The Annals of Statistics
1986, Vol. 14, No. 1, 188–205

MINIMAX MULTIPLE SHRINKAGE ESTIMATION

BY EDWARD I. GEORGE

University of Chicago

For the canonical problem of estimating a multivariate normal mean under squared-error-loss, this article addresses the problem of selecting a minimax shrinkage estimator when vague or conflicting prior information suggests that more than one estimator from a broad class might be effective. For this situation a new class of alternative estimators, called multiple shrinkage estimators, is proposed. These estimators use the data to emulate the behavior and risk properties of the most effective estimator under consideration. Unbiased estimates of risk and sufficient conditions for minimaxity are provided. Bayesian motivations link this construction to posterior means of mixture priors. To illustrate the theory, minimax multiple shrinkage Stein estimators are constructed which can adaptively shrink the data towards any number of points or subspaces.

1. Introduction. Consider the following canonical setup. From p independent experiments, we observe $Y = (Y_1, \dots, Y_p)'$, which has the p-dimensional multivariate normal distribution

$$Y|\theta \sim N_p(\theta, I), \tag{1.1}$$

with unknown mean $\theta = (\theta_1, \dots, \theta_p)'$ and the identity covariance matrix I. The problem is to find estimators $\delta \equiv \delta(Y)$: $R^p \to R^p$ of θ which yield small risk or expected squared-error-loss

$$R(\theta, \delta) = E_\theta(\theta - \delta)'(\theta - \delta) = E_\theta\|\theta - \delta\|^2, \tag{1.2}$$

where E_θ stands for averaging over the sample space with respect to the distribution (1.1) for fixed θ.

Beginning with the seminal work of Stein (1955) and James and Stein (1960), interest has focused on the use of minimax shrinkage estimators for this problem [see Berger (1983)]. Each of these estimators not only dominates the maximum likelihood estimator $\delta^{\text{MLE}}(Y) = Y$, but also yields substantially smaller risk in a certain region of the parameter space. By selecting an estimator for which θ happens to be close to its corresponding region of improvement, meaningful risk gains can be achieved in practice. However, because θ is unknown and an estimator must be selected before looking at the data, the selection of an estimator or equivalently the region of improvement is typically based on available prior information. As a result of this feature, a large number of minimax shrinkage estimators have been developed, offering a wide variety of regions of risk improvement corresponding to different types of prior information [see

Received March 1984; revised July 1985.
AMS 1980 *subject classifications.* Primary 62F10; secondary 62F15.
Key words and phrases. Bayes estimator, mixture, multivariate normal mean, Stein estimator, superharmonic function, unbiased estimate of risk.

Berger (1982) and Berger and Berliner (1984) for general discussions and references].

In this paper, we consider the general situation where conflicting or vague prior information suggests that more than one of a broad class of minimax shrinkage estimators may be effective. For this situation we present new minimax multiple shrinkage estimators which can incorporate this partial prior information by using the data to emulate the behavior and risk of the most effective estimators under consideration. These multiple shrinkage estimators enhance the practical potential of currently employed minimax shrinkage estimators by vastly broadening the region of the parameter space where meaningful risk reduction is available.

For example, suppose attention was restricted to using a Stein estimator of the form

$$\delta_v^S(Y) = Y - \left[1 \wedge \frac{p-2}{\|Y-v\|^2}\right](Y-v) \tag{1.3}$$

$[a \wedge b = \min(a, b)]$, which shrinks Y towards a target $v \in R^p$. (When $v = 0$, δ_v^S is the original positive-part Stein estimator which shrinks Y towards 0.) As is well known, when θ happens to be in a small neighborhood surrounding v, δ_v^S yields very small risk, and when θ is far from this neighborhood, δ_v^S is essentially indistinguishable from δ^{MLE}. Typically, v would be a prior guess as to the location of θ, perhaps the result of a previous experiment.

However, suppose prior information suggested several different choices for the target v. Denoting the corresponding choices for δ_v^S by $\delta_1^S, \dots, \delta_K^S$, use of a single δ_k^S would potentially forego important risk gains, especially if some of the target choices were far from each other. To avoid this limitation, we propose a multiple shrinkage Stein estimator for this situation. This estimator, which is described in greater generality in Section 3, is here of the form

$$\delta_*^S(Y) = \sum_{k=1}^{K} \rho_k^S(Y)\delta_k^S(Y), \tag{1.4}$$

where $\rho_1^S, \dots, \rho_K^S$ satisfy $\sum_{k=1}^K \rho_k^S(Y) \equiv 1$ and are adaptive functions of Y which place increasing weight on the δ_k^S which are shrinking most. Thus, δ_*^S is an adaptive convex combination of the δ_k^S which provides more shrinkage when Y is close to any of the targets. Unbiased estimates of risk and simulation results, also provided in Section 3, suggest that δ_*^S can offer meaningful risk reduction at each target. Moreover, it is shown that δ_*^S is minimax, and so possesses the same robustness quality as each δ_k^S with respect to misspecification of the targets.

In Section 2 general results on the construction, risk assessment, and Bayesian motivation of multiple shrinkage estimators are provided for the situation where a finite number of a broad class of minimax estimators are being contemplated. In Section 3 minimax multiple shrinkage Stein estimators are proposed and analyzed. In Section 4 the construction and assessment of multiple shrinkage estimators is indicated for the situation where a possibly infinite set of estimators is under consideration. In Section 5 it is shown that the main results of this paper generalize easily for the more realistic situation where $Y|\theta, \sigma \sim N_p(\theta, \sigma^2 I)$ with an available independent estimate of σ^2.

2. Multiple shrinkage estimators. The following definitions are required. A function m: $R^p \to R$ is said to be *almost differentiable* (a.d.) if there exists a function ∇m: $R^p \to R^p$ such that for all $z \in R^p$,

$$m(y+z) - m(y) = \int_0^1 z' \nabla m(y + tz)\, dt$$

for almost all $y \in R^p$. This definition implicitly defines ∇ be the vector differential operator

$$\nabla = (\nabla_1, \dots, \nabla_p)', \quad \text{where } \nabla_i = \partial/\partial y_i.$$

(Essentially an a.d. function is continuous and a.e. differentiable.) The function ∇m is said to be a.d. if each coordinate function $\nabla_i m$ is a.d. When both m and ∇m are a.d., m is *superharmonic* if for almost all $y \in R^p$,

$$\nabla^2 m(y) = \sum_{i=1}^{p} \nabla_i^2 m(y) \le 0.$$

See Helms (1975) for an introduction to more general superharmonic functions.

2.1. Constructing multiple shrinkage estimators. Throughout this section, we consider the general situation where vague or conflicting prior information suggests that small risk may be obtainable by any one of K shrinkage estimators of the form

$$\delta_k(Y) = Y + \nabla \log m_k(Y), \qquad k = 1, \dots, K, \tag{2.1}$$

where m_k: $R^p \to R^+ \cap \{0\}^c$ is such that m_k and ∇m_k are a.d. For each estimator δ_k, the function m_k determines the shrinkage component, $\nabla \log m_k(Y)$. The class of estimators of the form (2.1) includes all Bayes, formal Bayes, and admissible rules [see Brown (1971)], and some reasonable inadmissible rules such as the Stein estimator δ_v^S in (1.3) (see Section 3).

When the regions where each of $\delta_1, \dots, \delta_K$ offer especially small risk are very different, it may be preferable to consider using a multiple shrinkage estimator δ_* which we define to be

$$\delta_*(Y) = Y + \nabla \log m_*(Y), \qquad m_*(Y) = \sum_{k=1}^{K} w_k m_k(Y), \tag{2.2}$$

where $m_1, \dots, m_K$ are the functions corresponding to $\delta_1, \dots, \delta_K$ in (2.1), and

$$w_1, \dots, w_K, \qquad \left(\sum_{k=1}^{K} w_k = 1 \right) \tag{2.3}$$

are a fixed set of prespecified positive weights (scaled as probabilities for convenience), which we shall refer to as prior weights. In Section 2.3 it is shown that when $\delta_1, \dots, \delta_K$ are Bayes rules, δ_* is the Bayes rule for a mixture prior, and the prior weights arise naturally as prior probabilities.

The following reexpressions of δ_* illustrate the relationship between the behavior of δ_* and $\delta_1, \dots, \delta_K$, suggesting the description of δ_* as a multiple shrinkage estimator,

$$\delta_*(Y) = Y + \sum_{k=1}^{K} \rho_k(Y) \nabla \log m_k(Y) = \sum_{k=1}^{K} \rho_k(Y) \delta_k(Y), \tag{2.4}$$

where

$$\rho_k(Y) = w_k m_k(Y)/m_*(Y). \tag{2.5}$$

Since $\sum_{k=1}^{K} \rho_k(Y) \equiv 1$, the middle expression in (2.4) reveals the shrinkage component of δ_* to be an adaptive convex combination of the shrinkage components of $\delta_1, \dots, \delta_K$; the rightmost expression shows δ_* as an adaptive convex combination of the estimators $\delta_1, \dots, \delta_K$. We shall refer to $\rho_1, \dots, \rho_K$, which adaptively weight the shrinkage contribution of the combined estimators, as relevance functions, following the idea first introduced by Efron and Morris (1972, 1973b). Each relevance function ρ_k adaptively updates the prior weight w_k by the factor m_k/m_*. Because $\rho_1(Y), \dots, \rho_K(Y)$ are proportional to the terms $w_1 m_1(Y), \dots, w_K m_K(Y)$, the relevance functions put larger weight on those δ_k for which $w_k m_k(Y)$ is larger. For example, when $w_k m_k(Y) \gg w_j m_j(Y)$ for all $j \neq k$, $\rho_k(Y)$ will be close to 1, and $\delta_*(Y)$ will emulate $\delta_k(Y)$. Note that when $m_k(Y)$ and $\nabla \log m_k(Y)$ are large simultaneously, δ_* will incorporate more of the shrinkage of $\delta_1, \dots, \delta_K$.

2.2. Some risk results for multiple shrinkage estimators. In this section we establish some general results which link the risk properties of δ_* with those of the combined estimators $\delta_1, \dots, \delta_K$. Because δ_* and $\delta_1, \dots, \delta_K$ are of the form $\delta(Y) = Y + \nabla \log m(Y)$, we make use of the following results of Stein (1973, 1981), which provide unbiased estimates of risk and sufficient minimaxity conditions for such estimators.

THEOREM 1 (Stein). *Suppose $\delta(Y) = Y + \nabla \log m(Y)$ where m: $R^p \to R^+ \cap \{0\}^c$ is such that m and ∇m are a.d. If*

$$E_\theta |\nabla_i^2 m(Y)/m(Y)| < \infty, \qquad i = 1, \dots, p, \tag{i}$$

$$E_\theta \|\nabla \log m(Y)\|^2 < \infty, \tag{ii}$$

then the risk of δ may be expressed as

$$\begin{aligned} R(\theta, \delta) &= p - E_\theta D\delta(Y), \\ D\delta(Y) &= \|\nabla \log m(Y)\|^2 - 2\nabla^2 m(Y)/m(Y). \end{aligned} \tag{2.6}$$

The expression $D\delta(Y)$ above is an unbiased estimate of the amount of risk reduction offered by δ over δ^{MLE} $[R(\theta, \delta^{\text{MLE}}) \equiv p]$. $D\delta$ is used throughout to express unbiased estimates of risk reduction. Note that when θ is such that $D\delta(Y)$ is large with high probability, δ will yield especially small risk. Furthermore, because $D\delta(Y) \geq 0$ when $\nabla^2 m(Y) \leq 0$, the following sufficient condition for the minimaxity of δ is immediate.

COROLLARY 1 (Stein). *If* $\delta(Y) = Y + \nabla \log m(Y)$ *satisfies the conditions of Theorem* 1 *and* m *is superharmonic, then* δ *is minimax.*

Focusing now on the relationship between the risk properties of δ_* and $\delta_1, \ldots, \delta_K$, the following lemma shows when Theorem 1 and Corollary 1 may be applied to δ_*.

LEMMA 1. *If* $\delta_1, \ldots, \delta_K$ *satisfy the conditions of Theorem* 1, *then* δ_* *will satisfy the conditions of Theorem* 1.

PROOF. It is immediate from (2.2), that $m_*: R^p \to R^+ \cap \{0\}^c$, and that m_* and ∇m_* are a.d. Condition (i) follows by observing that

$$|\nabla_i^2 m_*(Y)/m_*(Y)| = \left| \sum_{k=1}^K \rho_k(Y) \nabla_i^2 m_k(Y)/m_k(Y) \right| \le \sum_{k=1}^K |\nabla_i^2 m_k(Y)/m_k(Y)|.$$

Condition (ii) follows from (2.4) and

$$\left\| \sum_{k=1}^K \rho_k(Y) \nabla \log m_k(Y) \right\|^2 \le \sum_{k=1}^K \rho_k(Y) \|\nabla \log m_k(Y)\|^2 \le \sum_{k=1}^K \|\nabla \log m_k(Y)\|^2.$$

□

The next result provides an easily verifiable sufficient condition for the minimaxity of δ_*; a condition that is somewhat stronger than the minimaxity of $\delta_1, \ldots, \delta_K$. Because of the potential complexity of the inputs for δ_*, the protection against misspecification provided by minimaxity is an especially appealing property here.

COROLLARY 2. *If* $\delta_1, \ldots, \delta_K$ *satisfy the conditions of Theorem* 1 *and if* $m_1, \ldots, m_K$ *are superharmonic, then* δ_* *is minimax.*

PROOF. Because $m_* = \sum_{k=1}^K w_k m_k$ will be superharmonic whenever $m_1, \ldots, m_K$ are superharmonic, the result is immediate from Lemma 1 and Corollary 1. □

To offer any practical advantage over δ^{MLE}, a minimax estimator must yield meaningful risk gains somewhere in the parameter space. The following result, which links the risk reduction estimate $D\delta_*$ to $D\delta_1, \ldots, D\delta_K$, suggests possible regions of improvement for δ_*.

COROLLARY 3. *If* $\delta_1, \ldots, \delta_K$ *satisfy the conditions of Theorem* 1, *then*

$$D\delta_*(Y) = \sum_{k=1}^K \rho_k(Y) \left[D\delta_k(Y) - \frac{1}{2} \sum_{l=1}^K \rho_l(Y) \|\delta_k(Y) - \delta_l(Y)\|^2 \right]. \tag{2.7}$$

PROOF. By Lemma 1 and Theorem 1,

$$D\delta_*(Y) = \|\nabla \log m_*(Y)\|^2 - 2\nabla^2 m_*(Y)/m_*(Y)$$
$$= \left\| \sum_{k=1}^{K} \rho_k(Y)\nabla \log m_k(Y) \right\|^2 - \sum_{k=1}^{K} \rho_k(Y)\big(2\nabla^2 m_k(Y)/m_k(Y)\big).$$

The desired result is obtained by substituting

$$\left\| \sum_{k=1}^{K} \rho_k(Y)\nabla \log m_k(Y) \right\|^2 = \sum_{k=1}^{K} \sum_{l=1}^{K} \rho_k(Y)\rho_l(Y)(\nabla \log m_k(Y))'(\nabla \log m_l(Y))$$
$$= \sum_{k=1}^{K} \rho_k(Y)\|\nabla \log m_k(Y)\|^2$$
$$-\frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \rho_k(Y)\rho_l(Y)\|\delta_k(Y) - \delta_l(Y)\|^2,$$

where the last equality follows from

$$\|\delta_k(Y) - \delta_l(Y)\|^2 = \|\nabla \log m_k(Y)\|^2 + \|\nabla \log m_l(Y)\|^2$$
$$-2(\nabla \log m_k(Y))'(\nabla \log m_l(Y)). \quad \square$$

Corollary 3 suggests when δ_* may offer meaningful risk gains in the same regions of the parameter space as any of $\delta_1, \ldots, \delta_K$. In (2.7) $D\delta_*$ is expressed as an adaptive convex combination of bracketed terms, each of which consists of the risk reduction estimate $D\delta_k$ penalized by a factor which weights the shrinkage conflict between δ_k and the other estimators. Note that when $\rho_k(Y) \approx 1$, $D\delta_*(Y) \approx D\delta_k(Y)$, since $\rho_l(Y) \approx 0$ for $l \neq k$. Thus, the size of $D\delta_*(Y)$ will be increased by sharply adaptive relevance functions which, for each Y, put most of their weight on the largest $D\delta_k(Y)$. Such behavior would yield $R(\theta, \delta_*) \approx \min_k R(\theta, \delta_k)$. Examples where this approximation is excellent are provided in Section 3.

2.3. Bayesian motivations. In this section multiple shrinkage estimators are shown to arise naturally as Bayes rules under mixture priors in the Bayesian context. More precisely, suppose $\delta_1, \ldots, \delta_K$ are Bayes rules corresponding to the prior densities $\pi_1, \ldots, \pi_K$, respectively. Using the well-known representation, see for example Stein (1981), each of these may be expressed as

$$\delta_k(Y) = E_{\pi_k}(\theta|Y) = Y + \nabla \log m(Y|\pi_k), \tag{2.8}$$

where

$$m(Y|\pi_k) = \int (2\pi)^{-p/2} e^{-\|Y-\theta\|^2/2}\, \pi_k(\theta)\, d\theta \tag{2.9}$$

is the marginal density of Y under π_k. Replacing $m_k(Y)$ by $m(Y|\pi_k)$ and $m_*(Y)$

by $m(Y|\pi_*)$, δ_* in (2.2) becomes

$$(2.10)\quad \delta_*(Y) = Y + \nabla \log m(Y|\pi_*), \quad \text{where } m(Y|\pi_*) = \sum_{k=1}^{K} w_k m(Y|\pi_k).$$

Because $m(Y|\pi_*)$ is the marginal density of Y under the mixture prior

$$(2.11)\qquad \pi_*(\theta) = \sum_{k=1}^{K} w_k \pi_k(\theta),$$

it follows that $\delta_*(Y) = E_{\pi_*}(\theta|Y)$ is the Bayes rule under π_*. The assumption that θ has the mixture prior π_*, being equivalent to the assumption that θ has the prior π_k with probability w_k, nicely expresses the vague or conflicting prior information that any of $\delta_1, \dots, \delta_K$ may be effective. This method of combining prior information through mixtures can also be motivated in the multi-Bayesian context [see Kempthorne (1985)].

The Bayesian motivation also provides a natural interpretation of each relevance function in (2.5) which here is,

$$(2.12)\qquad \rho_k(Y) = w_k m(Y|\pi_k)/m(Y|\pi_*) = P(\pi_k|Y),$$

the updated posterior probability that θ has the prior density π_k. The alternative representation of δ_* in (2.4),

$$(2.13)\qquad \delta_*(Y) = \sum_{k=1}^{K} \rho_k(Y)\delta_k(Y) = \sum_{k=1}^{K} P(\pi_k|Y)E_{\pi_k}(\theta|Y)$$

shows how the relevance functions here put increasing weight on the posterior mean $\delta_k(Y) = E_{\pi_k}(\theta|Y)$ which is supported by the data through $m(Y|\pi_k)$. The use of finite mixture distributions to obtain robustness properties in the Bayesian context has been used by Box and Tiao (1968), Abraham and Box (1978), and Zellner (1985).

Although these manipulations are carried through formally in Section 2.1, treating $m_1, \dots, m_K$ in (2.1) as arbitrary functions, the Bayesian character of δ_* suggests that desirable properties may be obtained when these functions are at least approximations to marginal densities. However, one drawback is that when $m_1, \dots, m_K$ are not marginal densities corresponding to bonafide priors, the weights $w_1, \dots, w_K$ lose their interpretation as prior probabilities in the mixture prior π_*. Nonetheless, it may be useful even in non-Bayes examples of δ_*, to consider calibrations of these weights which roughly reflect the statistician's prior probability or degree of belief in the potential effectiveness of the estimators $\delta_1, \dots, \delta_K$. Although the choice of prior weights ultimately corresponds to the choice of a risk function, such an interpretation may facilitate their specification in practice.

3. A multiple shrinkage Stein estimator. In this section, we consider the special case of δ_* in (2.1) obtained when $\delta_1, \dots, \delta_k$ in (2.2) are general positive-part Stein estimators. Other examples of multiple shrinkage estimators have been considered by the author in George (1986a, 1986b, 1986c). The following notation will be used throughout. Let

$$V_1, \dots, V_K$$

denote a set of (possibly affine) subspaces of R^p such that V_k has dimension $p - q_k$ where $q_k \geq 3$. For any $Y \in R^p$, let $P_k Y$ denote the projection of Y onto V_k, defined by $\|Y - P_k Y\| = \min_{v \in V_k} \|Y - v\|$. For convenience, let

$$s_k(Y) = \|Y - P_k Y\|^2$$

denote the squared distance from Y to V_k.

3.1. Construction of a multiple shrinkage Stein estimator. As a more general version of the example described in Section 1, suppose vague or conflicting prior information suggested that small risk might be obtainable by using one of the following K positive-part Stein estimators, $\delta_1^S, \dots, \delta_K^S$, which shrink Y towards the subspaces $V_1, \dots, V_K$, respectively,

$$\delta_k^S(Y) = Y - \left[1 \wedge \frac{q_k - 2}{s_k(Y)}\right](Y - P_k Y), \qquad k = 1, \dots, K, \tag{3.1}$$

where $a \wedge b = \min\{a, b\}$, see Sclove, Morris, and Radhakrishnan (1972). For example, the estimator δ_v^S in (1.3) is a special case of δ_k^S when $V_k = v \in R^p$, $q_k = p$, $P_k Y \equiv v$, and $s_k(Y) = \|Y - v\|^2$. Another common choice [see Lindley (1962) and Efron and Morris (1975)], is $V_k = [1_p]$, the subspace spanned by the vector $1_p = (1, \dots, 1)'$, in which case $q_k = p - 1$, $P_k Y = \bar{Y} 1_p$, and $s_k(Y) = \|Y - \bar{Y} 1_p\|^2$ where $\bar{Y} = \sum_{i=1}^p Y_i / n$.

Typically, the targets $V_1, \dots, V_K$ would correspond here to several guesses for the approximate location of θ. As distinct from the example in Section 1, this more general situation allows for overlapping targets; $V_1, \dots, V_K$ might even be a sequence of nested subspaces. As is well known [and is illustrated by (3.9)], each δ_k^S yields meaningful risk reduction over δ^{MLE} only when θ is close to V_k, and this reduction is larger when V_k has smaller dimension; indeed, when $\theta \in V_k$, $R(\theta, \delta_k^S)$ is slightly less than $p - q_k + 2$. Thus, when the prior information was correct that θ was close to one or more of $V_1, \dots, V_K$, some of the estimators $\delta_1^S, \dots, \delta_K^S$ could offer substantially smaller risk than others. Failure to choose a more effective δ_k^S would then result in foregoing large potential risk reduction.

To avoid the limitation of choosing a single Stein estimator for this situation, we construct a multiple shrinkage alternative. Generalizing the expression in Stein (1973) for the case $V_k = 0$, each of the estimators in (3.1) is of the form $\delta_k^S(Y) = Y + \nabla \log m_k^S(Y)$ as in (2.1), where

$$\nabla \log m_k^S(Y) = -\left[1 \wedge \frac{q_k - 2}{s_k(Y)}\right](Y - P_k Y) \tag{3.2}$$

when

$$m_k^S(Y) = \begin{cases} ((q_k - 2)/e s_k(Y))^{(q_k - 2)/2} & \text{if } s_k(Y) \geq (q_k - 2), \\ e^{-s_k(Y)/2} & \text{if } s_k(Y) < (q_k - 2). \end{cases} \tag{3.3}$$

Applying the construction in Section 2.1 to $\delta_1^S, \dots, \delta_K^S$, thus yields the multiple shrinkage Stein estimator

$$(3.4) \qquad \delta_*^S(Y) = Y + \nabla \log m_*^S(Y), \quad \text{where } m_*^S(Y) = \sum_{k=1}^{K} w_k m_k^S(Y),$$

a special case of δ_* in (2.2) where $m_* = m_*^S$, $m_k = m_k^S$, and $w_1, \dots, w_K$ are prior weights as in (2.3). Note that although each m_k^S is determined by (3.2) only up to a proportionality constant, to facilitate comparisons we have scaled $m_1^S, \dots, m_k^S$ in (3.3) to be equal when $s_1 = \cdots = s_K = 0$. It should be emphasized that $m_1^S, \dots, m_K^S$ are not real marginal densities so that $w_1, \dots, w_K$ will not be real prior probabilities here. Nonetheless, it may be useful to regard each m_k^S as an estimate of an unknown marginal (see Section 3.4). When $V_1, \dots, V_K$ are equidimensional, so that $q_1 = \cdots = q_K$, it may be reasonable to treat $w_1, \dots, w_K$ as prior probabilities; by symmetry considerations, the normalizing constants which would relate the m_k^S to real marginals would then be the same. However, when $q_1, \dots, q_K$ are unequal, the absence of an appropriate normalization of $m_1^S, \dots, m_K^S$ makes any such interpretation more tenuous.

As in (2.4) and (2.5), the following reexpressions show how δ_*^S is an adaptive convex combination of the estimators $\delta_1^S, \dots, \delta_K^S$,

$$(3.5) \quad \delta_*^S(Y) = Y - \sum_{k=1}^{K} \rho_k^S(Y)\left[1 \wedge \frac{q_k - 2}{s_k(Y)}\right](Y - P_k Y) = \sum_{k=1}^{K} \rho_k^S(Y)\delta_k^S(Y),$$

where

$$(3.6) \qquad \rho_k^S(Y) = w_k m_k^S(Y)/m_*^S(Y).$$

The behavior of δ_*^S is intuitively appealing. First of all, when Y is far from all the targets, δ_*^S behaves essentially like δ^{MLE} since the shrinkage provided by each δ_k^S is trivial. To describe the behavior of δ_*^S as Y approaches the targets, it is useful to begin with the special case of equidimensional targets, $q_1 = \cdots = q_K$, and uniform prior weights, $w_1 = \cdots = w_K$. In this case $w_1 m_1^S, \dots, w_K m_K^S$ are identical decreasing functions of $s_1, \dots, s_K$, so that $\rho_k^S > \rho_l^S$ iff $s_k < s_l$. Because $[1 \wedge (q_k - 2)/s_k]$ is also decreasing in s_k, δ_*^S puts more weight on those δ_k^S which are shrinking most. Effectively, δ_*^S shrinks Y in the direction of the closer targets, and the magnitude of shrinkage increases with the proximity of Y to these targets. Use of nonuniform prior weights proportionately changes the relative weighting of $\delta_1^S, \dots, \delta_K^S$, changing the magnitude and direction of shrinkage by δ_*^S accordingly. However, because $\rho_1^S, \dots, \rho_K^S$ are so sharply adaptive, especially when $q_1, \dots, q_K$ are large, δ_*^S will essentially emulate δ_k^S when Y is close to V_k and no other target, as long as the prior weights are not too disparate.

In the general case where $q_1, \dots, q_K$ are unequal, the functions m_k^S for which q_k is larger, decrease more rapidly. Unless w_k is chosen larger when q_k is larger, δ_*^S may fail to exploit very much of the shrinkage potential of the δ_K^S corresponding to the lower dimensional targets. For example, when $q_k > q_l$, ρ_k^S/ρ_l^S may drop off very quickly as s_k increases, especially if the targets were nested, $V_k \subset V_l$. Setting $w_k = w_l$ might result in $\rho_k^S \ll \rho_l^S$ even when $(q_k - 2)/s_k > (q_l - 2)/s_l$

and δ_k^S is shrinking more than δ_l^S. This behavior can be roughly avoided by using the calibration

$$w_k = (ce)^{(q_k-2)/2}, \qquad k = 1,\dots,K, \tag{3.7}$$

which for $c \geq 1$ forces $\rho_k^S > \rho_l^S$ when $(q_k - 2)/s_k > (q_l - 2)/s_l \geq 1/c$ and $q_k > q_l$. In the next section this calibration is seen to be reasonable from a risk perspective. Various choices of c are briefly examined in one of the simulations in Section 3.3.

3.2. The risk of a multiple shrinkage Stein estimator. The application of the results of Section 2.2, shows that δ_*^S inherits desirable risk properties from $\delta_1^S,\dots,\delta_K^S$. To begin with, δ_*^S is minimax. This property follows from Corollary 2 and the superharmonicity of $m_1^S,\dots,m_K^S$ which is demonstrated by

$$\nabla^2 m_k^S(Y)/m_k^S(Y) = \begin{cases} 0 & \text{if } s_k(Y) \geq (q_k - 2), \\ -(q_k - s_k(Y)) & \text{if } s_k(Y) < (q_k - 2). \end{cases} \tag{3.8}$$

Note that (3.8) and Corollary 1 provide an immediate verification of the well-known minimaxity of δ_k^S.

The following unbiased estimates of risk reduction provide some insight as to the regions of the parameter space where δ_*^S may potentially offer meaningful risk reduction. Inserting (3.2) and (3.8) into (2.6) in Theorem 1, yields the risk reduction estimate for δ_k^S,

$$D\delta_k^S(Y) = \begin{cases} (q_k - 2)^2/s_k(Y) & \text{if } s_k(Y) \geq (q_k - 2), \\ 2q_k - s_k(Y) & \text{if } s_k(Y) < (q_k - 2), \end{cases} \tag{3.9}$$

a slight generalization of the result in Stein (1973) for the case $V_k = 0$. By Corollary 3, the risk reduction estimate for δ_*^S may be expressed in terms of (3.9) as

$$D\delta_*^S(Y) = \sum_{k=1}^K \rho_k^S(Y)\left[D\delta_k^S(Y) - \frac{1}{2}\sum_{l=1}^K \rho_l^S(Y)\|\delta_k^S(Y) - \delta_l^S(Y)\|^2\right]. \tag{3.10}$$

We should point out that although $D\delta_k^S$ and $D\delta_*^S$ are useful for making risk comparisons, they are not always reasonable as estimates of risk. For example, $D\delta_k^S > p$, which occurs when $q_k = p$ and s_k is small, leads to a negative risk estimate, which is silly.

Comparison of (3.5) and (3.10) shows that $D\delta_*^S$ adaptively emulates the risk estimates $D\delta_1^S,\dots,D\delta_K^S$ much in the same way that δ_*^S adaptively emulates the estimators $\delta_1^S,\dots,\delta_K^S$. Consider first the equidimensional and uniformly weighted case where $\rho_k^S > \rho_l^S$ iff $s_k < s_l$. Because $D\delta_k^S$ is decreasing in s_k, ρ_k^S and $D\delta_k^S$ will be large simultaneously. Since $\rho_1^S,\dots,\rho_K^S$'are so sharply adaptive, $D\delta_*^S(Y) \approx \max_k D\delta_k^S(Y)$ whenever Y is close to some V_k, suggesting that $R(\theta,\delta_*^S) \approx \min_k R(\theta,\delta_k^S)$ whenever θ is close to $V_1 \cup \cdots \cup V_K$. Of course, we do not believe (although we have not been able to prove it) that δ_*^S will dominate δ_k^S when $\theta \in V_k$; intuitively, when $\theta \in V_k$, δ_k^S will always shrink in the correct direction, whereas δ_*^S will not. Although δ_*^S may not provide quite as much risk reduction as the most effective δ_k^S, the increased size of the region of improvement may be a very desirable trade-off. Indeed, the simulation results in the next section suggest that the approximation of $R(\theta, \delta_*^S)$ to $R(\theta, \delta_k^S)$ when θ is close to V_k can be excellent. Note that by increasing w_k, one can improve this approximation, although it would be at the expense of less risk improvement near some of the other targets.

In the general case where $q_1, \dots, q_K$ are unequal, the form of $D\delta_*^S$ suggests that uniform prior weights are less desirable because for larger q_k, ρ_k^S may drop off very quickly as s_k increases, especially when V_k was nested in a higher dimensional subspace. Instead, it seems desirable to choose $w_1, \dots, w_K$ so that $\rho_k^S(Y) \approx 1$ when $D\delta_k^S(Y) = \max_l D\delta_l^S(Y)$ and s_k is small. Analogously to the equidimensional case, such behavior would yield $D\delta_*^S(Y) \approx \max_k D\delta_k^S(Y)$ and consequently $R(\theta, \delta_*^S) \approx \min_k R(\theta, \delta_k^S)$, when Y or θ was close to $V_1 \cup \cdots \cup V_K$, respectively. The calibration suggested in (3.7) seems to roughly achieve this goal, as is borne out by the simulations in the next section.

3.3. Simulations of the multiple shrinkage Stein estimator. To gain some idea of the potential quality of the approximation of $R(\theta, \delta_*^S)$ to $\min_k R(\theta, \delta_k^S)$, we obtained Monte Carlo estimates for the case $p = 10$, of the risk of δ_*^S and the corresponding Stein estimators for simple examples of the equidimensional target case and the nested subspace target case. The risk of each estimator for each choice of θ was estimated by the average loss $\|\delta - \theta\|^2$ based on 10,000 independent samples of $Y \sim N_{10}(\theta, I)$. (The normal random deviates were generated from the IMSL routine GGNML.) In assessing the potential practical value of the estimates, recall that $R(\theta, \delta^{\mathrm{MLE}}) \equiv 10$ here.

In the equidimensional case, we simulated the risk of two Stein estimators δ_k^S with $V_k = v_k \in R^{10}$, $k = 1, 2$ and two choices of the corresponding multiple shrinkage estimator δ_*^S with $r = w_1/w_2 = 1$ and 9 ($K = 2$). Three choices of v_1 and v_2 were considered, corresponding to the separations $d^2 = \|v_1 - v_2\|^2 = 2.5, 10, 40$, obtained by changing each coordinate 0.5, 1, 2 standard deviations. For each separation, eight values of $\theta = (1 - \lambda)v_1 + \lambda v_2$ obtained by varying $\lambda = -0.5, 1.5\,(0.25)$ were considered. The risk estimates, which appear in Table 1, show that the risk reduction of δ_*^S is impressive. When $r = 1$, the performance of δ_*^S at the separation of $d^2 = 40$, is essentially indistinguishable from the best of δ_1^S and δ_2^S. For the smaller separations $d^2 = 2.5, 10$, the performance close to the targets deteriorates only slightly, although it improves between the targets. For the nonuniformly weighted case with $r = 9$, the performance of δ_*^S improves slightly when θ is close to v_2, and deteriorates slightly when θ is close to v_1, apparently the result of the strongly adaptive relevance functions.

In the case of nested subspace targets, we considered eight Stein estimators δ_k^S, $k = 1, \dots, 8$, for which $V_k = \{v \in R^{10}\colon v^{i'} = 0 \text{ if } i \geq k\}$ where v^i is the ith coordinate of v, and six choices of δ_*^S ($K = 8$), using calibrations of the prior weights given by (3.7) with $c = 1, 2, 3, 5, 10, 50$. The risk of these estimators was compared for $\theta = 0$ and for eight choices of $\|\theta\|^2 = \theta_i^2 = 40$, $i = 1, \dots, 8$. These values of θ were chosen because setting $\|\theta\|^2 = \theta_i^2 = 40$ effectively eliminates the

TABLE 1
The risk of δ_*^S when $Y \sim N_{10}(\theta, I)$—the equidimensional case

	$\theta = (1 - \lambda)v_1 + \lambda v_2$								
$\lambda =$	−0.50	−0.25	0.00	0.25	0.50	0.75	1.00	1.25	1.50
$d^2 = 40$									
$R(\theta, \delta_k^S)$, $k = 1$	6.2	3.2	1.3	3.2	6.1	7.8	8.6	9.1	9.4
2	9.4	9.1	8.7	7.8	6.2	3.2	1.3	3.2	6.1
$R(\theta, \delta_*^S)$, $r = 1$	6.2	3.2	1.3	3.6	6.1	3.6	1.3	3.2	6.1
9	6.2	3.2	1.6	4.3	6.1	3.3	1.3	3.2	6.1
$d^2 = 10$									
$R(\theta, \delta_k^S)$, $k = 1$	3.2	1.8	1.3	1.8	3.2	4.8	6.1	7.1	7.8
2	7.9	7.2	6.2	4.8	3.2	1.8	1.3	1.8	3.2
$R(\theta, \delta_*^S)$, $r = 1$	3.3	2.1	1.8	2.2	2.5	2.2	1.7	2.1	3.3
9	4.0	3.1	3.0	3.1	2.7	1.8	1.4	1.9	3.2
$d^2 = 2.5$									
$R(\theta, \delta_k^S)$, $k = 1$	1.8	1.4	1.3	1.4	1.8	2.4	3.1	4.0	4.8
2	4.7	3.9	3.1	2.4	1.8	1.4	1.3	1.4	1.8
$R(\theta, \delta_*^S)$, $r = 1$	2.3	1.9	1.6	1.5	1.4	1.5	1.6	1.9	2.3
9	3.6	3.0	2.5	2.0	1.6	1.4	1.3	1.5	1.9

Note: 10,000 replications. The standard error of each estimate is less than 0.04.

TABLE 2
The risk of δ_*^S when $Y \sim N_{10}(\theta, I)$—the nested case

			$\|\theta\|^2 = \theta_i^2 = 40$; $i =$							
		$\theta = 0$	1	2	3	4	5	6	7	8
$R(\theta, \delta_k^S)$, $k =$	1	1.3	8.6	8.6	8.6	8.6	8.6	8.6	8.6	8.6
	2	2.3	2.3	8.9	9.0	8.9	8.9	8.9	8.9	8.9
	3	3.3	3.3	3.3	9.2	9.2	9.2	9.2	9.2	9.2
	4	4.3	4.3	4.3	4.3	9.4	9.4	9.4	9.4	9.4
	5	5.4	5.4	5.4	5.4	5.4	9.6	9.6	9.6	9.6
	6	6.4	6.4	6.4	6.4	6.4	6.4	9.8	9.8	9.8
	7	7.5	7.5	7.5	7.5	7.5	7.5	7.5	9.9	9.9
	8	8.6	8.6	8.6	8.6	8.6	8.6	8.6	8.6	10.0
$R(\theta, \delta_*^S)$, $c =$	1	4.8	5.4	5.9	6.4	7.0	7.5	8.1	8.8	9.9
	2	2.3	3.3	4.3	5.3	6.3	7.3	8.2	9.0	9.6
	3	1.8	2.9	4.0	5.1	6.4	7.6	8.5	9.0	9.1
	5	1.5	2.6	3.9	5.5	7.1	8.3	8.7	8.8	8.8
	10	1.4	2.6	4.4	6.8	8.3	8.7	8.7	8.7	8.7
	50	1.3	3.1	7.0	8.6	8.6	8.7	8.6	8.6	8.7

Note: 10,000 replications. The standard error of each estimate is less than 0.05.

useful risk reduction of those δ_k^S for which $k \le i$. The risk estimates, which appear in Table 2, show that from a practical point of view, the approximation $R(\theta, \delta_*^S) \approx \min_k R(\theta, \delta_k^S)$, can be excellent when θ is close to any of the targets. Indeed, when $c = 2$, $\hat{R}(\theta, \delta_*^S) \le \min_k \hat{R}(\theta, \delta_k^S) + 1$ for each θ considered. As c is increased, the improvement at the smaller dimensional targets is improved, though at the expense of some deterioration at the other targets. The calibration of prior weights given by (3.7) seems to work quite well here, and yielded better results than other calibrations that we tried. Finally, to end on a cautious note, this second simulation explores a very small region of the parameter space. Before δ_*^S can be used with confidence in a nested subspace situation like this, a much more comprehensive simulation study would be needed.

3.4. An approximation for a family of mixture priors. Although δ_*^S is not a Bayes rule, it may be useful to regard it as an approximation to Bayes rules. Such an approximation is suggested by the empirical Bayes relationship of δ_k^S to the Bayes rule

$$E_{\pi_k}(\theta|Y) = Y - \left(\frac{1}{1 + a_k}\right)(Y - \mu_k), \tag{3.11}$$

$$\text{where } \pi_k(\theta) = (2\pi a_k)^{-p/2} e^{-\|\theta - \mu_k\|^2/2a_k},$$

when it is assumed only that π_k belongs to the family of conjugate priors

$$\Gamma_k = \{\pi_k(\theta) \colon \mu_k \in V_k \text{ and } a_k \ge 0\}. \tag{3.12}$$

δ_k^S is typically motivated as an empirical Bayes approximation to $E_{\pi_k}(\theta|Y)$ by inserting the estimates

$$\hat{\mu}_k = P_k Y \quad \text{and} \quad \hat{a}_k = \max\{0, (s_k(Y)/(q_k - 2)) - 1\} \tag{3.13}$$

into the left-hand expression in (3.11) [see e.g., Stein (1962), Efron and Morris (1973a), Zellner and Vandaele (1974), and Morris (1983)]. Because

$$E_{\pi_k}(\theta|Y) = Y + \nabla \log m(Y|\pi_k) = Y - \left(\frac{1}{1 + a_k}\right)(Y - \mu_k), \tag{3.14}$$

where

$$m(Y|\pi_k) = (2\pi(1 + a_k))^{-p/2} e^{-\|Y - \mu_k\|^2/2(1 + a_k)}, \tag{3.15}$$

$m_k^S(Y)$ may then be regarded as an estimate of the marginal density $m(Y|\pi_k)$ (up to a proportionality constant), implicitly determined by $\hat{\mu}_k$ and $\hat{a}_k$ (or equivalently $\delta_k^S(Y)$), through (3.14). Note that $m_k^S(Y)$ is not obtained by inserting the estimates $\hat{\mu}_k$ and $\hat{a}_k$ directly into $m(Y|\pi_k)$ in (3.15).

By treating $m_k^S(Y)$ and $\delta_k^S(Y)$ as estimates of $m(Y|\pi_k)$ and $E_{\pi_k}(\theta|Y)$, δ_*^S may then be regarded as an approximation to the Bayes rules for the family of mixtures of conjugate priors,

$$\Gamma_* = \left\{ \pi_*: \pi_*(\theta) = \sum_{k=1}^{K} w_k' \pi_k(\theta), \text{ where } \pi_k \in \Gamma_k \right\}, \tag{3.16}$$

since each of these Bayes rules may be expressed as

$$E_{\pi_*}(\theta|Y) = Y + \nabla \log m(Y|\pi_*), \quad \text{where } m(Y|\pi_*) = \sum_{k=1}^{K} w_k' m(Y|\pi_k) \tag{3.17}$$

or

$$E_{\pi_*}(\theta|Y) = \sum_{k=1}^{K} P(\pi_k|Y) E_{\pi_k}(\theta|Y), \tag{3.18}$$
$$\text{where } P(\pi_k|Y) = w_k' m(Y|\pi_k)/m(Y|\pi_*).$$

Because of the absence of meaningful norming constants for $m_1^S, \dots, m_K^S$, the prior probabilities $w_1', \dots, w_K'$ in (3.16)–(3.18) may differ from $w_1, \dots, w_K$.

The family Γ_* generalizes the family Γ_k in (3.11), allowing for much more flexibility in the specification of the location of the prior mean. Note that although the empirical Bayes approach of inserting parameter estimates has been used successfully with families of contaminated mixture priors by Berger and Berliner (1983, 1984), insertion of the estimators $\hat{\mu}_k$ and $\hat{a}_k$ directly into $E_{\pi_*}(\theta|Y)$ into (3.17) or (3.18) would not yield δ_*^S. Indeed, the resulting estimators appear not to be minimax in general [see George (1986c)].

It is interesting to contrast δ_*^S with the Bayes estimator $E_{\pi_*}(Y|\theta)$. Both the relevance function ρ_k^S and the posterior probability $P(\pi_k|Y)$ are adaptive and put increasing weight on the estimator which is supported by the data. However, each δ_k^S shrinks less when Y is further from V_k, in sharp contrast to $E_{\pi_k}(\theta|Y)$ which shrinks more. Only δ_*^S possesses the robust property of behaving like δ^{MLE} when Y is far from all the targets.

4. The general case. As a generalization of the situation in Section 2, suppose vague or conflicting prior information suggested that small risk might be obtainable by some member of a possibly infinite set of estimators,

$$\Delta_\Omega = \{\delta_\omega: \delta_\omega(Y) = Y + \nabla \log m_\omega(Y), \omega \in \Omega\}, \tag{4.1}$$

where for each ω in the indexing set Ω, $m_\omega: R^p \to R^+ \cap \{0\}^c$ is such that m_ω and ∇m_ω are a.d. Let W be a probability measure on Ω such that for a.e. $y \in R^p$, $m_\omega(y)$ is a measurable function of ω wrt W, and

$$m_*(Y) = \int_\Omega m_\omega(Y) W(d\omega) \tag{4.2}$$

exists and is such that ∇ and $\int$ may be interchanged to yield,

$$\nabla m_*(Y) = \int_\Omega \nabla m_\omega(Y) W(d\omega)$$
$$\text{and} \quad \nabla^2 m_*(Y) = \int_\Omega \nabla^2 m_\omega(Y) W(d\omega). \tag{4.3}$$

Note that any discrete finite probability measure W will always satisfy these conditions. With this setup, a multiple shrinkage estimator may be defined as

$$(4.4) \qquad \delta_* = Y + \nabla \log m_*(Y),$$

and may be reexpressed as

$$(4.5) \qquad \delta_*(Y) = Y + \int_\Omega \nabla \log m_\omega(Y)\rho(Y, d\omega) = \int_\Omega \delta_\omega(Y)\rho(Y, d\omega),$$

where

$$(4.6) \qquad \rho(Y, d\omega) = m_\omega(Y)W(d\omega)/m_*(Y).$$

The probability measure W generalizes the prior weights $w_1, \dots, w_K$ in (2.3), and the adaptive probability measure $\rho(Y, d\omega)$ generalizes the relevance functions $\rho_1(Y), \dots, \rho_K(Y)$ in (2.5). Indeed, when W is a discrete finite probability measure, δ_* in (4.4) reduces to δ_* in (2.2).

As in the discrete case, it is of interest to apply Stein's Theorem 1 and Corollary 1 to this general version of δ_*. The following analogues of Lemma 1 and Corollaries 2 and 3, which are proved similarly, depend on both Δ_Ω and W.

LEMMA 2. *If Δ_Ω and W are such that*

$$\text{(i)} \qquad E_\theta \int_\Omega |\nabla_i^2 m_\omega(Y)/m_\omega(Y)|\rho(Y, d\omega) < \infty, \qquad i = 1, \dots, p,$$

$$\text{(ii)} \qquad E_\theta \int_\Omega \|\nabla \log m_\omega(Y)\|^2 \rho(Y, d\omega) < \infty,$$

Then δ_* satisfies the conditions of Theorem 1.

COROLLARY 4. *If Δ_Ω and W are such that the conditions of Lemma* 2 *are satisfied and each $m_\omega \in \Delta_\Omega$ is superharmonic, then δ_* is minimax.*

COROLLARY 5. *If Δ_Ω and W are such that the conditions of Lemma* 2 *are satisfied, then*

$$(4.7) \quad D\delta_*(Y) = \int_\Omega \left[D\delta_\omega(Y) - \int_\Omega \|\delta_\omega(Y) - \delta_\eta(Y)\|^2 \rho(Y, d\eta) \right] \rho(Y, d\omega).$$

Also, note when each $\delta_\omega \in \Delta_\Omega$ is a Bayes rule with respect to a prior $\pi_\omega(\theta)$, then $\delta_* = E_{\pi_*}(\theta|Y)$ will be the Bayes rule corresponding to the mixture prior

$$(4.8) \qquad \pi_*(\theta) = \int_\Omega \pi_\omega(\theta) W(d\omega),$$

generalizing the motivation in Section 2.3.

EXAMPLE 1. *Shrinkage towards an arbitrary set.* Suppose interest was initially focused on using a Stein estimator of the form δ_v^S in (1.3), but vague prior information suggested only that θ was close to some set $A \subset R^p$. Instead of choosing an estimator from the set

$$\Delta_A = \{\delta_v^S\colon \delta_v^S(Y) = Y + \nabla \log m_v^S(Y),\ v \in A\}, \tag{4.9}$$

where $m_v^S(Y)$ is the special case of $m_k^S(Y)$ in (3.3) when $V_k = v$, a more desirable estimator may be a multiple shrinkage Stein estimator of the form

$$\delta_*^S(Y) = Y + \nabla \log m_*^S(Y), \qquad m_*^S(Y) = \int_A m_v^S(Y) W(dv), \tag{4.10}$$

where W is some probability measure on A such that $m_*^S(Y)$ exists and (4.3) holds. For example, if available prior information suggested only that $\|\theta\| \approx r > 0$, then appropriate choices for A and W would be $B_r = \{v \in R^p\colon \|v\| = r\}$ and the uniform measure on B_r. Alternative estimators which shrink Y towards B_r have been considered by Bock (1983) and George (1986c).

Although the conditions of Lemma 2 must in general be verified for each choice of Δ_A and W, it can be shown that these will hold whenever A is bounded. Thus, by Corollary 4, any choice of δ_*^S in (4.10) with $A = B_r$ will be minimax.

EXAMPLE 2. *Shrinkage towards a subspace measured with error.* Consider the situation where θ was thought to lie close to $[X]$, the subspace spanned by the columns of a $p \times n$ matrix X $(n \le p - 3)$, and interest was initially focused on using a Stein estimator of the form δ_k^S in (3.1) with $V_k = [X]$. However, suppose that these columns were covariates observed with error; that only $X_\xi = X + \xi$ was available, with ξ an unobservable $p \times n$ matrix of errors with distribution Ψ. Instead of choosing a Stein estimator from the set

$$\Delta_\Omega = \{\delta_\xi^S\colon \delta_\xi^S(Y) = Y + \nabla \log m_\xi^S(Y),\ \xi \in \Omega\}, \tag{4.11}$$

where $\Omega = R^{p \times n}$, and $m_\xi^S(Y)$ is the special case of $m_k^S(Y)$ in (3.3) with $V_k = [X_\xi]$, it may be more desirable to use a multiple shrinkage Stein estimator of the form

$$\delta_*^S(Y) = Y + \nabla \log m_*^S(Y), \qquad m_*^S(Y) = \int_\Omega m_\xi^S(Y) \Psi(d\xi) \tag{4.12}$$

when X and Ψ are such that $m_*^S(Y)$ exists and (4.3) holds. As in Example 1 above, to apply Corollaries 4 and 5, the conditions of Lemma 2 must in general be verified for each choice of X and Ψ.

5. The case of unknown variance. The multiple shrinkage estimator δ_* in (2.2) or (4.4) is easily extended to handle the more realistic situation where

$$Y|\theta, \sigma \sim N_p(\theta, \sigma^2 I), \tag{5.1}$$

and an independent estimate of σ^2 is available, namely

$$S \sim \sigma^2 \chi_d^2, \tag{5.2}$$

where χ_d^2 is the chi-square distribution with d degrees of freedom. Simply replace

$\delta_*(Y) = Y + \nabla \log m_*(Y)$ by the multiple shrinkage estimator

$$(5.3) \qquad \delta_*^{\sigma}(Y) = Y + \frac{S}{d+2} \nabla \log m_*(Y).$$

When δ_* satisfies the conditions of Theorem 1, it is easy to see from the main results of Stein (1981) (Section 8) that δ_*^{σ} has risk

$$(5.4) \qquad R(\theta, \sigma, \delta_*^{\sigma}) \equiv E_{\theta,\sigma}\|\theta - \delta_*^{\sigma}\|^2 = \sigma^2\left[p - \frac{d}{d+2} E_{\theta,\sigma} D\delta_*(Y/\sigma)\right],$$

where $D\delta_*(Y/\sigma)$ is given by (2.6). The generalization of the other results is straightforward. As Stein points out, the reduction in risk due to not knowing σ^2 is only reduced by a factor of $d/(d+2)$.

Acknowledgments. This research was supported by the Graduate School of Business of the University of Chicago. The author is extremely grateful to James Berger whose generous and insightful comments on earlier versions were essential to the crystallization of the main ideas in this paper.

REFERENCES

ABRAHAM, B. and BOX, G. E. P. (1978). Linear models and spurious observations. *Appl. Statist.* **27** 131–138.

BERGER, J. (1982). Selecting a minimax estimator of a multivariate normal mean. *Ann. Statist.* **10** 81–92.

BERGER, J. (1983). The Stein effect. In *Encyclopedia of Statistical Sciences.* (S. Kotz and N. L. Johnson, eds.). Wiley, New York.

BERGER, J. and BERLINER, L. M. (1983). Robust Bayes and empirical Bayes analysis with ε-contaminated priors, Technical Report 83-35, Statist. Dept., Purdue Univ.

BERGER. J. and BERLINER, L. M. (1984). Bayesian input in Stein-estimation and a new minimax empirical bayes estimator, *J. Econometrics* **25** 87–108.

BOCK, M. E. (1983). Minimax estimators that shift towards a hypersphere for location vectors of spherically symmetric distributions. Unpublished manuscript.

BOX, G. E. P., and TIAO, G. C. (1968). A Bayesian approach to some outliner problems. *Biometrika* **55** 119–129.

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

EFRON, B. and MORRIS, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators—Part II: The empirical Bayes case. *J. Amer. Statist. Assoc.* **67** 130–139.

EFRON, B. and MORRIS, C. (1973a). Stein's estimation rule and its competitors—An empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130.

EFRON, B. and MORRIS, C. (1973b). Combining possibly related estimation problems. *J. Roy. Statist. Soc. Ser. B* **35** 379–421.

EFRON, B. and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311–319.

GEORGE, E. I. (1986a). Combining minimax shrinkage estimators. To appear in *J. Amer. Statist. Assoc.*

GEORGE, E. I. (1986b). A formal Bayes multiple shrinkage estimator. To appear in *Comm. Statist. A —Theory Methods.*

GEORGE, E. I. (1986c). Multiple shrinkage generalizations of the James–Stein estimator. To appear in *Contributions to the Theory and Applications of Statistics.* Academic, New York.

HELMS, L. (1975). *Introduction to Potential Theory.* Wiley, New York.

JAMES, W. and STEIN, C. (1960). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 361–379.

KEMPTHORNE, P. J. (1985). Controlling risks under different loss functions: the compromise decision problem. Technical Report, Dept. of Statist., Harvard Univ.

LINDLEY, D. V. (1962). Discussion on Professor Stein's paper. *J. Roy. Statist. Soc. Ser. B* **24** 285–287.

MORRIS, C. N. (1983). Parametric empirical bayes inference: theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65.

SCLOVE, S. L., MORRIS, C. and RADHAKRISHNAN, R. (1972). Non-optimality of preliminary-test estimators for the multinormal mean. *Ann. Math. Statist.* **43** 1481–1490.

STEIN, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 197–206.

STEIN, C. (1962). Confidence sets for the mean of a multivariate normal distribution. *J. Roy. Statist. Soc. Ser. B* **24** 265–296.

STEIN, C. (1973). Estimation of the mean of a multivariate normal distribution. *Proc. Prague Symp. Asymptotic Statist.* 345–381.

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

ZELLNER, A. (1985). Bayesian statistics in econometrics. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith, eds.) 571–586. North-Holland, Amsterdam.

ZELLNER, A. and VANDAELE, W. (1974). Bayes-Stein estimators for K-means, regression and simultaneous equation models. In *Studies in Bayesian Econometrics and Statistics.* (S. Fienberg and A. Zellner, eds.). North-Holland, Amsterdam.

GRADUATE SCHOOL OF BUSINESS
UNIVERSITY OF CHICAGO
1101 EAST 58TH STREET
CHICAGO, ILLINOIS 60637