ satisfies the differential inequality*

$$(4.20) \qquad \gamma^2 - \frac{4}{k+2}\gamma \operatorname{div} g + \frac{4}{k+6}\operatorname{div}(\gamma g)$$
$$+ \frac{2}{(k+4)(k+6)}\triangle \gamma \leq 0.$$

PROOF. Since the distribution of (X, U) is spherically symmetric around θ, it suffices to obtain the result working conditionally on the radius. For $R > 0$ fixed, we can compute using the uniform distribution $U_{R,\theta}$ on the sphere $S_{R,\theta}$. Thus the conditional risk difference between δ_γ^g and δ_0^g, according to (4.19), equals

$$E_{R,\theta}[(\delta_\gamma^g(X, U) - \|\varphi(X, U) - \theta\|^2)^2]$$
$$- E_{R,\theta}[(\delta_0^g(X, U) - \|\varphi(X, U) - \theta\|^2)^2]$$
$$= E_{R,\theta}[\|U\|^8 \gamma^2(X)]$$
$$- E_{R,\theta}[2\|U\|^4 \gamma(X)$$
$$\cdot (\delta_0^g(X, U) - \|\varphi(X, U) - \theta\|^2)],$$

that is, expanding and separating the integrand terms depending on θ,

$$E_{R,\theta}\left[\|U\|^8 \gamma^2(X) - 2\frac{p}{k}\|U\|^6 \gamma(X) \right.$$
$$\left. - \frac{4}{k+2}\|U\|^8 \operatorname{div} g(X) \right]$$

$$+ E_{R,\theta}[4\|U\|^6(X-\theta)^t\gamma(X)g(X)]$$

$$+ E_{R,\theta}[2\|U\|^4\|X-\theta\|^2\gamma(X)],$$

according to (4.18) (note that the two terms involving $\|g(X)\|^2$ cancel). Now we have

$$E_{R,\theta}[4\|U\|^6(X-\theta)^t\gamma(X)g(X)]$$

$$= \frac{4}{k+6} E_{R,\theta}[\|U\|^8 \operatorname{div}(\gamma(X)g(X))]$$

according to Lemma A.2 and

$$E_{R,\theta}[2\|U\|^4\|X-\theta\|^2\gamma(X)]$$

$$= E_{R,\theta}\left[\frac{2p}{k+4}\|U\|^6\gamma(X)\right.$$

$$\left. + \frac{2}{(k+4)(k+6)}\|U\|^8\triangle\gamma(X)\right]$$

according to Corollary A.1. Therefore the above conditional risk difference is equal to

$$E_{R,\theta}\left[\|U\|^8\left(\gamma^2(X) - \frac{4}{k+2}\operatorname{div}g(X)\right.\right.$$

$$+ \frac{4}{k+6}\operatorname{div}(\gamma(X)g(X))$$

$$\left.\left. + \frac{2}{(k+4)(k+6)}\triangle\gamma(X)\right)\right]$$

$$+ E_{R,\theta}\left[2p\left(\frac{1}{k-4} - \frac{1}{k}\right)\|U\|^6\gamma(X)\right]$$

which is bounded above by the first expectation since the function γ is nonnegative. Hence, the sufficient condition for domination is (4.20) in order that the inequality $R(\delta^g, \theta, \varphi) \leq R(\delta_0^g, \theta, \varphi)$ holds. $\quad\square$

5. DISCUSSION

There are several areas of the theory of loss estimation that we have not discussed. Our primary focus has been on location parameters for the multivariate normal and spherical distributions. Loss estimation for exponential families is addressed in Lele [38, 39] and Rukhin [42]. In [38] and [39] Lele developed improved loss estimators for point estimators in the general setup of Hudson's [31] subclass of continuous exponential family. Hudson's family essentially includes distributions for which the Stein-like identities hold; explicit calculations and loss estimators are given for the gamma distribution, as well as for improved scaled quadratic loss estimators in the Poisson setting for the Clevenson–Zidek [13] estimator. Rukhin [42] studied the posterior loss estimator for a Bayes estimate (under

quadratic loss) for the canonical parameter of a linear exponential family.

As pointed out in the Introduction, in the known variance normal setting, Johnstone [32] used a version of Blyth's lemma to show that the constant loss estimate p is admissible if $p \leq 4$. Lele [39] gave some additional sufficient conditions for admissibility in the general exponential family and worked out the precise details for the Poisson model. Rukhin [42] considered loss functions for the simultaneous estimate of θ and $L(\theta, \varphi(X))$ and deduced some interesting admissibility results.

A number of researchers have investigated improved estimators of a covariance matrix, Σ, under the Stein loss, $L_S(\hat{\Sigma}, \Sigma) = \operatorname{tr}(\hat{\Sigma}\Sigma^{-1}) - \log|\hat{\Sigma}\Sigma^{-1}| - p$, using an unbiased estimation of risk technique. In the normal case, [15, 27, 45, 47], and [49] proposed improved estimators that dominate the sample covariance under $L_S(\hat{\Sigma}, \Sigma)$. In [36], it was shown that the domination of these improved estimators over the sample covariance estimator is robust with respect to the family of elliptical distributions. To date, there has not been any work on improving the unbiased estimate of $L_S(\hat{\Sigma}, \Sigma)$.

In addition to the theoretical ideas discussed in the previous sections there are very practical applications of loss estimation. The primary application of loss estimation ideas is to model selection. It was shown by Fourdrinier and Wells [24] that improved loss estimators give more accurate model selection procedures. Bartlett, Boucheron and Lugosi [3] studied model selection strategies based on penalized empirical loss minimization and pointed out the equivalence between loss estimation and data-based complexity penalization. It was shown that any good loss estimate may be converted into a data-based penalty function and the performance of the estimate is governed by the quality of the loss estimate. Furthermore, a selected model that minimizes the penalized empirical loss achieves an almost optimal trade-off between the approximation error and the expected complexity, provided that the loss estimate on which the complexity is based is an approximate upper bound on the true loss. The key point to stress is that there is a fundamental dependence on the notions of good complexity regularization and good loss estimation. The ideas in this review lay the theoretical foundation for the construction of such loss estimators and model selection rules as well as give a decision-theoretic analysis of their statistical properties.

In linear models the notion of degrees of freedom plays the important role as a model complexity measure in various model selection criteria, such as Akaike

information criterion (AIC) [1] , Mallow's C_p [41], and Bayesian information criterion (BIC) [44], and generalized cross-validation (GCV) [14]. In regression the degrees of freedom are the trace of the so-called "hat" matrix. Efron [18] pointed out that the theory of Stein's unbiased risk estimation is central to the ideas underlying the calculation of the degrees of freedom of certain regression estimators.

Specifically, let Y be a random vector having an n-variate normal distribution $\mathcal{N}(\theta, \sigma^2 I_n)$ with unknown p-dimensional mean θ and identity covariance matrix $\sigma^2 I_n$. Let $\hat{\theta} = \varphi(Y)$ be an estimate of θ. In regression one focuses on how accurate φ can be in predicting using a new response vector y^{new}. Under the quadratic loss, the prediction risk is $E\{\|Y^{\text{new}} - \theta\|^2\}/n$. Efron [18] noted that

$$
(5.1) \quad E\{\|\varphi - \theta\|^2\} = E\{\|Y - \varphi(Y)\|^2 - n\sigma^2\}
$$
$$
+ 2\sum_{i=1}^{n} \text{Cov}(\varphi_i, Y_i).
$$

This expression suggests a natural definition of the degrees of freedom for an estimator φ as $\text{df}(\varphi) = \sum_{i=1}^{n} \text{Cov}(\varphi_i, Y_i)/\sigma^2 = E_\theta[(Y - \theta)^t \varphi(Y)]/\sigma^2$. Thus one can define a C_p-type quantity

$$
(5.2) \quad C_p(\varphi) = \frac{\|Y - \varphi\|^2}{n} + \frac{2\,\text{df}(\varphi)}{n}\sigma^2
$$

which has the same expectations as the true prediction error but may not be an estimate if $\text{df}(\varphi)$ and σ^2 are unknown. However, if φ is weakly differentiable and $\hat{\sigma}^2$ is an unbiased estimate of σ^2, the integration by parts formula in Lemma 3.1 implies that $\text{df}(\varphi)\sigma^2 = E_\theta[\text{div}\,\varphi(Y)\hat{\sigma}^2]$, hence $\text{div}\,\varphi\hat{\sigma}^2$ is unbiased estimate for the complexity parameter term, $\text{df}(\varphi)\sigma^2$, in (5.2). Therefore an unbiased estimate for the prediction error is

$$
(5.3) \quad C_p^*(\varphi) = \frac{\|Y - \varphi\|^2}{n} + \frac{2\,\text{div}\,\varphi}{n}\hat{\sigma}^2.
$$

Note that, if φ is a linear estimator ($\varphi = \mathbf{S}y$ for some matrix $\mathbf{S}$ independent of Y), then it is easy to show that this definition coincides with the definition of generalized degrees of freedom given by Hastie and Tibshirani [28] since $\text{div}\,\varphi = \text{tr}(\mathbf{S})$. Note that, if φ also depends on $\hat{\sigma}^2$, then (5.1) needs to be augmented by additional derivative terms with respect to $\hat{\sigma}^2$ as in Theorem 3.1.

Other approaches for estimating the complexity term penalty involve the use of resampling methods [18, 52] to directly estimate the prediction error. A K-fold cross-validation randomly divides the original sample

into K parts, and rotates through each part as a test sample and uses the remainder as a training sample. Cross-validation provides an approximately unbiased estimate of the prediction error, although its variance can be large. Other commonly used resampling techniques are the nonparametric and parametric bootstrap methods.

A number of new regularized regression methods have recently been developed, starting with Ridge regression [29], followed by the Lasso [50], the Elastic Net [53], and LARS [19]. Each of these estimates is weakly differentiable and has the form of a general shrinkage estimate; thus the prediction error estimate in (5.3) may be applied to construct a model selection procedure. Zou, Hastie and Tibshirani [54] used this idea to develop a model selection method for the Lasso. In some situations verifying the weak differentiability of φ may be complicated.

Loss estimates have been used to derive nonparametric penalized empirical loss estimates in the context of function estimation, which adapt to the unknown smoothness of the function of interest. See Barron et al. [2] and Donoho and Johnstone [16] for more details.

In the previous sections, the usual quadratic loss $L(\theta, \varphi(x)) = \|\varphi(x) - \theta\|^2$ was considered to evaluate various estimators $\varphi(X)$ of θ. The squared norm $\|x - \theta\|^2$ was crucial in the derivation of the properties of the loss estimators in conjunction with its role in the normal density or, more generally, in a spherical density. Other losses are thinkable but, to deal with tractable calculations, it matters to keep the Euclidean norm as a component of the loss in use. Hence a natural extension is to consider losses which are a function of $\|x - \theta\|^2$, that is, of the form $c(\|x - \theta\|^2)$ for a nonnegative function c defined on $\mathbb{R}_+$. The problem of estimating a function c of $\|x - \theta\|^2$ was tackled by Fourdrinier and Lepelletier [21] to which we refer the reader for more details. In particular, they focused on the fact that estimating $c(\|x - \theta\|^2)$ can be viewed as an evaluation of a quantity which is not necessarily a loss. Indeed it includes the problem of estimating the confidence statement of the usual confidence set $\{\theta \in \mathbb{R}^p \mid \|x - \theta\|^2 \leq c_\alpha\}$ with confidence coefficient $1 - \alpha$: c is the indicator function $\mathbb{1}_{[0,c_\alpha]}$.

APPENDIX

A.1 Risk Finiteness Conditions

LEMMA A.1. 1. *Let $X \sim \mathcal{N}(\theta, I_p)$, where θ is unknown, and denote by E_θ the expectation with respect to the distribution of X. Consider an estimator of θ of*

the form $\varphi(X) = X + g(X)$ where g is a function from $\mathbb{R}^p$ into $\mathbb{R}^p$.

a. *If g is such that $E_\theta[\|g(X)\|^2] < \infty$, then the quadratic risk of $\varphi(X)$, that is, $R(\theta, \varphi) = E_\theta[\|\varphi(X) - \theta\|^2]$, is finite.*

b. *If, in addition, the function g is weakly differentiable so that $\delta_0(X) = p + 2 \operatorname{div} g(X) + \|g(X)\|^2$ is an unbiased estimator of the loss $\|\varphi(X) - \theta\|^2$, then the risk of $\delta_0(X)$ defined by $\mathcal{R}(\theta, \varphi, \delta_0) = E_\theta[(\delta_0(X) - \|\varphi(X) - \theta\|^2)^2]$ is finite as soon as $E_\theta[\|g(X)\|^4] < \infty$ and $E_\theta[(\operatorname{div} g(X))^2] < \infty$.*

2. *Let $X \sim \mathcal{N}(\theta, \sigma^2 I_p)$, where θ and σ^2 are unknown, let S be a nonnegative random variable independent of X and such that $S \sim \sigma^2 \chi_n^2$ and denote by E_{θ, σ^2} the expectation with respect to the joint distribution of (X, S). Consider an estimator of θ of the form $\varphi(X, S) = X + S g(X, S)$ where g is a function from $\mathbb{R}^p \times \mathbb{R}_+$ into $\mathbb{R}^p$.*

a. *If g is such that $E_{\theta, \sigma^2}[S^2 \|g(X, S)\|^2] < \infty$, then the quadratic risk of $\varphi(X)$, that is, $R(\theta, \sigma^2, \varphi) = E_{\theta, \sigma^2}[\|\varphi(X, S) - \theta\|^2 / \sigma^2]$, is finite.*

b. *If, in addition, the function g is weakly differentiable so that*

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

is an unbiased estimator of the loss $\|\varphi(X, S) - \theta\|^2 / \sigma^2$, then the risk of $\delta_0(X, S)$ defined by $\mathcal{R}(\theta, \sigma^2, \varphi, \delta_0) = E_{\theta, \sigma^2}[(\delta_0(X, S) - \|\varphi(X, S) - \theta\|^2 / \sigma^2)^2]$ is finite as soon as $E_{\theta, \sigma^2}[S^2 \|g(X, S)\|^4] < \infty$, $E_{\theta, \sigma^2}[(S \operatorname{div} g(X, S))^2] < \infty$ and $E_{\theta, \sigma^2}[(S^2 \frac{\partial}{\partial S} \|g(X, S)\|)^2]$.

PROOF. 1.a. The loss of $\varphi(X)$ can be expanded as

$$\|\varphi(X) - \theta\|^2 = \|X - \theta\|^2 + \|g(X)\|^2$$
$$\text{(A.4)}$$
$$+ 2(X - \theta)^t g(X).$$

Now we have $E_\theta[\|X - \theta\|^2] = p < \infty$. Hence, by Schwarz's inequality, it follows from (A.4) that $|E_\theta[(X - \theta)^t g(X)]| \leq (E_\theta[\|X - \theta\|^2])^{1/2} \cdot (E_\theta[\|g(X)\|^2])^{1/2}$. Therefore, as soon as $E_\theta[\|g(X)\|^2] < \infty$, we will have $|E_\theta[\|\varphi(X) - \theta\|^2]| < \infty$. This is the desired result.

b. Note that, under the usual domination condition, that is, $2 \operatorname{div} g(x) + \|g(x)\|^2 \leq 0$ for any $x \in R^p$, of $\delta_0(X)$ over X, the condition $E_\theta[(\operatorname{div} g(X))^2] <$

∞ implies that $E_\theta[\|g(X)\|^4] < \infty$. We will have $\mathcal{R}(\theta, \varphi, \delta_0) = E_\theta[(\delta_0(X) - \|\varphi(X) - \theta\|^2)^2)] < \infty$ as soon as $E_\theta[\delta_0^2(X)] < \infty$ and $E_\theta[\|\varphi(X) - \theta\|^4] < \infty$. Now $E_\theta[\delta_0^2(X)] = E_\theta[(p + 2 \operatorname{div} g(X) + \|g(X)\|)^2] < \infty$ since $E_\theta[(\operatorname{div} g(X))^2] < \infty$ and $E_\theta[\|g(X)\|^4] < \infty$. Also according to (A.4)

$$E_\theta[\|\varphi(X) - \theta\|^4] = E_\theta[(\|X - \theta\|^2 + \|g(X)\|^2$$
$$+ 2(X - \theta)^t g(X))^2]$$
$$< \infty$$

since $E_\theta[\|X - \theta\|^4] < \infty$ and $E_\theta[\|g(X)\|^4] < \infty$ and, consequently, since $|(X - \theta)^t g(X)| \leq \|X - \theta\| \|g(X)\|$ implies that

$$E_\theta[|(X - \theta)^t g(X)|^2]$$
$$\leq E_\theta[\|X - \theta\|^2 \|g(X)\|^2]$$
$$\leq (E_\theta[\|X - \theta\|^4])^{1/2} (E_\theta[\|g(X)\|^4])^{1/2}$$

by Schwarz's inequality.

2.a. Parallel to the case where the variance σ^2 is known, it should be noticed that the corresponding domination condition of $\delta(X, S)$ over $\delta_0(X, S)$, that is, for any $x \in R^p$ and any $s \in \mathbb{R}_+$, $(n+2)\|g(x, s)\|^2 + 2 \operatorname{div}_x g(x, s) + 2s \frac{\partial}{\partial s} \|g(x, s)\|^2 \leq 0$, entails that the two conditions $E_{\theta, \sigma^2}[(S \operatorname{div} g(X, S))^2] < \infty$ and $E_{\theta, \sigma^2}[(S^2 \frac{\partial}{\partial S} \|g(X, S)\|)^2]$ imply the condition $E_{\theta, \sigma^2}[S^2 \|g(X, S)\|^4] < \infty$. Also the derivation of the finiteness of $R(\theta, \sigma^2, \varphi)$ follows a similar way as in 1.a.

b. We will have $\mathcal{R}(\theta, \sigma^2, \varphi, \delta_0) = E_{\theta, \sigma^2}[(\delta_0(X, S) - \|\varphi(X) - \theta\|^2 / \sigma^2))^2] < \infty$ as soon as $E_{\theta, \sigma^2}[(\delta_0(X, S))^2 < \infty$ and $E_{\theta, \sigma^2}[\|\varphi(X) - \theta\|^4] < \infty$. Now $E_{\theta, \sigma^2}[(\delta_0(X, S))^2 = E_{\theta, \sigma^2}[p + S\{(n+2)\|g(X, S)\|^2 + 2 \operatorname{div}_X g(X, S) + 2S \frac{\partial}{\partial S} \|g(X, S)\|^2\}] < \infty$ since we assume that $E_{\theta, \sigma^2}[(S \operatorname{div}_X g(X, S))^2] < \infty$ and $E_{\theta, \sigma^2}[S^2 \|g(X, S)\|^4] < \infty$. Also $E_{\theta, \sigma^2}[\|\varphi(X, S) - \theta\|^4] = E_{\theta, \sigma^2}[(\|X - \theta\|^2 + S^2 \|g(X, S)\|^2 + 2S(X - \theta)^t g(X, S))^2)^2] < \infty$ since $E_\theta[\|X - \theta\|^4] < \infty$ and $E_{\theta, \sigma^2}[S^2 \|g(X, S)\|^4] < \infty$ (note that $|(X - \theta)^t g(X, S)| \leq \|X - \theta\| \|g(X, S)\|$ implies that

$$E_{\theta, \sigma^2}[|(X - \theta)^t S g(X, S)|^2]$$
$$\leq E_{\theta, \sigma^2}[\|X - \theta\|^2 S^2 \|g(X, S)\|^2]$$
$$\leq (E_{\theta, \sigma^2}[\|X - \theta\|^4])^{1/2} (E_{\theta, \sigma^2}[S^2 \|g(X, S)\|^4])^{1/2}$$

by Schwarz's inequality). □

A.2 Additional Technical Lemmas

This Appendix gives some technical results used in Section 4.2. The first two results deal with expectations conditioned on the radius of a spherically symmetric distribution in $\mathbb{R}^p \times \mathbb{R}^k$ centered at $(\theta, 0)$ where $\theta \in \mathbb{R}^p$. These expectations reduce to integrals with respect to the uniform distribution $U_{R,\theta}$ on the sphere

$$S_{R,\theta} = \{y = (x, u) \in \mathbb{R}^p \times \mathbb{R}^k |$$
$$(\|x - \theta\|^2 + \|u\|^2)^{1/2} = R\}.$$

If $E_{R,\theta}[\psi]$ is the expectation of some function ψ with respect to $U_{R,\theta}$, the expectation with respect to the entire distribution is given by $E_\theta[\psi] = E[E_{R,\theta}[\psi]]$ where E is the expectation with respect to the distribution of the radius.

When the spherical distribution has a density with respect to the Lebesgue measure, it is necessarily of the form $f(\|x - \theta\|^2 + \|u\|^2)$ for some function f. Then the radius has density $R \to \sigma_{p+k} f(R^2) R^{p+k-1}$ where $\sigma_{p+k} = 2\pi^{p+k}/\Gamma(\frac{p+k}{2})$. Therefore the expectation of any function ψ can be written as

$$E_\theta[\psi] = \int_0^\infty \left[\int_{S_{R,\theta}} \psi(y) U_{R,\theta}(dy) \right] f(R) \, dR.$$

Note that for a vector $y = (x, u) \in S_{R,\theta}$, we have $x = \pi(y)$ and $\|u\|^2 = R^2 - \|\pi(y) - \theta\|^2$ where π is the orthogonal projector from $\mathbb{R}^p \times \mathbb{R}^k$ onto $\mathbb{R}^p$. Under $U_{R,\theta}$, the distribution $\pi(U_{R,\theta})$ of this projector has a density with respect to the Lebesgue measure on $\mathbb{R}^p$ given by $x \to C_R^{p,k}(R^2 - \|x - \theta\|^2)^{k/2-1} \mathbf{1}_{B_{R,\theta}}(x)$ where $C_R^{p,k} = \Gamma(\frac{p+k}{2}) R^{2-p-k}/(\Gamma(\frac{k}{2}) \pi^{p/2})$ and $\mathbf{1}_{B_{R,\theta}}$ is the indicator function of the ball $B_{R,\theta} = \{x \in \mathbb{R}^p | \|x - \theta\| \le R\}$ of radius R centered at θ in $\mathbb{R}^p$.

According to the above, as a spherically symmetric distribution on $\mathbb{R}^p$ around θ, the radius of $\pi(U_{R,\theta})$ has density

$$r \to \sigma_p C_R^{p,k}(R^2 - r^2)^{k/2-1} \mathbf{1}_{]0,R[}(r) r^{p-1}$$
$$= \frac{2R^{2-p-k}}{B(p/2, k/2)} r^{p-1}(R^2 - r^2)^{k/2-1} \mathbf{1}_{]0,R[}(r).$$

We use repeatedly the fact that any such projection onto a space of dimension greater than 0 and less than $p + k$ is spherically symmetric with a density. Then we also often make use of its radial density.

LEMMA A.2. *For every twice weakly differentiable function* $g(\mathbb{R}^p \to \mathbb{R}^p)$ *and for every function*

$h(\mathbb{R}_+ \to \mathbb{R})$,

(A.5)
$$E_{R,\theta}[h(\|U\|^2)(X - \theta)^t g(X)]$$
$$= E_{R,\theta}\left[\frac{H(\|U\|^2)}{(\|U\|^2)^{k/2-1}} \operatorname{div} g(X) \right],$$

where H is the indefinite integral, vanishing at 0, of the function $t \to \frac{1}{2} h(t) t^{k/2-1}$.

PROOF. We have

$$E_{R,\theta}[h(\|U\|^2)(X - \theta)^t g(X)]$$
$$= C_R^{p,k} \int_{B_{R,\theta}} h(R^2 - \|x - \theta\|^2)(x - \theta)^t$$
$$\cdot g(x)(R^2 - \|x - \theta\|^2)^{k/2-1} \, dx$$
$$= C_R^{p,k} \int_{B_{R,\theta}} (\nabla H(R^2 - \|x - \theta\|^2))^t g(x) \, dx$$

since

$$\nabla H(R^2 - \|x - \theta\|^2)$$
$$= -2H'(R^2 - \|x - \theta\|^2)(x - \theta)$$
$$= h(R^2 - \|x - \theta\|^2)(R^2 - \|x - \theta\|^2)^{k/2-1}(x - \theta).$$

Then, by the divergence formula,

$$E_{R,\theta}[h(\|U\|^2)(X - \theta)^t g(X)]$$
$$= C_R^{p,k} \int_{B_{R,\theta}} \operatorname{div}(H(R^2 - \|x - \theta\|^2) g(x)) \, dx$$
$$- C_R^{p,k} \int_{B_{R,\theta}} H(R^2 - \|x - \theta\|^2) \operatorname{div} g(x) \, dx.$$

Now, if $\sigma_{R,\theta}$ denotes the area measure on the sphere $S_{R,\theta}$, the divergence theorem insures that the first integral equals

$$C_R^{p,k} \int_{S_{R,\theta}} (H(R^2 - \|x - \theta\|^2) g(x))^t \frac{x - \theta}{\|x - \theta\|} \sigma_{R,\theta}(dx)$$

and is null since, for $x \in S_{R,\theta}$, $R^2 - \|x - \theta\|^2 = 0$ and $H(0) = 0$. Hence, in terms of expectation, we have

$$E_{R,\theta}[h(\|U\|^2)(X - \theta)^t g(X)]$$
$$= C_R^{p,k} \int_{B_{R,\theta}} \frac{H(R^2 - \|x - \theta\|^2)}{(R^2 - \|x - \theta\|^2)^{k/2} - 1} \operatorname{div} g(x)$$
$$\cdot (R^2 - \|x - \theta\|^2)^{k/2-1} \, dx$$
$$= E_{R,\theta}\left[\frac{H(\|U\|^2)}{(\|U\|^2)^{k/2-1}} \operatorname{div} g(X) \right]$$

which is the desired result. $\square$

COROLLARY A.1. *For every twice weakly differentiable function $\gamma (\mathbb{R}^p \rightarrow \mathbb{R}_+)$ and for every integer q,*

$$E_{R,\theta}[\|U\|^q \|X - \theta\|^2 \gamma(X)]$$

$$= \frac{p}{k+q} E_{R,\theta}[\|U\|^{q+2} \gamma(X)]$$

$$+ \frac{1}{(k+q)(k+q+2)} E_{R,\theta}[\|U\|^{q+4} \triangle \gamma(X)].$$

PROOF. Take $h(t) = t^{q/2}$ and $g(x) = \gamma(x)(x - \theta)$ and apply Lemma A.2 twice. □

ACKNOWLEDGMENTS

REFERENCES

[1] AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) 267–281. Akadémiai Kiadó, Budapest. MR0483125

[2] BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028

[3] BARTLETT, P., BOUCHERON, S. and LUGOSI, G. (2002). Model selection and error estimation. *Machine Learning* **48** 85–113.

[4] BERGER, J. (1985). The frequentist viewpoint and conditioning. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. I* (*Berkeley, Calif.*, 1983) (L. Le Cam and R. Olshen, eds.) 15–44. Wadsworth, Belmont, CA. MR0822033

[5] BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York. MR0804611

[6] BERGER, J. O. (1985). In defense of the likelihood principle: Axiomatics and coherency. In *Bayesian Statistics*, 2 (*Valencia*, 1983) (D. V. Lindley, J. M. Bernardo, M. H. DeGroot and A. F. M. Smith, eds.) 33–66. North-Holland, Amsterdam. MR0862483

[7] BLANCHARD, D. and FOURDRINIER, D. (1999). Non-trivial solutions of non-linear partial differential inequations and order cut-off. *Rend. Mat. Appl.* (7) **19** 137–154. MR1710109

[8] BOCK, M. E. (1988). Shrinkage estimators: Pseudo-Bayes rules for normal mean vectors. In *Statistical Decision Theory and Related Topics*, *IV*, *Vol.* 1 (*West Lafayette, Ind.*, 1986) (S. S. Gupta and J. O. Berger, eds.) 281–297. Springer, New York. MR0927108

[9] BROWN, L. (1968). Inadmissiblity of the usual estimators of scale parameters in problems with unknown location and scale parameters. *Ann. Math. Statist.* **39** 29–48. MR0222992

[10] BROWN, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903. MR0286209

[11] CELLIER, D. and FOURDRINIER, D. (1995). Shrinkage estimators under spherical symmetry for the general linear model. *J. Multivariate Anal.* **52** 338–351. MR1323338

[12] CELLIER, D., FOURDRINIER, D. and ROBERT, C. (1989). Robust shrinkage estimators of the location parameter for elliptically symmetric distributions. *J. Multivariate Anal.* **29** 39–52. MR0991055

[13] CLEVENSON, M. L. and ZIDEK, J. V. (1975). Simultaneous estimation of the means of independent Poisson laws. *J. Amer. Statist. Assoc.* **70** 698–705. MR0394962

[14] CRAVEN, P. and WAHBA, G. (1978/79). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403. MR0516581

[15] DEY, D. K. and SRINIVASAN, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* **13** 1581–1591. MR0811511

[16] DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464

[17] DU PLESSIS, N. (1970). *An Introduction to Potential Theory*. *University Mathematical Monographs* **7**. Hafner, Darien, CT. MR0435422

[18] EFRON, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *J. Amer. Statist. Assoc.* **99** 619–642. MR0845884

[19] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

[20] FANG, K. T., KOTZ, S. and NG, K. W. (1990). *Symmetric Multivariate and Related Distributions*. *Monographs on Statistics and Applied Probability* **36**. Chapman & Hall, London. MR1071174

[21] FOURDRINIER, D. and LEPELLETIER, P. (2008). Estimating a general function of a quadratic function. *Ann. Inst. Statist. Math.* **60** 85–119. MR2400062

[22] FOURDRINIER, D. and STRAWDERMAN, W. E. (2003). On Bayes and unbiased estimators of loss. *Ann. Inst. Statist. Math.* **55** 803–816. MR2028618

[23] FOURDRINIER, D., STRAWDERMAN, W. E. and WELLS, M. T. (2003). Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix. *J. Multivariate Anal.* **85** 24–39. MR1978175

[24] FOURDRINIER, D. and WELLS, M. T. (1994). Comparaisons de procédures de sélection d'un modèle de régression: Une approche décisionnelle. *C. R. Acad. Sci. Paris Sér. I Math.* **319** 865–870. MR1300958

[25] FOURDRINIER, D. and WELLS, M. T. (1995). Estimation of a loss function for spherically symmetric distributions in the general linear model. *Ann. Statist.* **23** 571–592. MR1332582

[26] FOURDRINIER, D. and WELLS, M. T. (1995). Loss estimation for spherically symmetric distributions. *J. Multivariate Anal.* **53** 311–331. MR1336060

[27] HAFF, L. R. (1979). An identity for the Wishart distribution with applications. *J. Multivariate Anal.* **9** 531–544. MR0556910

[28] HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. Chapman & Hall, London. MR1082147

[29] HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

[30] HSIEH, F. S. and HWANG, J. T. G. (1993). Admissibility under the frequentist's validity constraint in estimating the loss of the least-squares estimator. *J. Multivariate Anal.* **44** 279–285. MR1219208

[31] HUDSON, H. M. (1978). A natural identity for exponential families with applications in multiparameter estimation. *Ann. Statist.* **6** 473–484. MR0467991

[32] JOHNSTONE, I. (1988). On inadmissibility of some unbiased estimates of loss. In *Statistical Decision Theory and Related Topics*, *IV*, *Vol.* 1 (*West Lafayette*, *Ind.*, 1986) (S. S. Gupta and J. O. Berger, eds.) 361–379. Springer, New York. MR0927112

[33] KIEFER, J. (1975). Conditional confidence approach in multi-decision problems. In *Multivariate Analysis* 4 (P. R. Krishnaiah, ed.). Academic Press, New York.

[34] KIEFER, J. (1976). Admissibility of conditional confidence procedures. *Ann. Statist.* **4** 836–865. MR0438543

[35] KIEFER, J. (1977). Conditional confidence statements and confidence estimators. *J. Amer. Statist. Assoc.* **72** 789–827. MR0518611

[36] KUBOKAWA, T. and SRIVASTAVA, M. S. (1999). Robust improvement in estimation of a covariance matrix in an elliptically contoured distribution. *Ann. Statist.* **27** 600–609. MR1714715

[37] LEHMANN, E. L. and SHEFFÉ, H. (1950). Completeness, similar regions and unbiased estimates. *Sankhyā* **17** 305–340.

[38] LELE, C. (1992). Inadmissibility of loss estimators. *Statist. Decisions* **10** 309–322. MR1200692

[39] LELE, C. (1993). Admissibility results in loss estimation. *Ann. Statist.* **21** 378–390. MR1212182

[40] LU, K. L. and BERGER, J. O. (1989). Estimation of normal means: Frequentist estimation of loss. *Ann. Statist.* **17** 890–906. MR0994274

[41] MALLOWS, C. (1973). Some comments on Cp. *Technometrics* **15** 661–675.

[42] RUKHIN, A. L. (1988). Estimated loss and admissible loss estimators. In *Statistical Decision Theory and Related Topics*, *IV*, *Vol.* 1 (*West Lafayette*, *Ind.*, 1986) (S. S. Gupta and J. O. Berger, eds.) 409–418. Springer, New York. MR0927116

[43] SANDVED, E. (1968). Ancillary statistics and prediction of the loss in estimation problems. *Ann. Math. Statist.* **39** 1756–1758. MR0231481

[44] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

[45] STEIN, C. Estimating the covariance matrix. Unpublished manuscript.

[46] STEIN, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Ann. Inst. Statist. Math.* **16** 155–160. MR0171344

[47] STEĬN, Č. (1977). Lectures on multivariate estimation theory. *Zap. NauČn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.* (*LOMI*) **74** 4–65, 146, 148. MR0501548

[48] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

[49] TAKEMURA, A. (1984). An orthogonally invariant minimax estimator of the covariance matrix of a multivariate normal population. *Tsukuba J. Math.* **8** 367–376. MR0767967

[50] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[51] WAN, A. T. K. and ZOU, G. (2004). On unbiased and improved loss estimation for the mean of a multivariate normal distribution with unknown variance. *J. Statist. Plann. Inference* **119** 17–22. MR2018447

[52] YE, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93** 120–131. MR1614596

[53] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327

[54] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967