*The Annals of Statistics*
1984, Vol. 12, No. 3, 864–879

# PARAMETRIC ROBUSTNESS: SMALL BIASES CAN BE WORTHWHILE[1]

BY P. J. BICKEL

*University of California, Berkeley*

We study estimation of the parameters of a Gaussian linear model $\mathcal{M}_0$ when we entertain the possibility that $\mathcal{M}_0$ is invalid and a larger model $\mathcal{M}_1$ should be assumed. Estimates are robust if their maximum risk over $\mathcal{M}_1$ is finite and the most robust estimate is the least squares estimate under $\mathcal{M}_1$. We apply notions of Hodges and Lehmann (1952) and Efron and Morris (1971) to obtain (biased) estimates which do well under $\mathcal{M}_0$ at a small price in robustness. Extensions to confidence intervals, simultaneous estimation of several parameters and large sample approximations applying to nested parametric models are also discussed.

**1. Introduction.** The basic aim of robust inference as developed by Huber, Hampel and others has been the production and study of statistical procedures which

(a) perform reasonably well when the parametric assumptions are perfectly satisfied; and

(b) are relatively insensitive to nonparametric departures from parametric assumptions which a given data set is believed to satisfy.

The main parametric model considered has been the Gaussian linear model and the departures, outliers and gross errors in the variables, have been modeled by assuming non-Gaussian error distributions and, where suitable, dependence between the independent and error variables.

An important aspect of this point of view is a focus on inference about parameters of interest rather than on deciding whether the parametric model provides an adequate fit. This is in contrast to the older approach of estimation and testing after a goodness of fit test or more generally rejection of outliers.

The same point of view makes sense in a purely parametric context. We have two possible parametric models in mind, $\mathcal{M}_0$, $\mathcal{M}_1$ with $\mathcal{M}_0 \subset \mathcal{M}_1$. Our primary interest is in estimating parameters which are identifiable in $\mathcal{M}_1$.

Again,

(i) we believe that $\mathcal{M}_0$ is adequate and want estimates or confidence regions based on estimates that perform well under that assumption. However

(ii) we wish to guard against the possible departures presented by $\mathcal{M}_1$.

---

Received April 1983; revised April 1984.

[1] Work performed with the partial support of Office of Naval Research Contract N00014-80-C-0163 and the Adolph and Mary Sprague Miller Foundation. Some of this material was presented at the 1980 Wald Lectures of the Institute of Mathematical Statistics.

*AMS* 1980 *classifications.* Primary 62F10; secondary 62F25.

*Key words and phrases.* Parametric robustness, pretesting, limited translation estimates, confidence intervals.

Here is the main situation we are thinking of with some specific examples.

*Nested linear models.* We observe $y_{n\times 1}$ where

$$y = \theta + e.$$

$e$ is an $n$-variate normal vector with mean 0 and covariance matrix $\Sigma$. $\theta$ ranges freely over an $r$-dimensional linear space $\Theta_0$ under $\mathscr{M}_0$ and over an $s$-dimensional linear space $\Theta_1 \supset \Theta_0$ under $\mathscr{M}_1$ where $r < s \leq n$. We suppose $\Sigma$ known. Our asymptotic analysis in Section 5 will permit us as usual to substitute a consistent estimate $\hat{\Sigma}$ for $\Sigma$. We are interested in inference about $\mu(\theta)$ where $\mu$ is a linear function of $\theta$. Special cases are:

1(a) *Pooling means* (Mosteller, 1948). We are given two samples $X_1, \cdots, X_m$ independent $\mathscr{N}(\mu, \sigma^2)$; $Y_1, \cdots, Y_n$ independent $\mathscr{N}(\mu + \Delta, \sigma^2)$. We want to estimate or set a confidence interval on $\mu$. We believe $\Delta = 0$ ($\mathscr{M}_0$) but want to guard against arbitrary $\Delta$ ($\mathscr{M}_1$). Plausible examples, e.g. measurements in a current and previous survey, are discussed by Mosteller.

1(b) *Additive effects with possible interactions.* Suppose $\mathscr{M}_1$ is an ANOVA model in the sense of Scheffé (1959), possibly including random effects, which contains some interaction terms as well as main effects, and $\mathscr{M}_0$ is purely additive specifying all interactions to be 0. We take the variances of all random effects as well as measurement errors to be known. We want to study some or all of the main effects. An interesting special case is the crossover design discussed by B. W. Brown (1980). Here two groups of subjects I and II which for simplicity we take of equal size $n/2$ are each administered two drugs A, B in succesion and responses measured. The second drug is administered after response to the first has been measured and a time deemed sufficient for the effect of the first to wear off has elapsed. The order of administration of the drugs is AB in group 1, BA in group 2. Model $\mathscr{M}_1$ here is that the response $Y_{ijk(u)}$ of the $j$th subject in group $i$ during period $k$ who is administered drug $u$ during that period is

$$Y_{ijk(u)} = \mu + \pi_k + \phi_u + \lambda_{uk} + \xi_{ij} + \varepsilon_{ijk}$$

where $\pi_k$, $k = 1, 2$, is the period effect, $\phi_u$, $u = \text{A, B}$ is the drug effect, and $\lambda_{uk}$ is the interaction of drug $u$ and period $k$ with $\lambda_{u1} = 0$. These are all fixed. As usual, identifiability requires further linear restrictions. On the other hand, $\xi_{ij}$, the effect of the $j$th subject in group $i$, is considered random $\mathscr{N}(0, \sigma_\xi)$, and $\varepsilon_{ijk}$, the within subject deviation for the $k$th period (including measurement error), is modeled as $\mathscr{N}(0, \sigma_\varepsilon^2)$. All are modeled as independent of each other. We assume $\sigma_\xi^2$, $\sigma_\varepsilon^2$ known. $\mathscr{M}_0$ specifies that, as we hope, there is no interaction, $\lambda_{uk} \equiv 0$. We are interested in estimating $\phi_b - \phi_a$, the difference in effectiveness of the drugs.

1(c) *Nested regression models.* Write $\theta = X\beta$, $\beta_{s\times 1}$, $X = (x_1, \cdots, x_s)$ an $n \times s$ matrix of rank $s$ and think of the $s$ columns of $X$ as corresponding to $s$ independent variables. Suppose $\beta$ ranges freely over $R^s$ under $\mathscr{M}_1$ but $s - r$ coordinates of $\beta$ are set equal to 0 under $\mathscr{M}_0$, i.e. $s - r$ of the independent variables are irrelevant. Various linear functions $\mu(\theta)$ are of interest, for instance the

vector of expectations $\theta$ itself or one or more predicted values $x\beta$, at various values $x$.

From this special case we will proceed (under regularity conditions) by an asymptotic analysis to the general case of

*Nested parametric models.* We observe $(X_1, \cdots, X_n)$ with joint density $p_n(x, \theta)$ (with respect to some measure $\nu_n$). Under $\mathcal{M}_1$, $\theta \in \Theta_1$, an open subset of $s$-dimensional space. Under $\mathcal{M}_0$, $\theta \in \Theta_0 \subset \Theta_1$, a (locally) $r$-dimensional subsurface of $\Theta_1$, and $\mu$ is a smooth vector-valued function of $\theta$. This of course covers all previous situations as well as many others including Example 1 with $\sigma^2$ unknown, nested loglinear models, etc.

Our point of view, essentially already suggested by Hodges and Lehmann (1952), page 402, is that procedures should be judged by their maximum risks under $\mathcal{M}_0$ and $\mathcal{M}_1$. So, in the context of nested parametric models, if $M(\theta, \delta)$ is the risk of a decision rule $\delta$ when $\theta$ is true we should look at

$$m(\delta) = \sup\{M(\theta, \delta): \theta \in \Theta_0\}, \quad M(\delta) = \sup\{M(\theta, \delta): \theta \in \Theta_1\}.$$

$M$ can be thought of as a measure of robustness of $\delta$ and we should be interested in procedures which make $m$ small subject to a bound on $M$.

In the basic linear model example the solutions we end up with are necessarily biased under $\mathcal{M}_1$. Robustness requires that the biases be bounded through $M$. The worthwhile gains are in reduction of $m$ over the unbiased minimax estimate.

In Section 2 we apply this theory to the linear model example for quadratic loss when $\mu$ is one dimensional. The optimal procedures are difficult to compute. We motivate a family of reasonable approximately optimal solutions, compare them numerically to the optimum and other competitors and also briefly discuss the crucial question of selection within the family.

In Section 3 we discuss confidence intervals based on these estimates. In Section 4, we derive, using results of Berger (1982) and Huber (1977), some procedures for the multivariate case. In Section 5, we show how these ideas generalize to yield reasonable procedures in nested parametric models and, finally, in Section 6, give conclusions and propose open questions.

## 2. The nested linear models: dim($\mu$) = 1, quadratic loss.

a) *Optimality theory.* We specialize to estimation of $\mu$ with quadratic loss. That is, we assume that $\mu$ is real, linear, and if $\delta(x)$ is an estimate

$$M(\theta, \delta) = E_\theta(\delta(X) - \mu(\theta))^2. \tag{2.1}$$

Since we assume $\Sigma$ known, we can, by taking $Y^* = Y\Sigma^{-1/2}$, $\mathcal{M}_i^* = \mathcal{M}_i\Sigma^{-1/2}$, reduce our problem to one in which the observation $Y^*$ has covariance matrix $\sigma^2 I$, the standard linear model.

Let $\hat{\mu}_i = \mu(\hat{\theta}_i)$, $i = 0, 1$, be the least squares estimates of $\mu$ under $\mathcal{M}_0$, $\mathcal{M}_1$ respectively. Then, for $i = 0, 1$, $\hat{\mu}_i$ has constant risk and is minmax under $\mathcal{M}_i$. Let

$\sigma_i^2$ be the variance of $\hat{\mu}_i$ so that

$$\inf_\delta M(\delta) = \sigma_1^2, \quad \inf_\delta m(\delta) = \sigma_0^2.$$

Let $\hat{\mu}_c^*$ minimize $m(\delta)$ subject to $M(\delta)/\sigma_1^2 \leq 1/c$ so that $\hat{\mu}_i^* = \hat{\mu}_i$, $i = 0, 1$. Note that $M(\hat{\mu}_0) = \infty$ and $\hat{\mu}_0$ is certainly not robust. Let

$$\rho = \operatorname{corr}(\hat{\mu}_0, \hat{\mu}_1) = \sigma_0/\sigma_1 \tag{2.2}$$

which is independent of the error variance $\sigma^2$,

$$\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0 \tag{2.3}$$

and

$$\sigma_\Delta^2 = \sigma_1^2(1 - \rho^2), \tag{2.4}$$

its variance.

PROPOSITION 1. *The estimate $\hat{\mu}_c^*$ may be written*

$$\hat{\mu}_c^* = \hat{\mu}_0 + \sigma_{\hat{\Delta}} w_q^*(\hat{\Delta}/\sigma_{\hat{\Delta}}) \tag{2.5}$$

*where*

$$q^2 = (1 - c)/c(1 - \rho^2) \tag{2.6}$$

$$\begin{aligned} &w_q^* \text{ is odd and obtained by minimizing } Ew^2(Z) \text{ subject to} \\ &\sup_\Delta E(w(Z + \Delta) - \Delta)^2 \leq 1 + q^2 \text{ for } Z \sim \mathcal{N}(0, 1). \end{aligned} \tag{2.7}$$

NOTE. Evidently $w_q^*$ is the solution of the special case $\mu = \theta$, $r = 0$, $s = 1$, $\sigma^2 = 1$. We call this problem $(P)$.

PROOF. By sufficiency reduce to $\hat{\theta}_1$ and without loss of generality choose a canonical basis so that $\hat{\theta}_0$ consists of the first $r$ components of $\hat{\theta}_1$ and all components of $\hat{\theta}_1$ are independent normal variables with variance $\sigma^2$. Moreover we can arrange that $\hat{\mu}_0/\sigma_0$ is the first component of $\hat{\theta}_1$ and $\hat{\Delta}/\sigma_1(1 - \rho^2)^{1/2}$ is the $(r+1)$st component. Note by Hodges and Lehmann (1952) that $\hat{\mu}_c^*$ is unrestrictedly minimax for the "mixed" model: for suitable $\lambda(c)$ and $\theta = (\theta^{(1)}, \cdots, \theta^{(s)})$, $\hat{\theta}_1$ has density $(1 - \lambda)p_1 + \lambda p_0$ where $p_1$ is the density of $\hat{\theta}_1$ under $\mathcal{M}_1$ and $\theta$, while $p_0$ is the density of $\hat{\theta}_1$ under $(\theta^{(1)}, \cdots, \theta^{(r)}, 0, \cdots, 0)$, i.e. under $\mathcal{M}_0$. We can reduce this unrestricted problem by invariance, using for instance Kiefer's (1957) general results. Since we want to estimate

$$\sigma_0\theta^{(1)} + (1 - \rho^2)^{1/2}\sigma_1\theta^{(r+1)},$$

the problem is invariant under arbitrary translations of $\theta^{(i)}$, $i \neq 1, r + 1$, and we can reduce to $\hat{\mu}_0$, $\hat{\Delta}$. The problem is also invariant under translations of $\hat{\mu}_0$, keeping $\hat{\Delta}$ fixed. Since $\hat{\mu}_c^*$ is unique it therefore must be of the form $\mu_0 + w(\hat{\Delta})$. Claims (2.7) and (2.6) follow by calculation. □

Unfortunately calculation of $w_q^*$ is difficult. See Bickel (1983) for its rather unpleasant qualitative features.

In view of these unpleasant features, it is natural to seek other families of robust estimates with more satisfactory behaviour. By invariance it seems reasonable to look for $\hat{\mu}$ of the form

$$\hat{\mu}_0 + \sigma_{\hat{\Delta}} w(\hat{\Delta}/\sigma_{\hat{\Delta}}). \tag{2.8}$$

For any such estimate

$$M(\hat{\mu}) = \sigma_1^2(\rho^2 + (1 - \rho^2)\sup_{\Delta} E(w(Z + \Delta) - \Delta)^2) \tag{2.9}$$

$$m(\hat{\mu}) = \sigma_1^2(\rho^2 + (1 - \rho^2)Ew^2(Z)). \tag{2.10}$$

Abusing notation, let us call the coefficients of $(1 - \rho^2)$ inside parentheses in these expressions $M_0(w)$, $m_0(w)$. They correspond to $M$ and $m$ in problem (P).

b) *"Approximate" optimality in problem* (P). From (2.9) and (2.10) reasonable $w$ in problem (P) correspond to reasonable $\hat{\mu}$. In problem (P) we observe $X = Z + \Delta$, $Z \sim \mathcal{N}(0, 1)$ and we want to minimize $m_0(w)$ subject to a bound on $M_0(w)$. Three approximate optimality principles lead to the same family, the limited translation estimates of Efron and Morris (1971) defined by

$$\begin{aligned} e_q(x) &= 0, && |x| \leq q \\ &= x - q \operatorname{sgn} x, && |x| > q, \end{aligned}$$

which leads to $M_0(e_q) = 1 + q^2$.

I. *Optimality in a related problem* (Bickel, 1983, Marazzi, 1980). Suppose $\pi$ is a prior distribution, $r(\pi)$ the Bayes risk, $w_\pi$ the Bayes estimate, and $G_\pi = \pi * \Phi$, where $*$ denotes convolution, is the marginal distribution of $X$. Then,

$$r(\pi) = 1 - I(G_\pi) \tag{2.11}$$

$$w_\pi(x) = x + (g_\pi'/g_\pi)(x) \tag{2.12}$$

where $g_\pi$ is the density of $G_\pi$, $I(G)$ is the Fisher information where

$$\begin{aligned} I(G) &= \int \frac{[g']^2}{g}(x)\, dx, && \text{if the integral is defined} \\ &= \infty && \text{otherwise.} \end{aligned}$$

By Hodges and Lehmann (1952) and (2.11), the optimal $w_q^*$ corresponds to $G_q^*$ which for some $\lambda(q)$ minimizes $I(G)$ over $\mathcal{G}_0 = \{G = (1 - \lambda)\Phi + \lambda\Phi * H, H \text{ arbitrary}\}$. If we "approximate" $\mathcal{G}_0$ by $\mathcal{G}_1 = \{G = (1 - \lambda)\Phi + \lambda H, H \text{ arbitrary}\}$ we arrive at Huber's (1964) problem with solution $G_1$ where

$$\begin{aligned} (g_1'/g_1)(x) &= -x, && |x| \leq q \\ &= -q \operatorname{sgn} x, && |x| > q. \end{aligned}$$

Substituting into (2.12), we get the Efron-Morris family.

II. *Bounding unbiased estimate of risk* (Berger, 1982). If

$$\psi(x) = x - w(x) \tag{2.13}$$

under mild conditions

$$M(\Delta, w) = 1 + E_\Delta(\psi^2(x) - 2\psi'(x))$$

so that $1 + \psi^2(x) - 2\psi'(x)$ is the UMVU estimate of $M(\eta, w)$. Berger (in a more general context) proposes minimizing $m_0(w)$ subject to $\psi^2(x) - 2\psi'(x) \le q^2$. The solution is easily seen to be $e_q$.

In fact Berger's approach must yield the same results as approach I both in our context and his more general restricted Bayes models. To see this in our model, note that

$$\begin{aligned}
&\inf_w\{(1 - \lambda)m_0(w) + \lambda \sup_x(1 + \psi^2(x) - 2\psi'(x))\} \\
&\quad = 1 + \inf_\psi \sup\left\{\int (\psi^2(x) - 2\psi'(x))G(dx): G \in \mathscr{G}_1\right\} \\
&\quad = 1 - \min\{I(G): G \in \mathscr{G}_1\}
\end{aligned}$$

by a minmax argument.

III. *Bounding unbiased estimate of bias.* Note that $\psi(X)$ is the UMVU estimate of the bias of $w(X)$. Thus it seems reasonable to minimize $m_0(w)$ subject to $\sup_x |\psi(x)| \le q$. This is the exact analogue of Hampel's robustness formulation. The solution is again $e_q$.

For further optimality properties of Efron-Morris estimates, see Bickel (1983).

c) *Performance of Efron-Morris* (E-M) *estimates and competitors.* We measure the relative performance of estimates $\hat{\mu}$ by their relative savings and losses in risk with respect to $\hat{\mu}_1$

$$S(\hat{\mu}) = 1 - m(\hat{\mu})/m(\hat{\mu}_1), \quad L(\hat{\mu}) = M(\hat{\mu})/M(\hat{\mu}_1) - 1.$$

For estimates of the form (2.8),

$$S(\hat{\mu}) = (1 - \rho^2)(1 - m_0(w)), \quad L(\hat{\mu}) = (1 - \rho^2)(M_0(w) - 1).$$

Table 1 gives $1 - m_0(w)$ as a function of $q^2 = M_0(w) - 1$ for the E-M estimates, for $w_q^*$ (calculated by Dr. A. Marazzi) and for some competitors which we now discuss.

*Pretesting estimates.* A type of procedure long advocated by Bancroft and others (see Bancroft and Han, 1977, for a review) are estimates

$$\begin{aligned}
\hat{\mu} &= \hat{\mu}_0, \quad |\hat{\theta}_1 - \hat{\theta}_0| \le c\sigma \\
&= \hat{\mu}_1, \quad \text{otherwise}
\end{aligned}$$

with $c$ chosen to produce an appropriate level for the test of $H$: $\mathscr{M}_0$ vs. $\mathscr{M}_1$ based

TABLE 1
*Gain at 0, $g = 1 - m_0(\omega)$, as a function of the increase in maximum risk $q^2 = M_0(w) - 1$.*

| $q^2$ | $g_e$ | $g_b$ | $g_s$ | $g_j$ | $q$ | $d(q)$ |
|---|---|---|---|---|---|---|
| .1 | .413 | .085 | — | — | .316 | .715 |
| .2 | .538 | .155 | — | .330 | .447 | .903 |
| .3 | .619 | .225 | — | .438 | .548 | 1.053 |
| .4 | .676 | .290 | — | .523 | .632 | 1.175 |
| .5 | .721 | .350 | .711 | .592 | .707 | 1.281 |
| .6 | .758 | .405 | .753 | .648 | .775 | 1.370 |
| .7 | .786 | .455 | .788 | .695 | .837 | 1.461 |
| .8 | .811 | .500 | .816 | .735 | .894 | 1.538 |
| .9 | .832 | .540 | .840 | .768 | .949 | 1.608 |
| 1.0 | .850 | .58 | .859 | .796 | 1.000 | 1.679 |

Note: $g_e$ is the increase for the E-M estimate, $g_b$ for the pretest, $g_s$ for the Sacks family, $g_j$ for Jeffreys' type of generalized Bayes estimate. $q$ and $d(q)$ are the critical values for the E-M and pretest estimates.

on $(|\hat{\theta}_1 - \hat{\theta}_0|)/\sigma$. If $|\hat{\mu}_1 - \hat{\mu}_0| \neq |\hat{\theta}_1 - \hat{\theta}_0|$, this estimate is not of the form (2.8). A version of that form can be based on testing $H$: $E\hat{\Delta} = 0$ vs. $E\hat{\Delta} \neq 0$ and is given by

$$\hat{\mu}_c^B = \hat{\mu}_0 + \sigma_{\hat{\Delta}} b_q\left(\frac{\hat{\Delta}}{\sigma_{\hat{\Delta}}}\right)$$

with

$$\begin{aligned} b_q(x) &= 0, \quad |x| \leq d(q) \\ &= x, \quad |x| > d(q) \end{aligned} \tag{2.14}$$

and $d$ chosen so that

$$M_0(b_q) = 1 + q^2.$$

The $\psi$ function corresponding to $b_q$ via (2.13) corresponds to hard rejection which is known not to work well. This seems true here too. The Bancroft-Han estimate is even worse. (See also Sclove et al. (1972).

Another interesting and desirable feature of the E-M family is monotonicity of $M(\Delta, e_q)$ as a function of $|\Delta|$, i.e. $M_0(e_q)$ is assumed at $|\Delta| = \infty$. This is not true of the pretest estimates and more generally estimates which correspond to redescending $\psi$ functions. Nevertheless we can expect smooth versions of such estimates to perform reasonably well. Motivated by Sacks and Ylvisaker (1978), J. Sacks has proposed a family of such $\psi$,

$$\psi_\gamma(x) = 2(2 + (|x| - \gamma)_+^2)^{-1}x.$$

Another natural family consists of the Jeffreys' type estimates which are generalized Bayes with respect to a prior distribution placing mass $p$ at 0 and corresponding to Lebesgue measure otherwise.

$$\delta_p(x) = x((1/p - 1)\varphi(x) + 1)^{-1}.$$

Table 1 shows very substantial gains in $m_0$ for small payments in $M_0$. Small

biases can be very worthwhile. The pretest estimates are clearly poor and the Jeffreys type estimates are inferior to both the E-M and Sacks estimates.

There is, of course, a serious question as to which E-M estimate to use. The natural way is to calibrate by the maximum $L(\hat{\mu})$ we are willing to tolerate. This of course depends both on $\rho^2$ and $M_0(w)$. For instance, if $n_1 = n_2$ in the pooling example $\rho^2 = \frac{1}{2}$. If we are willing to accept a 10% loss we would take $q = .2$ and obtain a gain of (.5) (.538) = 26.9%.

Another idea is to bound the maximum squared bias of $\hat{\mu}$ standardized by the variance of $\hat{\mu}_1$. For the E-M estimates this equals $L(\hat{\mu})$. The remaining approach of choosing $d$ according to a reasonable level for the test of $H$: $\Delta = 0$ based on $\hat{\Delta}$ yields unreasonably high values of $L(q)$ and is not recommended.

The performance of E-M is markedly better than that of the "Jeffreys" or pretest procedures for small $q^2$. This is in accordance with the asymptotic results of Bickel (1983). Since the Sacks' procedures which are on the whole comparable with E-M cannot be extended over the whole $q^2$ range, we are left with E-M as the candidate of choice.

The best we can do in terms of $m_0(w)$ for given $M_0(w)$ cannot be calculated exactly. However effective numerical procedures have been derived in Marazzi (1980, 1982). Here is a table of the optimal $g$ based on results he has supplied.

| $q$ | .06 | .12 | .19 | .29 | .44 | .70 |
|---|---|---|---|---|---|---|
| $g_0$ | .39 | .49 | .57 | .66 | .74 | .82 |

## 3. Nested linear models: $\mu$ univariate.

*Confidence intervals and other loss functions.* In univariate estimation problems, we usually want confidence intervals as well as point estimates. Since, given our assumed knowledge of $\sigma$, we can form fixed width confidence intervals based on $\hat{\mu}_1$, it seems reasonable to ask how intervals of the same width based on estimates $\hat{\mu}$ perform. This boils down to fixing a width $2z\sigma_1$ and using the loss function

$$\begin{aligned} \ell(\theta, d) &= 1 \text{ if } |d - \mu(\theta)| \geq z\sigma_1 \\ &= 0 \text{ otherwise} \end{aligned} \tag{3.1}$$

$$M(\theta, \hat{\mu}) = P[|\hat{\mu} - \mu(\theta)| \geq z\sigma_1] = 1 - P_\theta[\mu(\theta) \in \hat{\mu} + z\sigma_1]. \tag{3.2}$$

From the argument of Proposition 1 it is easy to see that for any loss function of the form $\ell(|\mu(\theta) - d|)$, equivariant estimates are of the form (2.8). Calculation of the optimal procedures is even more hopeless for this loss function. However, it is easy to see that approximate optimality approach III continues to yield the E-M estimate. More generally

PROPOSITION 2. *Suppose* $\ell(\theta, d) = \ell(|\mu(\theta) - d|)$ *and* $\ell$ *is nondecreasing. Then* $m(\hat{\mu})$ *is minimized among all equivariant* $\hat{\mu}$ *of the form* (2.8) *with* $|\psi(x)| \leq q$ *by an* E-M *estimate*

$$\hat{\mu}_c^e = \hat{\mu}_0 + \sigma_{\hat{\Delta}} e_q(\hat{\Delta}/\sigma_{\hat{\Delta}}). \tag{3.3}$$

PROOF. Without loss of generality, suppose $\sigma_{\hat{\Delta}} = 1$. If $\theta \in \Theta_0$ and $\hat{\mu}$ is given by (2.8)

$$m(\hat{\mu}) = E\ell(|U + w(V)|)$$

where $U$, $V$ are independent normal with mean 0. By Anderson's theorem (Anderson, 1955) $E(\tilde{\ell}(|U + w(V)|) \mid V)$ is monotone increasing in $|w(V)|$. The proposition follows. □

The risk of an E-M estimate (3.3) for a loss function $\ell(|\theta - d|)$ is given by

$$M(\theta, \hat{\mu}_c^e) = \int_{-\infty}^{\infty} \Big\{ \ell(\sigma_0 u - \Delta)[\Phi(d - \tilde{\Delta}) - \Phi(-q - \tilde{\Delta})]$$

$$+ \int_{q-\tilde{\Delta}}^{\infty} \ell(\sigma_0 u + \sigma_1(1 - \rho^2)^{1/2}(w - q))\phi(w)\, dw \tag{3.4}$$

$$+ \int_{-\infty}^{-q-\tilde{\Delta}} \ell(\sigma_0 u + \sigma_1(1 - \rho^2)^{1/2}(w + q))\phi(w)\, dw \Big\} \phi(u)\, du$$

where $\Delta = \mu(\theta) - \mu(\theta_0)$, $\tilde{\Delta} = \Delta/\sigma_1(1 - \rho^2)^{1/2}$. Evidently $M$ depends on $\theta$ through $\Delta$ only, as it must, and moreover,

PROPOSITION 3. *If $\ell$ is as in Proposition 2, then $M$ is a nondecreasing function of $|\Delta|$ for the estimator $\hat{\mu}_c^e$.*

PROOF. It is enough to consider $\ell$ such that $\ell'$ exists and is bounded since we can then obtain the general case by approximation. Differentiate $M$ with respect to $\Delta$ and interchange limits to get

$$\frac{\partial M}{\partial \tilde{\Delta}}(\theta, \hat{\mu}_c^e)$$

$$= \sigma_1(1 - \rho^2)^{1/2}[\Phi(q - \tilde{\Delta}) - \Phi(-q - \tilde{\Delta})] \int_{-\infty}^{\infty} \tilde{\ell}'(\sigma_0 u - \Delta)\phi(u)\, du \geq 0. \quad \square$$

NOTE. This establishes monotonicity of risk for an arbitrary monotone loss function in the original problem considered by Efron and Morris. Thus

$$m(\hat{\mu}_c^e) = \left( \int_{-\infty}^{\infty} \ell(\sigma_0 u)\phi(u)\, du \right)(2\Phi(q) - 1)$$

$$+ 2 \int_{-\infty}^{\infty} \int_{d}^{\infty} \ell(\sigma_0 u + \sigma_1(1 - \rho^2)^{1/2}(v - q))\phi(v)\phi(u)\, du\, dv \tag{3.5}$$

$$M(\hat{\mu}_c^e) = \int_{-\infty}^{\infty} \ell(\sigma_1(u - (1 - \rho^2)^{1/2}q))\phi(u)\, du. \tag{3.6}$$

TABLE 2
Minimum probabilities of coverage of fixed length intervals centered at E-M estimates: z = 1.960.

| $q^2$ | .2 | .4 | .6 | .8 |
|---|---|---|---|---|
| .2 | .982 | .978 | .972 | .962 |
| | .932 | .936 | .941 | .945 |
| .4 | .988 | .985 | .977 | .965 |
| | .912 | .922 | .932 | .941 |
| .6 | .992 | .989 | .980 | .966 |
| | .894 | .908 | .922 | .936 |
| .8 | .994 | .991 | .982 | .968 |
| | .874 | .894 | .913 | .932 |

Note: For each table, the first entry in each box is the minimum probability of coverage on $\mathscr{M}_0$ given by (3.7), the second the minimum on $\mathscr{M}_1$ given by (3.8).

If we specialize to confidence intervals as in (3.1), we obtained minimum probabilities of coverage,

$$1 - m(\hat{\mu}_c^e) = (2\Phi(z/\rho) - 1)(2\Phi(q) - 1) + 2P[-z - (1-\rho^2)^{1/2}q \le A \le z - (1-\rho^2)^{1/2}d,\ B \ge q] \tag{3.7}$$

where $(A, B)$ are bivariate standard normal with correlation $(1-\rho^2)^{1/2}$.

$$1 - M(\hat{\mu}_c^e) = \Phi(z - (1-\rho^2)^{1/2}q) + \Phi(z + (1-\rho^2)^{1/2}q) - 1. \tag{3.7a}$$

We give these probabilities for $z = 1.96$ (corresponding to a 95% confidence level) and selected $q$ in Table 2. The results are similar for the 90% and 99% levels. Again the cost benefit structure seems attractive.

Brown (1980) essentially uses pretest estimate based confidence intervals on a data set to illustrate the dangers of the crossover method. If we treat $\sigma_\xi^2$, $\sigma_\varepsilon^2$ as equal to their estimated values so that $\rho^2 = .48$ for these data and say select $q = .2$ in Table 1 so that $L(\hat{\mu}_q^e) \cong .10$ we obtain significant results for all ($\mathscr{M}_1$) confidence levels tabled and a fortiori all corresponding ($\mathscr{M}_0$) levels, which is consistent with an analysis of the data based on first period results only.

**4. Nested linear models: Quadratic loss in the multivariate case.** Suppose $\dim(\mu) = p$. Then $\hat{\mu}_1 \sim \mathscr{N}_p(\mu(\theta), \Sigma_1)$, $\hat{\mu}_0 \sim \mathscr{N}_p(\mu(\theta_0), \Sigma_0)$ where $\theta_0$ is the projection of $\theta$ on $\Theta_0$. If $\ell(\theta, d)$ is a function of $\mu(\theta) - d$, invariance considerations lead as before to estimates

$$\hat{\mu} = \hat{\mu}_0 + w(\hat{\Delta}) \tag{4.1}$$

where $\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0$ is independent of $\hat{\mu}_0$ with an $\mathscr{N}_p(\Delta, \Sigma_1 - \Sigma_0)$ distribution, $\Delta = \mu(\theta) - \mu(\theta_0)$. Specialize further to,

$$\ell(\mu(\theta) - d) = (\mu(\theta) - d)A(\mu(\theta) - d)^T, \quad A \text{ positive definite.}$$

Then,

$$m(\hat{\mu}) = \operatorname{tr}(A\Sigma_0) + \operatorname{tr}(AE_0(w^T w(\hat{\Delta})))$$

$$M(\hat{\mu}) = \operatorname{tr}(A\Sigma_0) + \sup_\Delta \operatorname{tr}(AE_\Delta((w(\hat{\Delta}) - \Delta)^T(w(\hat{\Delta}) - \Delta)))$$

and in minimizing $m(\hat{\mu})$ subject to a bound on $M$ we need only consider the second terms above. That is, it is enough to consider the special case $r = 0$, $s = p$. Exact solution is impossible. However we can attempt approximations. We can always reduce to the case $A = \| a_i^2 \delta_{ij} \|$ diagonal, $\Sigma_1 - \Sigma_0$ the identity. That is, we observe $X = \Delta + Z$, $Z \sim \mathcal{N}_p(0, I)$, $\Delta = (\Delta_1, \cdots, \Delta_p)$. The risk of an estimate $w = (w_1, \cdots, w_p) = x - \Psi(x)$ is

$$M(\Delta, w) = \sum_{i=1}^p a_i^2 E(w_i(X) - \Delta_i)^2$$

$$= \sum_{i=1}^p a_i^2 + E\left\{\sum_{i=1}^p a_i^2(\psi_i^2(X) - 2\frac{\partial \psi_i}{\partial x_i}(X))\right\}$$

under mild conditions. If $\pi$ is a Bayes prior distribution with Bayes risk $r(\pi)$, Bayes estimate $w_\pi$, and marginal density $g_\pi$ then

$$w_\pi(x) = x + \nabla \log g_\pi(x)$$

$$r(\pi) = \sum_{i=1}^p a_i^2 - I(G_\pi) \tag{4.2}$$

where $\nabla$ is the gradient $((\partial/\partial x_1), \cdots, (\partial/\partial x_p))$

$$I(G) = \sum a_i^2 \int \left(\frac{\partial g}{\partial x_i}(x)\right)^2 g^{-1}(x)\, dx \tag{4.4}$$

(and $= \infty$ if the quantity on the right is undefined). Again the original problem is to minimize $I(G)$ over $\mathcal{G}_0$ and approximation (I) is to minimize over $\mathcal{G}_1$ (with $\Phi$ now the $p$-variate standard normal). By the argument given for one dimension, this yields the same solution as does approximation (II) which minimizes $M(0, w)$ subject to a bound on $[\sum a_i^2(\psi_i^2(x) - 2(\partial\psi_i/\partial x_i)(x))] \leq q^2$, for suitable $q^2$. Unfortunately this approximation is also difficult to compute (but see Chen, 1983), unless all the $a_i^2$ are equal, say to $1/p$. In this case the solution is given for $p = 3$ by Huber (1977) and for general $p$ by Berger (1981), Theorem 3. Here

$$w(x) = 0 \qquad |x| \leq q$$

$$= \rho(|x|^2)x, \quad |x| > q \tag{4.5}$$

with $\rho$ a ratio of Bessel functions with parameters depending on $p$ and scale depending on $q^2$ and $\rho(|q|^2) = 0$. For $p \geq 3$ we can take $q = 0$, i.e., find the minimax estimate in this class which minimizes $M(0, w)$. The answer is the Stein positive part estimate, $q^2 = 2(p - 2)$,

$$\rho(r) = \left(1 - \frac{2(p-2)}{r}\right).$$

As Berger points out, $M(0, w)$ for this estimate drops very sharply from .296 when $p = 3$ to .07 for $p = 5$. Although ths solution is appealing we face the usual ambiguities of the multivariate case. For $p \geq 3$ we could, for instance, also reduce $M(\theta, \hat{\mu})$ for $|\mu(\theta_0)|$ small by applying Steinian shrinking to $\hat{\mu}_0$. Moreover, the effect of the choice of loss function on the suitability of the estimate is difficult to make precise.

For $a_i^2 = 1/p$, it seems reasonable to consider average squared bias and,

$$\text{minimize } E\{\textstyle\sum_{i=1}^p w_i^2(X)\} \text{ subject to } p^{-1} \textstyle\sum_{i=1}^p \psi_i^2 \leq q^2.$$

The solution is as in the one-dimensional case,

$$\begin{aligned} \tilde{w}(x) &= 0, && |x|^2 \leq q^2 \\ &= (1 - (q/|x|)x, && |x|^2 > q^2. \end{aligned} \tag{4.6}$$

If we define $M$ as in the introduction then for fixed $M(w) = 1 + q^2$, estimate (4.5) improves (4.6) at $\Delta = 0$. This follows since the estimates (4.6) also have, if $\tilde{\psi}$ corresponds to $\tilde{w}$,

$$M(\tilde{w}) = 1 + p^{-1} \sup_x \sum \left[\tilde{\psi}_i^2(x) - 2 \frac{\partial \tilde{\psi}_i}{\partial x_i}(x)\right] = 1 + q^2. \tag{4.7}$$

The difference is substantial and despite its attractive feature of computability for more general loss functions, this analogue to Hampel robustness seems unsatisfactory for this application.

**5. Nested parametric models: Asymptotics.** We extend the approaches of Sections 3 and 4 to general nested parametric models by using large sample approximations. Related results are given by Sen (1979) for pretesting estimates. For simplicity we consider estimation of $\mu(\theta)$ where $\mu$ is a smooth real-valued function of $\theta$.

Suppose $\Theta_1$, $\Theta_0$ are as we described previously, respectively an open subset of $R^s$ and a (locally) $r$-dimensional submanifold of $\Theta_1$. Suppose that the models are approximable locally in the sense of Le Cam, to scale $n^{-1/2}$, by nested Gaussian linear models and admit estimates $\hat{\theta}_{0n}$, $\hat{\theta}_{1n}$ (typically M.L.E.'s under $\mathcal{M}_0$, $\mathcal{M}_1$) which are efficient and locally sufficient uniformly on compact subsets of $\Theta_0$, $\Theta_1$ respectively. See Le Cam (1969), Chapters 3, 4 for a detailed description of these concepts and suitable conditions.

Fix $\theta_0 \in \Theta_0$ and reparametrize $\Theta$ by $\theta_0 + an^{-1/2}$ in Pitman form. Locally $\Theta$ permits arbitrary $a$ while $\Theta_0$ specifies $a \in V(\theta_0)$ an $r$-dimensional subspace of $R^s$. Also $\mu(\theta_0 + an^{-1/2}) = \mu(\theta_0) + a\dot{\mu}(\theta_0) + O(n^{-1/2})$ where $\dot{\mu}$ is the differential of $\mu$. Finally, $n^{1/2}\{(\hat{\theta}_{0n} - \theta_0), (\hat{\theta}_{1n} - \theta_0)\}$ is asymptotically normal uniformly on compact sets of $(\theta_0, a)$ with means $(a\Pi(\theta_0), a)$ and covariance matrix $\Sigma(\theta_0)$ where $\Pi(\theta_0)$ is the projection matrix of $V(\theta_0)$.

These approximations suggest that in order to minimize maximum M.S.E. of estimates of $\mu(\theta)$ over large Pitman neighbourhoods of $\theta_0$ in $\Theta_0$, subject to a bound on the maximum M.S.E. over large Pitman neighbourhoods of $\theta_0$ in $\Theta$, we

use asymptotically equivariant estimates as follows. Let

$$\hat{\Delta}_n = \mu(\hat{\theta}_{1n}) - \mu(\hat{\theta}_{0n}), \quad \sigma_\Delta^2(\theta_0) = \dot{\mu}^T(\theta_0)\begin{pmatrix}-1\\1\end{pmatrix}\Sigma(\theta_0)\begin{pmatrix}-1\\1\end{pmatrix}^T\dot{\mu}(\theta_0)$$

denote the asymptotic variance of $n^{1/2}\hat{\Delta}_n$ under $\theta_0 + an^{-1/2}$,

$$\Delta = a(I - \Pi(\theta_0))\dot{\mu}(\theta_0)$$

denote its asymptotic mean, and $\hat{\sigma}_{\Delta,n}$ be a consistent estimate of $\sigma_\Delta$, e.g.

$$\hat{\sigma}_{\Delta_n} = \sigma_\Delta(\hat{\theta}_{1n}).$$

Then, an asymptotically equivariant estimate is one of the form

$$\mu(\hat{\theta}_{0n}) + \hat{\sigma}_{\Delta_n} w(\hat{\Delta}_n/\hat{\sigma}_{\Delta_n}) \tag{5.1}$$

and $n$ times the M.S.E. at $\theta_0 + an^{-1/2}$ of such an estimate is (under mild conditions) approximated by

$$M(\theta_0, a, w) = \sigma_1^2(\theta_0)(\rho^2(\theta_0) + (1 - \rho^2(\theta_0))E(w(Z + \Delta) - \Delta)^2) \tag{5.2}$$

where $\sigma_i^2(\theta_0)$ is the asymptotic variance of $n^{1/2}\mu(\hat{\theta}_{in})$ and,

$$\rho^2(\theta_0) = \sigma_0^2(\theta_0)/\sigma_1^2(\theta_0).$$

From (5.2), given a bound $1/c$ on $\sup_a M(\theta_0, a, w)/\sigma_1^2(\theta_0)$, we minimize $\sup_{a\in V(\theta_0)} M(\theta_0, a, w)$ by taking $w = w_q^*$. As in Section 2, we obtain reasonable results by taking $w = e_q$, with $q$ related to $c$ via (2.6) and $\rho = \rho(\sigma_0)$. The asymptotic sufficiency and efficiency properties of $\hat{\theta}_{in}$, $i = 0, 1$, enable us to formulate asymptotic optimality and near optimality properties of these estimates in the class of all estimates. For simplicity, we omit these.

We give a simple illustration of this approach by applying it to the case of nested linear models with $\Sigma = \sigma^2 I$, $\sigma^2$ unknown, and $\mu$ a linear function of the mean $\theta$. Then our prescription is merely to replace $\sigma_{\hat{\Delta}}^2$ in (2.8) by

$$\hat{\sigma}_{\hat{\Delta}}^2 = \tau^2[\sigma^{-2}(\sigma_1^2 - \sigma_0^2)] \tag{5.2a}$$

where $\tau^2 = \| Y - \hat{\theta}_1 \|^2/(n - 2)$, the usual estimate of $\sigma^2$. The ratio in parentheses in (5.2) depends on the models only. For general $\Sigma$, given a consistent estimate $\hat{\Sigma}$ of $\Sigma$, we can calculate $\hat{\theta}_0$, $\hat{\theta}_1$ by generalized least squares using $\hat{\Sigma}$ and then plug $\hat{\Sigma}$ into $\sigma_{\hat{\Delta}}^2$ appropriately calculated.

As a second illustration, consider pooling two binomial samples. Let $\hat{p}_i = N_i/n_i$, $i = 1, 2$, where $N_i$ is $\text{bin}(n_i, p_i)$, $0 < p_i < 1$, $n_1/n_2 = \lambda$, $0 < \lambda < 1$. We want to estimate $p_1$. $\mathscr{M}_0$ prescribes $p_1 = p_2$. So, if we use $n = n_1 + n_2$ as an index,

$$\hat{\theta}_{1n} = (\hat{p}_1, \hat{p}_2), \quad \hat{\theta}_{0n} = (\hat{p}, \hat{p})$$

where

$$\hat{p} = (N_1 + N_2)/n = (\lambda\hat{p}_1 + \hat{p}_2)/(1 + \lambda).$$

If $\theta = (p, p)$,

$$\sigma_0^2(\theta) = p(1 - p), \quad \sigma_1^2(\theta) = p(1 - p)\frac{(1 + \lambda)}{\lambda}, \quad \rho^2(\theta) = \frac{\lambda}{1 + \lambda}.$$

Then if $\hat{r}_i = 1 - \hat{p}_i$, $i = 1, 2$, putting $w = e_q$ in (5.1),

$$\hat{\mu}_c^e = \hat{p} + \left(\frac{\hat{p}_1\hat{r}_1}{\lambda n}\right)^{1/2} e_q\left(\frac{(\lambda n)^{1/2}(\hat{p}_1 - \hat{p}_2)}{(1 + \lambda)(\hat{p}_1\hat{r}_1)^{1/2}}\right)$$

or

$$\begin{aligned}\hat{\mu}_c^e &= \hat{p} \text{ if } |(\lambda n)^{1/2}(\hat{p}_1 - \hat{p}_2)/(\hat{p}_1\hat{r}_1)^{1/2}(1 + \lambda)| \leq q \\ &= \hat{p}_1 - q \operatorname{sgn}(\hat{p}_1 - \hat{p}_2)(\lambda n)^{-1/2}(\hat{p}_1\hat{r}_1)^{1/2} \text{ otherwise.}\end{aligned} \tag{5.3}$$

This yields, by (5.1), for quadratic loss, a relative loss in risk of

$$\sigma_1^{-2}(\theta) \sup_a M(\theta, a, w) - 1 = q^2/(1 + \lambda) \tag{5.4}$$

while the relative savings in risk are

$$1 - \sigma_1^{-2}(\theta) \sup_{V(\theta)} M(\theta, a, w) = (1 - m_0(e_q))/(1 + \lambda). \tag{5.5}$$

Clearly we can extend this approach to confidence intervals and the $p$-variate case. What we are doing should be clear from the examples. We essentially interpolate between the M.L.E.'s of $\mu(\theta)$ under $\mathcal{M}_0$ and $\mathcal{M}_1$ using weights which are functions of Wald's form of the test statistic for $H$: $\mu(\theta) \in \mu(\Theta_0)$ vs. $K$: $\mu(\theta) \in \mu(\Theta_1)$.

When we consider the limit of ordinary risks $M(\theta, \{\delta_n\})$ we find that procedures (5.1) generally exhibit a discontinuity at points of $\Theta_0$, i.e. convergence of the risk is not uniform. This is reminiscent of Hodges' example of a super efficient estimate which is essentially a pretest estimate corresponding to a sequence of levels tending to 0. However the Hodges procedure has infinite relative loss in risk whereas we propose to pay a small price in the relative loss in exchange for improved behaviour on $\Theta_0$.

## 6. Conclusions: Open questions.

(1) We have applied robustness ideas to derive what we judge are useful biased estimates in the estimation of single parameters under a simple model $\mathcal{M}_0$ when we want to guard against deviations towards a larger model $\mathcal{M}_1$. The solutions involve both an approximation to the optimality principle and in general a large sample approximation. Tables 1 and 2 show that the first approximation is not serious for quadratic loss and the solutions give reasonable confidence intervals. The adequacy of the large sample approximation remains to be assessed in different models by obtaining approximate solutions of the Berger-Bickel type to the exact model, where possible.

(2) In the $p$-variate case, even approximate solutions can only be calculated in special cases and their structure depends on the loss function. It may be appropriate to apply Steinian "pulling in" within the simple model towards a yet simpler model as well as further "pulling in" towards the simple model itself. Alternatively, if we do not believe that losses from errors made in estimation of different components of $\mu$ should be combined it may still make sense to apply pulling in towards $\mathcal{M}_0$ on each component individually.

(3) This approach is applicable, in principle, to large sample problems when $\mathcal{M}_1$ is nonparametric. For example, suppose we want to estimate features of distributions such as medians, means, or even the whole distribution or its density. Our approach suggests reasonable ways of interpolating between estimates based on parametric assumptions and nonparametric estimates.

(4) Typically we have more than one simple candidate model $\mathcal{M}_0$. It would be very interesting to obtain reasonable estimates of $\mu(\theta)$ which do well at each member of a set of simple models while still performing adequately at a super model $\mathcal{M}_1$.

(5) This work is closely connected with the recent studies of Marazzi (1980) and Berger (1982) on robust Bayesian inference. See also the thesis of Y. Ritov (1982) and Masreliez and Martin (1977). Problem (P) is precisely of that form, minimize the Bayes risk for a prior degenerate at $\{0\}$ subject to a bound on the maximum risk—interpreted as the worst that misspecification of the prior can do. On the other hand, if in our original problem we replace the maximum risk over $\mathcal{M}_0$ by an average, we are again in the robust Bayesian framework. We prefer not to try to specify prior distributions. Our point is just that a possibly naive belief in a simpler model can be catered to with reasonable safety.

**Acknowledgement.** I am grateful to B. Efron and P. Huber for helpful conversations and A. Marazzi for the calculation of the lower bounds.

## REFERENCES

ANDERSON, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set. *Proc. Amer. Math. Soc.* **6** 170–176.

BANCROFT, T. A. and HAN, C. P. (1977). Inference based on conditional specification: a note and a bibliography. *Internat. Statist. Rev.* **45** 117–128.

BERGER, J. (1982). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. *Statistical Decision Theory and Related Topics III.* S. S. Gupta and J. Berger, eds. Academic, New York.

BICKEL, P. J. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. *Recent Advances in Statistics, Festschrift for H. Chernoff* 511–528. H. Rizvi and D. Siegmund, eds. Academic, New York.

BROWN, B. W. (1980). The crossover experiment for clinical trials. *Biometrics* **36** 69–80.

CHEN, S. Y. (1983). Restricted risk Bayes estimation for the mean of a multivariate normal distribution. Tech. Report 83-33, Purdue University.

HODGES, J. L. JR. and LEHMANN, E. L. (1952). The use of previous experience in reaching statistical decisions. *Ann. Math. Statist.* **23** 396–407.

HUBER, P. J. (1977). Robust covariances. *Statistical Decision Theory and Related Topics III.* S. S. Gupta and J. Berger, Eds. Academic, New York.

KIEFER, J. (1957). Invariance, minimax sequential estimation and continuous time processes. *Ann. Math. Statist.* **28** 573–601.

LECAM, L. (1969). Théorie asymptotique de la décision statistique. Presses de l'Université de Montréal.

MARAZZI, A. (1980). Robust Bayesian estimation for the linear model. Technical Report No. 27. E.T.H. Zurich.

MARAZZI, A. (1982). On constrained minimization of the Bayes risk for the linear model. Technical Report No. 34. E.T.H. Zurich.

MASRELIEZ, C. J. and MARTIN, R. D. (1977). Robust Bayesian estimation for the linear model and robustifying the Kalman filter. *I.E.E.E. Trans. Automat. Control* **AC-22** June 1977. 361–371.

MORRIS, C., RADHAKRISHNAN, R. and SCLOVE, S. L. (1972). Nonoptimality of preliminary test estimators for the mean of a multivariate normal distribution. *Ann. Math. Statist.* **43** 1481–1490.

MOSTELLER, F. (1948). On pooling data. *J. Amer. Statist. Assoc.* **43** 231–242.

RITOV, Y. (1982). Robust quasi Bayesian inference. Thesis, Hebrew University, Jerusalem.

SACKS, J. and YLVISAKER, D. (1978). Linear estimation for approximately linear models. *Ann. Statist.* **6** 1122–1137.

SCHEFFÉ, H. (1959). *The Analysis of Variance.* Wiley, New York.

SEN, P. K. (1979). Asymptotic properties of maximum likelihood estimators based on conditional specification. *Ann. Statist.* **7** 1019–1033.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720