Some Comments on Linear Programming Approach to Graduation

by

T. C. Wang

5th flr, No.6, Sec.1,Hsin-Hai Rd., Taipei, Taiwan 10089

ABSTRACT

This paper critiques and supplements Professor Schuette's "Linear Programming Approach to Graduation". The main points are:

(1) The graduation is a problem of estimating the mean vector of a multivariate variable under certain assumptions. Schuette and other discussants emphasized the robust-ness property of their proposed estimation. This paper points out their mistake and suggests some correcting directions based on Bayesian theory.

(2) This paper provides a more complete theoretical rationale for the parameter choosing in the graduation problem.

(3) This paper points out the incompleteness of a theorem's proof in Schuette's paper and supplements some interpretations and mathematical details for the proof of some other theorems from a different viewpoint.

1. INTRODUCTION

Let u''_x, x=1, 2, ..., n be a set of observed or ungraduated values, u_x, x=1, 2, …, n be a set of graduated values sought for the corresponding u''_x's.

Schuette (1978) proposed

(a) u_x should be calculated so as to minimize

$$\sum_{x=1}^{n} w_x|u_x-u''_x| + \theta \sum_{x=1}^{n-z} |\Delta^z u_x|$$

instead of the traditional Whittaker (1944) method of minimizing

$$\sum_{x=1}^{n} w_x(u_x-u''_x)^2 + \theta \sum_{x=1}^{n-z} (\Delta^z u_x)^2 .$$

(b) The minimization problem in (a) can be reformulated as a linear programming (LP) problem:

$$\text{minimize} \quad \sum_{x=1}^{n} w_x(P_x + N_x) + \theta \sum_{x=1}^{n-z} (R_x + T_x)$$

$$\text{s.t.} \quad \Delta^z(P_x - N_x) + R_x - T_x = \Delta^z u''_x \ ,$$

$$x = 1, \ldots, n-z$$

$$P_x \geqslant 0,\ N_x \geqslant 0,\ x = 1,\ldots,n,\ R_x \geqslant 0,\ T_x \geqslant 0,$$

$$x = 1, \ldots, n-z \ .$$

(c) There exist θ_L and θ_U such that a graduator should try various values of θ between these two values to solve the minimization problem until he feels the resulting graduated values satisfactory.

This paper will (I) discuss the statistical foundation problems surrounding (a) and (c), which were either vaguely or incorrectly stated in Schuette paper and its discussions; (II) point out an error about θ_L and supplement some explanations which were either omitted or chosen not to be presented surrounding (b) and θ_U in Schuette's paper; and (III) state some concluding remarks and suggest possible problems for further research.

Section 2, 3, 4 of this paper serve the purpose of the foregoing (I), (II) and (III) respectively.

The reader is assumed to have a copy of Schuette (1978) paper and discussions at hand.

2. STATISTICAL FOUNDATION PROBLEMS

2-1. Graduation Formulated as a Bayesian Estimation problem.

Let u denote the vector of u_x's and u" denote the vector of u_x's. Whittaker (1944) essentially assumed u''_x iid $N(u_x, w_x^{-2})$ with density

$$f(u''|\ u) \propto \exp(-\sum_{x=1}^{n} w_x(u''_x - u_x)^2),$$

u has a prior $f(u|\ \theta) \propto \theta \exp(-\ \theta \sum_{x=1}^{n-z} (\Delta^z u_x)^2)$.

The posterior density of u is then

$$f(u|\ u'',\ \theta) \propto \exp\{-\sum_{x=1}^{n} w_x(u''_x-u_x)^2 - \theta \sum_{x=1}^{n-z} (\Delta^z u_x)^2\}\ .$$

Under the utility function

$$U(u,\ d) = \begin{cases} 1 & \text{if} \quad d = u \\ 0 & \text{if} \quad d \neq u \end{cases} \qquad d = \text{ decision about } \ u\ ,$$

the Bayesian estimator, or the graduated values should be the mode of the posterior distribution of u, which is equivalent to the u that minimizes

$$\sum_{x=1}^{n} w_x(u''_x - u_x)^2 + \theta \sum_{x=1}^{n-z} (\Delta^z u_x)^2\ .$$

Noting that the prior density of u is singular, Hickman (1978) commented:"…while singularity is not a disastrous quality for a prior…, the justification for its use in this case is not immediately apparent."

Kimeldorf and Jones (1967) offered a revision to the foregoing model: Likelihood function and loss (utility) function remained the same, but the prior density is modified to

$$f(u|\ m,\ A) \propto \exp\{-\frac{1}{2}[(u-m)A^{-1}(u-m)']\}\ .$$

If A is assumed to be $(\rho^2 r^{|i-j|})_{n \times n}$ then

$$Std(u_x) \ = \rho$$

$$Std(u_x|\ u_{x+1}) = \rho\sqrt{1-r^2}\ .$$

When the graduator's prior opinion on the level of graduated values is vague, meaning $\rho \rightarrow \infty$, but his prior opinion on the shape of graduated values is strong, meaning $r \rightarrow 1$, Kimeldorf and Jones model will reduce to Whittaker's model under the assumption that $m_x = m$ for all x and both models use the same w_x. Hickman (1977) and Hickman (1979) provided some further refinements to the Kimdeldorf and Jones model.

2-2. The Problem of Robustness.

Schuette motivated his proposal (see section 1 of this paper) with preference to the "robustness" of absolute deviation vs. least square minimization procedures for solving regression problems in statistics. This section intends to clarify the "robustness" issue within the context of Bayesian Estimation as described in section 2-1 of this paper.

Realizing that Schuette and discussants of this paper were making their points with classical robust estimation in their mind, the following discussion will (1) provide a background of key concepts in the recently developed literature concerning classical robust estimation; (2) extract the main ideas from the literature just beginning to appear concerning Bayesian robust estimation; (3) pinpoint the vagueness and incorrectness of Schuette and the discussants; and will in Section 2-3 show that based on (2): models described in 2-1 can not be considered robust, Schuette’s model as described in 1 possesses certain part but not all of the robust properties, and a model suggested by Ramsay (1980) can be adapted to a robust version of the graduation models in 2-1.

2-2-(1). Key Concepts in Classical Robust Estimation.

The key concepts in classical robustness are itemized as (C1), (C2), (C3), (C3'), (C4), (C5), (C6), (C7) and described in Appendix I.

2-2-(2): Ideas on Bayesian Robust Estimation.

The main ideas in Bayesian robustness analogous to the classical ones are itemized as (B1), (B1'), (B5), (B6), (B6'), (B7), (B8) and described in Appendix II.

2-2-(3). Criticism of Schuette and Discussants.

i. Hickman touched the issue of robustness by merely pointing out that robustness considerations existed as early as Edgeworth (1888). From the foregoing descriptions in 2-2-(1) and (2), we can see modern research on robustness is much more concrete and systematic than a mere ad hoc consideration.

ii Klugman (1978) suggested an estimate developed by Huber (1964), referred as H_1 in the literature, be used. This estimation procedure would replace the

$|u_x - x''_x|$ in 1-(a) by $(u_x - u''_x)^2/\mathrm{Var}(u''_x)$ if

$|u_x - u''_x| \leq c$, by $2c \left(\frac{u_x - u''_x}{\mathrm{Std}(u''_x)} \right) - c^2$ otherwise, c

is some value between 1 and 2, $w_x = 1/\mathrm{Var}(u''_x)$.

We have to note that:

(a) H_1 was developed to minimize the maximal asymptotic variance over a restricted neighborhood of the standard normal. as described in (C3). It may not be optimal in the sense of (B3).

(b) There are other robust estimators performing better than H_1 with respect to (C1). (C2), (C3'), (C4), …, (C7) and nearly as good as H_1 with respect to (C3). Hampel (1974) compared various robust estimators for a standard normal mean.

(c) Huber (1977) pp.36-37 showed that for any robust estimators to work for regression problems, the ratio of the number of parameters to the number of observations must vanish rapidly as the latter goes to infinity. In our case of graduation, this ratio is always 1.

(d) Klugman (1979) used H_1 to estimate the crude mortality rate. It is a completely different problem from a graduation or smoothing problem that we are discussing about.

iii. Klugman (1978) seemed to indicate an inconsistency between robustness and smoothing data with some large true underlying $\Delta^z u_x$ values. But (B5), (B6) and (B7) are exactly to deal with "some small residual or hedging probability for parameter values considerably removed from those considered most probable" (Ramsay (1980), p.902).

iv. Greville (1978) stated: "He (Schuette) has given cogent reasons why the criterion of least absolute values should be considered as an alternative to least squares ...". Actually, Schuette merely vaguely mentioned this criterion in the regression problem and had not put the robustness consideration within the proper framework as described in II-1. We will show later in 2-3-(1) that the Whittaker's least square criterion (see 2-1) is indeed not robust in the proper context of Bayesian estimation.

2-3. Robust Graduation as Robust Bayesian Estimation.

2-3-(1). Whittaker's model and Kimdeldorf & Jones' model as described in 2-1 can not be considered robust in the sense that they do not satisfy (B5), (B5'), (B7), (B7') and (B8);

they satisfy (Bl').

2-3-(2). Schuette version (l-(a)) satisfies (Bl'), (B5), (B5') but not (B6), (B6'), (B7), (B7') and (B8).

2-3-(3). Ramsay (1980) proposed a robust version satisfying all of (B1'), (B5)-(B7') for the regression problem:

$$y_{n \times 1} = x_{n \times p} \beta_{p \times 1} + e_{n \times 1}$$

$$m(y|\beta, \sigma) = N(x\beta, I\sigma^2), \quad p(\beta|\sigma) = N(\beta_0, \Sigma\sigma^2)$$

$p(\sigma)$ =inversed chi-square.

It can be easily adapted as a robust version for Kimeldorf and Jones graduation model (2-1) by making n=p, X=I, $y_x = u''_x$, $\beta_x = u_x$, $\beta_{0_x} = m_x$, $\Sigma\sigma^2 = A$. Due to the lack of mathematical rigor, this proposal is meant to be a tentative one.

2-4. The choice of θ.

As indicated by Hickman (1978), θ as a parameter defining the prior density as described in 2-1 should be fixed and known a priori by the estimator. If the procedure in l-(c) means θ is unknown or vaguely known, then θ itself should have a density h(θ), the posterior density of u then is

$$\int f(u''|u)p(u|\theta)h(\theta)d\theta \quad .$$

This is also the underlying principle used by Lindley (1981) for his "Bayes Empirical Bayes" approach.

However, if the purpose of trying different θ is to test the robustness of the procedure near the neighborhood of assumed fixed θ, it is comparable to the sensitivity analysis in many decision models. As Kadane (1978) argued:" A well-known principle of personalistic Bayesian theory is that no one can tell someone else what loss function to have or what opinion to hold. Having said that, the reasons for looking into properties of particular choices of loss functions and opinions might be obscure. The standard of personalistic Bayesian theory may be too severe for many of us. Generally when a personalistic Bayesian tells you his loss function and opinion, he means them only approximately. He hopes that his loss function and opinion, he means them only approximately. He hopes that his approximation is good, and that whatever errors he may have made will not lead to decisions with loss substantially greater than he would have obtained had he been able to write down his true loss function and opinion."

Ramsay (1980) argued for thick tailed prior:"… even if the analyst's own beliefs are properly represented as short-tailed, dialogue with other investigators who have significant prior probability for values for which the analyst's own prior density is nearly zero may be desirable. … Thus in a spirit of compromise and to avoid accusations of too much subjective bias, the analyst may be led to choose thicker tails. Berger (1981) and DeGroot (1974) provided an operational procedure for a group or team of analysts to reach consensus on the prior distribution of parameters. These can be used to argue for doing sensitivity analysis on θ which is supposed to be fixed and known a priori. However, Klugman (1980)'s approach is questionable.

3. PROBLEMS RELATED TO LINEAR PROGRAMMING THEORY

Theorem numbers, relationship statement numbers and notations in Schuette's paper will be referred without explanation.

3-1. $\theta \leqslant \theta_L$ is a necessary but has not been proved to be sufficient condition for the graduated values to be the same as the observed values.

The statement of Theorem 1 that $\theta \leqslant \theta_L$ is sufficient is incorrect. The following diagram indicates that necessity but not sufficiency has been proved in the proof of Theorem 1: (see Figure 1)

$u_x = u_x''$ for all x implies

(17),(18),(19),(20).Also (17),(18) implies (19),(20).

The fact that (19)=(21),

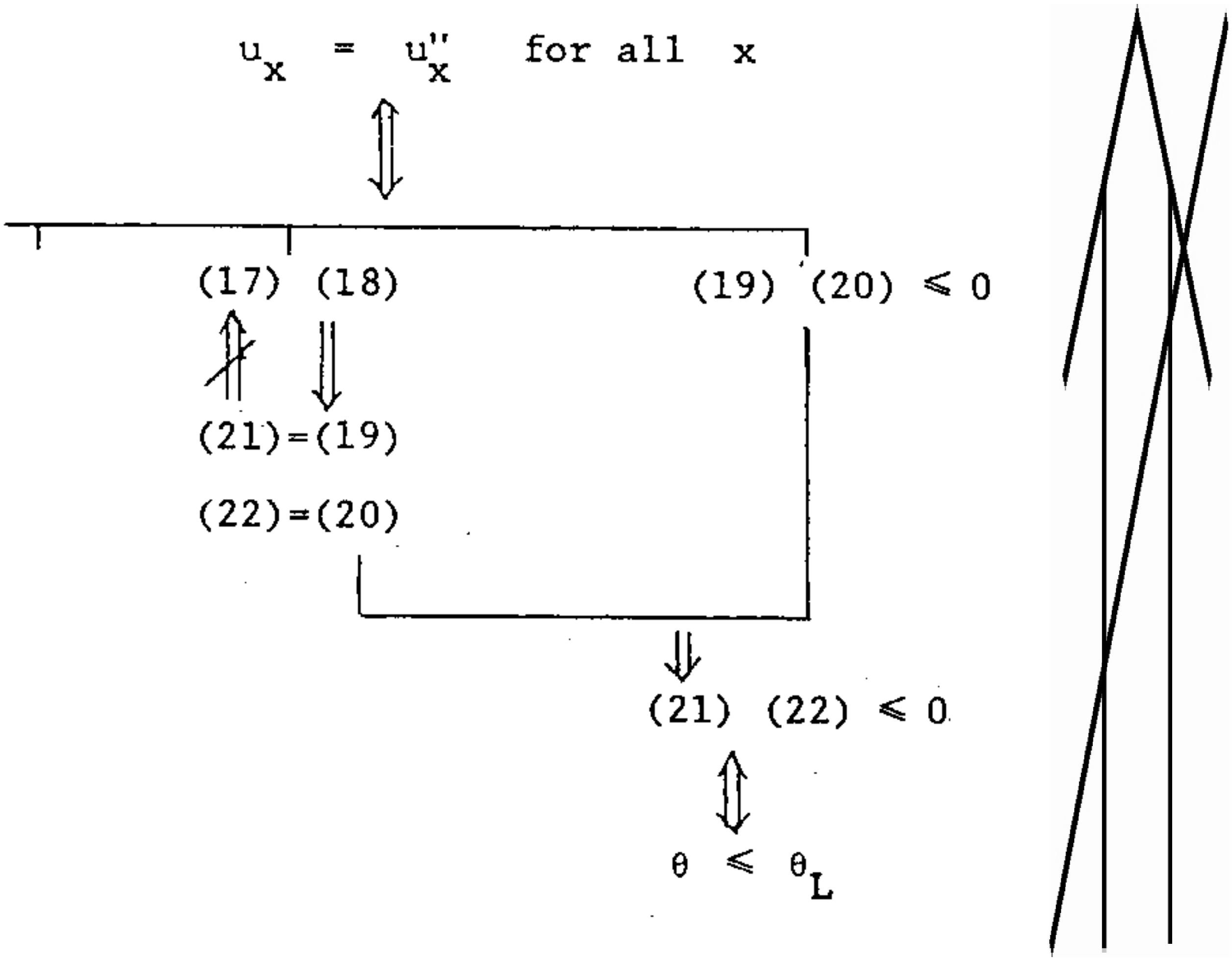
(20)=(22) makes $u_x = u_x''$ for all x

imply (21),(22), which implies θ≤θ$_L$ .Therefore $u_x = u_x''$ for all x

implies θ≤θ$_{L.}$ The foregoing has been proved.

$\theta \le \theta_L$ implies (21), (22), but (21)=(19), (22)=(20) has not been proved to imply (17), (18) and hence not been proved to imply (17), (18), (19), (20), and hence not been proved to imply $u_x = u_x''$ for all x. Therefore $\theta \le \theta_L$ has not been proved to imply $u_x = u_x''$ for all x. (The right $\not\Uparrow$)

The “not been proved” means there is no proof rather than the proof is invalid.



(Figure 1)

Both the primal and dual problem are assumed to have no degeneracy.

3-2. Interpretation about Theorem 2 and "linear programming graduation strategy" in terms of duality and optimal conditions: Schuette noted: " ... some of the later sections of this (Schuette's) paper could have been presented in terms of the dual problem and its variables, but the author has chosen not to do so." The following will supplement what he chose not to do.

Step 1. For a sufficiently large θ^*. presumably (not proved) we can find an optimal solution such that $\sum_{x=1}^{n-z} |\Delta^z u_x| = 0$, i.e., $R_j = T_j = 0$ for all $j = 1, \ldots,$ n-z. Section 3-3 will prove the meaning of this solution.

Step 2. Find the shadow prices $y_1, y_2, \ldots, y_{n-z}$ of the optimal solution found in step 1.

Step 3. Define $\theta_U = \max\{y_1, \ldots, y_{n-z}\}$ found in Step 2.

Step 4. For any $\theta > \theta_U$ the optimal solution found in Step 1., i.e. the u_x that minimizes

$$\sum_{x=1}^{n} w_x |u''_x - u_x| + \theta^* \sum_{x=1}^{n-z} |\Delta^z u_x|$$

will also be minimizing

$$\sum_{x=1}^{n} w_x |u''_x - u_x| + \theta \sum_{x=1}^{n-z} |\Delta^z u_x|$$

i.e. be the optimal solution of using θ in the objective function rather than the θ^* used in Step 1.

Proof of Step 4:

We will prove that the solution found in Step 1 will be an optimal solution for the dual problem & the primal problem. The primal problem is (1-(b)), $\theta > \theta_U$.

Primal Feasibility:

Since $R_j = T_j = 0, \ \forall j = 1, \ldots, n-z$, the primal constraints are satisfied as the primal constants in step 1. θ appears in the constraints; its value does not matter.

Dual Feasibility:

(i) Since $\theta > \theta_U$, the constraints dual to the primal variables R_j, T_j, $j = 1, \ldots, n\text{-}z$, are satisfied. (cf. last four 1ines of the proof of theorem 2 in Schuette's)

(ii) Constraints dual to the primal variables P_j, N_j have been satisfied as the dual constraints for optimal P_j, N_j in the primal problem of step 1, in which θ does not appear at all. $j = 1, \ldots, n$.

Complementary Slackness:

(i) Complementary Slackness for the variable y_j has been satisfied because the jth constraint in the primal problem is equality for all $j = 1, 2, \ldots, n\text{-}z$ . . θ appears in the constraints, but its value does not matter.

(ii) Complementary slackness for the variable P_j, N_j is satisfied as the complementary slackness for the complementary slackness for the Step 1 primal problem, in which θ does not appear. $j = 1, 2, \ldots, n.$

(iii) Complementary slackness for R_j, T_j is satisfied because $R_j = T_j = 0$. θ appears in the slackness but its value does not matter. $j = 1, 2, \ldots, n\text{-}z$.

The foregoing arguments are based on Bradley et al. (1980).

3-3. The meaning of the optimal solution found in step 1.3-2.

Meaning: the solution represents a polynomial of degree lower than z-l that

(i) minimizes $\sum_{x=1}^{n} w_x |u''_x - u_x|$,

(ii) $u_x = u''_x$ for at least z_x's ,

(iii) $u_x = q(x) = c_0 + c_1 x + \ldots + c_{z-1} x^{z-1}$,
$x = 1, 2, \ldots, n$

where $q(y) = \sum_{i=0}^{z-1} c_i y^i$ is the polynomial.

Proof:

(i) Since the LP problem is equivalent to and is a reformulated version of the problem:

$$\min \sum_{x=1}^{n} w_x |u''_x - u_x| + \theta \sum_{x=1}^{n-z} |\Delta^z u_x|$$

now the solution reduces the second part of the second part of the sum to zero, (i) follows.

(ii) Since the rank of constraint matrix ((9) of Schuette's paper) is n-z, there can be at most n-z positive P_j's or N_j's, $i = 1, 2, \ldots, n$, hence there must be at least z zero $P_j = N_j$'s. (ii) follows.

(iii) The solution satisfies

$$\sum_{x=1}^{n-z} |\Delta^z u_x| = 0$$

and hence satisfies the linear difference equations homogeneous of order z:

$$\Delta^z u_x = 0, \quad x = 1, \ldots, n-z .$$

It will be shown below that solutions to this linear difference equation has the general form q(x).

Lemma 1. If $(LU)_n = u_n + a_1 u_{n-1} + \ldots + a_z u_{n-z}$

$$p(v) = v^n + a_1 v^{n-1} + \ldots + a_z v^{n-z}$$

and if w is such that $p(w) = p'(w) = \ldots = p^{(k)}(w) = 0$, then

$$U^{(1)} = \{u_x = xw^{x-1}, \ x = n-z, \ldots, n\}$$

is a solution to $(LU)_n = 0$.

$$U^{(m)} = \{u_x = x(x-1) \ldots (x-k+1)w^{x-m}, \ x = n-z, \ldots, n\}$$

is a solution to $(LU)_n = 0$, $m = 0,1,2,\ldots,k$.

Proof: $(LU^{(m)})_n = (w^{n-z}p(w))^{(m)} = 0. \ \because \ p^{(m)}(w) = 0$, m =0,1,2, ..., k.

Lemma 2. $U^{(m)} = \{u_x = u^m w^x, \ x = n-z, \ldots, n\}$ is a solution to $(LU)_n = 0$ as defined in

Lemma 1.

Proof: x^m can be expressed as a linear combination of x, x(x-1), …, x(x-1) … (x-m+1).

Lemma 3. $\Delta^z u_{n-z} = u_n - \binom{z}{1} u_{n-1} + \dots + (-1)^z \binom{z}{z} u_{n-z} =$ 0 solutions of general from ,

$u_x = \sum_{m=0}^{z-1} c_m x^m = q(x)$,x=n-z, …, n.

Proof: Treated as a special case of Lemma 2, p(v)=$(v-1)^z$, $w = 1$, $k = z-1$, $U^{(m)} = \{u_x = x^m$, x=n-z, …, n}, m=0,1, …, z-1 is a set of solutions for $\Delta^z u_{n-z} = 0$. They are linearly independent because the Wronskian.

$$\begin{vmatrix} 1 & n & n^2 & \dots & n^{z-1} \\ 1 & n-1 & (n-1)^2 & \dots & (n-1)^{z-1} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & n-z+1 & (n-z+1)^2 & \dots & (n-z+1)^{z-1} \end{vmatrix} \neq 0 .$$

Therefore any solution can be expressed as a linear combination of these z solutions, the lemma follows.

Our desired result (iii) follows by applying Lemma 3 to each $\Delta^z u_x = 0,\ x = 1, \dots, n-z$ and finding out that their general solutions are consistent. Lemma 1 to 3 are adapted from Henrici (1964).

4. CONCLUSION

McGill (1982) discussed construction of 1980 CSO Mortality tables and the effect of graduated mortality rates on various life insurance business and public policy decisions. Cummins (1981) mentioned estimation of mortality for group life insurance. Therefore it is important to use graduation method with solid statistical ground and being robust to the gross errors in the crude data. Cohen and Fisher (1982) emphasized the importance of and insights given by duality and optimal conditions. This paper reaffirmed their emphasis.

This paper pinpointed some vagueness and incorrectness of an important part of actuarial/insurance/statistical theories literature so that future readers would not be misled. Further clarifications based on future theories possible.

Appendix I. The Key Concepts in Classical Robust Estimation

According to Huber (1977), "robustness signifies insensitivity against small deviations from the assumptions." Box and Tiao (1973) gave some philosophical interpretation.

Tukey (1960) gave an illuminating example:

If the assumption that X_i iid $N(\mu, \sigma^2)$ is slightly deviated to a mixture of $(1-\varepsilon)N(\mu, \sigma^2)$ and $\varepsilon\ N(\mu, q\sigma^2)$ either due to $100\ \varepsilon\ \%$ gross errors in the data or due to the longer tail of the true underlying distribution, then the asymptotic relative efficiency of estimator for $s_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ to $d_n = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$ will change from .876 $\varepsilon = 0$ when to 2.035 when $\varepsilon = 0.5$ .Therefore s_n is very sensitive to small deviations from the assumptions and is not a robust estimator.

The following mathematical descriptions are synthesized from Huber (1972, 1977, 1981) and Hampel (1971, 1974). A Monte Carlo experiment on these theories was reported in Andrew et al.(1972). Hogg (1979) provided a recipe for some classical robust estimation procedures.

Let $(\Omega, \mathcal{A})$ be a measurable space such that Ω is a complete separable metric space with metric d and $\mathcal{A}$ denotes the σ-algebra generated by the topology. Let $\mathcal{F}$ denote the set of all probability measures. defined on $(\Omega, \mathcal{A})$. For any $A \subset \Omega$, let A^{ε} denote the set . $\{w \in \Omega \mid \exists a \in A \text{ such that } d(w, a) \leqslant \varepsilon\}$. The Prohorov distance between F and $G \in \mathcal{F}$, denoted by $\pi(\mathcal{A}, G)$,is defined to be $\inf\{\varepsilon \mid \text{ for all } A \in \mathcal{A},\ F(A) \leqslant G(A^{\varepsilon}) + \varepsilon\ \&\ G(A) \leqslant F(A^{\varepsilon}) + \varepsilon\}$ A ε-neighborhood ·of $F \in \mathcal{F}$ denoted by $\mathcal{P}_{\varepsilon}(F)$, is defined to be

$$\{G \in \mathcal{F} \mid \pi(F, G) < \varepsilon\}.$$

We are primarily interested in $\Omega = R$, the real line and we will use the notation F for both a distribution function and the probability measure induced by it.

$\mathcal{F}_n \subset \mathcal{F}$ is defined to $\{F_n \mid F_n$ is the empirical distribution of a sample $X_1, \ldots, X_n$: iid $F\}$.

A sequence of estimators $\{T_n\}$ is defined to be a sequence of measurable mappings $T_n : \mathcal{F}_n \to R^k$, e.g. sample mean .= $\int_0^1 F_n^{-1}(t)\,dt$ The distribution of $T_n(F_n)$ is denoted by $\mathcal{L}_F(T_n)$.

(C1) $\{T_n\}$ is said to be qualitatively robust at F_0 iff $\forall \varepsilon > 0, \exists \delta$
$F \in \mathcal{P}_{\delta}(F_0) \Rightarrow \mathcal{L}_F(T_n) \in \mathcal{P}_{\varepsilon}(\mathcal{L}_{F_0}(T_n))$.

Intuitively, (Cl) means that a small deviation of the true F from the assumed F_0 will have only a small effect on the functional value of T_n.

In fact, a mapping $T: \mathcal{F} \to R^k$ is continuous everywhere if and only if,

(i) $T_n(F_n) \equiv T(F_n)$ is consistent for T(F), i.e.

$T_n(F_n) \to T(F)$ in probability $\forall F$.

& (ii) $\{T_n\}$ is qualitatively robust at all .$F \in \mathcal{F}$

Thus, for a continuous T, we are further interested in the bias of T_n under F against $T(F_0)$ where F_0 is the assumed distribution: $M(F, T_n)$= median of $\mathcal{L}_F(T_n - T(F_0))$

And "variance" of T_n under F:

$$Q_t^2(F, T_n), \; Q_t(F, T_n) = \text{normalized t-quantile range of } \mathcal{L}_F(\sqrt{n}\, T_n)$$

$$\equiv \frac{\mathcal{L}_F^{-1}(\sqrt{n}\, T_n).(1-t) - \mathcal{L}_F^{-1}(\sqrt{n}\, T_n)(t)}{\Phi^{-1}(1-t) - \Phi^{-1}(t)}, \; \Phi = \text{standard normal.}$$

The asymptotic maximal bias and variance are defined respectively as:

$$\text{(C2)} \quad b(\varepsilon) = \lim_{n\to\infty} \sup_{F \in \mathcal{P}_\varepsilon(F_0)} |M(F, T_n)|$$

$$\text{(C3)} \quad v(\varepsilon) = \lim_{n\to\infty} \sup_{F \in \mathcal{P}_\varepsilon(F_0)} Q_t^2(F, T_n) \; .$$

The maximal asymptotic bias and variance are defined respectively as:

$$b_1(\varepsilon) = \sup_{\mathcal{P}_\varepsilon(F_0)} T(F) - T(F_0)$$

$$= \sup_{F \in \mathcal{P}_\varepsilon(F_0)} \lim_{n\to\infty} |T_n(F_0) - T(F_0)|$$

$$v_1(\varepsilon) = \sup_{\mathcal{P}_\varepsilon(F_0)} V(T, F) = \sup_{\mathcal{P}_\varepsilon(F_0)} \lim_{n\to\infty} \mathrm{Var}(\mathcal{L}_F(\sqrt{n}\, T_n - T(F_0)))$$

In general, we have

$$b(\varepsilon) \geqslant b_1(\varepsilon), \ b(\varepsilon) \geqslant v_1(\varepsilon) .$$

Usually, the equality holds under "regular" conditions. Other than (Cl), robustness requires small (C2) and (C3). Moreover,

(C3') $V(T, F_0)$ defined above should not be too far from $\inf_{\tilde{T}} V(\tilde{T}, F_0)$.

(C4) Breakdown point of T at F_0

$$\varepsilon^* = \sup\{\varepsilon | \ b(\varepsilon) < b(1)\}$$

should be high. Intuitively, ε^* gives the maximum fraction of gross errors that the estimation still make some sense. For example, the α-trimmed mean

$$= \frac{1}{1-2\alpha} \int_{\alpha}^{1-\alpha} F^{-1}(t)dt$$

has a breakdown point α at standard normal.

The influence curve of T at F_0 is defined as

$$IC(x, T_0, T) = \lim_{\varepsilon \to 0} \frac{\{T((1-\varepsilon)F_0 + \varepsilon\delta_x) - T(F_0)\}}{\varepsilon}$$

where δ_x is distribution with mass 1 at x.

Intuitively, this measures the effect of one additional observation x in a very large sample under distributions F_0.

(C5) Gross error sensitivity of T at F_0

$$r^* = \sup_x |IC(x; \ T, \ F_0)|$$

is to be small for a robust T at F_0 Under certain conditions,

$$b(\varepsilon) = b_1(\varepsilon) = \varepsilon * r^* + 0(\varepsilon) \ .$$

(C6) Local shift sensitivity of T at F_0

$$\lambda^* = \sup_{x \neq y} \frac{IC(x;\ T,\ F_0) - IC(y;\ T,\ F_0)}{x - y}$$

is to be small. It measures the effect of wriggling (rounding or grouping) data. Wriggling can be thought as taking out x and throwing a nearby y in a sample.

(C7) Rejection point $\rho^* = \inf\{x |\ IC(x,\ T,\ F_0) = 0\}$ of T at F_0 should be set. Beyond ρ^*, an observation will implicitly rejected as an outlier. Note $V(T,\ F) = \int IC(x,\ F)^2 dF$.

Appendix II. The Main Ideas in Bayesian Robust Estimation.

Let $m(y|\theta)$ denote the density of observations $y = (y_1,\ \ldots,\ y_n)$ cond. on $\theta = (\theta_1,\ \ldots,\ \theta_p)$ parameters. $\ell(\theta,\ y) = m(y|\theta)$ viewed as a function of θ. Both ℓ and m are referred as likelihood function of observations. Let p(θ) denote the prior density of θ, $U(t,\ \theta)$ denote the utility of estimator t about θ.

The posterior density of θ given y is then

$$f(\theta|y) \propto \ell(\theta,\ y)p(\theta) \ .$$

Bayes estimator is to minimize

$$U_{p,\ell}(t|y) = U_f(t|y) = \int u(\theta,\ t)f(\theta|y)d\theta$$

over all possible t.

Thus the couple $(U,\ f)$ or the triple $(U,\ p,\ \ell)$ uniquely defines a t for any given y.

Kadane (1978), refining the ideas of Edwards, Lindeman and Savage (1963), Fishburn et al. (1967), Pierce and Folks (1970) and Dempster (1975), defined

(B1) (U, f) to be stable if and only if for any y for every $\{f_n\} \to f$ in distribution and every $\{U_n\} \to U$ uniformly in t and θ, if $U_f(t(\varepsilon)|y) \leq \inf_t U_f(t|y) + \varepsilon$ then

$$\lim_{\varepsilon \to 0} \overline{\lim_{n \to \infty}}\ U_{n_{f_n}}(t(\varepsilon)|y) - \inf_t u_{n_{f_n}}(t|y) = 0 \ .$$

Intuitively, (B1) means a small change in U and f will have a small effect on the average utility of the Bayes estimators.

Kadane further showed that if $U(\theta, t)$ is bounded and continuous in t uniformly in θ, then (U, f) stable with any f.

$$\text{(B1')} \quad U^*(\theta, t) = \begin{cases} 1 & \text{if} \quad t = \theta \\ 0 & \text{if} \quad t \neq \theta \end{cases}$$

is bounded and continuous in t uniformly in U except at t= θ.

The Bayes estimate defined by (U^*, f) is the mode of f and its Bayes utility is sup $f(\theta|y)$ for any given y.

(B1) is comparable with (C1).

The concept "bias" does not seem to apply in Bayesian estimation. In fact, Britney and Winkler (1974) showed cases that true value of a parameter was far from optimal if the loss function (utility function) was asymmetric.

(B3) Concept comparable to (C3) seems to require the minimum of $U(t|y)$ over a neighborhood of assumed (U, p, λ) be as large as possible for a robust t, Section 10.11 of DeGroot (1970) showed that under certain conditions, for r_i iid any posterior density is asymptotically proportional to

$$\exp - \frac{n}{2}\{(\theta - \hat{\theta}_n)I(\hat{\theta}_n)(\theta - \hat{\theta}_n)'\}$$

as $n \to \infty$ where $\hat{\theta}_n$ is the maximum likelihood estimator and $I_{ij}(\theta) = - \int (\frac{\partial \ln \ell(\theta|x)}{\partial \theta_i \theta_j}) dx$, the "Fisher information matrix" $I(\theta) = (I_{ij}(\theta))_{p \times p}$' .

(B3') Concept comparable to (C3') would require that an estimate determined by the robust version of (U, p, ℓ) wouldn't be too far from optimal under the assumed version.

(B3) and (B3') have not been used as a starting point to derive Bayesian robust estimators in the literature, although they are examined numerically after some estimators have been derived. Concepts comparable to (C5)-(C7) are developed in Ramsay (1980) but much less rigorously. y_i assumed iid.

$$\text{(B5)} \quad \{- \frac{\partial \ln p(\theta)}{\partial \theta_j}, j = 1, \ldots, p\} = P \text{ be bounded.}$$

$$\text{(B5')} \quad \{- \frac{\partial \ln \ell(\theta, y_i)}{\partial \theta_j} - \frac{\partial \ln r_\theta(y_i|\theta)}{\partial y_i}\} = L \text{ be bounded}$$

$$i = 1, \ldots, n, \quad j = 1, \ldots, p$$

(B6) Functions in P

(B6') Functions in L .

(B7) Functions in P $\to 0$ as $\theta_j \to \infty$.

(B7') Function in L $\to 0$ as $\theta_j \to \infty$, $y_i \to \infty$.

Heuristically, they mean "thick tail" densities, or probability vanishes very slowly over values with probabilities (extreme values) and eventually remain constant. They are supposed less sensitive to one additional extreme value. Berger (1984) and Berger (1980) interprets Rubin (1977) that one way to develop robust prior is to have

(B8) $p(\theta)$ of much flatter tail than $\ell(\theta, y)$ (e.g. $p \propto e^{-c|\theta-\mu|}$, $\ell \propto e^{-c(y-\theta)^2}$) .

REFERENCES

Andrew, D. F. , Bickel, P.J.,Hampel, F.R., Huber,P.J., Rogers,W.H. and Tukey,J.W. (1972), Robust Estimates of Location: Survey and Advances. Princeton University Press, Princeton, NJ.

Berger, J. (1980), Statistical Decision Theory: Foundations, Concepts and Methods. Springer, NY, p.141.

Berger, J. (1984), The Robust Bayesian Viewpoint. In Robustness of Bayesian Analysis, Kadane, I.B. (Editor), Elsevier, Amsterdam, pp.63-124.

Berger, R. L. (1981), A necessary and sufficient condition for reading a consensus using DeGroot's method, JASA, 76, pp.415-418.

Box, GEP & Tiao, GC (1973), Bayesian Inference in Stat Analysis. Addison-Wesley, Reading.

Bradley, S., Hax, A.C. and Magnanti, T.L. (1980), Applied Mathematical Programming. Addison-Wesley, Reading, MA.

Britney. R. and Winkley, R. (1974), Bayesian point estimation and prediction. Ann. Inst. Statist. Math., 26, pp.15-34.

Cohen M. and Fisher M. (1982), DS903 class notes. University of Pennsylvania.

Cummins, J. D. (1981), INS 832 class notes. University of Pennsylvania.

DeGroot, M. (1970), Optimal Statistical Decisions. McGraw-Hill, NY.

DeGroot, M. (1974), Reaching consensus. JASA, 69, pp.118-121.

Dempster, A. (1975), A Subjective Look at Robustness. Research Report S-33, Dept. of Statistics, Harvard University.

Edwards, W., Lindeman Y. and Savage, L. J. (1963), Bayesian statistical inference for psychological research. Psychological Review, 70, pp.193-242.

Fishburn, P. ,Murphy, A.H. and Isaacs, H.H.(1967), Sensitivity of decision to probability estimation errors: a reexamination. Operations Research, 15, pp.254-267.

Greville, T.N.E. (1978), Discussion on Schuette's paper. TSA, XXX, p.442.

Hampel, F. (1971), A general qualitative definition of robustness. Ann. Math. Statist,42, pp.1887-1896.

Hampel, F. (1974), The influence curve and its role in robust estimation. JASA, 69, pp.383-393.

Henrici, P. (1964), Elements of Numerical Analysis. Wiley, NY.

Hickman, J. (1977), Notes on Bayesian graduation. TSA, XXIX, pp.7-21.

Hickman, J. (1978), Discussions on Schuette's paper. TSA, XXX, pp.433-436.

Hickman, J. (1979), Bivariate Bayesian graduation. American Statist. Assoc., Business and Eco Stat. Section, Proceedings, pp.59-63.

Hogg, R. (1979), Statistical robustness. American Statistician, 33, pp.l08-115.

Huber, P. (1964), Robust estimation of a location parameter. Ann. Math. Statist., 35, pp.73-101.

Huber, P. (1972), Robust statistics: a review. Ann. Math. Statist., 47, pp.l041-1067.

Huber, P. (1977), Robust Statistical Procedures. SIAM, Philadelphia.

Huber, P. (1981). Robust Statistics. Wiley, NY.

Kadane, J. (1978), Stable decision problems. Ann. of Statist., 6, pp.1095-1110.

Kimeldorf, G. and Jones, D. (1967), Bayesian graduation. TSA, XIX, pp.66-112.

Klugman, S. (1978), Discussion of Schuette's paper. TSA, XXX, pp.436-439.

Klugman, S. (1979), Robust mortality estimation. ARCH, 1979 Issue 3, pp.61-75.

Klugman, S. (1980), Mortality estimation and graduation. American Statist. Assoc., Social Stat.

Section, Proceedings, p.34.

Lindley, D. (1981), Bayes empirical Bayes. JASA, 76, pp.833-841.

McGill, Dan May (1982), INS 831 class notes. University of Pennsylvania.

Pierce, D. and Folks, J. (1970), Sensitivity of Bayes procedures to the prior distribution. Operations Research, 18, pp.344-350.

Ramsay, J. and Novick, M. (1980), PLU robust Bayesian decision theory: Point estimation. JASA, 75, pp. 901-907.

Rubin, M. (1977), Robust Bayesian estimation. In Statistical Decision Theory and Related Topics, Ⅱ S. S. Gupta and D. S. Moore (eds). Academic Press, NY.

Schuette, D. (1978), A linear programming approach to graduation. TSA, XXX, pp.407-432. Discussion, pp.443-445.

Tukey, J. W. (1960), A Survey of Sampling from Contaminated Distributions. In Contributions to Probability and statistics, I. Olkin (ed), Stanford Univ. Press, Stanford.

Whittaker, E. T. (1944), The Calculus of Observation, 4th ed. London and Glasgow: Blackie & Son, Ltd.