77

BARD

FINAL REPORT

PROJECT NO. I-216-80

Statistical Methods in Many Parameters Problems and Applications in Agricultural Economics

J.A. Yahav, Y. Mundlak, P.J. Bickel

1985

BARD
P.O.B. 6
Bet Dagan, 50250
ISRAEL

BARD Research Project No. 1-216-80

Title of Research Project

Statistical Methods in Many Parameters Problems and Applications
in Agricultural Economics

Name and Address of Investigators

Principal Investigators: Joseph A. Yahav

Yair Mundlak

Cooperating Investigator: Peter J. Bickel

Name and Address of Affiliated Institutions

Principal Institution: The Center for Agricultural Economic Research
P.O. Box 12, Rehovot, Israel

Cooperating Institution: Department of Statistics, University of
Berkeley, CA. U.S.A.

Project's Commencement date 1 Sept. 1981.

Type of Report: Final

Signature

Principal Investigator

Institution's Authorized Official

B. Table of Contents

Abstract	3
Objective - See Abstract	
Body of the report	
Preface	4
A. Statistical Methodology: Overview	5
B. Statistical Methodology: Detailed Reports	12
On Empirical Bayes: The Normal Case	13
Introduction and Summary	13
Linear Empirical Bayes and Empirical Bayes: An Overview	50
Introduction	50
On Clustering Points in R^1 : Results and Anomalies	59
Abstract	60
Introduction	61
List of Publications	84

C.

A B S T R A C T

This research was concerned with many parameters and many observation models in statistical analysis. Such methods are applied frequently in economics in general and in agricultural economics in particular. The main results show that the frequently used maximum likelihood methods are inefficient for such models. Under some reasonable assumptions the non-parametric Bayes methods are proved to be asymptotically optimal.

Some specific problems related to clustering were also considered. In these cases, under some assumptions clustering observations optimally, yields good results for clustering parameters.

The theoretical work was supplemented by empirical analysis.

Preface

The report contains three parts:

- (A) An overview of the results of the investigation of the statistical methodology. This subject constituted the core of the proposal.
- (B) More detailed results, discussed in manuscripts by (1) Robbins and Yahav, (2) Yahav and (3) Bickel, Krieger and Yahav.
- (C) A brief description of the empirical work. This will be discussed in forthcoming research reports of the Center for Agricultural Economic Research.

A. Statistical Methodology: Overview

1. Introduction.

Many problems in economics in general and in agricultural economics in particular can be described in the following way: We are interested in a parameter A (possibly a multidimensional parameter) that governs the relationship between one set of variables say X and another set of variables say Y through a known functional relation say

$$(1) \quad f_A(X) = Y.$$

If we knew X and Y exactly we could solve (1) so as to get A

$$(2) \quad A = g(X, Y).$$

Unfortunately, in many situations X, Y are observed with some error, i.e., we observe X^* and Y^* where,

$$(3) \quad X^* = X + \varepsilon, \quad Y^* = Y + \delta$$

In practical problems it is often customary to replace X by X^* and Y by Y^* and solve (2) to get:

$$(4) \quad A = g(X^*, Y^*).$$

This is usually done under the assumption that ϵ and δ are small in some relevant sense and A is continuous. The properties of such a solution are investigated under some simplified assumptions. The model that we worked on and the solutions that we have can easily be extended to many practical situations.

Consider X_i to be normally distributed with mean μ_i and variance σ_i^2 and given μ_i, σ_i^2 the X_i are independent. Formally:

$$(5) \quad X_i = \mu_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_i^2) \quad i = 1, 2, \dots, n.$$

ϵ_i are independent. A is assumed to depend on parameters μ_i .

$$(6) \quad A = A(\mu_1, \mu_2, \dots, \mu_n).$$

We are interested in A , A is known whenever $\mu = (\mu_1, \mu_2, \dots, \mu_n)$ is known. However, μ is unknown and instead we observe x . In practice we use the solution:

$$(7) \quad \hat{A} = A(X_1, X_2, \dots, X_n).$$

which is consistent under the assumptions $\sigma_i^2 \rightarrow 0$ for all $i = 1, 2, \dots, n$ and A continuous in μ . If we assume A to be continuous and differentiable and we assume $\sigma_i^2 \rightarrow 0$ then we can expand $A(\mu)$ around the true μ and $\text{Var}(\hat{A}(\mu))$ will depend on $\|\hat{\mu} - \mu\|^2$. As is well known X as an estimator of μ does not

minimize $\|\hat{\mu} - \mu\|^2$ asymptotically as $n \rightarrow \infty$. However, $\|\hat{\mu} - \mu\|^2$ is minimized by an empirical Bayes estimator. In this case, $\hat{\mu}_i$ can be constructed easily when σ_i^2 are known and $\hat{\mu}_i(X_1, \dots, X_n) \rightarrow X_i$ as $\sigma_i^2 \rightarrow 0$. This justifies to some extent the use of X_i in (7). A simple example for consideration is

$$(8) \quad A(\mu_1, \dots, \mu_n) = (\mu_1, \mu_2, \dots, \mu_n).$$

Another case of interest is the case where $A(\mu)$ is symmetric in μ in the sense that A is invariant under permutation of $(\mu_1, \mu_2, \dots, \mu_n)$. In this case when $n \rightarrow \infty$ we need to construct consistent estimator of the limit distribution of the μ 's which we need to assume to exist. Again, $A(X_1, \dots, X_n)$ will be a consistent estimator under the assumption $\sigma_i^2 \rightarrow 0$ uniformly in i but this estimator will have efficiency zero compared to the estimator based on the consistent estimator of the limiting distribution of the μ_i 's. A simple example for such a problem is

$$(9) \quad A(\mu_1, \dots, \mu_n) = \frac{1}{n} \sum_{i=1}^n I(\mu_i \leq 0).$$

2. $A(\mu)$ is Differentiable

If $A(\mu)$ is differentiable at μ_0 and $\hat{\mu}$ is close to μ_0 we have

$$(10) \quad A(\hat{\mu}_0) = A(\mu_0) + A'(\mu_0)(\hat{\mu} - \mu_0)$$

and

$$(11) \quad \|A(\hat{\mu}) - A(\mu_0)\|^2 = \|A'(\mu)\|^2 \|\hat{\mu} - \mu_0\|^2$$

Hence the investigation is reduced to find estimators $\hat{\mu}$ so as to minimize $\sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2$. This is an "Empirical Bayes" problem and the solution for the normal case is, for σ_i^2 known

$$(12) \quad \hat{\mu}_i = X_i + \sigma_i \frac{h'(X_i/\sigma_i)}{h(X_i/\sigma_i)}$$

where

$$(13) \quad h_n(t) = .8n^{4/5} \sum_{i=1}^n \phi(.8n^{1/5}(t - \frac{X_i}{\sigma_i}))$$

$$\text{and } \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

These estimators behave well for $n \geq 10$; they are not recommended for $n < 10$. For $3 < n < 10$ we prefer

$$(14) \quad \hat{\mu}_i = \frac{(S^2 - \sigma_i^2)^+}{S^2} X_i + \frac{\sigma_i^2}{S^2} \bar{X}$$

where $\bar{X} = \frac{1}{n} \sum X_i$ and $S^2 = \frac{1}{n-3} \sum_{i=1}^n (X_i - \bar{X})^2$

and for $n = 1, 2, 3$ $\hat{\mu}_i = X_i$ would suffice.

If σ_i^2 are not known and repetitions are available so that σ_i^2 can be reasonably estimated one can use $\hat{\sigma}_i^2$ in the above formulas.

If σ_i^2 are not known and reasonable estimators for σ_i^2 are not available then it seems that $X_i = \hat{\mu}_i$ is the only reasonable solution. The above solution is optimal provided $n \rightarrow \infty$. We did carry extensive Monte-Carlo experimentation which invariably showed the advantage of the above solution compared with $\hat{\mu}_i = X_i$. This approach to the problem is the reasonable one provided the dimensionality of the data, i.e., n . If A is of considerable smaller dimension than n , then other type of analysis is adequate. No work has been done yet for that case.

3. $A(\mu)$ is symmetric in μ .

In many statistical problems $A(\mu)$ is symmetric in μ and of low dimension compared with n . For a symmetric A , A depends on μ through the empirical distribution of $(\mu_1, \mu_2, \dots, \mu_n)$ and hence is a function of $F_n(\cdot)$, the empirical distribution. In some cases n is comparatively large and under normal conditions we are able to approximate F_n fairly reasonably. The estimator that we did investigate is the inverse of the Fourier transform for the measure convoluted with a normal variable. This estimator is very slow to converge and is useful for very large n , so it seems that it is more appropriate to look for estimators of F_n that would agree with F_N on the first few moments. This approach seems to be useful and works well whenever A does depend on these moments only. For instance

$$(15) \quad \hat{\mu}_i = \frac{S}{\sqrt{\sigma_i^2 + S^2}} X_i + \left(1 - \frac{S}{\sqrt{\sigma_i^2 + S^2}}\right) \bar{X} \quad \text{where}$$

$$\bar{X} = \frac{\sum_{i=1}^n \frac{S}{\sigma_i^2 + S^2} X_i}{\sum_{i=1}^n \frac{S}{\sigma_i^2 + S^2}} \quad \text{and}$$

$$S^2 = \frac{1}{n-3} \sum_{i=1}^n (X_i - \bar{X})^2. \quad \text{Then the empirical distribution of } \hat{\mu}_i$$

will match the first two moments of the distribution of the μ_i for large n and σ_i^2 uniformly bounded. So if A depends on the empirical distribution through its two first moments only then (15) will provide a reasonable solution. Generally A should be investigated to find out moment relationships. One interesting example that was investigated extensively is

$$(16) \quad A(\mu) = \frac{1}{n} \sum_{i=1}^n I(\mu_i = 0).$$

This problem occurs in the regression problem where we would like to determine those coefficients which are relatively negligible.

Several different methods were compared with the conclusion that for n large all coefficients that are of smaller order than $\frac{1}{n}$ should be ignored, i.e.

$$\mu_i \begin{cases} = 0 & |X_i| \leq \frac{t}{n} \\ \neq 0 & |X_i| > \frac{t}{n} \end{cases}$$

where t can be determined by significance requirements or prior probability. Here again Monte Carlo experiments were used heavily to determine t in specific examples.

B. Statistical Methodology: Detailed Reports

Introduction

The Robbins-Yahav draft relates to the construction of empirical Bayes estimators in the normal case. It contains analytic work regarding the asymptotic behavior of these estimators and Monte Carlo work comparing the performance of these estimators with the Stein-type estimators for moderate size n .

The Yahav draft gives an overall look at empirical Bayes estimators.

The Bickel-Krieger and Yahav draft deals with the problem of clustering, where the loss function is symmetric and invariant under permutation of the parameters for fixed estimators.

Introduction and Summary

Stein, (1955) considering the problem of estimating the means μ_i of n normal populations simultaneously, proved that the conventional estimator, the sample means, is not admissible with respect to a loss function that is the sum of the squared errors. James and Stein (1961) exhibited a simple estimator that always has a smaller risk.

Efron and Morris (1975) showed that the James-Stein estimator can be viewed as an Empirical Bayes estimator for a normal prior of μ . We show in Section 1 that what Efron and Morris call E.B. (Empirical Bayes) estimators might better be called Linear E.B. estimators. In Section 2 we derive the General Empirical Bayes (G.E.B.) estimator. In Section 3 we introduce an estimator that behaves as $n \rightarrow \infty$ like L.E.B. whenever the conditional expectation of μ given X is linear in X , and like G.E.B. otherwise. In Section 4 we present numerical Monte Carlo results for a moderate number of observations. In Section 5 we give three examples using the different estimators. In Section 6 we consider the problem of estimating the quantity $E(\mu|X>t)$. Section 7 concludes the paper with a discussion.

1. Framework

Let (μ_i, X_i) be i.i.d. random vectors distributed like (μ, X) , and let $f_\mu(x)$ denote the conditional density of X given μ . We consider here the case where

$$f_\mu(x) = \frac{1}{(2\pi)^{1/2}} e^{-(x-\mu)^2/2}$$

We assume the distribution of μ to be unknown and denote this distribution by G .

We observe the sequence $\{X_i; i=1,2,\dots,n\}$ and consider the problem of simultaneously estimating each of the μ_i , with loss proportional to the sum of the squared errors. The best linear estimator of μ

$$(1.1) \quad a^*X + b^*$$

is that for which

$$(1.2) \quad a^* = \frac{\text{Cov}(X, \mu)}{V(X)}$$

$$(1.3) \quad b^* = E(\mu) - a^*E(X).$$

Since G is unknown, $\text{Cov}(X, \mu)$, $V(X)$, $E(\mu)$, $E(X)$ are unknown parameters. These parameters can be estimated from the X_i sequence in the following manner. Since $X = \mu + \epsilon$; where ϵ is independent of X and ϵ is normal with mean zero and variance one,

$$(1.4) \quad \text{Cov}(\mu, X) = V(\mu)$$

$$(1.5) \quad V(X) = V(\mu) + 1$$

$$(1.6) \quad E(\mu) = E(X).$$

We can therefore estimate a^* and b^* by

$$(1.7) \quad \hat{a} = \frac{s^2 - 1}{s^2},$$

$$(1.8) \quad \hat{b} = \frac{1}{s^2} \bar{X},$$

where

$$\bar{X} = \frac{\sum X_i}{n}, \quad s^2 = \frac{\sum (X_i - \bar{X})^2}{n-3}.$$

(The use of $n-3$ in the denominator conforms with (7.1) of Efron and Morris (1973).)

Combining (1.1), (1.7), and (1.8) we get the L.E.B. estimator of μ_i ,

$$(1.9) \quad \hat{\mu}_i = X_i + \frac{1}{s^2} (\bar{X} - X_i) = \bar{X} + (1 - \frac{1}{s^2}) (X_i - \bar{X}).$$

The best, not necessarily linear, estimator of μ in this context is $E(\mu|X)$, where

$$(1.10) \quad E(\mu|X) = \frac{\int_{-\infty}^{\infty} \mu \phi(X-\mu) dG(\mu)}{\int_{-\infty}^{\infty} \phi(X-\mu) dG(\mu)}$$

and

$$\phi(x) = \frac{1}{(2\pi)^{1/2}} e^{-x^2/2}.$$

A simple computation yields

$$(1.11) \quad E(\mu|X) = X + \frac{h'(X)}{h(X)}$$

where

$$h(x) = \int_{-\infty}^{\infty} \phi(x-\mu) dG(\mu)$$

is the marginal density of X . The G.E.B. for this problem is therefore

$$(1.12) \quad \hat{\mu}_i = X_i + \frac{\hat{h}'(X_i)}{\hat{h}(X_i)}$$

for consistent estimators $\hat{h}$ and $\hat{h}'$ of h and h' respectively; see Robbins (1964). We remark that (1.1) and (1.11) are equal if G itself is normal.

2. Estimation of $h(t)$ and $h'(t)$

Following Rosenblatt (1971) we use a normal kernel estimator to estimate $h(t)$,

$$\hat{h}_n(t) = \frac{n^\epsilon K}{n} \sum_{i=1}^n \phi(Kn^\epsilon(t-X_i)),$$

where K and ϵ are positive constants.

We note that

$$(2.1) \quad E[\phi(a_n X + b)] = \int_{-\infty}^{\infty} \frac{1}{(1+a_n^2)^{1/2}} \phi\left(\frac{b+a_n \mu}{(1+a_n^2)^{1/2}}\right) dG(\mu)$$

and

$$(2.2) \quad E[\phi^2(a_n X + b_n)] = \int_{-\infty}^{\infty} \frac{1}{[2\pi(1+2a_n^2)]^{1/2}} \phi\left[\left(\frac{2}{1+2a_n^2}\right)^{1/2}(b_n + a_n \mu)\right] dG(\mu).$$

Setting $a_n = n^\epsilon K$, $b_n = -ta_n$, we get

$$(2.3) \quad \lim_n \hat{h}_n(t) = h(t), \text{ a.s.},$$

$$(2.4) \quad \lim_n n^{1-\epsilon} V(\hat{h}_n(t)) = \frac{K}{2\sqrt{\pi}} h(t),$$

$$(2.5) \quad \lim_n K^2 n^{2\epsilon} (E[\hat{h}_n(t)] - h(t)) = (1/2) \int_{-\infty}^{\infty} [(\mu-t)^2 - 1] \phi(t-\mu) dG(\mu).$$

(2.3) follows from Nadaraya (1964); (2.4) and (2.5) can be verified by direct computation, using (2.1) and (2.2). Hence the asymptotic squared error for $\hat{h}_n(t)$ behaves like

$$\frac{K h(t)}{2\sqrt{\pi}} \cdot \frac{1}{n^{1-\epsilon}} + \left(\frac{1}{2}\right)^2 \left[\int_{-\infty}^{\infty} [(\mu-t)^2 - 1] \phi(t-\mu) dG(\mu) \right]^2 \frac{1}{K^4 n^{4\epsilon}},$$

and this expression is asymptotically minimized by taking

$\epsilon = 1/5$. Since G is unknown, we do not have a universal K to minimize

the above expression. One method for choosing K is that of Silverman (1978). In our Monte Carlo simulations we used $K = 0.8$. We have then as an estimate of $h(t)$ based on $X_1, \dots, X_n$

$$(2.6) \quad \hat{h}_n(t) = .8n^{-4/5} \sum_{i=1}^n \phi(.8n^{1/5}(t-X_i)).$$

Theorem 2.1: As $n \rightarrow \infty$

$$(a) \quad \hat{h}_n(t) \rightarrow h(t) \text{ a.s.}$$

$$(b) \quad n^{4/5} E(\hat{h}_n(t) - h(t))^2 \rightarrow \frac{.4}{\sqrt{\pi}} h(t) + \frac{1}{1.6384} b^2(t),$$

where

$$b(t) = \int_{-\infty}^{\infty} [(\mu - t)^2 - 1] \phi(t - \mu) dG(\mu)$$

$$(c) \quad n^{2/5} (\hat{h}_n(t) - h(t)) \xrightarrow{D} N\left(-\frac{1}{1.28} b(t), \frac{.4}{\sqrt{\pi}} h(t)\right).$$

Proof: (a) follows from (2.3); (b) follows from (2.4) and (2.5). To prove (c) let

$$Z_{in} = .8 \frac{[\phi(.8n^{1/5}(t-X_i)) - E\phi(.8n^{1/5}(t-X_i))]}{n^{2/5}}$$

By a classical theorem,

$$\sum_{i=1}^n Z_{in} \xrightarrow{D} N(0, \sigma^2),$$

where

$$\lim_n \sum_{i=1}^n V(Z_{in}) = \sigma^2.$$

It follows from (2.4) that

$$\sigma^2 = \frac{.4h(t)}{\sqrt{\pi}}.$$

For the expectation we have

$$\frac{n(.8)E\phi(.8n^{1/5}(t-X_i))}{n^{2/5}} = \frac{n(.8)}{n^{2/5}} \int_{-\infty}^{\infty} \frac{1}{(1+.8^2 n^{1/5})^{1/2}} \phi\left[\frac{.8^{1/5}(t-u)}{(1+.8^2 n^{2/5})^{1/2}}\right] dG(u) \\ = n^{2/5} (h(t) + \frac{1}{n^{2/5}} \frac{b(t)}{1.28}),$$

which yields the expectation in the limit.

We could proceed now to estimate $h'(t)$ by the formal derivative of $\hat{h}_n(t)$, as suggested by Schuster (1969); namely,

$$(2.7) \quad \hat{h}_n'(t) = -(.8)^3 n^{-2/5} \sum_{i=1}^n (t-X_i) \phi(.8n^{1/5}(t-X_i)).$$

This estimate is strongly consistent uniformly in t . However, we can consider the estimation of $h'(t)$ separately. We note that

$$(2.8) \quad (1+a^2)^{3/2} E[(t-X)\phi(a(t-X))] = \int_{-\infty}^{\infty} (t-u) \phi\left(\frac{a}{(1+a^2)^{1/2}}(t-u)\right) dG(u),$$

(2.9)

$$(2\pi)^{1/2} (1+2a^2)^{3/2} E[(t-X)^2 \phi^2(t-X)] = \int_{-\infty}^{\infty} \left\{1 + \frac{(t-u)^2}{1+2a^2}\right\} \phi\left(\left[\frac{2a^2}{1+2a^2}\right]^{1/2}(t-u)\right) dG(u).$$

Hence, we consider estimators of the form

$$(2.10) \quad S_n(K, t) = - \frac{(Kn^\epsilon)^3}{n} \sum_{i=1}^n (t-X_i) \phi(Kn^\epsilon(t-X_i)).$$

Using (2.8) and (2.9) we get

$$(2.11) \quad V(S_n(K, t)) \sim \frac{K^3}{n^{1-3\epsilon}} \frac{1}{4\sqrt{\pi}} h(t)$$

and

$$(2.12) \quad E[S_n(K, t)] - h'(t) \sim \frac{-1}{2K^2 n^{2\epsilon}} \int_{-\infty}^{\infty} [(t-u)^3 - 3(t-u)] \phi(t-u) dG(u).$$

To minimize the asymptotic squared error, considering (2.11) and (2.12), we take $\epsilon = 1/7$ and use $K = .5$. Hence, our estimator of $h'(t)$ is

$$(2.13) \quad \hat{h}'_n(t) = \frac{(.5)^3 n^{3/7}}{n} \sum_{i=1}^n (t-X_i) \phi(.5n^{1/7}(t-X_i)),$$

and, finally, our estimator of $E(\mu|X=t)$ is

$$(2.14) \quad m_1(t|X_1, \dots, X_n) = t - \frac{n^{8/35} (.5)^3}{(.8)} \frac{\sum_{i=1}^n (t-X_i) \phi(.5n^{1/7}(t-X_i))}{\sum_{i=1}^n \phi(.8n^{1/5}(t-X_i))}.$$

For this estimator we have the following theorem.

Theorem 2.2: For every $-\infty < t < \infty$, as $n \rightarrow \infty$

$$(2.15) \quad |m_1(t|X_1, \dots, X_n) - E(\mu|X=t)| \rightarrow 0 \text{ a.s.,}$$

(2.16)

$$n^{4/7} E(m_1(t) - E(\mu|X=t))^2 \rightarrow \frac{1}{h(t)} \left\{ \frac{1}{32\sqrt{\pi}} + \frac{4}{h(t)} \left(\int_{-\infty}^{\infty} [(t-u)^3 - 3(t-u)] \phi(t-u) dG(u) \right)^2 \right\},$$

(2.17)

$$n^{2/7} (m_1(t|X_1 \dots X_n) - E(\mu|X=t)) \overset{D}{\rightarrow} N(\theta, V),$$

$$\text{where } \theta = \frac{-2}{h(t)} \int_{-\infty}^{\infty} [(t-u)^3 - 3(t-u)] \phi(t-u) dG(u) \text{ and } V = \frac{1}{h(t) 32\sqrt{\pi}}.$$

Proof: We have

$$(2.18) \quad |m_1(t|X_1 \dots X_n) - E(\mu|X=t)| = \left| \frac{\hat{h}'_n(t)}{\hat{h}_n(t)} - \frac{h'(t)}{h(t)} \right|$$

$$= \left| \frac{1}{h(t)} \left(\frac{h(t)}{\hat{h}_n(t)} \hat{h}'_n(t) - h'(t) \right) \right|.$$

Since $h'(t)/\hat{h}_n(t) \rightarrow 1$ a.s., $h(t) > 0$, and $|\hat{h}'_n(t) - h'(t)| \rightarrow 0$ a.s., we get (2.15). Using (2.18) again we get

$$(2.19) n^{4/7} E[(m_1(t|X_1 \dots X_n) - E(\mu|X=t))^2] = n^{4/7} \frac{1}{h^2(t)} E\left[\frac{h(t)}{\hat{h}_n(t)} \hat{h}'_n(t) - h'(t)\right]^2.$$

Since $h'(t)/\hat{h}_n(t)$ is $1 + O(n^{-2/5})$ and

(2.20)

$$n^{4/7} E[\hat{h}'_n(t) - h'(t)]^2 \rightarrow \frac{(.5)^3}{4\sqrt{\pi}} h(t) + \frac{1}{4(.5)^4} \left[\int_{-\infty}^{\infty} [(t-\mu)^3 - 3(t-\mu)] \phi(t-\mu) dG(\mu) \right]^2,$$

we get (2.16).

To prove (2.17) we note that the classical Central Limit Theorem applies to (2.13). Using (2.18), $h'(t)/\hat{h}_n(t) \sim 1 + O(n^{-2/5})$, (2.11), and (2.12), we get (2.17).

We consider now two other consistent estimators for $E(\mu|X=t)$. Using the formal derivative (2.7) of $\hat{h}_n(t)$ we get

$$(2.21) m_2(t|X_1, X_2, \dots, X_n) = t - \frac{n^{2/5} (.8)^2 \sum_{i=1}^n (t-X_i) \phi(.8n^{1/5}(t-X_i))}{\sum_{i=1}^n \phi(.8n^{1/5}(t-X_i))}.$$

The strong consistency of m_2 follows from Schuster (1969).

Another possible estimator follows from considering $E(\mu|t-\delta < X < t+\delta)$. We have

$$(2.22) E(\mu|t-\delta < X < t+\delta) = \frac{E[\mu I_{\{t-\delta < X < t+\delta\}}]}{E[I_{\{t-\delta < X < t+\delta\}}]},$$

where $I_{\{A\}}$ is the indicator function of the set A , and we estimate (2.22) by

$$(2.23) \quad m_3(t|X_1, \dots, X_n) =$$

$$\frac{\sum_{i=1}^n X_i I_{\{t-\delta < X_i < t+\delta\}} + \sum_{i=1}^n .8n^{1/5} \{\phi(.8n^{1/5}(t+\delta-X_i)) - \phi(.8n^{1/5}(t-\delta-X_i))\}}{\sum_{i=1}^n I_{\{t-\delta < X_i < t+\delta\}}}$$

where $\delta = 2n^{-1/5}$. The motivation and proof of consistency for this estimator appear in Section 6.

One can use other kernels for estimating the density; however, it appears from the simulation results that the normal kernel works well in this case. Miyasawa (1968) gives a histogram estimator for $h(t)$.

3. An Adaptive Estimator

If G is such that

$$(3.1) \quad E(\mu|X=t) = a^*t + b^* = t + \frac{1}{V(\mu)+1} (E(\mu) - t)$$

then we can estimate $E(\mu|X=t)$ by

$$(3.2) \quad M(t) = t + \frac{1}{S^2} (\bar{X} - t),$$

and it can be shown that

$$(3.3) \quad \sqrt{n}[M(t) - E(\mu|X=t)] = \sqrt{n}\left[\frac{1}{S^2} (\bar{X} - t) - \frac{1}{V(\mu)+1} (E(\mu) - t)\right] \xrightarrow{D} N(0, A),$$

where A depends on $V(\bar{X})$, $V(S^2)$, and $\text{Cov}(\bar{X}, S^2)$. We see, then, that under the assumption (3.1), if $E[\mu^4] < \infty$, (3.2) is $n^{-\epsilon}$ consistent for $0 < \epsilon < 1/2$.

If G is such that (3.1) does not hold, then (3.2) is inconsistent and the estimator m_1 of (2.14), which is consistent, should be used. However, m_1 is only $n^{-\delta}$ consistent for $\delta < 2/7$, which means that whenever (3.1) holds the estimator M is preferable. Following Robbins (1980) we give here an adaptive estimator that behaves asymptotically like M whenever (3.1) holds and like m_1 whenever (3.1) does not hold. One such estimator is

$$(3.4) \quad Z(t) = \frac{M(t) + |m_1(t) - M(t)|^\gamma c n^\epsilon m_1(t)}{1 + c n^\epsilon |m_1(t) - M(t)|^\gamma}.$$

For $\epsilon > 2/7$, $\gamma > \frac{7}{4} + \frac{7}{2} \epsilon$, c a positive constant, we have whenever (3.1) holds, $|m_1(t) - M(t)| \sim n^{-2/7}$ and

$$(3.5) \quad \sqrt{n}|m_1(t) - M(t)|^\gamma c n^\epsilon \rightarrow 0,$$

in probability, and hence $Z(t) \sim M(t)$. Whenever (3.1) does not hold we have $|m_1(t) - M(t)| \sim K(t)$ where $K(t) > 0$, and hence

$$(3.6) \quad n^{2/7} c n^{-\epsilon} |m_1(t) - M(t)|^{-\gamma} \rightarrow 0$$

in probability. This yields $Z(t) \sim m_1(t)$. We shall use $c=1$, $\epsilon = \frac{9}{28}$, $\gamma=3$ for the numerical evaluation of (3.4). We do not know how to choose ϵ , γ , c in an optimal way.

We note that $E(\mu|X=t)$ is a monotone function of t , and hence we consider here a monotone version of $m_1(t)$. The isotonizing procedure is a nonparametric one following Barlow et al. (1972).

4. Numerical Monte-Carlo Evaluations

In this section we summarize some Monte Carlo evaluations of the performance of the maximum likelihood estimator (MLE), the L.E.B. estimator (1.9), the three G.E.B. procedures (m_1 , m_2 , m_3) given in section 2, a monotonized version of m_1 (m_1^{ISO}), and the adaptive procedure of section 3 (m_1^{ADA}). The five different distributions of μ that were considered are given in Table 1. Sample sizes considered were $n = 6, 10, 25, 100, 1000$. For each combination of sample size and distribution, fifty repetitions of the experiment were conducted; i.e., n observations μ_i were taken from the distribution and for each one, X_i was selected from $N(\mu_i, 1)$. In each experiment the mean squared error was computed, and Tables 2 and 3 summarize

$$(4.1) \quad \frac{1}{50} \sum_{r=1}^{50} \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2$$

The standard error is given below the number. For comparison, the appropriate Bayes estimator was also computed and included in the tables. In order to display the distance between these estimators in terms of the standard error, Table 4 was constructed. If we denote the entries in a given row of Table 2 or 3 by M_i (with SE_i the standard error) and let M_{Be} be the least of these (not including the Bayes estimator), then the entries in Table 4 are given by

$$(4.2) \quad 2 \frac{M_{Best} - M_i}{SE_{Best} + SE_i}$$

rounded to the nearest integer. (The "Best" entry in each row is marked with an asterisk). This may be interpreted as the distance from the best

estimator in units of the standard error.

The first impression one gets from these tables is yet another demonstration of the inadmissibility of the maximum likelihood estimator. Only for $n = 6$ does the MLE perform acceptably, and even there, for the uniform distribution, it is 6 standard errors away from the best. The next impression is that none of the Empirical Bayes procedures is uniformly best. Selection of the appropriate estimator must be made with care and consideration of special characteristics of the data. Even the m_2 procedure, although never the best, is not dramatically worse than m_1 and seems to provide better estimates of extreme points. (When applied to extreme points, m_1 seems to go astray due to the different window sizes used for $\hat{h}_n$ and $\hat{h}'_n$.) Also, when applied to extreme points the m_3 procedure may well have only one point in the 2δ interval about X_i (X_i itself). In this case the procedure is quite inaccurate and should not be used. (In several experiments this situation was responsible for 20 to 40 percent of the sum of squared errors.)

The choice between L.E.B. and one of the G.E.B. procedures should be based on the linearity of the conditional expectation over the relevant range. Figures A to E present curves for $E(\mu|t)$, $m_1^{ISO}(t)$, and L.E.B. For distribution 1, where $E(\mu|t)$ is a linear function of t , and distribution 3, where $E(\mu|t)$ is well approximated by a linear function over the relevant range, the L.E.B. estimator is superior because of its better rate of convergence. However, in the strongly bimodal distributions 2 and 5, the G.E.B. procedures are best. For distribution 4, where a linear approximation provides only a so-so fit, the adaptive procedure exhibits its superiority.

5. Three Examples

The batting averages data of Efron and Morris (1975) can be analyzed using the G.E.B. estimator m_1 . Here the data apparently forms a single cluster, so we would expect L.E.B. to be the appropriate procedure. In fact, the squared error loss for L.E.B. is 5.00 and for m_1 it is 10.17 (for MLE, it is 17.58). The detailed predictions are in Table 5, which may be compared with Table 2 of Efron and Morris (1975).

For a second example, Empirical Bayes procedures were used to predict the results of measurements. Data from Mandel (1964) as cited by Lehman (1975) give five measurements by each of 15 laboratories of the smoothness of a certain kind of paper. The problem we posed was: given the first three measurements by each laboratory, predict the average of the last two measurements. For this situation we assume that the five measurements made by the i^{th} laboratory are drawn from $N(\mu_i, \sigma^2)$. An estimate of the average of the last two measurements is provided by μ_i .

We first estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{15} \sum_{i=1}^{15} \frac{1}{2} \sum_{j=1}^3 (X_{ij} - \bar{X}_i)^2$$

where $\bar{X}_i = \frac{1}{3} (X_{i1} + X_{i2} + X_{i3})$. Using $\tau^2 = \sigma^2/3$ (the variance of $\bar{X}_i$) we estimate μ_i by

$$\hat{\mu}_i = \tau M(\bar{X}_i/\tau)$$

where M is the estimation procedure under consideration.

Denoting by $\bar{\mu}_i = (X_{i4} + X_{i5})/2$ the quantity to be estimated, we use the loss function

$$\sum_{i=1}^{15} (\hat{\mu}_i - \bar{\mu}_i)^2$$

The estimates $\hat{\mu}_i$ from the various procedures are given in Table 6, and the

values of the loss function in Table 7.

The third example is a constructed set of data. Nine values μ_i were selected and for each, two values X_{i1} and X_{i2} were chosen from $N(\mu_i, 2)$. However, we assume, as in the second example, that the variance is constant but unknown and estimate it from the data. The procedure is analogous to the second example except that we have only two observations. The data and the various estimations are summarized in Table 8. This procedure was repeated with a different set of values for X_{i1} and X_{i2} , and the values of the loss are presented in Table 9.

the values μ_i were
chosen from
that the variance
was estimated from
the data
was repeated with
values for
Table 9.

6. Estimating $E(\mu|X>t)$

Consider the problem of estimating the expected true ability of persons who were selected by passing a test with a grade of t or higher. If we call the true ability of the i^{th} person μ_i and his grade on the test X_i , and if we assume that $X_i = \mu_i + \epsilon_i$ where the ϵ_i are i.i.d. $N(0,1)$, then we are interested in estimating $E(\mu|X>t)$.

In this section we shall construct a G.E.B. estimator of the parameter

$$(6.1) \quad \theta(t) = E[\mu|X>t] = \frac{\int_t^{\infty} E(\mu|x) h(x) dx}{P(X>t)},$$

where t is any fixed value. We note that from (1.11) the numerator is

$$(6.2) \quad \int_t^{\infty} x h(x) dx - h(t)$$

and this is equal to

$$(6.3) \quad E[XI_{\{X>t\}}] - h'(t)$$

Hence, a possible estimator for $\theta(t)$ is

$$(6.4) \quad \hat{\theta}_n(t) = \hat{\theta}(t) = \frac{\sum_{i=1}^n X_i I_{\{X_i>t\}} - \sum_{i=1}^n .8n^{1/5} \phi(.8n^{1/5}(t-X_i))}{\sum_{i=1}^n I_{\{X_i>t\}}}$$

Theorem 4.1: As $n \rightarrow \infty$

$$\hat{\theta}(t) \rightarrow \theta(t) \quad \text{a.s.,}$$

$$(6.6) \quad n^{4/5} E(\hat{\theta}(t) - \theta(t))^2 \rightarrow$$

$$\left[\frac{.4}{\sqrt{\pi}} h(t) + \frac{1}{1.6384} b^2(t) \right] / \left\{ \int_{-\infty}^{\infty} (1 - \phi(t - \mu)) dG(\mu) \right\}^2$$

where

$$b(t) = \int_{-\infty}^{\infty} [(t - \mu)^2 - 1] \phi(t - \mu) dG(\mu),$$

and

$$(6.7) \quad n^{2/5} (\hat{\theta}_3(t) - \theta(t)) \rightarrow N(B, W),$$

where

$$B = - \frac{1}{1.28} / \int_{-\infty}^{\infty} \{1 - \phi(t - \mu)\} dG(\mu),$$

$$W = .4 h(t) / \sqrt{\pi} \left(\int_{-\infty}^{\infty} \{1 - \phi(t - \mu)\} dG(\mu) \right)^2.$$

Proof: $\hat{\theta}(t)$ can be written as

$$(6.8) \quad \hat{\theta}(t) = \hat{\lambda}_1(t) - \hat{\lambda}_2(t)$$

where

$$(6.9) \quad \hat{\lambda}_1(t) = \sum_{i=1}^n X_i I_{\{X_i > t\}} / \sum_{i=1}^n I_{\{X_i > t\}},$$

$$(6.10) \quad \hat{\lambda}_2(t) = \sum_{i=1}^n .8 n^{1/5} \phi(.8 n^{1/5} (t - X_i)) / \sum_{i=1}^n I_{\{X_i > t\}}.$$

We write

$$(6.11) \quad \lambda_1(t) = \theta(t) + h(t)/P(X > t),$$

$$(6.12) \quad \lambda_2(t) = h(t)/P(X > t).$$

Then

$$(6.13) \quad \hat{\lambda}_1(t) \rightarrow \lambda_1(t) \text{ a.s.}, \quad \hat{\lambda}_2(t) \rightarrow \lambda_2(t) \text{ a.s.},$$

which yields (6.5).

To prove (6.6) it is enough to recall that

$$(6.14) \quad n^{4/5} E[(\hat{\lambda}_1(t) - \lambda_1(t))^2] \rightarrow 0,$$

and that by (2.4), (2.5) with $K=.8$ we have

$$(6.15) \quad n^{4/5} E[(\hat{h}_n(t) - h(t))^2] \rightarrow \frac{.4}{\sqrt{\pi}} h(t) + \frac{1}{1.6384} b^2(t).$$

To prove (6.7), the asymptotic normality of $\hat{h}_n(t)$ yields the asymptotic normality of $\hat{\lambda}_2(t)$. Noticing that $n^{2/5}(\hat{\lambda}_1(t) - \lambda_1(t)) \rightarrow 0$ in probability, we get (6.7).

In the following tables we summarize Monte Carlo results for the average squared error of three estimators of $\theta(0)$. We let

$$(6.16) \quad \bar{\theta} = \sum_{i=1}^n \mu_i I_{\{X_i > 0\}} / \sum_{i=1}^n I_{\{X_i > 0\}}$$

$$(6.17) \quad \hat{\theta}_1 = \sum_{i=1}^n X_i I_{\{X_i > 0\}} / \sum_{i=1}^n I_{\{X_i > 0\}}, \quad (\text{naive}),$$

$$(6.18) \quad \hat{\theta}_2 = \sum_{i=1}^n \hat{\mu}_i I_{\{X_i > 0\}} / \sum_{i=1}^n I_{\{X_i > 0\}}, \quad (\text{L.E.B.}),$$

$$(6.19) \quad \hat{\theta}_3 = \hat{\theta} = \frac{\sum_{i=1}^n X_i I_{\{X_i > 0\}} - \sum_{i=1}^n (.8)n^{1/5} \phi(.8n^{1/5} X_i)}{\sum_{i=1}^n I_{\{X_i > 0\}}} \quad (\text{G.E.B.}),$$

where $\hat{\mu}_i$ is given by (1.9).

For each repetition r of the experiment we compute

$$v_{ir} = (\hat{\theta}_{ir} - \bar{\theta}_r)^2 \text{ and } \bar{V}_i = \frac{1}{50} \sum_{r=1}^{50} v_{ir} \quad (i=1,2,3)$$

Table 10 gives $\bar{V}_i$ for the distributions of Table 1 and $n = 6, 20, 100, 1000$, with (s.e.).

We can see that $\hat{\theta}_3$ performs well over all distributions and

especially well for distributions 2, 4, and 5. We compute here the squared error with respect to $\bar{\theta}_r$, which is a good approximation to the expected squared error.

7. Discussion

G.E.B. and density estimation are considered to be somewhat questionable for moderate sample sizes. The Monte Carlo simulations for sample sizes 6, 10, 25 show that in our normal problem, the G.E.B. and the normal kernel method of estimation work reasonably well even for small and moderate sample sizes.

One can profitably use G.E.B., especially m_1^{ISO} , in problems where the data indicate several clusters of observations. When it is not easy to decide if the data indicate one or several clusters, one should use the adaptive estimator (m_1^{ADA}). If it is clear that there is only one cluster, one should use L.E.B.

Two problems remain open. The first is the determination of the band width K , in the normal kernel estimation. The second is the determination of c , ϵ and γ for the adaptive estimator. Both problems are analytically difficult.

1. Distributions Used in Experiments

Index	Distribution of μ	Relevant Range of X
1	$N(0,4)$	$[-8,7]$
2	$P(\mu = -2) = 1/2 = P(\mu = 2)$	$[-5,5]$
3	$U(0,1)$	$[-3,4]$
4	$\exp(1)$	$[-5,8]$
5	$N(-3,1)$ with probability $1/2$ and $N(3,1)$ with probability $1/2$	$[-7,9]$

Table 2. For n=6, 10, 25

Distributions 1, 2, 3, 4, 5. $\frac{1}{50} \sum_{r=1}^{50} \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_{ri} - \mu_{ri})^2$
(st. error)

n	Dist.	MLE	Bayes	L.E.B.	M ₁	M ₂	M ₃	M ₁ ISO	M ₁ ADA
n=6	1	1.058 (.095)	.831 (.071)	.970 (.086)	.983 (.089)	1.029 (.092)	1.163 (.096)	.983 (.089)	.967* (.086)
	2	1.082 (.115)	.332 (.130)	.986 (.105)	.984 (.112)	.911 (.130)	.907* (.136)	.984 (.112)	.985 (.105)
	3	.993 (.096)	.082 (.005)	.536 (.076)	.697 (.081)	.582 (.088)	476* (.082)	.697 (.081)	.477 (.075)
	4	1.067 (.112)	.670 (.137)	.931 (.121)	.888 (.107)	.900 (.109)	.943 (.113)	.888 (.107)	.868* (.117)
	5	1.122 (.102)	.624 (.070)	1.081 (.095)	1.048 (.104)	1.057 (.111)	1.128 (.124)	1.048* (.104)	1.079 (.095)
n=10	1	.987 (.075)	.789 (.042)	.854* (.062)	.935 (.069)	1.051 (.081)	1.175 (.087)	.935 (.069)	.867 (.062)
	2	.931 (.079)	.198 (.086)	.894 (.074)	.799 (.082)	.685 (.099)	.677* (.109)	.799 (.082)	.890 (.075)
	3	1.052 (.066)	.077 (.003)	.359* (.046)	.632 (.055)	.505 (.061)	.394 (.055)	.632 (.055)	.389 (.050)
	4	1.061 (.079)	.422 (.071)	.651 (.074)	.734 (.065)	.692 (.069)	.652 (.064)	.734 (.065)	.614* (.069)
	5	1.090 (.069)	.670 (.070)	1.042 (.066)	.976* (.069)	1.035 (.074)	1.083 (.078)	.976 (.069)	1.033 (.067)
n=25	1	1.062 (.044)	.873 (.034)	.901* (.038)	.958 (.045)	1.087 (.049)	1.127 (.055)	.958 (.045)	.913 (.042)
	2	1.052 (.041)	.344 (.061)	.868 (.033)	.766 (.045)	.657 (.057)	.596* (.060)	.766 (.045)	.856 (.038)
	3	1.026 (.046)	.075 (.002)	.192* (.020)	.450 (.036)	.354 (.039)	.273 (.034)	.449 (.036)	.251 (.028)
	4	1.072 (.040)	.482 (.039)	.594* (.042)	.693 (.032)	.671 (.037)	.663 (.050)	.693 (.032)	.584 (.034)
	5	1.030 (.045)	.635 (.036)	.949 (.040)	.833* (.043)	.902 (.044)	.903 (.048)	.833 (.043)	.922 (.042)

Table 3. For $n=100, 1000, \infty$
Distributions 1, 2, 3, 4, 5: $\frac{1}{50} \sum_{r=1}^{50} \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_{ri} - \mu_{ri})$
(st. error)

n	Dist.	MLE	Bayes	L.E.B.	M ₁	M ₂	M ₃	M ₁ ISO	M ₁ ADA
100	1	1.018 (.022)	.798 (.018)	.801* (.017)	.839 (.018)	.914 (.019)	.907 (.017)	.838 (.018)	.811 (.018)
	2	1.001 (.023)	.295 (.026)	.813 (.016)	.548 (.022)	.459 (.025)	.437* (.025)	.544 (.022)	.708 (.022)
	3	.982 (.022)	.077 (.001)	.105* (.004)	.276 (.014)	.201 (.012)	.164* (.011)	.260 (.014)	.130 (.008)
	4	.997 (.022)	.408 (.013)	.507 (.019)	.519 (.015)	.511 (.016)	.497 (.015)	.511 (.015)	.474* (.014)
	5	.991 (.022)	.629 (.017)	.924 (.020)	.713 (.019)	.765 (.020)	.747 (.020)	.711* (.019)	.834 (.021)
1000	1	.990 (.007)	.789 (.006)	.790* (.006)	.803 (.006)	.847 (.007)	.825 (.006)	.802 (.006)	.794 (.006)
	2	1.006 (.007)	.296 (.008)	.806 (.005)	.401 (.007)	.334 (.007)	.323* (.007)	.373 (.007)	.479 (.008)
	3	.998 (.007)	.076 (.000)	.080* (.000)	.179 (.004)	.120 (.003)	.116 (.004)	.126 (.003)	.081 (.001)
	4	1.004 (.006)	.413 (.003)	.502 (.004)	.469 (.004)	.450 (.004)	.449 (.003)	.446* (.004)	.459 (.004)
	5	1.007 (.005)	.624 (.005)	.912 (.005)	.650 (.005)	.683 (.007)	.658 (.006)	.644* (.005)	.711 (.005)
∞	1	1	.8	.8	.8	.8	.8	.8	.008
	2	1	.8	~.3	~.3	~.3	~.3	~.3	~.003
	3	1	.08	~.077	~.077	~.077	~.077	~.077	~.077
	4	1	.542	~.42	~.42	~.42	~.42	~.42	~.42
	5	1	.9	~.63	~.63	~.63	~.63	~.63	~.63

4. Separation of Estimates

n	MLE	Bayes	L.E.B.	m1	m2	m3	m1 ISO	m1 ADA
			a.	Distribution 1				
6	1	-1	0	0	1	2	0	0*
10	2	-0	0*	1	3	4	1	0
25	4	-0	0*	1	4	5	1	0
100	11	0	0*	2	6	6	2	1
1000	31	0	0*	2	9	6	2	1
			b.	Distribution 2				
6	1	-3	1	1	0	0*	1	1
10	3	-4	2	1	0	0*	1	2
25	9	-3	6	3	1	0*	3	5
100	23	-5	18	5	1	0*	4	11
1000	97	-2	80	11	2	0*	7	21
			c.	Distribution 3				
6	6	-8	1	3	1	0*	3	0
10	12	-11	0*	5	3	1	5	1
25	25	-10	0*	9	6	3	9	2
100	66	-10	0*	19	12	8	17	4
1000	254	-7	0*	48	27	15	30	2
			d.	Distribution 4				
6	2	-1	1	0	0	1	0	0*
10	6	-2	1	2	1	1	2	0*
25	13	-2	0	3	2	2	3	0*
100	29	-4	2	3	2	2	3	0*
1000	111	-8	14	6	4	1	0*	4
			e.	Distribution 5				
6	1	-4	0	0*	0	1	0	0
10	2	-3	1	0*	1	1	0	1
25	4	-4	3	0*	2	2	0	2
100	14	-4	11	0	3	2	0*	6
1000	72	-3	56	1	7	3	0*	14

5. Batting Averages and Their Estimates

i	Batting Average for Season remainder	Retransform of Stein's estimator L.E.B.	Retransform of G.E.B.
	p_i		m_1
1	.346	.290	.351
2	.298	.286	.340
3	.276	.281	.326
4	.222	.277	.310
5	.273	.273	.295
6	.270	.273	.295
7	.263	.268	.279
8	.210	.264	.264
9	.269	.259	.247
10	.230	.259	.247
11	.264	.254	.230
12	.256	.254	.230
13	.303	.254	.230
14	.264	.254	.230
15	.226	.254	.230
16	.285	.249	.214
17	.316	.244	.202
18	.200	.239	.196

6. Measurement Predictions

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\bar{\mu}$	142	110	106	164	133	147	153	125	148	140	133	139	141	148	147
MLE	166	148	165	161	111	131	129	143	111	140	157	116	138	143	149
LEB	132	131	133	137	138	140	140	142	142	143	143	146	147	148	149
m_1	120	120	124	133	134	139	141	143	143	147	147	153	156	158	159
m_2	122	122	124	137	139	142	143	144	144	146	150	146	151	152	152
m_3	130	130	131	135	142	144	144	144	144	146	148	147	146	147	146

7. Loss of Measurement Predictions for Example 2

<u>Estimator</u>	<u>Loss Function</u>
MLE	642
L.E.B	185
m_1	224
m_2	173
m_3	192

8. Data and Results for Third Example

μ_i	X_{i1}	X_{i2}	MLE	L.E.B.	m_1	m_2	m_3	m_1^{ADA}
-1.0	-1.11	-1.88	-1.49	-1.10	-.80	-.74	.05	-1.09
-0.5	.29	1.27	.31	1.0	.77	.83	1.13	1.00
0.0	2.18	-1.56	.78	.57	.41	.76	.91	.57
0.5	.88	1.20	1.04	1.25	.98	.87	1.04	1.24
-1.0	4.27	.460	2.37	2.47	1.90	1.30	.80	2.32
7.0	5.97	10.65	8.25	7.98	8.20	7.90	7.67	7.99
7.5	7.10	6.72	8.31	6.68	7.21	8.02	8.29	6.80
8.0	7.48	8.51	6.91	7.69	7.97	7.97	7.81	7.70
8.5	9.04	7.47	8.00	7.93	8.16	7.92	7.69	7.93

9. Values of Loss Function for Example 3.

Estimator	Initial Data Set	Second Data Set
MLE	.70	.88
L.E.B.	.82	.88
m_1	.50	.73
m_2	.45	.89
m_3	.74	.83
m_1^{ADA}	.75	.87

Table 10

n	Dist.	V_1	V_2	V_3
6	1	4.27 (.09)	.350 (.07)	.380 (.07)
	2	.432 (.07)	.453 (.06)	.494 (.07)
	3	.415 (.10)	.502 (.14)	.408 (.09)
	4	.258 (.06)	.226 (.06)	.222 (.06)
	5	.370 (.06)	.328 (.06)	.352 (.06)
20	1	.277 (.04)	.080 (.02)	.084 (.02)
	2	.129 (.032)	.197 (.032)	.140 (.032)
	3	.318 (.036)	.090 (.015)	.083 (.014)
	4	.152 (.018)	.054 (.01)	.048 (.009)
	5	.108 (.027)	.115 (.019)	.108 (.024)
100	1	.156 (.014)	.020 (.004)	.023 (.004)
	2	.034 (.006)	.093 (.012)	.022 (.005)
	3	.254 (.011)	.012 (.002)	.014 (.003)
	4	.129 (.011)	.024 (.004)	.017 (.003)
	5	.016 (.004)	.055 (.006)	.012 (.003)
1000	1	.130 (.004)	.0018 (.0003)	.0031 (.0006)
	2	.012 (.0013)	.092 (.0039)	.0027 (.0005)
	3	.253 (.004)	.0014 (.0003)	.0021 (.0004)
	4	.122 (.003)	.009 (.0009)	.002 (.0004)
	5	.007 (.001)	.044 (.003)	.002 (.0004)

REFERENCES

1. Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972), "Statistical inference under order restrictions". John Wiley and Sons, N.Y., 72-73.
2. Efron, B. and Morris, C. (1973) "Stein's Estimation Rule and Its Competitors - An Empirical Bayes Approach". Journal of the American Statistical Association 68 (March 73), 117-130.
3. Efron, B. and Morris, C. (1975) "Data Analysis Using Stein's Estimator and Its Generalizations". Journal of the American Statistical Association 70, 311-319.
4. James, W. and Stein, C. (1961) "Estimation with Quadratic Loss". Proceedings of the Fourth Berkeley Symposium on Math. Stat. and Prob., University of California Press 361-379.
5. Lehmann, E.L. (1975) "Nonparametrics". Holden-Day, Inc., San Francisco, 1975, 251.
6. Mandel (1964) "The Statistical Analysis of Experimental Data". Interscience Publishers, New York, 1964, Table 3.5.
7. Miyasawa, K. (1968) "An Empirical Bayes Estimator of the Mean of a Normal Population". Review of the International Statistical Institute.
8. Nadaraya, E. A. (1964) "On Non Parametric Estimates of Density Functions and Regression Curves". Theory of Probability and Applications 10, 186-190.
9. Robbins H. (1956) "An Empirical Bayes Approach to Statistics". Proceedings of the Third Berkeley Symposium on Math. Stat. and Prob., University of California Press, 157-163.

10. Robbins, H. (1964) "The Empirical Bayes Approach to Statistical Design Problems". Annals of Mathematical Statistics 35, 1-16.
11. Robbins, H. (1980) "An Empirical Bayes Estimation Problem", Proc. Natl. Acad. Sci. USA.
13. Schuster, E. (1969) "Estimation of a Probability Density Function and Its Derivatives". Annals of Mathematical Statistics, 40, 1187- 1195.
14. Silverman, B.W. (1978) "Choosing the window width when estimating a density". Biometrika 65, 1-11.15.
15. Stein, C. (1956) "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution". Proceedings of the Third Berkeley Symposium on Math. Stat. and Prob., Vol. 1, University of California Press, 197-206.

Statistical

35, 1-16.

Proc

Probability Function

40, 1187-1195.

1978

the

to

1

1956

1

1956

FIGURE A

FIGURE B

FIGURE C

FIGURE D

FIGURE E

Introduction: Morris [4] wrote a survey of parametric empirical Bayes (PEB) inference, followed by comments and by a rejoinder. We hope here to shed some more light on the history of the problem and on the PEB approach, restricting ourselves to the problem of estimating many means simultaneously assuming quadratic loss. Our approach will justify to some extent the use of PEB solutions and will allow us to compare them with other solutions.

The general empirical Bayes (GEB) approach for estimating many parameters was introduced in 1956 [8], in a continuation of basic ideas introduced in 1951 [7]. (Recently Forcina [3] pointed out that some of these ideas were anticipated by Gini). The work of Stein [12] had a tremendous impact on classical statistics. This work warned statisticians against considering one parameter at a time when many parameters were to be estimated simultaneously.

The problem, concerns N pairs (X_i, θ_i) where the X_i 's are observed, the distribution of the X_i 's given the θ_i 's is known or partially known, and the statistician's task is to estimate the unknown θ_i 's. In the GEB approach we assume the θ_i 's to be i.i.d. random variables with some unknown prior distribution. Our task is then to estimate the functional $E[\theta_i | X_i]$ using the X_i 's.

The Efron and Morris [2] approach to this problem and that presented by Morris [4] is that of GEB but considers only the possibility of a parametric family of priors, hence the name PEB.

This approach leads to useful results, but at least two questions arise. The first relates to the robustness of the solution with respect to a change in the family of priors. (This question might even be raised by a non-orthodox Bayesian; orthodox Bayesians do not have any use for GEB). The question can be rephrased by the non-Bayesian statistician as to how sensitive the procedure is with respect to changes in the empirical distribution of the θ_i 's? The second question relates to the fact that in order to phrase the problem in PEB one has to assume a parametric structure and it is not clear how to deal with problems where the parametric structure is missing.

From the classical point of view the θ_i 's are parameters, i.e.

fixed numbers that were not chosen at random. However, in many problems it is reasonable to require solutions to be symmetric in the θ_i 's. This leads us to look for solutions under the theoretical assumption that the order of the θ_i 's is uniformly distributed over all the $N!$ permutations. This approach will lead to solutions that are only slightly different from GEB solutions. Under mild conditions on the asymptotic behavior of the empirical distribution of the θ_i 's one can use the GEB analysis.

We believe that PEB solutions should then be investigated in the framework of GEB.

In a technical report [1] the PEB and GEB solutions were compared via Monte Carlo experimentation for the standard normal problem with one observation per population. It was shown there that, even for $N=10$, when the empirical distribution of the θ_i 's is strongly bi-modal the GEB solution is preferable to the PEB solution. As $N \rightarrow \infty$ the PEB is efficient in the normal case if and only if the conditional expectation of the θ_i 's given the X_i 's is linear in the X_i 's. Morris' examples of PEB are all natural exponential families with conjugate priors. These, as is shown in [5], satisfy the above condition namely $E[\theta_i | X_i] = a_i X_i + b_i$ where a_i and b_i are constants. This leads to a somewhat different approach where we look for estimators of θ_i that are linear in X_i and investigate their properties. We call this approach linear empirical Bayes (LEB). We start by asking what is the optimal linear procedure if the joint distribution of (θ_i, X_i) is known and then proceed to estimate the unknown constants. This approach may enable us to consider different functionals of the X_i 's depending on the problem, the number of populations and the ability to solve the problem. In some cases an adaptive procedure that is a convex combination of LEB and GEB might be useful and efficient [1].

In section 2 we take up the standard normal problem. We will show that LEB and PEB agree and we will summarize some results of [1].

In section 3 we deal with the normal problem with unknown variances and varying numbers of observations per population, and indicate some different possible LEB solutions. No PEB solution has been given for this problem.

In section 4 we take up the nonparametric problem of estimation of many means from populations with unknown variances with varying number of observations.

2. The Normal Case with Unit Variance and One Observation per Population

Let X_i given θ_i be independent r.v., each distributed as $N(\theta_i, 1)$, $i=1, 2, \dots, N$. We assume the θ_i to be i.i.d. distributed a priori as $N(\xi, \tau^2)$. The Bayesian estimate of θ_i under a quadratic loss function is then

$$(2.1) \quad \hat{\theta}_i = E[\theta | X_i] = \xi + \frac{\tau^2}{1+\tau^2} (X_i - \xi)$$

The PEB approach assumes that the θ prior is normal but that we do not know the parameters ξ and τ^2 , and use estimates of these parameters from the data. Efron and Morris [2] estimates are (for $N > 3$):

$$(2.2) \quad \hat{\xi} = \frac{1}{N} \sum_{i=1}^N X_i, \quad \hat{\tau}^2 = \left\{ \frac{1}{N-3} \sum_{i=1}^N (X_i - \bar{X})^2 - 1 \right\}^+$$

where $Y^+ = \begin{cases} Y & \text{if } Y > 0 \\ 0 & \text{if } Y < 0 \end{cases}$

The LEB approach starts with the formula

$$(2.3) \quad \hat{\theta}_i = E[\theta_i] + \frac{\text{Cov}(X_i, \theta_i)}{V(X_i)} (X_i - E[X_i])$$

Since

$$(2.4) \quad E(X_i) = E(\theta_i) = \xi$$

and

$$(2.5) \quad \text{Cov}(X_i, \theta_i) = V(\theta_i) = \tau^2$$

it follows that

$$(2.6) \quad V(X_i) = \tau^2 + 1$$

Thus (2.3) and (2.1) agree and hence PEB and LEB agree.

As for GEB, it is easy to verify that in the standard normal case

$$(2.7) \quad E[\theta_i | X_i] = X_i - \frac{h'(X_i)}{h(X_i)}$$

where

$$(2.8) \quad h(x) = \int \phi(x-\theta) dG(\theta)$$

Here $G(\cdot)$ is the prior of θ and $\phi(\cdot)$ is the standard normal density, so that $h(\cdot)$ is the marginal density of X_i . Under the EB approach $G(\cdot)$ is not known and hence $h(\cdot)$ and $h'(\cdot)$ have to be estimated from the data.

In [1] different methods for the estimation of h and h' were introduced using kernel estimators. These estimators were compared via Monte Carlo experimentation with LEB and with the true Bayes solution. The experimentation was done under five different priors with $N = 6, 10, 25, 100, 1000$. The priors were

2. $1/2$ of the θ_i 's = -2 and $1/2$ of the θ_i 's = 2.
3. $\theta_i \sim U(0,1)$
4. $\theta_i \sim \exp(1)$
5. $1/2$ of the θ_i 's $\sim N(-3,1)$ and $1/2$ of the θ_i 's $\sim N(3,1)$
6. $\theta_i \sim N(0,4)$

It is interesting to note that already for $N=10$ with priors 5 and 2 the GEB average loss was less than the average LEB loss. For $N=25$ or more the LEB performed better for priors 1 and 3 only. This indicates strongly that LEB should not be used if the empirical distribution of the θ_i 's is suspected not to be unimodal. One can use an adaptive procedure that is a convex combination of LEB and GEB. In [1] it is shown that such a procedure does behave asymptotically like LEB whenever $E[\theta_i | X_i]$ is linear and behaves asymptotically like GEB when $E[\theta_i | X_i]$ is not linear.

3. The Normal Case with Unequal Variances and Unequal Number of Observations

Let X_{ij} , $j = 1, 2, \dots, n_i$, $i = 1, 2, \dots, N$, conditionally on (θ_i, V_i) be independent and distributed $N(\theta_i, V_i)$.

We denote

$$(3.1) \quad \bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2.$$

We start like Morris [4] and assuming the V_i 's to be known. Then if LEB and PEB agree and we have

$$(3.2) \quad \hat{\theta}_i = \frac{V(\theta_i)}{V(X_i)} (X_i - E[X_i]) + E[\theta_i]$$

There are various ways of estimating $E[\theta]$, $E[X]$, $V(\theta)$ and $V(X_i)$. One reasonable way is

$$(3.3) \quad E[X_i] = E[\theta] \hat{=} \frac{1}{\sum_{i=1}^N \frac{n_i}{V_i}} \sum_{i=1}^N \frac{n_i}{V_i} X_i$$

$$(3.4) \quad V(\theta) \hat{=} \left\{ \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2 - \frac{1}{N} \sum_{i=1}^N \frac{V_i}{n_i} \right\}^+$$

and

$$(3.5) \quad V(X_i) \hat{=} \left\{ \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2 - \frac{1}{N} \sum_{i=1}^N \frac{V_i}{n_i} \right\}^+ + \frac{V_i}{n_i}$$

If the V_i 's are not known PEB approach should consider the conjugate prior on (θ_i, V_i) solve for $E[\theta_i | X_i]$ and estimate the parameters in the solution. Morris [4] does not suggest this, he says "We will take the variances V_i to be known here, in order to concentrate on the main issues. Usually, however, they would have to be estimated, perhaps by the within groups sums of squares". This means that Morris adheres to LEB and suggests using $V_i \hat{=} S_i^2$ in formulas (3.4) and (3.5). However, if the variances are not known it might be more appropriate to use LEB estimators for the variances.

As is shown in [9] we can take,

$$(3.6) \quad v_i = \bar{S}^2 + \left[1 - \frac{R^4 + (\bar{S}^2)^2}{R^4 + (\bar{S}^2) + r_i \{R^4 - \bar{C}(\bar{S}^2)^2\}^+} \right] (S_i^2 - \bar{S}^2)$$

where

$$(3.7) \quad r_i = \frac{n_i - 1}{2}, \quad \bar{S}^2 = \frac{1}{N} \sum_{i=1}^N S_i^2, \quad R^4 = \frac{1}{N-1} \sum_{i=1}^N (S_i^2 - \bar{S}^2)^2, \quad \bar{C} = \frac{1}{N} \sum_{i=1}^N \frac{1}{r_i}.$$

4. Estimating Many Means in General

A classical statistician who is not willing to put a prior probability on the parameter space might be willing to accept the requirement that the method of estimation should be symmetric in the populations. This requirement will lead to formulas that are similar to those derived under the EB approach. We will take here a non-parametric approach. Consider $\{F_i: i=1,2,\dots,N\}$ a finite collection of distribution functions (representing the populations). We assume $\int_{-\infty}^{\infty} t^2 dF_i(t) < \infty$ for $i=1,2,\dots,N$. We sample X_{ij} , $j=1,2,\dots,n_i$, from population i . i.e. Given F_i , X_{ij} i.i.d. random variables with distribution F_i . Since we want a symmetric analysis in F_i , we assume the order of the F_i is uniformly distributed over all N_i permutations.

We define

$$(4.1) \quad E[X_{ij} | F_i] = \theta_i, \quad E[\theta_i] = \frac{1}{N} \sum_{i=1}^N \theta_i = \bar{\theta}, \quad V(\theta) = \frac{1}{N} \sum_{i=1}^N (\theta_i - \bar{\theta})^2.$$

$$V(X_{ij} | F_i) = \sigma_i^2, \quad E(\sigma_i^2) = \frac{1}{N} \sum_{i=1}^N \sigma_i^2 = \bar{\sigma}^2$$

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N \bar{X}_i, \quad S^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{X}_i - \bar{X})^2$$

We recall that the best linear predictor of θ_i given X_i is:

$$(4.2) \quad t(X_i) = E[\theta_i] + \frac{\text{Cov}(X_i, \theta_i)}{V(X_i)} (X_i - E[X_i]).$$

We note,

$$(4.3) \quad E[X_i] = \bar{\theta}, \quad V(X_i | F_i) = \frac{\sigma_i^2}{n_i}, \quad E[V(X_i | F_i)] = \frac{\overline{\sigma^2}}{n_i}$$

$$\begin{aligned} \text{Cov}(X_i, \theta_i) &= V(\theta), \quad V(X_i) = E[V(X_i | F_i)] + V(E[X_i | F_i]) = \\ &= \frac{\overline{\sigma^2}}{n_i} + V(\theta) \end{aligned}$$

Hence there is need to estimate $\bar{\theta}$, $\overline{\sigma^2}$ and $V(\theta)$. We can use:

$$(4.4) \quad \bar{\theta} \hat{=} \bar{X}$$

$$(4.5) \quad \overline{\sigma^2} \hat{=} \frac{1}{N} \sum_{i=1}^N S_i^2 = \overline{S^2}$$

As for $V(\theta)$ we can verify that

$$\begin{aligned} (4.6) \quad E[S^2] &= \frac{1}{N} \sum_{i=1}^N V(X_i) + \frac{1}{N-1} V(\theta) \\ &= \frac{1}{N} \sum_{i=1}^N E[V(X_i | F_i)] + \frac{1}{N-1} V(\theta) \end{aligned}$$

Hence we can use

$$(4.7) \quad V(\theta) \hat{=} \frac{N-1}{N} \left\{ S^2 - \frac{\overline{S^2}}{N} \sum_{i=1}^N \frac{1}{n_i} \right\}^+$$

and thus

$$(4.8) \quad t(X_i) \hat{=} \bar{X} + \frac{\frac{N-1}{N} \left\{ S^2 - \frac{\overline{S^2}}{N} \sum_{i=1}^N \frac{1}{n_i} \right\}^+}{\frac{N-1}{N} \left\{ S^2 - \frac{\overline{S^2}}{N} \sum_{i=1}^N \frac{1}{n_i} \right\}^+ + \frac{\overline{S^2}}{n_i}} (X_i - \bar{X})$$

For large N the reduction in the risk is by the factor

of $\left(1 - \frac{V(\theta)}{\overline{\sigma^2} + V(\theta)}\right)$. This means that when $\frac{\overline{\sigma^2}}{n_i}$ is not negligible

compared with $V(\theta)$ this method of estimation is useful. We do not know how to approach this problem via the PEB approach. If the LEB approach is used assuming the F_i to be i.i.d. from some unknown

collection of distribution functions the only difference would be that in (4.7) the factor $\frac{N}{N-1}$ will be deleted. i.e. no difference in the solution for large N. This approach works nicely also for the estimation of many variances simultaneously [6].

It is interesting to note that in (4.8) $\hat{a}_{iN}$ and $\hat{b}_{iN}$ converge as $N \rightarrow \infty$ a.s. to a_i and b_i respectively, where

$$(4.9) \quad \hat{a}_{iN} = \frac{\frac{N-1}{N} \{ S^2 - \frac{S^2}{N} \sum_{i=1}^N \frac{1}{n_i} \}^+}{\frac{N-1}{N} \{ S^2 - \frac{S^2}{N} \sum_{i=1}^N \frac{1}{n_i} \}^+ + \frac{S^2}{n_i}}$$

$$\hat{b}_{iN} = (1 - \hat{a}_{iN}) \bar{X}$$

and

$$(4.10) \quad a_i = \frac{V(\theta)}{V(X_i)}$$

$$b_i = (1 - a_i) E(\theta)$$

under the assumption $\bar{\theta}_N \rightarrow E[\theta]$ and $\frac{\sum_{i=1}^N (\theta_i - \bar{\theta})^2}{N} \rightarrow V(\theta)$ as $N \rightarrow \infty$.

While replacing $V(X_i)$ by its estimator (3.5) where V_i are not known and estimated via in group sum of squares or by the LEB variance estimates, leaves $\hat{a}_{iN}$ and $\hat{b}_{iN}$ random variables as $N \rightarrow \infty$.

References

- [1] Campbell, G., Robbins, H., Yahav, J. (1980) "Linear and General Empirical Bayes Estimation, the Normal Case", Brookhaven National Laboratory Technical Report 29247.
- [2] Efron B. and Morris C. (1973) "Stein's Estimation Rule and Its Competitors - an Empirical Bayes Approach" J. Amer. Statist. Assoc. 68 117-130.
- [3] Forcina, A. (1982) "Gini's Contribution to the Theory of Inference", Int. Statist. Rev. 50 65-70.
- [4] Morris, C. (1983) "Parametric Empirical Bayes Inference: Theory and Applications". J. Amer. Statist. Assoc. 78 47-55.
- [5] Morris, C. (1983) "Natural exponential families with quadratic variance functions: Statistical Theory", Annals of Statistics 515-529.
- [6] Peierls, R. and Yahav, J. (1983) "Sequential allocation of observations for the purpose of estimating a linear combination of means", Brookhaven National Laboratory Technical Report.
- [7] Robbins, H. (1951) "Asymptotically subminimax solutions of compound statistical decision problems", Proc. of 2nd Berkeley Symp. Math. Statist. Prob., Univ. of Calif. Press.
- [8] Robbins, H. (1956) "An empirical bayes approach to statistics", Proc. 3rd Berkeley Symp. Math. Statist. Prob. 157-163, Univ. of Calif. Press.
- [9] Robbins, H. (1982) "Estimating many variances", Statist. Decision Theory and Related Topics III Vol. 2 218-226. Academic Press, N.Y.
- [10] Robbins, H. (1983) "Some thoughts on empirical Bayes estimation", Annals of Statistics.
- [11] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, Proc. 3rd Berkeley Symp. Math. Statist. Prob. 192-206. Univ. of Calif. Press.

On Clustering Points in R^1 : Results and Anomalies

Peter J. Bickel

Abba M. Krieger

Joseph Yahav

Preliminary Draft

Abstract

We assume that we are given points $x_1 < x_2 < \dots < x_n$ on the real line and a distance function $d: \mathbb{R}^2 \rightarrow \mathbb{R}^+ \cup 0$. The problem is to divide the n points into k clusters. Two variations of the problem are considered. In one case, the objective function is composed of all pairwise distances between points in the cluster, and in the second case, distances between points in the cluster to optimally chosen centers of the clusters are used. The thrust in each case is to find conditions under which clusters are formed by consecutive points.

Key words: Cluster analysis; k-median problem; connected sets; convex functions.

Introduction

There are many applications which require that individuals or items be divided into groups commonly called clusters. For example, data files are stored on one of several disks for future access; the files that are assigned to each of the disks represent the clusters. Meteorological satellites or stations are positioned to obtain data on the weather; the points in the atmosphere that are assigned to each station define the clusters. In a statistical framework, observations are sometimes clustered for descriptive purposes or often as a means for tailoring analyses to subsets of the observations.

Although the governing principle in cluster analysis is to achieve homogeneity within clusters and heterogeneity among clusters, the objective or loss function in most applications is vague. Nevertheless, we carve out a well-defined, albeit limited in scope, mathematical problem from this setting and attempt to prove theorems and provide insights. In particular, we assume that the items or individuals can be represented by points on the real line and the objective or loss function can be described by a distance function between pairs of points. We focus on the following property of a clustering scheme.

Conjecture: Given a set S of points on the real line, the clusters are formed by connected subsets of S .

Where

Definition: A subset $A \subset S$ is connected in S if there do not exist $x_1, x_2 \in A$ and $y \in S - A$ such that $x_1 < y < x_2$.

Our aim is to determine the conditions on the distance function and the representation of the objective function in terms of the distance function so that the above conjecture holds.

The applications above suggest two different formulations of the objective function in terms of the distance function. In the data file

problem, the total distance within a cluster is comprised of the pairwise distances of the points in the cluster. We term this problem the interpoint model. It turns out that the conjecture is true under limited conditions for this model. We defer the mathematical development of this problem to the third section. In the meteorological problem, the total distance within a cluster is comprised of the distances from the points in the cluster to the center of the cluster. The positions of the centers in conjunction with the assignments of points to clusters are chosen in order to minimize the total within cluster distance. This is commonly called the gravity model in the Engineering literature and the k-median problem in the Operations Research literature and is discussed in the next section. It is interesting to note that the k-median problem is known to be NP hard. This means that the time required by an algorithm to obtain the optimal solution increases exponentially in the parameters of problem. If the above conjecture is true, then the number of possible solutions decreases from $\sum_{j=0}^{k-1} (-1)^j \binom{k}{j} (k-j)^n / k!$ to $\binom{n-1}{k-1}$ assuming that no cluster is empty and the clusters are indistinguishable (i.e., the issue is to determine which points are in the same clusters; labelling of the clusters is irrelevant) where n and k denote the number of points and clusters respectively. For example, if k=3 and n=19, then the number of possible solutions is reduced from 9330 to 36.

2. Gravity Model

In this section, we show that the conjecture stated in the introduction applies to the gravity model for general distance functions. To this end, we denote the centers by $c_1, \dots, c_k$ and let

$$y_{ij} = \begin{cases} 1 & \text{if } x_i \text{ is assigned to cluster } j \\ 0 & \text{otherwise} \end{cases}$$

Remember that x_i , $i = 1, \dots, n$ are in nondecreasing order. The problem is then to choose $c_1, \dots, c_k$ and y_{ij} ; $i = 1, \dots, n$ and $j = 1, \dots, k$ in order to minimize

$$L(c, y) = \sum_{i=1}^n \sum_{j=1}^k d(x_i, c_j) y_{ij} \quad (2.1)$$

We need to place the following modest regularity condition on d ,

$$d(x_2, y_2) \leq d(x_1, y_1) \quad \text{if } x_1 \leq x_2 \leq y_2 \leq y_1. \quad (2.2)$$

We also assume throughout that d is reflexive.

Proposition 2.1 If (2.2) is satisfied, then the clusters (i.e. choice of y_{ij}) that minimize (2.1) are connected subsets of $S = \{x_1, \dots, x_n\}$.

Proof We assume that $c_1, \dots, c_k$ are fixed. If all clusters are not connected subsets of S , that implies there exists x_1 and x_2 in one cluster and y in another such that $x_1 < y < x_2$. Let c_1 denote the center of the cluster that contains x_1 and x_2 and c_2 the center that contains y . If $c_2 \in [x_1, x_2]$ and $c_1 < c_2$, x_2 could be assigned to c_2 , since $d(x_2, c_2) \leq d(x_2, c_1)$. If $c_2 \in [x_1, x_2]$ and $c_1 > c_2$, x_1 could be assigned to c_2 , since $d(x_1, c_2) \leq d(x_1, c_1)$. Hence $c_2 \notin [x_1, x_2]$. Furthermore, if $c_2 > x_2$, then $c_1 < y$ since y is assumed to be assigned to c_2 . But, it follows that c_1 is the center for y from $d(y, c_2) > d(x_2, c_2) > d(x_2, c_1) > d(y, c_1)$.

which is a contradiction. Similarly, if $c_2 < x_1$, then $c_1 > y$ since c_2 is assumed to be the center for y . But, it follows that c_1 is the center for y from $d(y, c_2) > d(x_1, c_2) > d(x_1, c_1) > d(y, c_1)$ which is also a contradiction.

3.1 Introduction

We restrict the problem to two clusters to simplify notation. Also, the extension to an arbitrary number of clusters is relatively easy once results for two clusters are available.

The problem is to find a subset A of $S = \{x_1, \dots, x_n\}$; this defines the first cluster and hence implicitly the second cluster. The objective is then to minimize over A

$$\psi(A) \equiv \sum_{x_i \in A} \sum_{x_j \in A} d(x_i, x_j) / f(a) + \sum_{x_i \in S-A} \sum_{x_j \in S-A} d(x_i, x_j) / f(n-a) \quad (3.1.1)$$

where d is the distance function, a is the number of points in A and $f(a)$ is the normalizing function that weights the sum of intracluster distances by the size of the cluster. Intuitive choices of $f(a)$ include:

(i) $f(a) = a^2$: since there are a^2 pairwise distances in each cluster

(ii) $f(a) = a$: then $\psi(A)/n$ equals the expected distance between two randomly chosen points within a cluster assuming that clusters are chosen proportional to size. Furthermore, if $d(x_i, x_j) = (x_i - x_j)^2$, then

$$\sum_{x_i \in A} \sum_{x_j \in A} d(x_i, x_j) / a = 2 \sum_{x_i \in A} (x_i - \bar{x}_A)^2 \quad \text{where } \bar{x}_A = (\sum_{x_i \in A} x_i) / a$$

Substitution of this equation and an equivalent expression for the second cluster into (3.1.1) shows that the result for the interpoint model with squared distance is the same as Analysis of Variance and hence the gravity model.

The issue is then to find conditions on d and f such that the set A that minimizes $\psi(A)$ is connected in S . We divide the analysis into two parts. In section 3.2, we present counterexamples in order to limit the scope of choices of d to convex functions in $|x_j - x_i|$ and $f(a)$ to a . In section 3.3

we then prove that for the above choices for a and 1 (3.1.1) is minimized when A is connected.

3.2 Counterexamples

We first show that the conjecture is not true in general by considering:

$$d(x, y) = \begin{cases} 0 & \text{if } |x-y| \leq 1/4 \\ 1 & \text{if } |x-y| > 1/4 \end{cases}$$

Example 1 $f(a) = a$

and

$$S = \{ x_1 = 0, x_2 = 3/8, x_3 = 5/8 \text{ and } x_4 = 1 \}$$

For this example $A = \{x_1, x_4\}$ minimizes $\psi(A)$.

It is clear that d can be modified to be continuous without altering example

1. The condition is then not continuity but rather convexity of d . This observation will follow from the mathematical development in section 3.3.

Since example 1 does not provide a counterexample to the conjecture when $f(a) = a^2$ we consider

Example 2 $d(x, y)$ as in example 1

$$f(a) = a^2$$

and

$$S = \{ x_1 = 0, x_2 = x_3 = 1/2, \text{ and } x_4 = x_5 = x_6 = 1 \}$$

For this example $A = \{x_1, x_4, x_5, x_6\}$

We next consider Euclidean distance and demonstrate that A need not be connected if $f(a) = a^2$

Example 3 $d(x, y) = |x-y|$

$$f(a) = a^2$$

and

$$S = \{ x_1 = 0, x_2 = x_3 = .4, x_4 = \dots = x_{13} = 1 \}$$

For this example $A = \{x_1, x_4, \dots, x_{13}\}$ - 67-

The above examples suggest that we focus on $d(x, y) = g(|x - y|)$ where g is convex and $f(a) = a$. Furthermore, it follows from example 3 that even if Euclidean distance is assumed and $f(a) = a$, it is still necessary to minimize (3.1.1) over sets of differing sizes a ; if a is fixed, for example $a = 11$ in example 3, then $\psi(A)$ is minimized by the set $A = \{x_1, x_4, \dots, x_{13}\}$ which is not connected in S .

3.3 Mathematical Results

We consider the problem of finding the subset A of S that minimizes

$$\psi(A) = \sum_{i \in A} \sum_{j \in A} d(x_i, x_j) / a + \sum_{i \in S-A} \sum_{j \in S-A} d(x_i, x_j) / (n-a) \quad (3.3.1)$$

where $S = \{x_1, \dots, x_n\}$ with $x_1 < x_2 < \dots < x_n$ and a denotes the number of elements in A . In particular, we show that if d is convex in $|x_j - x_i|$ and without loss of generality $d(x_i, x_i) = 0$, then $\psi(A)$ is minimized by a subset A of S so that A and $S-A$ are connected. Implicit in (3.3.1) is that the n points are divided into two clusters. The extension to an arbitrary number of clusters follows straightforwardly.

Remark: Technically $\psi(S)$ (or equivalently $\psi(\emptyset)$ where $\emptyset$ denotes the null set) is not well defined, although it is clear from the application that the second (first) term in (3.3.1) is zero. It is interesting to note that (3.3.1) cannot obtain a unique minimum at $A=S$. To see that this is the case let x_m be the value from among $x_i, i=1, \dots, n$ that maximizes $\sum_{j=1}^n d(x_i, x_j)$.

Then

$$\begin{aligned} \psi(S) &= \sum_{i=1}^n \sum_{j=1}^n d(x_i, x_j) / n > \frac{n}{n-1} \sum_{i \neq m} \sum_{j=1}^n d(x_i, x_j) / n > \sum_{i \neq m} \sum_{j \neq m} d(x_i, x_j) / (n-1) \\ &= \psi(S - x_m). \end{aligned}$$

The main results are more easily derived from an equivalent but different

form of (3.3.1). To this end, we let

$$f(i; A) = \begin{cases} 1/a & \text{if } x_i \in A \\ 0 & \text{otherwise} \end{cases} \quad i = 1, \dots, n$$

$$\text{and } g(i; A) = f(i; S-A)$$

We now have

$$\psi(A) = a \sum_{i=1}^n \sum_{j=1}^n d(x_i, x_j) f(i; A) f(j; A) + (n-a) \sum_{i=1}^n \sum_{j=1}^n d(x_i, x_j) g(i; A) g(j; A)$$

This implies that minimizing (3.3.1) over A is equivalent to minimizing

$$\psi_1(A) = p_a \sum_{i=1}^n \sum_{j=1}^n d(x_i, x_j) f(i; A) f(j; A) + q_a \sum_{i=1}^n \sum_{j=1}^n d(x_i, x_j) g(i; A) g(j; A) \quad (3.3.2)$$

where $p_a = a/n$ and $q_a = 1-p_a$. It is important to note that

$$p_a f(i; a) + q_a g(i; a) = 1/n \quad \forall i. \quad (3.3.3)$$

We present and prove Abel's formula, the discrete analogue to integration by parts, since it is used several times in modifying $\psi_1(A)$ into a more workable form.

Lemma 1

Let $z_1 < \dots < z_n$. Then

$$\sum_{i=1}^n h(z_i) f(z_i) = h(z_n) F(z_n) - \sum_{i=1}^{n-1} h'(z_i) F(z_i) \quad (3.3.4)$$

where $F(z_i) = \sum_{j=1}^i f(z_j)$ and $h'(z_i) = h(z_{i+1}) - h(z_i)$.

$$\begin{aligned}
 &= \sum_{i=1}^n h(z_i) F(z_i) - \sum_{i=1}^{n-1} h(z_{i+1}) F(z_i) \\
 &= h(z_n) F(z_n) - \sum_{i=1}^{n-1} h'(z_i) F(z_i).
 \end{aligned}$$

We repeatedly use lemma 1 to show

Lemma 2 Minimizing $\psi_1(A)$ over A is equivalent to maximizing over A

$$\psi_2(A) = (p_a/q_a) \sum_{1 \leq i < j \leq n-1} \alpha_{ij} (F(i;A) - i/n)(F(j;A) - j/n) \quad (3.3.5)$$

where

$$F(i;A) = \sum_{j=1}^i f(j;A)$$

and

$$\alpha_{ij} = \begin{cases} d(x_{i+1}, x_i) & \text{if } i=j \\ d(x_{j+1}, x_i) + d(x_i, x_{j+1}) - d(x_i, x_j) - d(x_{i+1}, x_{j+1}) & \text{if } i < j \end{cases}$$

Proof Consider

$$T = \sum_{1 \leq i < j} d(x_i, x_j) f(i;A) f(j;A) = \sum_{i=1}^n \left[\sum_{j=1}^n d(x_i, x_j) f(j;A) \right] f(i;A) \quad (3.3.6)$$

From lemma 1, the term in brackets in (3.3.6) can be written as

$$\begin{aligned}
 B_i &= d(x_i, x_n)(1 - F(i-1;A)) - \sum_{j=1}^{n-1} (d(x_i, x_{j+1}) - d(x_i, x_j))(F(j;A) - F(i-1;A)) \\
 &= d(x_i, x_n) - \sum_{j=1}^{n-1} (d(x_i, x_{j+1}) - d(x_i, x_j)) F(j;A); \quad i=1, \dots, n-1
 \end{aligned} \quad (3.3.7)$$

where $F(0;A) \equiv 0$.

Substitution of (3.3.7) into (3.3.6) yields

$$T = \sum_{i=1}^n d(x_i, x_n) f(i;A) - \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} (d(x_i, x_{j+1}) - d(x_i, x_j)) F(j;A) f(i;A) \quad (3.3.8)$$

We now use lemma 1 on the first term in (3.3.8) to obtain

$$\begin{aligned} T_1 &= d(x_n, x_n) - \sum_{i=1}^{n-1} (d(x_{i+1}, x_n) - d(x_i, x_n)) F(i; A) \\ &= \sum_{i=1}^{n-1} (d(x_i, x_n) - d(x_{i+1}, x_n)) F(i; A) \end{aligned} \quad (3.3.9)$$

We next consider the second term in (3.3.8)

$$T_2 = \sum_{j=1}^{n-1} \left[\sum_{i=1}^j (d(x_i, x_{j+1}) - d(x_i, x_j)) f(i; A) \right] F(j; A) \quad (3.3.10)$$

More specifically, the term in brackets in (3.3.10) can be rewritten using lemma 1 as

$$C_j = \alpha_{jj} F(j; A) + \sum_{i=1}^{j-1} \alpha_{ij} F(i; A); \quad j=1, \dots, n-1 \quad (3.3.11)$$

where the sum in (3.3.11) is zero when $j=1$.

By replacing the term in brackets in (3.3.10) with (3.3.11)

$$T_2 = \sum_{1 \leq i < j \leq n-1} \alpha_{ij} F(i; A) F(j; A) \quad (3.3.12)$$

The expression for $T = T_1 - T_2$ replaces the first double summation in (3.3.2) and the same argument leads to a similar expression for the second double summation so

(3.3.13)

$$= p_a \sum_{i=1}^n \sum_{j < n-1} \alpha_{ij} F(i;A) F(j;A) - q_a \sum_{i=1}^n \sum_{j < n-1} \alpha_{ij} G(i;A) G(j;A).$$

To simplify (3.3.13) we use (3.3.3) which implies that $p_a F(i;a) + q_a G(i;a) = 1/n \forall i$. Hence the first two terms in (3.3.13) when combined do not depend on A and

$$p_a F(i;A) F(j;A) + q_a G(i;A) G(j;A) =$$

$$p_a F(i;A) F(j;A) + (1/n - p_a F(i;A))(j/n - p_a F(j;A))/q_a = \quad (3.3.14)$$

$$p_a/q_a (F(i;A) - 1/n) (F(j;A) - j/n) + 1j/n^2.$$

Finally,

$$\psi_1(A) = \sum_{i=1}^{n-1} (d(x_i, x_n) - d(x_{i+1}, x_n)) 1/n - \sum_{i=1}^n \sum_{j < n-1} \alpha_{ij} 1j/n^2 - \psi_2(A)$$

Since only the last term depends on A the lemma is proved.

Ostensibly what we have done is to use indicator functions to separate out the interpoint distances, α_{ij} , from the choice of the set A, $F(i;A)$. As a result, the representation of interpoint distances within clusters by $\psi_2(A)$, aside from a constant, provides interesting insights.

(i) The conjecture is equivalent to showing that $\psi_2(A)$ is maximized when $F(i;A)$ is as large as possible (i.e. $A = \{x_1, \dots, x_a\}$) or as small as possible

(i.e., $A = \{x_{n-a+1}, \dots, x_n\}$).

(ii) The structure on d is imbedded in α_{ij} . If d is convex in $|x_j - x_i|$ then $\alpha_{ij} > 0$ and if we use Euclidean distance then

α_{ii} is $x_{i+1} - x_i > 0$ and $\alpha_{ij} = 0$ if $i \neq j$. It suffices to show that

$\psi_2(A)$ when viewed as a quadratic form in $F(i;A) = 1/n$, $i=1, \dots, n$ is maximized when $F(i;A)$ is as large as possible for non-negative weights.

(iii) An immediate consequence is that if we restrict A to have half of the elements of S and d is convex in $|x-y|$ then $\psi_2(A)$ is maximized for $A = \{x_1, \dots, x_{n/2}\}$ assuming n is even.

The proofs of the results that lead to the basic theorem - if d is convex in $|x-y|$ then $\psi(A)$ is minimized when A and $S-A$ are connected in S - are somewhat complicated and better described graphically rather than algebraically. In order to modify the problem into graphical form we translate $F(i;A)$ to $F_A(x): [0,1] \rightarrow [0,1]$ as follows:

$$(i) \quad F_A(i/n) = F(i;A); \quad i = 0, 1, \dots, n$$

$$(ii) \quad F_A(x) = n[F_A(i/n) - F_A((i-1)/n)](x - (i-1)/n) + F_A(i/n) \text{ for } (i-1)/n < x < i/n$$

(i.e. linear interpolation between $F_A((i-1)/n)$ and $F_A(i/n)$) Similarly, $1/n$ is

replaced by $F(x)=x \quad \forall x \in [0,1]$ so that $F(i;A) - 1/n = F_A(i/n) - F(i/n)$. Note that

any A gives rise to a function $F_A(x)$ which has a slope of zero or n/a for

$(i-1)/n < x < i/n$ depending on whether $x_i \notin A$ or $x_i \in A$ respectively. For

example, $F_A(x)$ corresponding to $n=5$ and $A = \{x_1, x_4\}$ appears in figure 3.1.

Figure 3.1

We first focus on maximizing $|F_A(x) - F(x)|$ uniformly in x over all sets

A with a elements. Clearly if a maximum is obtained when $F_A(x)$ has slope n/a

for $x \leq a/n$ and $F_A(x) = 1$ for $x > a/n$ (i.e. $A = \{x_1, \dots, x_a\}$) or $F_A(x) = 0$

for $x \leq 1-a/n$ and $F_A(x)$ has slope n/a for $x > 1-a/n$ we are done.

Unfortunately, this is not the case. What is true, however, is that A must be of the form $\{x_1, \dots, x_i, x_j, \dots, x_n\}$ where

$i \leq j$ and $i+n+1-j = a$. To prove this we first observe that if

$F_A(x) > F(x)$ for $(i-1)/n \leq x \leq (i+1)/n$ and $F_A(x)$ has slope zero from $(i-1)/n$ to i/n and slope n/a from i/n to $(i+1)/n$ (i.e. $x_i \notin A$ and $x_{i+1} \in A$) the $|F_A(x) - F(x)| < |F_{A'}(x) - F(x)|$ where A' is identical to A except that x_i is included in A' and x_{i+1} is removed from A' (see figure 3.2). A similar argument follows if $F_A(x) < F(x)$ (see figure 3.2).

Figure 3.2

Lemma 3 For fixed a , $|F_A(x) - F(x)|$ is maximized uniformly in x if $F_A(x)$ crosses $F(x)$ at most once in $x \in (0,1)$.

Proof Since $\psi(A) = \psi(S-A)$ we can assume $a \leq n/2$. Let $F_A(x)$ be a function which crosses $F(x)$ at least twice. This implies that $F_A(x)$ is at least four-piecewise linear (see figure 3.3). Let x_i denote the right breakpoints of the last four pieces; $i=1,2,3,4$. For purposes of discussion we assume that the slopes are $n/a, 0, n/a$ and 0 since the proof is similar if the slopes are $0,$

n/a , 0 and n/a . we draw two parallelograms. The first parallelogram replaces the last two linear pieces with one with slope 0 and the other with slope n/a and the second parallelogram replaces the second and third pieces with lines with slopes n/a and 0. In either case,

$F_{A_1}(x)$ and $F_{A_2}(x)$ crosses $F(x)$ one less time than $F_A(x)$ crosses $F(x)$. We show that either

$|F_A(x) - F(x)| < |F_{A_1}(x) - F(x)| \forall x$ or $|F_A(x) - F(x)| < |F_{A_2}(x) - F(x)| \forall x$. This proves the lemma since we can apply this argument successively until only three linear pieces remain. Assume that there exists an x such that

$|F_{A_1}(x) - F(x)| < |F_A(x) - F(x)|$. This implies that $F_A(x_3) - F(x_3) > F(x_3) - F_{A_1}(x_3)$ since $F_A(x) = F_{A_1}(x) \forall x < x_2$ and $(F_A(x) - F(x)) - (F(x) - F_{A_1}(x))$ is maximum at $x = x_3$ (see figure 3.3). Similarly, if there exists an x such that

$|F_{A_2}(x) - F(x)| < |F_A(x) - F(x)|$ this implies $F_{A_2}(x_2) - F(x_2) < F(x_2) - F_{A_2}(x_2)$.

It is apparent from figure 3.3 that both of these inequalities cannot hold simultaneously since

$$F_{A_2}(x_2) - F(x_2) > F_A(x_3) - F(x_3) \text{ and } F(x_3) - F_{A_1}(x_3) > F(x_2) - F_{A_2}(x_2).$$

Figure 3.3

Remarks

1. The lemma states that in terms of maximizing $(F_A(x)-F(x))$ we need only consider sets A that include only the smallest x_1 or only the largest x_1 or some of the smallest x_1 and largest x_1 . For fixed a we cannot do better in the sense of limiting A further as is apparent from example 3 in section 3.2.
2. In terms of Euclidean distance, since $a_{ij} = 0$ if $i \neq j$ then the set A for fixed a must either be connected in S or S-A must be connected in S.
3. If for example,
 $A = \{x_1, x_2, \dots, x_1, x_j, \dots, x_k\}$ where $1 < j < k < n$ then the proof of the lemma says that for Euclidean distance A can be replaced by A_1 which substitutes for $x_j, \dots, x_k$ the last $k-j+1$ items in S or by A_2 which substitutes for $x_j, \dots, x_k$ the elements $x_{i+1}, \dots, x_{i+k-j+1}$.

We showed that for any subset A of S with a elements we can find a set B of the form $\{x_1, \dots, x_b, x_{n-a+b+1}, \dots, x_n\}$ such that $|F_A(x)-F(x)| < |F_B(x)-F(x)| \forall x$. Note that for the remainder of this section, without loss of generality, we assume $x_1 \in A$ since $\psi(A) = \psi(S-A)$. If $b=a$ (i.e. B consists of the smallest a elements in S) we are done. Our only concern is when $b < a$ (see figure 3.4).

Figure 3.4

Up to this point, we compared $\psi(A)$ over sets A with the same number of elements a . We now prove that there is a set C of necessarily a different number of elements c such that $(p_a/q_a) |F_B(x) - F(x)|^2 < (p_c/q_c) |F_C(x) - F(x)|^2$ where $C = \{x_1, \dots, x_c\}$.

Lemma 4 Let $B = \{x_1, \dots, x_b, x_{n-a+b+1}, \dots, x_n\}$

$$B_1 = \{x_1, \dots, x_b\}$$

$$\text{and } B_2 = \{x_{n-a+b+1}, \dots, x_n\}$$

$$\begin{aligned} \text{Then } (a/(n-a)) |F_B(x) - F(x)|^2 &< \omega(b/(n-b)) |F_{B_1}(x) - F(x)|^2 \\ &+ (1-\omega)((a-b)/(n-a+b)) |F_{B_2}(x) - F(x)|^2 \end{aligned}$$

provided that

$$(b/a)(1-(a-b)/(n-a)) < \omega < (b/a)(1+(a-b)/(n-a)). \quad (3.3.15)$$

Proof It is apparent from figure 3.4 that we need to divide the analysis into four parts:

$$(i) \quad x < b/n,$$

$$(ii) \quad b/n < x < x_0$$

$$(iii) \quad x_0 < x < (n-a+b)/n$$

and

$$(iv) \quad x > (n-a+b)/n$$

where x_0 is the point at which $F_B(x) = F(x)$. To simplify the exposition we let

$$D_A(x) = (a/(n-a)) |F_A(x) - F(x)|^2 \text{ for any set } A \text{ with } a \text{ elements.}$$

$$(i) \quad D_B(i/n) = (a/(n-a))(1/a - i/n)^2 = ((n-a)/a)i^2/n^2,$$

$$\text{in a similar manner } D_{B_1}(i/n) = ((n-b)/b)i^2/n^2$$

$$\text{and } D_{B_2}(i/n) = ((a-b)/(n-a+b))i^2/n^2.$$

So

$$\omega D_{B_1}(i/n) + (1-\omega) D_{B_2}(i/n) > D_B(i/n) \text{ is equivalent to}$$

$$\omega(n-b)/b + (1-\omega)(a-b)/(n-a+b) > (n-a)/a.$$

Simplifying the above expression yields the first inequality in (3.3.15).

(ii) We consider $R(x) = \omega D_{B_1}(x) + (1-\omega)D_{B_2}(x) - D_B(x)$ for $b/n \leq x \leq x_0$.

Since $F_B(x)$, $F_{B_1}(x)$ and $F_{B_2}(x)$ are constant for $b/n \leq x \leq x_0$ and $F(x)=x$, we have

$$\begin{aligned} \frac{dR(x)}{dx} &= -2\omega(b/(n-b))(F_{B_1}(x)-F(x)) - 2(1-\omega)((a-b)/(n-a+b))(F_{B_2}(x)-F(x)) \\ &\quad + 2(a/(n-a))(F_B(x)-F(x)) \end{aligned}$$

and

$$\frac{d^2R(x)}{dx^2} = 2\omega b/(n-b) + 2(1-\omega)((a-b)/(n-a+b)) - 2a/(n-a) < 0.$$

This implies that $R(x)$ is minimized at $x=b/n$ or $x=x_0$. Since $R(b/n) > 0$ from case (i) and $R(x_0) > 0$ because $D_B(x_0) = 0$, then $R(x) > 0$ for $b/n \leq x \leq x_0$ as desired.

(iii) This case follows from (iv) just as (ii) follows from (i)

(iv)

$$\begin{aligned} D_B(1/n) &= (a/(n-a))\left(\frac{1}{n} - \frac{b}{a} - \frac{1-n+a-b}{a}\right)^2 = (a/(n-a))((n-1)/a - (n-1)/n)^2 \\ &= ((n-a)/a)(n-1)^2/n^2 \end{aligned}$$

Also, $D_{B_1}(1/n) = (b/(n-b))(n-1)^2/n^2$

and $D_{B_2}(1/n) = ((a-b)/(n-a+b))(1/n - (1-n+a-b)/(a-b))^2$

$$= ((a-b)/(n-a+b))((n-1)/(a-b) - (n-1)/n)^2 = ((n-a+b)/(a-b))(n-1)^2/n^2$$

So $\omega D_{B_1}(1/n) + (1-\omega)D_{B_2}(1/n) > D_B(1/n)$ implies

$$\omega(b/(n-b)) + (1-\omega)((n-a+b)/(a-b)) > (n-a)/a$$

Hence $\omega[b/(n-b) - (n-a+b)/(a-b)] > (n-a)/a - (n-a+b)/(a-b)$

Simplifying the above expression yields the second inequality in (3.3.15).

(iv) and (iii) follow from (i) and (ii) by replacing $1-\omega$ for ω and interchanging the roles of b and $a-b$. From the first inequality in (3.3.15), we have $((a-b)/a)(1-b/(n-a)) < 1-\omega$. This leads directly to the second inequality in (3.3.15).

In particular $\omega=b/a$ works. Furthermore, since $\psi(B_2) = \psi(B'_2)$ where $B'_2 = \{x_1, \dots, x_{n-a+b}\}$, then either $\psi(B_1) < \psi(B)$ or $\psi(B'_2) < \psi(B)$ and hence there exists C such that $\psi(C) < \psi(B)$ where $d=|x-y|$ so that $\alpha_{ij} = 0$ if $i \neq j$. We now use the above lemma to show that if d is a convex function in $|x-y|$ so that $\alpha_{ij} > 0 \forall i, j$, then for any A there exists C of the form $\{x_1, \dots, x_c\}$ such that $\psi(C) < \psi(A)$. To show this, it is sufficient to prove

Lemma 5 For any set A there exists a set C such that $\psi(A) < \psi(C)$ and C is of the form $\{x_1, \dots, x_c\}$.

Proof

We use lemma 3 to find the set B of the form $\{x_1, \dots, x_b, x_{n-a+b+1}, \dots, x_n\}$ such that $|F_B(x) - F(x)| > |F_A(x) - F(x)| \forall x$. We again consider B_1 and B_2 as in the previous lemma. We now have ten cases depending on whether each of $x < y$ belongs to case (i), (ii), (iii) and (iv) as indicated in lemma 4. Since (ii) reduces to (i) and (iii) to (iv) as in the previous lemma, we restrict attention to

- a): $x < y < b/n$,
- b): $(n-a+b)/n < x < y$ and
- c): $x < b/n, y > (n-a+b)/n$.

We only need to consider $x=i/n$ and $y=j/n$.

- a): This reduces to case (i) in the previous lemma since

$$(a/(n-a))(F_B(i/n)-i/n)(F_B(j/n)-j/n) = ((n-a)/a)ij/n^2,$$

$$(b/(n-b))(F_{B_1}(i/n)-i/n)(F_{B_1}(j/n)-j/n) = ((n-b)/b)ij/n^2 \text{ and}$$

$$((a-b)/(n-a+b))(F_{B_2}(i/n)-i/n)(F_{B_2}(j/n)-j/n) = ((a-b)/(n-a+b))ij/n$$

b): In a similar manner this case reduces to (iv).

c): If $(F_A(i/n)-F(i/n)) > 0$ and $(F_A(j/n)-F(j/n)) > 0$ then

$$(a/(n-a))(F_A(i/n)-F(i/n))(F_A(j/n)-F(j/n)) < (a/(n-a))(i/a-i/n)(1-j/n) = i(n-j)/n^2.$$

$$\text{Also, } (b/(n-b))(F_{B_1}(i/n)-i/n)(F_{B_1}(j/n)-j/n) < (b/(n-b))(1/b-i/n)(1-j/n) = i(n-j)/n^2$$

and

$$((a-b)/(n-a+b))(F_{B_2}(i/n)-i/n)(F_{B_2}(j/n)-j/n)$$

$$= ((a-b)/(n-a+b))(i/n)(n-j)/n-(n-j)/(a-b) = i(n-j)/n^2.$$

A similar argument shows that if

$$(F_A(i/n)-F(j/n)) < 0 \text{ and } (F_A(i/n)-F(j/n)) < 0 \text{ then}$$

$$\begin{aligned} & (a/(n-a))(F_A(i/n)-F(i/n))(F_A(j/n)-F(j/n)) \\ & < (b/(n-b))(F_{B_1}(i/n)-F(i/n))(F_{B_1}(j/n)-F(j/n)) \\ & = ((a-b)/(n-a+b))(F_{B_2}(i/n)-F(i/n))(F_{B_2}(j/n)-F(j/n)). \end{aligned}$$

Hence from a), b) and c) so long as ω satisfies (3.3.15) then

$$\omega\psi(B_1) + (1-\omega)\psi(B_2) < \psi(A). \quad \text{This implies that at least one of}$$

$\psi(B_1)$ and $\psi(B_2)$ is no greater than $\psi(A)$.

Now that we have shown that if d is convex in $|x-y|$ then $\psi(A)$ is minimized if A and $S-A$ are connected in S we can extend the result to the following:

Theorem: Let $A_1, \dots, A_k$ be mutually exclusive and collective exhaustive subsets of $S = \{x_1, \dots, x_n\}$. Let

$$Z = \sum_{\ell=1}^k \left(\sum_{i=1}^n \sum_{j=1}^n d(x_i, x_j) y_{i\ell} y_{j\ell} \right) / a_{\ell} \quad (3.3.16)$$

where $y_{i\ell} = \begin{cases} 1 & \text{if } x_i \in A_{\ell} \\ 0 & \text{if } x_i \notin A_{\ell} \end{cases}$

and a_{ℓ} denotes the number of elements in A_{ℓ} , $\ell=1, \dots, k$. Then Z is minimized when $A_1, \dots, A_k$ are connected in S provided that $d(x, y)$ is convex in $|x-y|$.

Proof Assume there exists a subset A_1 which is not connected in S . This implies there exists x_i and x_j in A_1 and y not in A_1 with $x_i < y < x_j$. Let A_h denote the subset that contains y . This implies A_1 is not connected in $A_1 \cup A_h$. From our previous results we can then modify A_1 and A_h so that A_1 and A_h are each connected in $A_1 \cup A_h$, thereby lowering the contribution to Z from A_1 and A_h with the contribution to Z from other subsets remaining unchanged.

4. Extensions

We found conditions on d (i.e., d is convex in $|x-y|$), and $f(a)$ (i.e., $f(a) = a$) such that the interpoint distances within clusters weighted by $1/f(a)$ is minimized when the clusters consist of consecutive points in $S = \{x_1 < \dots < x_n\}$. The dual problem is not to place restrictions on d and f but rather on the points, so that the conjecture stated in the introduction holds. In particular, it is easy to show by induction that if the points are equally spaced then the conjecture is true.

The above extension suggests the following conjecture:

If $X_1, \dots, X_n$ are independent and identically distributed random variables then $E(Z)$ is minimized for any distance function d where we replace x_1 by the corresponding random variable X_1 in Z in (3.3.16).

Another extension is to consider $x_i \in \mathbb{R}^p$. This complicates the problem since the points, $x_1, \dots, x_n$, are no longer well ordered. Perhaps a reasonable generalization to the conjecture in this article is to show that the clusters must be determined by the points in convex subregions of $\mathbb{R}^p$.

C. Empirical Analysis

The Demand for Major Food Items in Israel (Y. Mundlak)

The Central Bureau of Statistics (CBS) time series on expenditure reports expenditures in current and fixed prices of ten food groups. These data were used to estimate a system of demand analysis, allowing for the effect of demographic factors.

Supply and Demand of Fruits in Israel (Y. Mundlak and D. Guy)

Previous work on the demand for fruits was extended to build a complete model of the fruit branch. The system was estimated, using time series techniques and predictions for the several years outside the sample period was obtained for the major fruits.

The Demand for Quality and Quantity of Food in the Process of Growth

In order to understand the nature of the demand for food, and more important, the demand for the agricultural product that goes into food, it is recognized that food consists of agricultural and nonagricultural components. The nonagricultural component includes devices, packaging, refrigeration, processing, etc. All these are aggregated and the quality of food is measured by the ratio of the nonagricultural component to that of agriculture. The statistical analysis, using USDA data for the U.S. indicates that the quality increases with income but it also depends on the price ratio of agriculture to nonagriculture. That explains why in low income countries food contains an important nonfood component. The price of services (nonagricultural component) then is low relative to the price of the agricultural component which is determined largely in the world market.

This work was read as an invited paper in the Congress of the International Association of Agricultural Economists (1985), will appear in the book containing selected papers of this meeting, and circulated as a working paper of the Center for Agricultural Economic Research (CFAER).

The Structure of the Production Sector of Israeli Agriculture

(Y. Mundlak, D. Guy, R. Ben Zvi)

Work on the formulation and estimation of the production structure has continued with a considerable progress, partly with additional sources of finance. Some of the ideas are reported in working papers of the CFAER. The statistical work is in progress and will be continued under the new BARD grant.

H. List of Publications

Mundlak, Yair. Endogenous Technology and the Measurement of Productivity, Working Paper No. 8410. The Center for Agricultural Economic Research The Hebrew University of Jerusalem, P.O.B. 12, Rehovot, Israel.

Mundlak, Yair. Agricultural Growth and the Price of Food, Working Paper No. 8505. The Center for Agricultural Economic Research, The Hebrew University of Jerusalem, P.O.B. 12, Rehovot, Israel.