AN ABSTRACT OF THE THESIS OF

LEONARD WILLIAM DEATON for the DOCTOR OF PHILOSOPHY
(Name) (Degree)

in STATISTICS presented on 7-31-73
(Major) (Date)

Title: A PRIOR DISTRIBUTION FOR SMOOTH REGRESSION

Abstract approved: Redacted for Privacy
H. D. Brunk

Let the random vector $\tilde{y}$ have an N-variate normal distribution with mean $Q\theta$ and covariance matrix $\sigma^2 I$. That is, $\tilde{y} \sim N_N(Q\theta, \sigma^2 I)$ where Q is an $Nx(m+1)$ matrix such that $Q'Q = I$, I is the appropriate dimensioned identity matrix and $\sigma^2 > 0$. A method for estimation of θ is proposed which is Bayesian in the sense that the components θ_i of θ are assumed to be fixed values of the random components $\tilde{\theta}_i$ of $\tilde{\theta}$ where $\tilde{\theta}_i \sim N_1(\mu_i, \sigma_i^2)$ and independent for $i = 0,1,\ldots,m$. The μ_i's are taken as known (usually zero) and σ_i^2's are taken as unknown (or known) for $i = 0,1,\ldots,m$.

The proposed method for estimation of θ allows one to incorporate in a prior distribution (and hence in his estimate of θ) a variety of restrictions on σ_i^2 for $i = 0,1,\ldots,m$. One set of restrictions, namely

$$\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_m^2 ,$$

leads to a method for estimating smooth regression with a polynomial which appears in Monte Carlo studies to be an improvement on popular classical methods.

Another set of restrictions,

$$\sigma_0^2 = \sigma_1^2 = \cdots = \sigma_m^2 ,$$

leads to a class of estimates which dominate the least squares estimator when using squared error loss.

A Prior Distribution for Smooth Regression

by

Leonard William Deaton

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

June 1974

APPROVED:

Redacted for Privacy

Professor of Statistics

in charge of major

Redacted for Privacy

Chairman of Department of Statistics

Redacted for Privacy

Dean of Graduate School

Date thesis is presented 7-31-73

Typed by Wendy Deaton and Merla Harris for

Leonard William Deaton

ACKNOWLEDGMENT

The author would like to express his appreciation to the members of the Faculty of Oregon State University with whom he has had course work or discussions which have contributed to the development of this thesis. This would include Dr. John Lee, Dr. Justus Seely and Dr. Donald Pierce.

The author is especially indebted to Dr. H.D. Brunk who suggested the problem and directed its investigation. Without his encourgement and valuable suggestions this thesis would not have been completed.

The author recognizes the financial assistance of the Department of Statistics, the computer center of the university, and the Public Health Service.

The author is grateful to his wife and children who made numerous sacrifices to provide an atmosphere conducive to the attainment of the educational goals of the author.

Final thanks go to the typists, Wendy (his wife) and Merla Harris, for undertaking the herculean task of deciphering his notes and making them intelligible on a very short notice.

TABLE OF CONTENTS

A PRIOR DISTRIBUTION FOR SMOOTH REGRESSION

1. INTRODUCTION

1.1 The Problem

This thesis is concerned with the general linear regression problem. That is, we have an N dimensional vector y, which is the observed value of the random vector $\tilde{y}$. It is also assumed that $\tilde{y}$ has a multivariate normal distribution with mean $X\beta$ and covariance matrix V_1. This will be denoted by writing $\tilde{y} \sim N_N(X\beta, V_1)$. It is also assumed that X is a known $Nx(m+1)$ dimensional matrix of rank $m+1$; $V_1 = \sigma^2 V$, where V is a known symmetric positive definite $N \times N$ matrix and $\sigma^2 > 0$ is known or unknown; β is an unknown $m+1$ dimensional vector. The problem is to estimate β.

The problem will be modified (without loss of generality). We shall reparameterize to obtain "orthogonality". Let X_i denote the i^{th} column of X for $i = 0,1,\ldots,m$. We may apply the Gram-Schmidt process to the X_i's since they are linearly independent. The inner product to be used in this process is $\langle \cdot, \cdot \rangle$, defined by

$$\langle u, w \rangle = u'V^{-1}w$$

for all N dimensional vectors u and w (u' denotes

the transpose of u). If the process is performed in the order of the X-subscript, we obtain orthonormal vectors $\bar{Q}_o, \bar{Q}_1, \ldots, \bar{Q}_m$ where

$$\bar{Q}_o = \bar{Q}_o(X_o), \quad \bar{Q}_1 = \bar{Q}_1(X_o, X_1), \ldots, \bar{Q}_m = \bar{Q}_m(X_o, X_1, \ldots, X_m) .$$

That is, $\bar{Q}_i$ is a function of X_k for $k = 0,1,\ldots,i$. Let $Q = (\bar{Q}_o, \bar{Q}_1, \ldots, \bar{Q}_m)$. Let δ_{ij} denote the Kronecker delta function. That is:

$$\delta_{ij} = \begin{cases} 0 \text{ if } i \neq j \\ 1 \text{ if } i = j \end{cases}$$

Then we have

$$\langle \bar{Q}_i, \bar{Q}_j \rangle = \delta_{ij}$$

for $i,j = 0,1,\ldots,m$,

$$Q'V^{-1}Q = I$$

where I is the appropriate dimensioned identity matrix, and

$$X\beta = Q\theta$$

where θ is defined by

$$\theta = Q'V^{-1}X\beta$$

We now wish to estimate θ, since $\beta = (Q'V^{-1}X)^{-1}\theta$. Thus we have

$$\tilde{y} \sim N_N(Q\theta, V_1) .$$

The approach to the problem will be Bayesian, in

that we shall assume a prior distribution on the parameter θ. Our assumptions are

1. $\tilde{\theta} \sim N_{m+1}(\mu, V_2)$
   where the element of the i^{th} row and the j^{th} column of V_2 is $\sigma_i^2 \delta_{ij}$
   for $i,j = 0,1,\ldots,m$.
2. $\tilde{y} | \tilde{\theta} = \theta \sim N_N(Q\theta, V_1)$,
   where $V_1 = \sigma^2 V$.

That is, the conditional distribution of $\tilde{y}$ given $\tilde{\theta} = \theta$ is N-variate normal. Throughout this dissertation, V, Q, and μ are assumed known. For most of the results we obtain, we also assume $Q'V^{-1}Q = I$ and $N \geq m+1$. We shall consider estimation of θ under conditions that vary from $(\sigma^2, \sigma_o^2, \sigma_1^2, \ldots, \sigma_m^2)$ all known to $(\sigma^2, \sigma_o^2, \sigma_1^2, \ldots, \sigma_m^2)$ all unknown. The exact nature of these conditions will be discussed later. We shall first digress in order to consider Assumption 1. In particular, we consider the assumption that the components of $\tilde{\theta}$, $\tilde{\theta}_i$ for $i = 0,1,\ldots,m$ are assumed independent. Let $u = (u_o, u_1, \ldots, u_m)$, and suppose one is concerned with estimating a real valued function P defined by

$$P(u) = \sum_{j=0}^{m} \beta_j g_j(u_j)$$

where g_j (a real valued function of the real variable u_j) is known for $j = 0,1,\ldots,m$ and β_j for $j = 0,1,\ldots,m$ is unknown. Define the function f_j by

$$f_j(u) = g_j(u_j)$$

for $j = 0,1,\ldots,m$. Then we may write

$$P(u) = \sum_{j=0}^{m} \beta_j f_j(u).$$

The reason for defining the f_j's is to enable us to view P as a sum of functions all of which have the same domain as P, yet f_j depends only on the j^{th} coordinate of u, u_j for $j = 0,1,\ldots,m$.

When considering the estimation of P, it is useful to have a distance defined on the class of functions which would include P and its possible estimates. Define the set D by

$$D = \{u^{(1)}, u^{(2)}, \ldots, u^{(N)}\},$$

where for $i = 1,2,\ldots,N$, $u^{(i)}$ is a fixed $m+1$ dimensional vector. Suppose for each $u \in D$ we are given

a real number Y which is assumed to be the observed value of a random variable $\tilde{Y}$ where $E(\tilde{Y}) = P(u)$. That is the expectation of $\tilde{Y}$ is $P(u)$. More precisely, assume that

$$\tilde{y} \sim N_N(X\beta, V_1)$$

where the element of the i^{th} row and the j^{th} column of X is

$$f_j(u^{(i)})$$

for $i = 1,2,\ldots,n$ and $j = 0,1,\ldots,m$, $\beta = (\beta_0, \beta_1, \ldots, \beta_m)'$, $V_1 = \sigma^2 V$, V is symmetric positive definite and $\sigma^2 > 0$. We shall now restrict the domain of the function P and its possible estimates to the set D. The class of all real valued functions on D will be denoted by S. For each h in S, define $\bar{h}$ by

$$\bar{h} = [\, h(u^{(1)}),\ h(u^{(2)}),\ \ldots,\ h(u^{(N)})]'.$$

Then, if h and H are in S, we have $h = H$ if and only if $\bar{h} = \bar{H}$. This was the only reason for restricting the domain of P to D. Now we may define an inner product $\langle \cdot, \cdot \rangle$ on S by

$$\langle h, H \rangle = \bar{h}' V^{-1} \bar{H}$$

for all h, H in S. This leads to the definition of distance between any pair of functions h, H on D. We define this distance as $\|h-H\|$, where

$$\| h-H \|^2 = (\bar{h}-\bar{H})' V^{-1} (\bar{h}-\bar{H}).$$

This definition seems consistent with the classical least squares theory [Searle, 1971], since in that spirit, one looks for a vector β which minimizes

$$(y-X\beta)' V^{-1} (y-X\beta)$$

After "orthogonalizing" in the manner described earlier, we may write P as

$$P(u) = \sum_{j=0}^{m} \theta_j Q_j(u)$$

for all $u \in D$ where Q_j is defined by the relation

$$\bar{Q}_j = [Q_j(u^{(1)}), Q_j(u^{(2)}), \ldots, Q_j(u^{(N)})]'$$

That is, the set of linearly independent vectors $\{\bar{f}_o, \bar{f}_1, \ldots, \bar{f}_m\}$ has been replaced by the set of orthonormal vectors $\{\bar{Q}_o, \bar{Q}_1, \ldots, \bar{Q}_m\}$.

If A is a subset of $\{Q_o, Q_1, \ldots, Q_m\}$, we write "S(A)" to denote the subspace of S spanned by the members of A. For example, $S = S(\{Q_o, Q_1, \ldots, Q_m\})$. If we wish to approximate P with a member of $S(A)$ for some A, the classical projection theorem [Luenberger, 1969] implies that the closest function in $S(A)$ to P is

$$\sum_{j=0}^{m} \theta_j Q_j(u) \, 1_A(Q_j)$$

where 1_A is the indicator function defined by

$$1_A(x) = \begin{cases} 0 & \text{if } x \notin A \\ 1 & \text{if } x \in A \end{cases}$$

For example, if we wish to use a member of $S(\{Q_0, Q_1, \ldots, Q_{m-1}\})$ to estimate P, the "best" approximation to P is

$$\sum_{j=0}^{m-1} \theta_j Q_j(u).$$

The point of this discussion is "the coefficient θ_k of Q_k (in the expression for P) is unchanged in the expression for the "best" approximation to P if Q_k is a member of the subspace the approximation is taken from." This results, of course, from the orthogonality of the Q_i's.

This is the justification for the assumption of independence in the distribution of the θ's. If one is prepared to estimate P with

$$\bar{P}_1(u) = \sum_{j=0}^{m} \bar{\theta}_j Q_j(u)$$

and then one was told that $\bar{\theta}_m = 0$, one should be satisfied with the estimate

$$\bar{P}_2(u) = \sum_{j=0}^{m-1} \bar{\theta}_j Q_j(u)$$

provided one still felt that $\bar{P}_1$ was close to P.

Hence, his opinions on the values of $\bar{\theta}$, for $j = 0,1,\ldots,m-1$ will be unchanged with knowledge of the value of $\bar{\theta}_m$. The most familiar example of such reasoning is perhaps in the context of orthogonal polynomials.

We may also note, that if

$$P = \sum_{j=0}^{m} \theta_j Q_j$$

and one believes that P can be "adequately" approximated by a member of $S(\{Q_0, Q_1, \ldots, Q_n\})$, then one believes that the distance from P to $\bar{P}$, where

$$\bar{P} = \sum_{j=0}^{n} \theta_j Q_j \ , \quad n < m,$$

is small. Thus, one believes that

$$\|P-\bar{P}\|^2 = \Big\|\sum_{j=n+1}^{m} \theta_j Q_j\Big\|^2 = \sum_{j=n+1}^{m} \theta_j^2$$

is small. Equivalently, one believes that θ_j for $j = n+1, n+2, \ldots, m$ is near zero.

The method of estimating θ proposed in this dissertation was designed for the purpose of exploiting one's prior opinions that θ_i for some specified values of i are near zero. The nature of such prior opinions

which may occur in applications is illustrated in the following examples.

Example 1: Let g be a continuous real valued function on a closed interval I, $I \subseteq (-\infty, \infty)$. Suppose we wish to approximate g. By the Weirstrass Theorem (see [Rudin, 1964]), if given $\epsilon > 0$, there exists a polynomial P_n of degree n, n depends on ϵ, such that

$$|P_n(x) - g(x)| < \epsilon \quad \text{for all} \quad x \in I.$$

Thus, suppose we wish to approximate g with a polynomial, $\bar{g}$, where

$$\bar{g} = \sum_{i=0}^{m} \beta_i x^i$$

for some $\beta_0, \beta_1, \ldots, \beta_m$. That is, our prior knowledge enables us to be sure that m is a sufficiently high degree for $\bar{g}$ to be an adequate approximation. We may also be sure that $\bar{g}$ should be of degree k, or higher for some $k \in \{1,2,\ldots,m\}$. Suppose we have obtained observations y_i of $g(x_i)$ with an error ϵ_i, which is assumed to be $N(0, \sigma^2)$ for $i = 1,2,\ldots,N$. We assume σ^2 is unknown and that there are at least $m+1$ distinct x_i's. Hence, $(y_1, y_2, \ldots, y_N)'$ is an observation of the random vector $\tilde{y}$ where

$$\tilde{y} \sim N_N(X\beta, \sigma^2 I)$$

where the element of the i^{th} row and the j^{th} column of X is

$$x_i^j$$

for $i = 1,2,\ldots,N$, $j = 0,1,\ldots,m$, $\beta = (\beta_0, \beta_1, \ldots, \beta_m)'$, I is the $N \times N$ identity matrix and $\sigma^2 > 0$. Thus, an estimate of β will determine $\bar{g}$. Since there are at least $m+1$ distinct x_i's, we may "orthogonalize" as described earlier to obtain

$$\bar{g} = \sum_{i=0}^{m} \theta_i Q_i \ , \quad \text{and}$$

$$\tilde{y} \mid \tilde{\theta} = \theta \sim N_N(Q\theta, \sigma^2 I)$$

We add the assumption that

$$\tilde{\theta} \sim N_{m+1}(0, V_2)$$

where the element of the i^{th} row and the j^{th} column of V_2 is

$$\sigma_i^2 \delta_{ij}$$

for $i,j = 0,1,\ldots,m$.

Now, one method of expressing our prior opinion that the degree of $\bar{g}$ should be between k and m is to assume

$$\sigma_i^2 = \infty \quad \text{for} \quad i = 0,1,\ldots,k \quad \text{and}$$

$$\sigma_{k+1}^2 \geq \sigma_{k+2}^2 \geq \cdots \geq \sigma_m^2 \ .$$

Roughly speaking, this could be interpreted as the belief

that we are not at all sure of the values of θ_i for $i = 0,1,\ldots,k$, but we believe there is a positive probability that θ_{k+1} is near (within a distance d of zero, $d > 0$) zero. There is a larger probability that θ_i (within a distance d of zero) is near zero as i increases from k+2 to m. This also expresses a belief that $\bar{g}$ is smooth. If the values of σ_i^2 for $i = k+1,\ldots,m$ are not known, the method proposed in this thesis allow us to estimate σ_i^2 with $\bar{\sigma}_i^2$ in such a way that

$$\bar{\sigma}_{k+1}^2 \geq \bar{\sigma}_{k+2}^2 \geq \cdots \geq \bar{\sigma}_m^2 .$$

It should be noted, that the computational aspects of the methods proposed in this thesis for obtaining $\bar{\sigma}_i^2$ for $i = k+1,\ldots,m$ and the estimate $\bar{\theta}$ are relatively simple. That is, we use closed form formulas. No iterations, or approximations of the estimates are needed. In fact, the computations are only slightly more (if any) involved than those needed for the classical procedures of obtaining a least squares estimate of θ followed by a sequence of F-tests of the hypothesis that $\theta_i = 0$ for $i = k+1,k+2,\ldots,m$.

A Monte Carlo study was done for the case $k = 1$, $m = 6$ and $N = 14$. This method for estimation of θ (called the IR rule) was compared to three popular

classical methods and one other method (called the GMP rule) which is derived in Section 3.2.3 of this dissertation. The GMP rule is a "classical-like" method which was derived within the basic structure we have imposed. The results are summarized in Chapter 4 of this thesis. It will be seen that the IR rule performed almost uniformly better than the others.

The purpose of deriving the GMP rule was to get a clearer understanding of the relationship between the classical type rules and the IR rule. The GMP rule may be considered a generalization of the optimal classical rule given in [Anderson, 1971]. The main distinction between the GMP and the IR rules is that the quantity

$$\frac{\sigma^2}{\sigma^2 + \sigma_i^2}$$

for $i = 0,1,\ldots,m$ is assumed to have a value of either zero or one in the GMP rule, while in the IR rule it is assumed to have any possible value in the interval $(0,1]$. It is believed that the IR rule, derived under more realistic assumptions, would be an improvement over the GMP rule, and hence, an improvement over the classical rule given in [Anderson, 1971].

Example 2: Suppose we wish to establish a linear regression equation for a particular response Y in terms of the "independent" predictor variables $X_1, X_2, \ldots, X_6$. Consider the problem of selecting the "best" regression equation. Suppose we are willing to assume

$$\tilde{y} \sim N_N(X\beta, \sigma^2 I), \quad \sigma^2 \text{ unknown}$$

where the first column of X is a vector of ones and the $(i+1)^{th}$ column of X has components which are values of the predictor variable X_i for $i = 1,2,\ldots,6$. After "orthogonalization" we have

$$\tilde{y} \mid \tilde{\theta} = \theta \sim N_N(Q\theta, \sigma^2 I) .$$

If the Gram-Schmidt process is applied in the ordinary manner we will essentially replace the predictor variables $X_1, X_2, \ldots, X_6$ with $Q_1, Q_2, \ldots, Q_6$ where Q_k is a function of $X_1, X_2, \ldots, X_k$. We also assume

$$\tilde{\theta} \sim N_7 (0, V_2)$$

where the element of the i^{th} row and the j^{th} column of V_2 is

$$\sigma_i^2 \delta_{ij}$$

for $i,j = 0,1,\ldots,6$. Suppose one has the following prior opinions:

1. A fit based on $Q_2(X_1, X_2)$ is preferred to a fit based on $Q_4(X_1, X_2, X_3, X_4)$.
2. A fit based on Q_3 is preferred to a

fit based on Q_4.

3. A fit based on Q_4 is preferred to a fit based on Q_5.

Then an appropriate set of restrictions may be

$$\sigma_4^2 \leq \sigma_2^2, \quad \sigma_4^2 \leq \sigma_3^2 , \quad \sigma_5^2 \leq \sigma_4^2 .$$

If the σ_i^2's are unknown, the method for estimating θ, proposed in this dissertation, will do so subject to the desired restrictions. The computation involved in this example is more complicated than that of Example 1, however an algorithm is available which leads to the exact estimate.

Example 3: Suppose we assume

$$\tilde{y} \mid \tilde{\theta} = \theta \sim N_N(Q\theta, \sigma^2 I)$$

and

$$\tilde{\theta} \sim N_{m+1}(0, \sigma_o^2 I), \qquad m \geq 2.$$

That is, the $\tilde{\theta}_i$'s are independent as well as identically distributed for $i = 0,1,\ldots,m$. We will see in Section 3.2.2 that when σ^2 is assumed to be known and equal to one, the methods for estimation of θ proposed in this thesis lead to a class of estimators which uniformly dominate the least squares estimator. One

member of this class is a rule which is a uniform improvement on the James-Stein estimator. The loss function, in this discussion is taken as the usual squared error loss.

The assumption that $\tilde{\theta} \sim N_{m+1}(0, \sigma_o^2 I)$ may not seem realistic in most applications, however, it is used in [Efron and Morris, 1973] to derive a class of estimators which dominate the least squares estimator. One member of the class Efron and Morris derived, was the James-Stein estimator.

The property of dominating the least squares estimator with the methods proposed in this thesis is not limited to the case when σ^2 is known. We shall see in Section 3.2.2 that when σ^2 is unknown, the methods proposed in the dissertation lead to a class of rules which are given as an example in [Baranchik, 1970] of rules known to dominate the least squares estimator with squared error loss.

In view of the fact that the methods proposed for estimation of θ in this dissertation lead to such excellent results in very special cases, as in this example, it seems likely that the methods are good for more general (and seemingly more realistic) cases as well.

1.2 Related Research

To the best of this author's knowledge, there has been very little research on this problem that is closely related to the approach taken in this thesis.

Brunk and Pierce [1965] used methods similar to those proposed in this thesis in connection with density estimation.

The work done by Efron and Morris [1973] was already mentioned in Example 3 of Section 1.1. They were primarily concerned with obtaining estimators which would dominate the least squares estimator. The question of elimination of parameters or smooth regression was not considered.

Halpern [1973] has investigated polynomial regression from a point of view that places prior probabilities on the degree of a polynomial where the degree is assumed to be one of some finite set of consecutive positive integers. He does not work with orthogonal polynomials and hence requires a prior distribution on β for each given degree assumed. A Monte Carlo study indicates that he obtained good results for determining the degree of a polynomial, however there was no indication as to the usefulness

of his methods in prediction. His methods would lead to a polynomial of maximum degree (whenever a positive prior was placed on the maximum degree) if one used a loss function that was quadratic. It appears that the computations involved in actually obtaining the estimation of β would be rather cumbersome compared to methods currently in use.

Lindley and Smith [1972] have investigated linear regression under the assumptions

$$\tilde{y} \mid \tilde{\beta} = \beta \sim N_N(X_1\beta,\ V_1)$$

$$\tilde{\beta} \mid \tilde{\gamma} = \gamma \sim N_{m+1}(X_2\gamma\ ,\ V_2)$$

$$\tilde{\gamma} \sim N_n(\mu,\ V_3)$$

with the additional assumption that the distribution of $\tilde{\beta}$ be exchangeable or at least that the components of $\tilde{\beta}$ fall into classes in which the elements of any given class are assumed to have an exchangeable distribution. As pointed out by M. R. Novick [Lindley and Smith, 1972], one weakness with their methods is that of computational difficulties in actually getting the estimate of β .

Lindley and Smith also mention that the assumption of exchangeability would not be appropriate in many applications.

2. DISTRIBUTIONAL RESULTS

In this chapter some distributional results are provided for later use. Parts a and b of the first theorem are well known. Those results as well as the first equation of part c of the first theorem may be found in [Lindley and Smith, 1972]. The results are presented here for the sake of completeness.

Theorem 1

Let V_1 and V_2 be positive definite covariance matrices. Let $\tilde{y}|\tilde{\theta} = \theta \sim N_N(Q\theta, V_1)$ and $\tilde{\theta} \sim N_{m+1}(\mu, V_2)$. Then

a. $\tilde{\theta}|\tilde{y} = y \sim N_{m+1}(\bar{\mu}, \bar{V})$, where
$\bar{\mu} = \bar{V}(V_2^{-1}\mu + Q'V_1^{-1}y)$ and $\bar{V} = (V_2^{-1} + Q'V_1^{-1}Q)^{-1}$

b. $\tilde{y} \sim N_N(Q\mu, V_1 + QV_2Q')$

c. $(V_1 + QV_2Q')^{-1} = V_1^{-1} - V_1^{-1}Q\bar{V}Q'V_1^{-1}$,
$V_2^{-1} - V_2^{-1}\bar{V}V_2^{-1} = Q'(V_1 + QV_2Q')^{-1}Q$, and
$|\bar{V}V_2^{-1}| \cdot |V_1^{-1}| = |(V_1 + QV_2Q')^{-1}|$, where $|A|$ is the determinant of the matrix A and $\bar{V}$ is defined in part a.

d. If $\tilde{\theta}$ is a least squares estimate of θ, then
$$\bar{\mu} = \bar{V}V_2^{-1}\mu + (I - \bar{V}V_2^{-1})\tilde{\theta}$$

Proof. We begin by proving the second equation in part c.

From the definition of $\bar{V}$ we have

$$\bar{V}(V_2^{-1} + Q'V_1^{-1}Q) = I$$

which implies

$$\bar{V}V_2^{-1} = I - \bar{V}Q'V_1^{-1}Q$$

From the first equation in part c, we have

$$Q'(V_1 + QV_2Q')^{-1}Q = Q'(V_1^{-1} - V_1^{-1}Q\bar{V}Q'V_1^{-1})Q$$

$$= Q'V_1^{-1}Q(I - \bar{V}Q'V_1^{-1}Q)$$

$$= Q'V_1^{-1}Q\bar{V}V_2^{-1}$$

$$= Q'V_1^{-1}Q\bar{V}V_2^{-1} + V_2^{-1}\bar{V}V_2^{-1} - V_2^{-1}\bar{V}V_2^{-1}$$

$$= (Q'V_1^{-1}Q + V_2^{-1})\bar{V}V_2^{-1} - V_2^{-1}\bar{V}V_2^{-1}$$

$$= V_2^{-1} - V_2^{-1}\bar{V}V_2^{-1}$$

[]

The symbol [] indicates the conclusion of the proof of part a.

We now prove the third equation in part c. Using the fact that $|A| = |A^{-1}|^{-1}$ for any invertable matrix A, we see that it suffices to prove

$$|V_2\bar{V}^{-1}| \cdot |V_1| = |(V_1 + QV_2Q')| \ .$$

After multiplying both sides by $|V_1^{-1}|$, we see that it suffices to prove

$$|V_2\bar{V}^{-1}| = |(I + V_1^{-1}QV_2Q')|$$

which is equivalent to

$$|(I + V_2Q'V_1^{-1}Q)| = |(I + V_1^{-1}QV_2Q')| .$$

To see this, we use the result from matrix algebra that

$$|E||(B - CE^{-1}D)| = \begin{vmatrix} B & C \\ D & E \end{vmatrix} = |B||E - DB^{-1}C|$$

for appropriately dimensioned matrices B, C, D, and E.

By putting

$$B = I_N, \quad C = V_1^{-1}Q, \quad D = -V_2Q', \quad E = I_{m+1},$$

where I_N indicates the n x n identity matrix, we obtain the desired result. (This proof was suggested by Justus Seely.) []

Now we prove part d. Recall [Searle, 1971] that $\bar{\theta}$ is a least squares estimate of θ if and only if θ satisfies

$$Q'V_1^{-1}Q\bar{\theta} = Q'V_1^{-1}y . \qquad (1)$$

From part a, we have

$$\bar{\mu} = \bar{V}(V_2^{-1}\mu + Q'V_1^{-1}y) = \bar{V}V_2^{-1}(\mu + V_2Q'V_1^{-1}y).$$

Thus, using (1) we get

$$\bar{\mu} = \bar{V}V_2^{-1}[(\mu - \bar{\theta}) + (\bar{\theta} + V_2Q'V_1^{-1}Q\bar{\theta})]$$

$$= \bar{V}V_2^{-1}(\mu - \bar{\theta}) + \bar{V}V_2^{-1}(I + V_2Q'V_1^{-1}Q)\bar{\theta}$$

$$= \bar{V}V_2^{-1}(\mu - \bar{\theta}) + \bar{V}(V_2^{-1} + Q'V_1^{-1}Q)\bar{\theta}.$$

Using the definition of $\bar{V}$, we have

$$\bar{\mu} = \bar{V}V_2^{-1}(\mu - \bar{\theta}) + \bar{\theta}$$

$$= \bar{V}V_2^{-1}\mu + (I - \bar{V}V_2^{-1})\bar{\theta}. \qquad []$$

The next theorem deals with partitioning the exponent of the density function of $\tilde{y}$ given in Theorem 1b.

Theorem 2

Let V_1 and V_2 be positive definite covariance matrices. Let Q be an $Nx(m+1)$ matrix with rank $m+1$. Let $\tilde{y}|\tilde{\theta} = \theta \sim N_N(Q\theta, V_1)$ and $\tilde{\theta} \sim N_{m+1}(\mu, V_2)$.

a. If $\bar{\theta}$ denotes the usual least squares estimate of θ (i.e. $\bar{\theta} = [Q'V_1^{-1}Q]^{-1}Q'V_1^{-1}y$, [Searle, 1971]), then

$$(y - Q\mu)'(V_1 + QV_2Q')^{-1}(y - Q\mu) =$$

$$R + (\bar{\theta} - \mu)' [(Q'V_1^{-1}Q)^{-1} + V_2]^{-1}(\bar{\theta} - \mu)$$

where R is defined by

$$R = (y - Q\bar{\theta})'V_1^{-1}(y - Q\bar{\theta}).$$

b. If $N > m+1$, then $\tilde{R} = (\tilde{y} - Q\tilde{\theta})'V_1^{-1}(\tilde{y} - Q\tilde{\theta})$ has a central chi-square distribution with $N-(m+1)$ degrees of freedom.

c. If $[(Q'V_1^{-1}Q)^{-1} + V_2]^{-1}$ is a diagonal matrix, say $A = (a\delta_{ij})$, where $a_i > 0$ $i = 0,1,\ldots,m$, then

$$(\tilde{\bar{\theta}}_i - \mu_i)^2 a_i$$

has a central chi-square distribution with one degree of freedom and is stochastically independent of

$$\tilde{R} \text{ and } (\tilde{\bar{\theta}}_j - \mu_j)^2 a_j$$

when $i \neq j$, $i,j = 0,1,\ldots,m$.

Proof (a) Since $\bar{\theta} = (Q'V_1^{-1}Q)^{-1}Q'V_1^{-1}y$, we may rewrite R as

$$R = y'V_1^{-1}y - \bar{\theta}'Q'V_1^{-1}y. \qquad (2)$$

Recall that $\bar{V}$ of Theorem 1 was defined by

$$\bar{V} = (V_2^{-1} + Q'V_1^{-1}Q)^{-1}.$$

Thus

$$I = \bar{V}V_2^{-1} + \bar{V}Q'V_1^{-1}Q .$$

Multiplying by $(Q'V_1^{-1}Q)^{-1}$ gives

$$\bar{V} = (Q'V_1^{-1}Q)^{-1} - \bar{V}V_2^{-1}(Q'V_1^{-1}Q)^{-1}.$$

From Theorem 1c,

$$(V_1 + QV_2Q')^{-1} = V_1^{-1} - V_1^{-1}Q\bar{V}Q'V_1^{-1}$$

$$= V_1^{-1} - V_1^{-1}Q(Q'V_1^{-1}Q)^{-1}Q'V_1^{-1} + V_1^{-1}Q\bar{V}V_2^{-1}(Q'V_1^{-1}Q)^{-1}Q'V_1^{-1}$$

$$= V_1^{-1} - V_1^{-1}Q(Q'V_1^{-1}Q)^{-1}Q'V_1^{-1} + V_1^{-1}Q(Q'V_1^{-1}Q)^{-1}(Q'V_1^{-1}Q)\bar{V}V_2^{-1}(Q'V_1^{-1}Q)^{-1}Q'V_1^{-1}.$$

Thus,

$$(y - Q\mu)'(V_1 + QV_2Q')^{-1}(y - Q\mu) =$$

$$y'V_1^{-1}y - y'V_1^{-1}Q(Q'V_1^{-1}Q)^{-1}Q'V_1^{-1}y$$

$$- 2\mu'Q'[V_1^{-1} - V_1^{-1}Q(Q'V_1^{-1}Q)^{-1}Q'V_1^{-1}]y$$

$$+ \mu'Q'[V_1^{-1} - V_1^{-1}Q(Q'V_1^{-1}Q)^{-1}Q'V_1^{-1}]Q\mu$$

$$+ (y - Q\mu)'V_1^{-1}Q(Q'V_1^{-1}Q)^{-1}(Q'V_1^{-1}Q)\bar{V}V_2^{-1}(Q'V_1^{-1}Q)^{-1}Q'V_1^{-1}(y-Q\mu)$$

$$= R$$

$$- 2\mu'[Q'V_1^{-1}y - Q'V_1^{-1}y]$$

$$+ \mu'[Q'V_1^{-1}Q - Q'V_1^{-1}Q]\mu$$

$$+ (\bar{\theta} - \mu)'(Q'V_1^{-1}Q)\bar{V}V_2^{-1}(\bar{\theta} - \mu),$$

by using (2).

Since

$$(Q'V_1^{-1}Q)\bar{V}V_2^{-1} = [V_2(V_2^{-1} + Q'V_1^{-1}Q)(Q'V_1^{-1}Q)^{-1}]^{-1}$$

$$= [(Q'V_1^{-1}Q)^{-1} + V_2]^{-1} ,$$

(a) is true. []

(b) The validity of (b) when R is replaced by $\tilde{R}|\tilde{\theta} = \theta$ is a well known classical result. Let $f(R|\theta)$ denote the density of $\tilde{R}|\tilde{\theta} = \theta$ and $g(\theta)$ denote the density of $\tilde{\theta}$. Then

$$\int f(R|\theta)g(\theta)d\theta = f(R|\theta),$$

since $f(R|\theta)$ does not depend on θ. But the integral gives the density of $\tilde{R}$. Thus, $\tilde{R}$ and $\tilde{R}|\tilde{\theta} = \theta$ have the same density. []

(c) First, observe (from Theorem 1b and the definition of $\bar{\theta}$) that the covariance of $(\tilde{\bar{\theta}} - \mu)$ is

$$(Q'V_1^{-1}Q)^{-1}Q'V_1^{-1}[V_1 + QV_2Q']V_1^{-1}Q(Q'V_1^{-1}Q)^{-1}$$

$$= (Q'V_1^{-1}Q)^{-1} + V_2$$

$$= A^{-1} ,$$

by the definition of A. Hence,

$$(\tilde{\bar{\theta}} - \mu) = (Q'V_1^{-1}Q)^{-1}Q'V_1^{-1}\tilde{y} - \mu \sim N_{m+1}(0, A^{-1}).$$

Therefore, the components of $\tilde{\tilde{\theta}} - \mu$, $\tilde{\tilde{\theta}}_i - \mu_i$, are stochastically independent and $N_1(0, a_i^{-1})$. Thus, $(\tilde{\tilde{\theta}}_i - \mu_i)^2 a_i$ has a central chi-square distribution with one degree of freedom.

To see independence of $\tilde{Z} = (\tilde{\tilde{\theta}}_i - \mu_i)^2 a_i$ and $\tilde{R}$, we may again appeal to the corresponding classical result. It is known that $\tilde{Z}|\tilde{\theta} = \theta$ and $\tilde{R}|\tilde{\theta} = \theta$ are independent. Let $h(Z|\theta)$ and $f(R|\theta)$ denote the densities of $\tilde{Z}|\tilde{\theta} = \theta$ and $\tilde{R}|\tilde{\theta} = \theta$ respectively. Let $g(\theta)$ denote the density of θ.

Then the joint density of $\tilde{Z}$ and $\tilde{R}$ given $\tilde{\theta} = \theta$ is given by $h(Z|\theta)f(R|\theta)$. The density of $\tilde{Z}$, $\tilde{R}$ and $\tilde{\theta}$ is then given by $h(Z|\theta)f(R|\theta)g(\theta)$. We may integrate with respect to θ to obtain the joint density of $\tilde{Z}$ and $\tilde{R}$. But this is $h(Z|\theta)f(R|\theta)$ since neither of these functions depends on θ. Since the joint density of $\tilde{Z}$ and $\tilde{R}$ factors into a function of Z alone and a function of R alone, $\tilde{Z}$ and $\tilde{R}$ are independent. []

In this thesis the full generality of Theorems 1 and 2 is not needed. A special case of sufficient importance is to be stated next.

Corollary 1

Let $V_2 = (\sigma_i^2 \delta_{ij})$, where $\sigma_i^2 > 0$ for $i = 0,1,\ldots,m$,

and $V_1 = \sigma^2 V$, where V is a symmetric positive definite matrix. Let Q be an $N \times (m+1)$ matrix with rank $m+1$ with the property that $Q'V^{-1}Q = I$. Also, let $\tilde{y}|\tilde{\theta} = \theta \sim N_N(Q\theta, V_1)$ and $\tilde{\theta} \sim N_{m+1}(\mu, V_2)$.

a. $\tilde{\theta}|\tilde{y}=y \sim N_{m+1}(\bar{\mu}, \bar{V})$, where

$$\bar{\mu}_i = z_i\mu_i + (1 - z_i)\bar{\theta}_i,$$

$$\bar{V} = \sigma^2([1 - z_i]\delta_{ij}), \text{ and}$$

$$z_i = \frac{\sigma^2}{\sigma^2+\sigma_i^2} \qquad i = 0,1,\ldots,m.$$

b. The density of the random vector $\tilde{y}$ whose distribution is given in Theorem 1b may be written as

$$(2\pi)^{-\frac{1}{2}N}|V^{-1}|^{\frac{1}{2}}(1/\sigma^2)^{\frac{1}{2}N}e^{-(r/2\sigma^2)}\left[\prod_{i=0}^{m} z_i^{\frac{1}{2}}e^{-(1/2\sigma^2)z_i(\bar{\theta}_i-\mu_i)^2}\right]$$

where r is defined by

$$r = \sigma^2 R$$

and R is defined as in Theorem 2a.

Proof (a) $Q'V_1^{-1}Q = \frac{1}{\sigma^2}Q'V^{-1}Q = \frac{1}{\sigma^2}I.$

Now, $\bar{V} = (V_2^{-1} + Q'V_1Q)^{-1} = \left(\frac{1}{\sigma_i^2}\delta_{ij} + \frac{1}{\sigma^2}\delta_{ij}\right)^{-1}$

$$= \sigma^2\left(\frac{\sigma_i^2}{\sigma^2+\sigma_i^2}\,\delta_{ij}\right) = \sigma^2\left([1-z_i]\delta_{ij}\right) .$$

So

$$\bar{V}V_2^{-1} = \sigma^2\left(\frac{\sigma_i^2}{\sigma^2+\sigma_i^2}\,\delta_{ij}\right)\left(\frac{1}{\sigma_i^2}\delta_{ij}\right) = (z_i\delta_{ij}) \quad (3)$$

The result follows from Theorem 1d. []

(b) From Theorem 1b, we know the density of $\tilde{y}$ is given by

$$(2\pi)^{-N/2}\left|(V_1+QV_2Q')^{-1}\right|^{\frac{1}{2}}\exp\left[-\tfrac{1}{2}(y-Q\mu)'(V_1+QV_2Q')^{-1}(y-Q\mu)\right]$$

$$= (2\pi)^{-N/2}\left|\bar{V}V_2^{-1}\right|^{\frac{1}{2}}\left|V_1^{-1}\right|^{\frac{1}{2}}\exp\left[-\tfrac{1}{2}\left\{R+(\bar{\theta}-\mu)\left(\frac{1}{\sigma^2+\sigma_i^2}\delta_{ij}\right)(\bar{\theta}-\mu)\right\}\right],$$

by Theorems 1c and 2a. The result now follows from (3) and the hypothesis. []

3. ESTIMATION OF θ

3.1 V_1 and V_2 known (The Basic Rule)

In this chapter, we consider the estimation of θ under the assumptions

$$\tilde{y}|\tilde{\theta} = \theta \sim N_N(Q\theta, V_1),$$

$$\tilde{\theta} \sim N_{m+1}(\mu, V_2),$$

where V_1 and V_2 are positive definite and Q and μ are known.

In this section we shall assume V_1 and V_2 are known. In section 3.2 this assumption will be dropped. We shall take the posterior mean (which is also the posterior mode) as the estimate of θ.

From Theorem 1a and 1d we have

$$\bar{\mu} = \bar{V}(V_2^{-1}\mu + Q'V_1^{-1}y)$$

and

$$\bar{\mu} = \bar{V}V_2^{-1}\mu + (I - \bar{V}V_2^{-1})\bar{\theta},$$

where

$$\bar{V} = (V_2^{-1} + Q'V_1^{-1}Q)^{-1} \quad \text{and}$$

$\bar{\theta}$ is a least squares estimate of θ.

As an estimate of θ, $\bar{\mu}$ has several appealing properties. We shall list some of them.

Properties of $\bar{\mu}$

1. It is a "weighted average" of the prior mean μ and the least squares estimate $\bar{\theta}$, in the sense that the weighting matrices sum to the identity.

2. As our prior knowledge becomes vague, our estimate approaches the least squares estimate. More precisely, for fixed V_1, $\bar{\mu} \to \bar{\theta}$ as $V_2^{-1} \to 0$.

3. As sampling becomes less precise the estimate approaches the prior mean. That is, for fixed V_2, $\bar{\mu} \to \mu$ as $V_1^{-1} \to 0$.

4. The estimate, $\bar{\mu}$, is unique (even though $\bar{\theta}$ may not be).

   Define a loss function L by

   $$L(\theta, a) = (\theta - a)'C(\theta - a)$$

   where C is a symmetric non-negative definite matrix. Then we have the additional properties:

5. The estimate, $\bar{\mu}$, is a Bayes rule. (See [DeGroot, 1970])

6. If C is positive definite, then $\bar{\mu}$ is admissible. (This is immediate from the fact that $\bar{\mu}$ is unique and a theorem in [Ferguson, 1967]).

Perhaps the loss function, L, merits some discussion. We note that it gives the usual squared error loss if C is taken as the identity. In case V_1 is unknown, C is often taken as V_1^{-1}. L has also been used in another way.

Suppose we wished to estimate the polynomial

$$\sum_{i=0}^{m} \theta_i x^i$$

over the interval [a, b] and used the estimate

$$\sum_{i=0}^{m} a_i x^i$$

We may wish to define our loss as the "average squared error over the interval [a, b]". That is we define L by

$$L(\theta,a) = \frac{1}{b-a} \int_a^b \left[\sum_{i=0}^{m} (\theta_i - a_i) x^i \right]^2 dx$$

Then the positive definite matrix C would be defined by

$$C = (c_{ij})$$

where

$$c_{ij} = \frac{b^{i+j-1} - a^{i+j-1}}{(i+j-1)(b-a)} \quad , \quad i,j = 0,1,\ldots,m.$$

This is the loss function that was used in the Monte Carlo

study the results of which are given in Chapter 4.

In this dissertation, we are particularly interested in the special case in which

$$V_1 = \sigma^2 V, \quad Q'V^{-1}Q = I, \text{ and}$$

$$V_2 = (\sigma_i^2 \delta_{ij}), \quad \text{where}$$

$$\sigma^2 > 0, \sigma_i^2 > 0, \quad \text{for} \quad i = 0,1,\ldots,m.$$

With these assumptions we have (from Corollary 1)

$$\bar{\mu}_i = z_i \mu_i + (1-z_i)\bar{\theta}_i$$

where

$$z_i = \frac{\sigma^2}{\sigma^2 + \sigma_i^2} = \frac{1/\sigma_i^2}{1/\sigma_i^2 + 1/\sigma_i^2}$$

for $i = 0,1,\ldots,m$.

This estimate, $\bar{\mu}$, will henceforth be referred to as <u>the basic rule</u>.

We note that the i^{th} component of $\bar{\mu}$ is a weighted average of the i^{th} component of the prior mean with the weight of its precision, $1/\sigma_i^2$ and the i^{th} component of the least squares estimate with the weight of its precision, $1/\sigma^2$.

It is also interesting to note that

$\bar{\mu}_i \to \mu_i$ as $\sigma^2 \to \infty$ (σ_i^2 fixed) or as $\sigma_i^2 \to 0$ (σ^2 fixed)

and

$$\bar{\mu}_i \to \bar{\theta}_i \quad \text{as} \quad \sigma_i^2 \to \infty \ (\sigma^2 \text{fixed}) \quad \text{or as} \quad \sigma^2 \to 0 \ (\sigma_i^2 \text{ fixed}).$$

3.2 V_1 and V_2 unknown

3.2.0 Introduction

In this section we continue with the problem of estimating θ under the assumptions:

$$\tilde{y}|\tilde{\theta} = \theta \sim N_N(Q\theta,\ V_1), \quad \tilde{\theta} \sim N_{m+1}(\mu,\ V_2),$$

$$V_1 = \sigma^2 V, \quad Q'V^{-1}Q = I, \quad V_2 = (\sigma_i^2 \delta_{ij}),$$

$$\sigma^2 > 0, \quad \sigma_i^2 > 0 \quad \text{for} \quad i = 0,1,\ldots,m.$$

The estimate considered in the previous section was

$$\bar{\mu}_i = z_i \mu_i + (1 - z_i)\bar{\theta}_i \qquad (1)$$

where

$$\bar{\theta} = Q'V^{-1}y \quad \text{(the least squares estimate of } \theta),$$

and

$$z_i = \frac{\sigma^2}{\sigma^2 + \sigma_i^2}$$

for $i = 0,1,2,\ldots,m$.

We now assume that V is known and that σ^2 is unknown. In addition, we shall assume that some of the z_i's are unknown. Without loss of generality, we assume there is a $k \in \{0,1,\ldots,m\}$ such that when $i \in \{k, k+1,\ldots,m\}$ z_i is unknown.

The assumptions we are making are not unusual. In most applications V is taken as the identity matrix. The assumption that Q is orthogonal, is made without loss of generality if one starts with a full rank design matrix. One would know z_i, for some particular i, if he knew σ_i^2/σ^2. If one thought σ_i^2/σ^2 was very large, he might (in view of (1)) wish to act as though $z_i = 0$. This would have the effect of estimating θ_i with the least squares estimate $\overline{\theta}_i$. This was done for $i = 0$ in the Monte Carlo study given in Chapter 4. Likewise, if one believed σ_i^2/σ^2 was very small, he may wish to act as though $z_i = 1$. When $\mu_i = 0$, as it was in the Monte Carlo study, taking $z_i = 1$ has a smoothing effect, or in other words, it has the effect of eliminating the parameter θ_i from the model.

We shall consider using the basic rule, (1) as the estimate of θ, with unknown z_i's being replaced by their estimates. We shall use a "maximum likelihood" procedure. That is, we shall choose the z_i's to maximize the marginal density of $\tilde{y}$ when viewed as a function of the z_i's. It is mathematically convenient to estimate σ^2 along with the unknown z_i's. The appropriate density is given in Corollary 1b. Thus,

the quantity to be maximized is

$$(1/\sigma^2)^{\frac{1}{2}N} e^{-(r/2\sigma^2)} \left[\prod_{i=0}^{m} z_i^{\frac{1}{2}} e^{-(1/2\sigma^2) z_i (\bar{\theta}_i - \mu_i)^2}\right] .$$

If we let

$$v_{m+1} = 1/\sigma^2 \quad \text{and}$$

$$v_i = \frac{1}{\sigma^2} z_i$$

for $i = k,k+1,\ldots,m$, then we wish to choose v_i for $i = k,k+1,\ldots,m+1$ to maximize

$$v_{m+1}^{\frac{1}{2}n} \exp(-\tfrac{1}{2} v_{m+1} w_{m+1}) \left[\prod_{i=k}^{m} v_i^{\frac{1}{2}} \exp(-\tfrac{1}{2} v_i w_i)\right] \qquad (2)$$

where

$$n = N-(m+1-k) ,$$

$$w_i = (\bar{\theta}_i - \mu_i)^2 \quad \text{for} \quad i = 0,1,\ldots,m ,$$

$$w_{m+1} = r + \sum_{i=0}^{k-1} z_i w_i \quad \text{if} \quad k \neq 0$$

$$= r \quad \text{if} \quad k = 0, \quad \text{and}$$

$$r = (y - Q\bar{\theta})' V^{-1} (y - Q\bar{\theta})$$

(r is the error sum of squares).

It is easily seen that the quantity in (2) is maximized by taking $v_i = \bar{v}_i$, where

$$\bar{v}_i = 1/w_i \quad i = k,k+1,\ldots,m \quad \text{and}$$

$$\bar{v}_{m+1} = n/w_{m+1}$$

Thus,

$$\bar{z}_i = \frac{\bar{v}_i}{\bar{v}_{m+1}} = \frac{w_{m+1}}{nw_i} \quad i = k,k+1,\ldots,m \quad (3)$$

So far, we have ignored the obvious fact that $0 < z_i < 1$, $i = 0,1,\ldots,m$. In practice, $\bar{z}_i$ [as defined in (3)] could be larger than 1. It would seem that maximizing (2) subject to the restriction $0 < z_i < 1$, $i = 0,1,\ldots,m$ should lead to improved estimates. This will be done in the next section. However, the procedure considered so far is very closely related with the standard classical procedures as we shall see next.

The usual classical procedure uses the least squares estimate of θ only as a starting point. Once $\bar{\theta}$ is obtained, some sort of procedure is usually used to eliminate some of the parameters θ_i from the model. If normality is assumed, most procedures are based on some sort of F-test or sequence of F-tests.

We shall proceed along similar lines. First, note that the conditions we have assumed satisfy the conditions of Theorem 2C. The a_i of that theorem is

now $z_i \sigma^{-2}$. Thus, as a result of that theorem, we see that

$$\frac{z_i}{\tilde{z}_i} = \frac{n z_i \tilde{w}_i}{\tilde{w}_{m+1}} \qquad i = k, k+1, \ldots, m$$

has an F-distribution with 1 and n degrees of freedom. Hence, we could obtain confidence intervals or perform F-tests on the z_i's.

We note that our basic rule (1), will estimate θ_i to be μ_i if and only if z_i is estimated to be one (since $P[\tilde{\theta}_i = \mu_i] = 0$). In fact, if we take $k = 0$, then $(\tilde{z}_i)^{-1}$ (which is an appropriate statistic for testing $H_o: \theta_i = \mu_i$ or $H_o^1: z_i = 1$) is the same statistic used in the classical F-test of $H_o: \theta_i = \mu_i$ (recall Theorem 2). When H_o is true, $(\tilde{z}_i)^{-1}$ is distributed as it is in the classical situation.

3.2.1. Estimating the hyperparameters: flat prior

In this section, we shall continue with the same problem that was introduced in the previous section. We make the same assumptions. The only difference is that now we consider maximizing the quantity given in (2) subject to the restrictions that $0 < z_i < 1$ for

$i = 0,1,\ldots,m$. In terms of the v_i's, this means that we require $0 < v_i < v_{m+1}$ for $i = k, k+1,\ldots,m$.

Hence, our problem now is to find v_i's which minimize

$$\sum_{i=k}^{m+1} [v_i - g_i \log(v_i)] w_i \qquad (4)$$

Subject to $0 < v_i < v_{m+1}$ for $i = k,k+1,\ldots,m$ where

$$g_i = \frac{1}{w_i}$$

for $i = k,k+1,\ldots,m$ and

$$g_{m+1} = \frac{n}{w_{m+1}} .$$

For the remainder of this thesis, we shall assume that an observed value, X, of a random variable, $\tilde{X}$, is positive and finite whenever $\tilde{X}$ is positive and finite almost surely. Thus, we shall assume $(\tilde{\theta}_i - \mu_i)^2$ and its inverse is positive and finite for $i = 0,1,\ldots,m$. This implies that w_{m+1} is positive and finite whenever $n > 0$, since if $N > (m+1)$, the error sum of squares $\tilde{r}$ is almost surely positive and finite. Hence, when $n > 0$, g_i for $i = k,k+1,\ldots,m+1$ is assumed positive and finite.

This problem may not have a solution since the

inequalities on the v_i's define an open set. However, in most practical situations, one may be satisfied to have a solution which minimizes the quantity in (4) subject to $\epsilon \leq v_i \leq v_{m+1}$, for $i = k, k+1, \ldots, m$; when $\epsilon > 0$ is sufficiently small. In fact, we are usually able to minimize the quantity in (4) subject to $0 < v_i \leq v_{m+1}$ for $i = k, k+1, \ldots, m$ as stated in the next theorem. But, first we need a lemma.

Lemma 1

Let g_i and w_i be positive and finite for $i = k, k+1, \ldots, m$. The problem of finding v_i's to minimize

$$\sum_{i=k}^{m+1} [v_i - g_i \log(v_i)] w_i$$

subject to a given set of restrictions is equivalent to the problem of finding v_i's to minimize

$$\sum_{i=k}^{m+1} \Delta(g_i, v_i) w_i$$

subject to the same set of restrictions, where

$$\Delta(a,b) = \Phi(a) - \Phi(b) - (a-b)\Phi'(b).$$

$$\Phi(x) = x\log(x), \quad \text{and}$$

Φ' is the derivative of Φ. (It is assumed that the restrictions on the v_i's will in include $v_i > 0$ for $i = k,k+1,\ldots,m$.

Proof: Since $\Delta(g,v) = [v-g\log(v)] + g[\log(g)-1]$

We have

$$\sum_{i=k}^{m+1}\Delta(g_i,v_i)w_i = \sum_{i=k}^{m+1}[v_i-g_i\log(v_i)]w_i + \sum_{i=k}^{m+1}g_i[\log(g_i)-1]w_i. \quad (5)$$

Since the second term of the right hand side of (5) does not depend on v_i, it may be ignored for the purpose of minimizing in the v_i's. []

The next theorem will not only give us the solution to the minimization of (4) with the restrictions

$$0 < v_i \leq v_{m+1}$$

for $i = k,k+1,\ldots,m$ but will also give the solution for a more general set of restrictions. Let the set X be defined by

$$X = \{k,k+1,\ldots,m+1\}.$$

Let $\preceq$ be a quasi-order on X (A quasi-order is a binary relation that is reflexive and transitive. See the appendix of this thesis.) The set of restrictions

for which the next theorem applies is defined by requiring the function v on X to be isotonic (i.e. for $i, j \in X$, $i \precsim j$ implies $v_i \leq v_j$).

To obtain the restrictions $0 < v_i \leq v_{m+1}$ for $i = k, k+1, \ldots, m$ we may define the quasi-order $\precsim$ by

$$i \precsim i \quad \text{and} \quad i \precsim m+1 \quad \text{for all} \quad i \in X.$$

This only requires $v_i \leq v_{m+1}$ for all i in X but we shall see that this is sufficient for our purpose.

Theorem 3

Let g_i and w_i be positive and finite for $i = k, k+1, \ldots, m+1$. The sum

$$\sum_{i=k}^{m+1} [v_i - g_i \log(v_i)] w_i$$

is minimized over the set of isotonic functions, v, by taking v_i to be the isotonic regression of g_i with weights w_i for $i = k, k+1, \ldots, m+1$. The minimizing function is unique. If $\bar{v}_i$ is the isotonic regression of g_i with weights w_i for $i = k, k+1, \ldots, m+1$, then $\bar{v}_i > 0$ for $i = k, k+1, \ldots, m+1$. (See the appendix for definitions and key theorems on the subject of isotonic

regression.)

Proof: The first two assertions follow immediately from Lemma 1 and Theorem A2 (Theorem A2 is in the appendix). The last assertion follows from Theorem A1 of the appendix by noting that $0 < \min \{g_k, g_{k+1}, \ldots, g_{m+1}\} \leq g_i$ for $i = k, k+1, \ldots, m+1$. []

For the quasi-order which lead to the restrictions $v_i \leq v_{m+1}$ for $i = k, k+1, \ldots, m$. a very simple algorithm for computing the isotonic regression is available. The algorithm is called the "Maximum Violator Algorithm" and is given in the appendix of this dissertation.

Example

Let us consider Example 1 of the Introduction to this thesis. We wished to obtain a polynomial estimate of a continuous function. The estimate was to be of the form

$$\sum_{i=0}^{m} \theta_i Q_i$$

where the Q_i's were orthonormal polynomials of

degree i for $i = 0,1,\ldots,m$. We wished to get an estimate that would be of degree k or higher. It is slightly simpler (notation-wise) to assume we want the degree to be $k-1$ or higher. This would be expressed by taking

$$\sigma_i^2 = \infty \quad \text{for} \quad i = 0,1,\ldots,k-1.$$

We also wished to express the prior opinion of smooth regression by requiring

$$\sigma_k^2 \geq \sigma_{k+1}^2 \geq \cdots \geq \sigma_m^2 .$$

These restrictions can be rewritten in terms of the z_i's,

$$\text{where} \quad z_i = \frac{\sigma^2}{\sigma^2 + \sigma_i^2} ,$$

as $z_i = 0$ for $i = 0,1,\ldots,k-1$

$$0 < z_k \leq z_{k+1} \leq \cdots \leq z_m < 1$$

or in terms of the v_i's

$$(v_i = \frac{1}{\sigma^2} v_i \ , \ i = k,k+1,\ldots,m, \ v_{m+1} = \frac{1}{\sigma^2})$$

as $z_i = 0$ for $i = 0,1,\ldots,k-1$

$$0 < v_k \leq v_{k+1} \leq \cdots \leq v_m < v_{m+1} .$$

The basic rule, with $\mu = 0$, is

$$\bar{\mu}_i = (1 - z_i)\bar{\theta}_i \quad \text{for} \quad i = 0,1,\ldots,m.$$

Thus for $z_i = 0$, we would use the least squares estimate of θ_i. Our theory at this point does not allow for taking $z_i = 0$, but suppose (temporarily) we are content with taking $z_i = \epsilon$ for $i = 0,1,\ldots,k-1$ for some ϵ, $0 < \epsilon < 1$ suffiently small. Also, suppose we are willing to replace the restriction $v_m < v_{m+1}$ with $v_m \leq v_{m+1}$. Our theory includes the solution to this problem. The quasi-order we assume on X is the usual simple order (i.e. for $i,j \in X$, $i \succeq j$ if and only if $i \leq j$). For v to be isotonic on X, with this order, means simply for v to be nondecreasing.

If $N > m+1$, then $n > 0$ (for all $k \geq 0$), where $n = N - (m+1-k)$. Thus, the g_i's and w_i's as defined in (2) are positive and finite so we may apply Theorem 3 to the problem of minimizing the quantity in (4). We obtain estimates $\bar{v}_i(\epsilon)$ (we use the notation $\bar{v}_i(\epsilon)$ instead of $\bar{v}_i$ to indicate the dependence on the known values of $z_i = \epsilon$ for $i = 0,1,\ldots,k$) of v_i and use the relation

$$\bar{z}_i(\epsilon) = \frac{\bar{v}_i(\epsilon)}{\bar{v}_{m+1}(\epsilon)} \quad \text{for} \quad i = k,k+1,\ldots,m,$$

and

$$z_i = \epsilon \quad \text{for} \quad i = 0,1,\ldots,m.$$

Recall from (2), that

$$w_{m+1} = r + \epsilon \sum_{i=0}^{k-1} w_i$$

$$g_{m+1} = \frac{n}{w_{m+1}}$$

Since the w_i's and the g_i's are all well defined for $\epsilon = 0$, we may speak of the isotonic regression of g_i with weights w_i for $i = k,k+1,\ldots,m+1$ when $\epsilon = 0$. We denote this by $\bar{v}_i(0)$. It is an immediate result of the definition of isotonic regression and Theorem A3 of the appendix that $\bar{v}_i(\epsilon) \to \bar{v}_i(0)$ as $\epsilon \to 0$.

Hence, for the problem we wish to solve (i.e. with restrictions $\sigma_i^2 = \infty$ for $i = 0,1,\ldots,k-1$), the recommended solution is

$$\bar{\mu}_i = (1 - \bar{z}_i)\bar{\theta}_i \quad \text{for} \quad i = 0,1,\ldots,m$$

where

$$\bar{z}_i = 0 \quad \text{for} \quad i = 0,1,\ldots,k-1$$

$$\bar{z}_i = \frac{\bar{v}_i(0)}{\bar{v}_{m+1}(0)} \quad \text{for} \quad i = k,k+1,\ldots,m.$$

3.2.2 Estimating the hyperparameters: gamma prior

In this section, we generalize the results of the previous section. We put a truncated gamma prior distribution on the unknown v_i's. We will see that the estimator obtained in the previous section is a limit of the estimator derived in this section. We also include an example in which the estimator derived is known to dominate the least squares estimator.

At this point, it may be useful to restate the assumptions. They are:

1. $\tilde{y}|\tilde{\theta} = \theta \sim N_N(Q\theta, V_1)$,

   $V_1 = \sigma^2 V$, $Q'V^{-1}Q = I$,

   $\sigma^2 > 0$, V is positive definite,

   Q is known, V is known, σ^2 is unknown;

2. $\tilde{\theta} \sim N_{m+1}(\mu, V_2)$,

   μ is known, $V_2 = (\sigma_i^2 \delta_{ij})$,

   $\sigma_i^2 > 0$, $z_i = \dfrac{\sigma^2}{\sigma^2 + \sigma_i^2}$, $i = 0,1,\ldots,m$,

   $k \in \{0,1,\ldots,m\}$ and z_i is $\begin{cases} \text{unknown for } k \leq i \leq m \\ \text{known otherwise} \end{cases}$.

3. We use the posterior mean, $\bar{\mu}$, of the distribution of $\tilde{\theta}|\tilde{y} = y$ to estimate θ.

The components of $\bar{\mu}$ are $\bar{\mu}_i$, where

$$\bar{\mu}_i = z_i\mu_i + (1 - z_i)\bar{\theta}_i$$

and where μ_i is the i^{th} component of the prior mean μ and $\bar{\theta}_i$ is the i^{th} component of the least squares estimate $\bar{\theta}$.

4. When z_i is unknown, we shall replace it with its estimate, $\bar{z}_i$, in the expression for $\bar{\mu}_i$.

5. $v_i = \frac{1}{\sigma^2} z_i \qquad i = k,k+1,\ldots,m$

$$v_{m+1} = \frac{1}{\sigma^2}$$

We now add the following assumptions.

6. Prior knowledge imposes a set of restrictions on the function v defined on $X = \{k,k+1,\ldots,m+1\}$ such that those restrictions are satisfied if and only if v is isotonic with respect to some quasi-order on X. We let B denote the set of $(v_k,v_{k+1},\ldots,v_{m+1})$ which satisfy the given restrictions. Then $(v_k,v_{k+1},\ldots,v_{m+1}) \in B$ if and only if v is isotonic.

7. The prior distribution of $\tilde{v}_k,\tilde{v}_{k+1},\ldots,\tilde{v}_{m+1}$ is

$$\propto \left[\prod_{i=k}^{m+1} v_i^{\gamma_i - 1} e^{-v_i/\beta_i} \right] 1_B(v_k, v_{k+1}, \ldots, v_{m+1})$$

where "$\propto$" means "proportional to", and where

$$\gamma_i > 0 \quad \text{and} \quad \beta_i > 0 \quad \text{for} \quad i = k, k+1, \ldots, m+1;$$

$$1_B(x) = \begin{cases} 1 & \text{if } x \text{ in } B \\ 0 & \text{otherwise} \end{cases}$$

In some applications, it may be easier for one to express his prior opinions with a prior distribution on σ^2, σ_i^2 for $i = k, k+1, \ldots, m$ or their inverses, but it is mathematically more convenient to put a prior on the v_i's or the z_i's. When using $\mu = 0$ and desiring to express smoothness, it seems appropriate to put a prior on the z_i's. This is particularly true in view of the nature of the basic rule. An opinion that θ_i should be eliminated from the model is expressed as an opinion that $z_i = 1$. To express a belief that θ_i is small and perhaps should be eliminated is to believe that z_i is near 1. In terms of σ_i^2, we would need to express the fact that σ_i^2 is near ∞. It seems somehow easier to this author to think of something being near 1 than to think of something being near ∞.

Next, we proceed to finding estimates of the v_i's.

We shall use a posterior mode. The posterior distribution of $\tilde{v}_k,\tilde{v}_{k+1},\ldots,\tilde{v}_{m+1}$ given $\tilde{y} = y$ is proportional to the product of the density given in Corollary 1b and the prior density given above. We get

$$v_{m+1}^{\frac{1}{2}n} \cdot \exp(-\tfrac{1}{2}v_{m+1}w_{m+1})\Big[\prod_{i=k}^{m} v_i^{\frac{1}{2}(2\gamma_i-1)} e^{-\frac{1}{2}v_i w_i}\Big] \qquad (6)$$

for $(v_k,v_{k+1},\ldots,v_{m+1})$ in B and zero elsewhere, where

$$n = N-(m+1-k) + 2(\gamma_{m+1}-1)$$

$$w_i = (\bar{\theta}_i - \mu_i)^2 \quad \text{for} \quad i = 0,1,\ldots,k-1 ,$$

$$= (\bar{\theta}_i - \mu_i)^2 + 2/\beta_i \quad \text{for} \quad i = k,k+1,\ldots,m ,$$

$$w_{m+1} = r + \sum_{i=0}^{k-1} z_i w_i + 2/\beta_{m+1} \quad \text{if } k > 0,$$

$$= r + 2/\beta_{m+1} \quad \text{if } k = 0.$$

If we define g_i by

$$g_i = \frac{2\gamma_i - 1}{w_i} \qquad i = k,k+1,\ldots,m$$

$$= \frac{n}{w_{m+1}} \qquad i = m+1,$$

and take the negative of the logarithm of the quantity in (6), then we see that we wish to minimize

$$\sum_{i=k}^{m+1}[v_i - g_i \log(v_i)]w_i \qquad (7)$$

subject to $(v_k, v_{k+1}, \ldots, v_{m+1}) \in B$.

The quantity in (7) is the same as that in (4), except for the definitions of the w_i's and g_i's . Thus, if $g_i > 0$ and $w_i > 0$ for $i = k,\ldots,m+1$; we obtain $\bar{v}_i$ as the isotonic regression of g_i with weights w_i , as in Theorem 3.

We observe that if $\gamma_i = 1$ for $i = k,k+1,\ldots,m+1$, then

(i) the prior distribution of the v_i's approximates an improper distribution that is uniform over B as $\beta_i \to \infty$ for $i = k,\ldots,m+1$ and

(ii) the $\bar{v}_i$'s of this section become the $\bar{v}_i$'s of the previous section when $\beta_i = \infty$ for $i = k,\ldots,m+1$.

To prove (ii), we note that the g_i's and w_i's of this section then differ from those of the last section only in the β_i's and do not differ at all when the β_i's are all equal to ∞ .

To determine appropriate values for the γ_i's and β_i's in practical applications it may be useful to know what prior distribution Assumption 7 places on the joint distribution of $(\tilde{z}_k,\ldots,\tilde{z}_m,\tilde{v}_{m+1})$. The transformation ψ, defined by $\psi(v_k,v_{k+1},\ldots,v_{m+1}) = (z_k,z_{k+1},\ldots,z_m,v_{m+1})$ where

$$z_i = v_i/v_{m+1} \quad \text{for} \quad i = k,k+1,\ldots,m$$

maps B onto $\psi(B)$ in a 1-1 fashion. The Jacobian is v_{m+1}^{m+1-k}.

Thus, the density of $(\tilde{z}_k,\ldots,\tilde{z}_m,\tilde{v}_{m+1})$ is

$$\propto v_{m+1}^{m+1-k}\Big[\prod_{i=k}^{m}(z_i v_{m+1})^{\gamma_i-1} e^{-z_i v_{m+1}/\beta_i}\Big] v_{m+1}^{\gamma_{m+1}-1} e^{-v_{m+1}/\beta_{m+1}}$$

for $(z_k,\ldots,z_m,v_{m+1})$ in (B) and zero elsewhere

or

$$\propto v_{m+1}^{m+1-k+\sum(\gamma_i-1)} e^{-v_{m+1}/\beta_{m+1}}\Big[\prod_{i=k}^{m} z_i^{\gamma_i-1} e^{-z_i v_{m+1}/\beta_i}\Big]$$

for $(z_k,\ldots,z_m,v_{m+1})$ in (B) and zero elsewhere, where the sum is taken from i=k to m+1.

Apart from the factor "$1_{(B)}(\cdot)$", the density converges to a product of of independent beta distributions on the $\tilde{z}_i$'s and a gamma distribution on $\tilde{v}_{m+1}$ as $\beta_i \to \infty$ for $i = k,k+1,\ldots,m$. The beta distributions would be uniform distributions if the γ_i's were equal to one, but the distribution on $\tilde{v}_{m+1}$ is not flat, unless $m+1-k = 0$, even as $\beta_{m+1} \to \infty$.

Thus, one may not want to use the prior distribution given in Assumption 7 if he wanted to express vague knowledge about v_{m+1} and knowledge about the z_i's in terms of beta distributions. It seems that in practical applications, such a situation would be rare, but one can circumvent this state of affairs by taking a prior distribution on $(\tilde{z}_k,\ldots,\tilde{z}_m,\tilde{v}_{m+1})$ which is

$$\propto v_{m+1}^{\gamma_{m+1}-1} e^{-v_{m+1}/\beta_{m+1}} \left(\prod_{i=k}^{m} z_i^{\gamma_i-1}\right) 1_B(v_k,\ldots,v_{m+1}). \qquad (8)$$

Now, one way vague knowledge on v_{m+1} can be obtained is by taking $\gamma_{m+1} = 1$ and letting $\beta_{m+1} \to \infty$. We shall next derive the posterior distribution of $(\tilde{z}_k,\ldots,\tilde{z}_m,\tilde{v}_{m+1})$ given $\tilde{y} = y$ under the assumption that prior distribution is proportional to that in (8). We may rewrite the quantity in (8) in terms of the v_i's as

$$v_{m+1}^{\gamma_{m+1}-1-\sum(\gamma_i-1)} e^{-v_{m+1}/\beta_{m+1}} \left(\prod_{i=k}^{m} v_i^{\gamma_i-1}\right) 1_B(v_k,\ldots,v_{m+1}),$$

where the summation is taken from $i=k$ to m. The desired posterior density is then proportional to the product of the above quantity and the density given in Corollary 1b. We get

$$v_{m+1}^{\frac{1}{2}n} e^{-\frac{1}{2}v_{m+1}w_{m+1}} \left(\prod_{i=k}^{m} v_i^{\frac{1}{2}(2\gamma_i-1)} e^{-\frac{1}{2}v_i w_i}\right) 1_B(v_k,\ldots,v_{m+1}) \qquad (9)$$

where

$$n = N-(m+1-k) + 2[\gamma_{m+1}-1 - \sum_{i=k}^{m}(\gamma_i-1)],$$

$$w_i = (\bar{\theta}_i-\mu_i)^2 \quad \text{for} \quad i = 0,1,\ldots,m; \quad \text{and}$$

$$w_{m+1} = 2/\beta_{m+1} + r = \sum_{i=0}^{k-1} z_i w_i \quad \text{if} \quad k > 0$$

$$= 2/\beta_{m+1} + r \quad \text{if} \quad k = 0$$

If we define g_i by

$$g_i = \frac{2\gamma_i-1}{w_i} \quad \text{for} \quad i = k,k+1,\ldots,m$$

$$= \frac{n}{w_{m+1}} \quad \text{for} \quad i = m+1,$$

We see that the problem of finding the posterior mode is the same as the problem given in (7). The Monte Carlo study, for which results are given later in this thesis,

was based in part on a prior of the form given in (8).

Using a prior proportional to the quantity given in (8), instead of that of Assumption 7, has a possible computational disadvantage since one may make g_{m+1} negative by taking sufficiently large values for the γ_i for $i = k,k+1,\ldots,m$.

With either prior, it is possible to get negative values for some of the g_i's. If this happens, a posterior mode may or may not exist. It depends on the region B. When some of the g_i's are negative and a posterior mode does exist, it may not be the isotonic regression of something. Thus, other computational methods may be needed to compute the posterior mode. Such computational methods will not be discussed in this thesis.

Hence, with either prior, one may encounter computational difficulties if one or more of the γ_i's was less than $\frac{1}{2}$. This would be the case if one wished to express strong opinions that the corresponding z_i or v_i was near zero.

First, consider the case in which one believes some particular z_i is near zero. If one feels very strongly about this, he may essentially take z_i as known to be zero. That is, one may place z_i in the group of known z_i's where $i \leq k$ (this would require relabeling the

z_i's and a new value for k). One may then obtain the estimate $\bar{v}_j(z_i)$ which is the isotonic regression of $g_j(z_i)$ with weights $w_j(z_i)$ for $j = k,k+1,\ldots,m+1$ (We use $\bar{v}_j(z_i)$, $g_i(z_i)$ and $w_j(z_j)$ to indicate that these quantities depend on the value of z_i .). The limiting estimate $\bar{v}_j(0) = \lim_{z_i \to 0} \bar{v}_j(z_i)$ is the isotonic regression of $g_j(0)$ with weights $w_j(0)$ for $j = k,k+1,\ldots,m+1$. (The validity of the limit statement follows from the definition of isotonic regression and Theorem A3 of the appendix.)

Next, suppose one wished to express a strong opinion that v_i is near zero. Since $v_i = 1/(\sigma^2 + \sigma_i^2)$, one must feel that either σ^2 or σ_i^2 is very large. If one is willing to act as though $\sigma_i^2 = \infty$ and σ^2 is finite, then one is willing to act as though $z_i = \sigma^2 v_i$ is near zero. Thus one may proceed as described in the previous paragraph. If one is willing to act as though $\sigma^2 = \infty$ and σ_i^2 is finite, then one may be willing to act as though $z_i = \dfrac{1}{1 + \sigma_i^2/\sigma^2}$ is known to be 1 .

In view of the preceding discussion, it seems that the possible computational difficulties caused by wanting to choose γ_i less than $\frac{1}{2}$, can often be overcome.

We now carry out the promise made in Example 3 of

the Introductory chapter of this thesis for the case in which σ^2 is unknown.

Example: We make the Assumptions 1 through 7 along with the following assumptions:

(i) $k = 0$

(ii) $V = I$

(iii) $\sigma_0^2 = \sigma_1^2 = \cdots = \sigma_m^2$, (unknown)

(iv) $\mu = 0$

(v) $N > m+1 > 2$

(vi) $N-(m+1) + 2(\gamma_{m+1}-1) > 0$

(vii) $\gamma_i > \frac{1}{2}$ for $i = 0,1,\ldots,m$

(viii) $$\frac{\sum_{i=0}^{m}(2\gamma_i-1)}{n} \leq \frac{2(m-1)}{N-(m+1)+2}$$

The assumption (iii) implies that

$$B = \{(v_0,\ldots,v_{m+1});\ 0 < v_0=v_i \leq v_{m+1} \text{ for } i = 1,2,\ldots,m\}.$$

Assumption (iii) and (iv) imply that

$$\bar{\mu}_i=(1-\bar{z}_0)\bar{\theta}_i \quad \text{for} \quad i = 0,1,\ldots,m.$$

In vector notation we have

$$\bar{\mu} = (1-\bar{z}_0)\bar{\theta}$$

We must minimize the quantity given in (7) which is

$$\sum_{i=0}^{m+1}[v_i-g_i\log(v_i)]w_i$$

$$= v_0\bar{w}_0 - \log(v_0)\bar{g}_0\bar{w}_0 + v_{m+1}w_{m+1}-\log(v_{m+1})g_{m+1}w_{m+1}$$

where

$$\bar{w}_o = \sum_{i=0}^{m} w_i \ , \quad \text{and} \quad \bar{g}_o = (1/\bar{w}_o) \sum_{i=0}^{m} (2\gamma_i - 1)$$

subject to the restrictions that $v_o \le v_{m+1}$.

By the Maximum Violator Algorithm in the appendix we have

$$\bar{v}_o = \begin{cases} \bar{g}_o & \text{if} \quad \bar{g}_o \le g_{m+1} \\ (\bar{w}_o + w_{m+1})^{-1}(\bar{g}_o \bar{w}_o + g_{m+1} w_{m+1}) & \text{otherwise} \end{cases}$$

and

$$\bar{v}_{m+1} = \begin{cases} g_{m+1} & \text{if} \quad \bar{g}_o \le g_{m+1} \\ (\bar{w}_o + w_{m+1})^{-1}(\bar{g}_o \bar{w}_o + g_{m+1} w_{m+1}) & \text{otherwise.} \end{cases}$$

Therefore

$$\bar{z}_o = \bar{v}_o / \bar{v}_{m+1} = \begin{cases} \bar{g}_o / g_{m+1} & \text{if} \quad \bar{g}_o \le g_{m+1} \\ 1 & \text{otherwise} \end{cases}$$

We now let $\beta_i \to \infty$ for $i = 0,1,\ldots,m+1$;

Then

$$\bar{z}_o = \begin{cases} ar/n\bar{\theta}'\bar{\theta} & \text{if} \quad a/n \le \bar{\theta}'\bar{\theta}/r \\ 1 & \text{otherwise} \end{cases}$$

where $a = \sum_{i=0}^{m} (2\gamma_i - 1)$, $n = N-(m+1) + 2(\gamma_{m+1}-1)$ and r is the usual residual sum of squares defined in Corollary 1b.

Now define F by $F = \frac{\bar{\theta}'\bar{\theta}}{r}$.

Then

$$\bar{z}_o = \begin{cases} \frac{a}{n}\cdot\frac{1}{F} & \text{if } F \geq \frac{a}{n} \\ 1 & \text{otherwise} \end{cases}$$

Thus

$$\bar{\mu} = \begin{cases} (1-\frac{a}{n}\cdot\frac{1}{F})\bar{\theta} & \text{if } F \geq \frac{a}{n} \\ 0 & \text{otherwise} \end{cases}$$

We can now easily see that the rule $\bar{\mu}$ is obtained by performing a classical F-test of the hypothesis H_o: $\theta=0$, against the alternative H_1:$\theta\neq 0$. The critical region is given by $\{F;\ F \geq a/n\}$. If H_o is rejected, then $\bar{\mu} = (1-\frac{a}{n}\cdot\frac{1}{F})\bar{\theta}$, while if H_o is not rejected, then $\bar{\mu} = 0$. Also, we see that an increase in the value of a, which is increased sureness in the prior opinion that z_o is near 1, lowers the level of significance of the test.

The interpretation of this rule under the assumptions from which it was derived would be to perform an F-test of the hypothesis that $z_o = z_1 = \ldots = z_m = 1$. The critical region and degrees of freedom are the same under either interpretation. We may rewrite $\bar{\mu}$ as

$$\bar{\mu} = [\max\{0,\ 1 - \frac{a}{n}\cdot\frac{1}{F}\}]\bar{\theta}.$$

In [Baranchik, 1970], the rules

$$\delta_c = (1-c/F)\bar{\theta},$$

where $0 \le c \le 2(m-1)/[N-(m+1)-2]$,

are cited as being the James-Stein estimates. Baranchik, in that same reference remarks that these estimates may be improved by replacing $(1-c/F)$ by $\max\{0, 1-c/F\}$. Therefore, $\tilde{\mu}$ is an improvement of the James-Stein estimates and dominates the least squares estimate $\bar{\theta}$.

3.2.3 Estimating the hyperparameters: Bernoulli prior.

Consider the basic rule when $\mu = 0$. We have

$$\tilde{\mu}_i = (1-z_i)\bar{\theta}_i , \quad i = 0,1,\ldots,m$$

If the values of z_i were restricted to being zero or one for $i = 0,1,\ldots,m$; the basic rule would give estimates resembling the usual (i.e. least squares or modifications of least squares) estimates. This section is devoted to estimation of the z_i's when they are assumed to have independent Bernoulli prior distributions.

We make Assumptions 1 through 5 of section 3.2.2. We shall also assume $k = 0$. Thus, by Corollary 1b the distribution of $\tilde{y}$ given $(\tilde{z}_0,\tilde{z}_1,\ldots,\tilde{z}_m,\tilde{v}_{m+1}) = (z_0,z_1,\ldots,z_m,v_{m+1})$ is

$$\propto v_{m+1}^{\frac{1}{2}N} e^{-\frac{1}{2}v_{m+1}w_{m+1}} \left(\prod_{i=0}^{m} z_i^{\frac{1}{2}}\right) \qquad (10)$$

where

$$w_i = \bar{\theta}_i^{\,2}, \quad i = 0,1,\ldots,m$$

$$w_{m+1} = r + \sum_{i=0}^{m} z_i w_i .$$

The quantity in (10) is based on the assumption that $0 < z_i < 1$ for $i = 0,1,\ldots,m$. Thus, we cannot put an ordinary Bernoulli prior on the z_i's.

Let $0 < \gamma < \frac{1}{2}$, $0 < \epsilon < \frac{1}{2}$ and $0 < p_i < 1$ for $i = 0,1,\ldots,m$. Let

$$P[\tilde{z}_i = z_i] = \begin{cases} p_i & \text{if } z_i = 1-\gamma \\ q_i & \text{if } z_i = \epsilon \end{cases}$$

where $q_i = 1-p_i$ for $i = 0,1,\ldots,m$. Assume that $\tilde{z}_0,\tilde{z}_1,\ldots\tilde{z}_m,\tilde{v}_{m+1}$ are independent and that $\tilde{v}_{m+1}$ has a gamma distribution with parameters $z > 0$ and $\beta > 0$. Then the posterior density of $(\tilde{z}_0,\ldots,\tilde{z}_m,\tilde{v}_{m+1})$ given $\tilde{y} = y$ is

$$\propto v_{m+1}^{\frac{1}{2}N + z-1} e^{-v_{m+1}[1/B+(\frac{1}{2})w_{m+1}]} \left(\prod_{i=0}^{m} z_i^{(\frac{1}{2})} P[z_i = z_i]\right)$$

for $v_{m+1} > 0$ and zero elsewhere.

After integrating with respect to v_{m+1}, we obtain the prior density of $(\tilde{z}_0,\tilde{z}_1,\ldots,\tilde{z}_m)$. It is

$$[(1/\beta) + \tfrac{1}{2}w_{m+1}]^{-(\frac{1}{2}N+z)}\left(\prod_{i=0}^{m} z_i^{\frac{1}{2}P[z_i=z_i]}\right). \qquad (11)$$

We denote the posterior density by $f_{\gamma,\epsilon}$. Then $f_{\gamma,\epsilon}$ is proportional to the quantity in (11). The domain of $f_{\gamma,\epsilon}$ has 2^{m+1} points. We could obtain the posterior mode, $(\bar{z}_0,\bar{z}_1,\ldots,\bar{z}_m)_{(\gamma,\epsilon)}$, by evaluating $f_{\gamma,\epsilon}$ at each point of its domain.

In view of our original objectives, it would seem that this procedure should become more desirable (in the sense of being nearer Bernoulli) as γ and ϵ approach zero. Unfortunately there is a danger in taking ϵ too small. Letting $\gamma \to 1$ presents no difficulty, we obtain $f_{1,\epsilon}$ for each ϵ. But,

$$\lim_{\epsilon \to 0} (\bar{z}_0,\ldots,\bar{z}_m)_{(1,\epsilon)} = (1,1,\ldots,1),$$

since if for some $j \in \{0,1,\ldots,m\}$, $z_j = \epsilon$, then

$$\lim_{\epsilon \to 0} f_{1,\epsilon}(z_0,\ldots,z_m) = f_{1,0}(z_0,\ldots,z_m) = 0.$$

However,

$$\lim_{\epsilon \to 0} f_{1,\epsilon}(1,\ldots,1) = f_{1,0}(1,\ldots,1) > 0.$$

Therefore

$$\lim_{\epsilon \to 0} (\bar{z}_0,\ldots,\bar{z}_m)_{(i,\epsilon)} = (1,\ldots,1).$$

The preceding limits we obtained under the

assumption that p_j did not depend on ϵ for $j = 0,1,\ldots,m$. We would like to choose values of z_i from the set $\{0,1\}$, but we are given a choice of values of z_i from the set $\{\epsilon,1\}$. As ϵ approaches zero, we may become more willing to increase q_i for $i = 0,1,\ldots,m$. That is, the choice of ϵ, as opposed to 1, becomes more appealing as $\epsilon \to 0$.

We shall now assume that there exists C_i's such that

$$p_i = C_i\sqrt{\epsilon} \quad \text{and} \quad q_i = 1-p_i$$

While it is recognized that the preceding discussion perhaps provides a rather "weak" motivation for assuming $p_i = C_i\sqrt{\epsilon}$ for $i = 0,1,\ldots,m$, this assumption does lead to the desired goal of obtaining a rule which resembles those currently used. Then

$$z_i^{\frac{1}{2}}\, P[z_i=z_i] = \begin{cases} \sqrt{\epsilon}\,(1-C_i\sqrt{\epsilon}) & \text{if } z_i = \epsilon \\ C_i\sqrt{\epsilon} & \text{if } z_i = 1 \end{cases}$$

Therefore we may rewrite the quantity in (11) as

$$(1/\beta + \tfrac{1}{2}w_{m+1})^{-\frac{1}{2}N-}\, z(\sqrt{\epsilon})^{m+1}\Big[\prod_{i=0}^{m}(1-C_i\sqrt{\epsilon})^{1_{\{\epsilon\}}(z_i)}\, C_i^{1_{\{1\}}(z_i)}\, 1_{\{\epsilon,1\}}(z_i)\Big]. \tag{12}$$

Now, for each fixed ϵ, the quantity $(\sqrt{\epsilon})^{m+1}$ may

be ignored when comparing the values of $f_{1,\epsilon}$ at the points of its domain to find the mode. Hence, we will get a nontrivial limiting mode as ϵ approaches zero. It would be given by computing the mode of the function

$$[1/\beta + \tfrac{1}{2}(r + \sum_{i=0}^{m} z_i w_i)]^{-\frac{1}{2}N - z} [\prod_{i=0}^{m} C_i^{z_i} 1_{\{0,1\}}(z_i)], \tag{13}$$

which is proportional to the limit of the posterior densities, $f_{1,\epsilon}$ as ϵ approaches zero, where $p_i = \sqrt{\epsilon} C_i$ for $i = 0,1,\ldots,m$. We shall denote this limit by f. Thus, the limiting posterior density, f, is proportional to the quantity given in (13).

This is the sort of posterior density that was desired. It does not depend on ϵ. It is a function of the z_i's which gives positive probability only when each $z_i \in \{0,1\}$. The relative magnitude of the prior probabilities, p_i, is reflected in the relative magnitudes of C_i for $i = 0,1,\ldots,m$. The function, f, has a unique mode (almost surely). Even though f is not a well known density, its use leads to an optimal rule derived by T. W. Anderson [1971] for determining the degree of a polynomial.

In using f to duplicate Anderson's rule, we

do not compute its mode. We take $\beta = \infty$. Next we restrict the domain of f to points x of the form

$$0 = (0,0,\ldots,0), \quad \text{and}$$

$$x_k = (z_0,\ldots,z_m)$$

where

$$z_i = \begin{cases} 1 & \text{for } k \le i \le m \\ 0 & \text{otherwise} \end{cases}$$

for $k \in \{0,1,\ldots,m\}$. In Anderson's procedure we would begin by comparing $f(0)$ with $f(x_m)$. If $f(0) > f(x_m)$, then we fit a polynomial of degree m. If $f(0) < f(x_m)$, then we compare $f(x_m)$ with $f(x_{m-1})$. If $f(x_m) > f(x_{m-1})$, then we fit a polynomial of degree $m-1$. If not, we compare $f(x_{m-1})$ with $f(x_{m-2})$. The procedure continues until we find a point, x_j, such that $f(x_j) > f(x_{j-1})$ and fit a polynomial of degree $j-1$ or until we find $f(x_1) < f(x_0)$ and then fit a constant.

To see that the described procedure is that given by Anderson's rule, we note that

$$f(x_j) > f(x_{j-1})$$

if and only if

$$[\tfrac{1}{2}(r+\sum_{i=j}^{m} w_i)]^{-\frac{1}{2}N-z}(\prod_{i=j}^{m} C_i) > [\tfrac{1}{2}(r+\sum_{i=j-1}^{m} w_i)]^{-\frac{1}{2}N-z}(\prod_{i=j-1}^{m} C_i)$$

if and only if

$$w_{j-1}/(r + \sum_{i=j}^{m} w_i) > c_{j-1}^{2/(N+2z)} - 1$$

In Chapter 4, T. W. Anderson's rule is compared to the rule of taking $(\bar{z}_0,\ldots,\bar{z}_m)$ as the mode of f when $\beta = \infty$ as well as other rules.

3.3 V_1 known, V_2 unknown

3.3.0 Introduction

The sections 3.2.1, 3.2.2, and 3.2.3 differ from 3.3.1, 3.3.2, 3.3.3 primarily in just one respect. In the latter sections, it is assumed that σ^2 is known. We shall make Assumptions 1 through 5 as given in 3.2.2 except that we now assume σ^2 is known.

We again wish to estimate z_i for $i = k,k+1,\ldots,m$. If we wish to obtain "maximum likelihood estimates", the appropriate density function is the density of $\tilde{y}$. Thus, by Corollary 1b, we wish to find $v_k,v_{k+1},\ldots,v_m$ which maximize

$$\prod_{i=k}^{m} v_i^{(\frac{1}{2})} e^{-(\frac{1}{2})v_i w_i} \qquad (14)$$

where $w_i = (\bar{\theta}_i-\mu_i)^2$ for $i = k,k+1,\ldots,m$.

First, we consider maximizing the quantity in (14) with no restrictions. We may consider the problem as that of finding $v_k,v_{k+1},\ldots v_m$ which minimize

$$\sum_{i=k}^{m} [v_i-g_i\log(v_i)]w_i \qquad (15)$$

where $g_i = 1/w_i$ for $i = k,k+1,\ldots,m$,

Clearly

$$\bar{v}_i = 1/w_i$$

for $i = k,k+1,\ldots,m$.

Hence,

$$\bar{z}_i = \bar{v}_i/v_{m+1} = \sigma^2/w_i$$

for $i = k,k+1,\ldots,m$.

From Theorem 2C, we see that

$$z_i\tilde{w}_i/\sigma^2$$

for $i = k,k+1,\ldots,m$; has a central chi-square distribution with one degree of freedom. We maximize the quantity in (15) without imposing the restriction that $v_i < v_{m+1}$ for $i = k,k+1,\ldots,m$. Thus, the estimate $\bar{z}_i$ could exceed one.

If for some $i \in \{k,k+1,\ldots,m\}$, we wished to test the hypothesis H_o: $z_i=1$ against the alternative that H_1: $z_i<1$, it seems that a reasonable procedure might be to reject H_o: if w_i/σ^2 were sufficiently large. When $z_i = 1$, the basic rule gives us $\bar{\mu}_i = \mu_i$.

If a classical statistician were to test the hypothesis H_o^1: $\theta_i=\mu_i$ against the alternative H_1^1: $\theta_i\neq\mu_i$,

he would most likely reject H_o^1: if w_i/σ^2 were suffiently large.

In either of the above cases, the test statistic has a central chi-square distribution with one degree of freedom when the null hypothesis is true.

Hence, after comparing these results with those of 3.2.0, it appears that regardless of the knowledge of σ^2, the procedure of taking the estimates of z_i to be the maximizing values of the density of $\tilde{y}$, yields results which closely resemble those of the classical procedure. It also seems that in both cases we could improve our estimates by imposing the restriction that $v_i \leq v_{m+1}$ for $i = k,k+1,\ldots,m$. This is the primary goal of the next section.

3.3.1 Estimating the hyperparameters: flat prior, σ^2 known

In this section, we make Assumptions 1 through 5 as given in 3.2.2, except that we assume σ^2 is known. We add the following assumption:

6': Let $X = k,k+1,\ldots,m$ and let "$\preceq$" denote a quasi-order on X. There is a set, B, such that $(v_k,v_{k+1},\ldots,v_m) \in B$. We require the estimate of $(v_k,v_{k+1},\ldots,v_m)$, $(\bar{v}_k,\bar{v}_{k+1},\ldots,\bar{v}_m)$, to be an

element of B. We also require B to be defined in such a way that $g = (g_k, g_{k+1}, \ldots, g_m) \in B$ if and only if g is isotonic with respect to $\leq$, and $0 < g \leq v_{m+1}$.

Example

Consider again Example 1 of the introduction to this dissertation. Suppose that now σ^2 is known. Suppose we wish to obtain a polynomial estimate

$$\sum_{i=0}^{m} \theta_i Q_i$$

with a degree of k-1 or higher. Then we may take

$$\sigma_i^2 = \infty \text{ for } i = 0,1,\ldots,k-1.$$

Suppose we wish to express the opinion that the regression is smooth by requiring

$$\sigma_k^2 \geq \sigma_{k+1}^2 \geq \cdots \geq \sigma_m^2 ,$$

or

$$0 < z_k \leq z_{k+1} \leq \cdots \leq z_m < 1$$

or

$$0 < v_k \leq v_{k+1} \leq \cdots \leq v_m < v_{m+1} .$$

Suppose also, we are willing the replace the restriction

$v_m < v_{m+1}$ with $v_m \leq v_{m+1}$. Then we define the quasi-order $\precsim$ on $X = \{k,k+1,\ldots,m\}$ as the usual total order $\leq$. Then v is isotonic if and only if

$$v_k \leq v_{k+1} \leq \cdots \leq v_m.$$

We add the restrictions that $0 < v_i \leq v_{m+1}$ for $i = k,k+1,\ldots,m$. Then we are requiring v to be isotonic with bounds 0 and v_{m+1}. The set B of Assumption 6' would be defined as

$$B = \{(v_k,\ldots,v_m) \; ; \; 0 < v_k \leq v_{k+1} \leq \cdots \leq v_m \leq v_{m+1}\}.$$

The main purpose of this section is to describe the estimate $\bar{v} = (\bar{v}_k,\bar{v}_{k+1},\ldots,\bar{v}_m)$ with Assumption 6'. We wish to minimize the quantity in (15) subject to v being isotonic and $0 < v \leq v_{m+1}$.

Theorem 4

Let g_i and w_i be positive and finite for $i = k,k+1,\ldots,m$. The sum

$$\sum_{i=k}^{m} [v_i - g_i \log(v_i)] w_i$$

is minimized over isotonic v for which $0 < v \leq v_{m+1}$ by taking $v = \bar{v}$, where $\bar{v}_i$ is the bounded isotonic regression of g_i with weights w_i for $i = k,k+1,\ldots,m$ and bounds 0 and v_{m+1}. The minimizing function is

unique. (See the appendix of this thesis for definitions and key results on bounded isotonic regression.)

Proof: Without loss of generality, we may assume that the bounds are 0 and 1. To see this, define $\bar{g} = g/v_{m+1}$ and $\bar{v} = v_{m+1}\underline{v}$, where $\underline{v}$ minimizes

$$\sum [\underline{v}_i - \bar{g}_i \log(\underline{v}_i)]w_i$$

over all isotonic $\underline{v}$ such that $0 < \underline{v} \leq 1$. (Note: All summations in this proof will be taken from i=k to m. Hence, we shall omit showing the limits.)

Then

$$\sum [\bar{\underline{v}}_i - \bar{g}_i \log(\bar{\underline{v}}_i)]w_i \leq \sum [\underline{v}_i - \bar{g}_i \log(\underline{v}_i)]w_i$$

for all isotonic $\underline{v}$ such that $0 < \underline{v} \leq 1$

if and only if

$$\sum \left\{ \left(\frac{\bar{v}_i}{v_{m+1}} - \frac{g_i}{v_{m+1}} \right) [\log(\bar{v}_i) - \log(v_{m+1})] \right\} w_i \leq \sum \left\{ \left(\frac{v_i}{v_{m+1}} - \frac{g_i}{v_{m+1}} \right) [\log(v_i) - \log(v_{m+1})] \right\} w_i$$

for all isotonic v_i such that $0 < v_i \leq v_{m+1}$

if and only if

$$\frac{1}{v_{m+1}} \cdot \sum [\bar{v}_i - g_i \log(\bar{v}_i)]w_i + \frac{1}{v_{m+1}} \cdot \log(v_{m+1}) \sum g_i w_i$$

$$\frac{1}{v_{m+1}} \cdot \sum [v_i - g_i \log(v_i)]w_i + \frac{1}{v_{m+1}} \cdot \log(v_{m+1}) \sum g_i w_i$$

for all isotonic v_i such that $0 < v_i \leq v_{m+1}$.

We shall now assume the bounds are 0 and 1. By Lemma 1, in section 3.2.1, it suffices to minimize

$$\sum \Delta(g_i, v_i) w_i \tag{16}$$

subject to v being isotonic and $0 < v \leq 1$. By Theorem A2 in the appendix of this thesis, the bounded isotonic regression $\bar{v}$ uniquely minimizes the quantity in (16) in the class of isotonic functions v such that $0 \leq v \leq 1$. By Theorem A1 in the appendix of this thesis, $\bar{v} > 0$. []

Thus, for a given quasi-order $\overline{\leq}$ on X, we may apply Theorem 4 to minimize the quantity in (15) subject to v being isotonic and $0 < v \leq v_{m+1}$. The minimizing function $\bar{v}_i$ is the bounded isotonic regression of g_i with weights w_i for $i = k, k+1, \ldots, m$. It will be seen in the next section that this estimate of the v_i's is a limit of the estimates derived in that section.

If the quasi-order, $\overline{\leq}$, on X is a simple order (i.e. $i, j \in X$; $i \overline{\leq} j$ if and only if $i \leq j$, method for the computation of the bounded isotonic regression is given in the appendix of this thesis. For methods of computation of bounded isotonic regression with other orders, see [Barlow, Bartholomew, Bremmer, and Brunk, 1972].

3.3.2 Estimating the hyperparameters: gamma prior, σ^2 Known

This section contains a method for estimating the unknown v_i's when they are assumed to have "truncated" gamma distributions. The exact distribution will be specified in Assumption 7'. We shall see that with an appropriate choice of the parameters of the priors, we may obtain the rule (the rule obtained by minimizing (15) subject to v being isotonic and bounded) given in the preceding section. This section is concluded with an example. In the example, a particular quasi-order is assumed. The resulting rule, for several choices of the prior parameters, dominates the least squares estimator when using squared error loss. For a particular choice of the prior parameters, the resulting rule is seen to be a plus-rule version of the James-Stein estimator which uniformly improves on the James-Stein rule.

Again, we shall make Assumptions 1 through 5 as stated in 3.2.2 except that we assume σ^2 is known. We shall also assume 6' as given in 3.3.1. In addition we assume

7'. The prior distribution of $(\tilde{v}_k,\ldots,\tilde{v}_m)$ is

$$\propto \Big[\prod_{i=k}^{m} v_i^{\gamma_i - 1} e^{-(v_i/\beta_i)}\Big] \cdot 1_B(v_k,\ldots,v_m) \qquad (17)$$

where $\gamma_i > 0$ and $\beta_i > 0$ for $i = k,k+1,\ldots,m$.

The posterior distribution of $(\tilde{v}_k,\tilde{v}_{k+1},\ldots,\tilde{v}_m)$ given $\tilde{y} = y$ is proportional to the product of the quantities given in (14) and (17). The product is

$$\Big[\prod_{i=k}^{m} v_i^{(\frac{1}{2})(2\gamma_i - 1)} e^{-(\frac{1}{2})v_i w_i}\Big] \cdot 1_B(v_k,\ldots,v_m) \qquad (18)$$

where we have redefined w_i as

$$w_i = (\bar{\theta}_i - \mu_i)^2 + 2/\beta_i$$

for $i = k,k+1,\ldots,m$.

If we define g_i by

$$g_i = (2\gamma_i - 1)/w_i \qquad (19)$$

for $i = k,k+1,\ldots,m$, then we see that finding the v which maximizes (18) is equivalent to finding v which minimizes

$$\sum_{i=k}^{m} [v_i - g_i \log(v_i)]w_i \qquad (20)$$

Subject to v being isotonic and $0 < v \leq v_{m+1}$.

We may apply Theorem 4 to minimize the quantity in (20) when $g_i > 0$ for $i = k,k+1,\ldots,m$. Thus when $g_i > 0$ for $i = k,\ldots,m$, $\bar{v}$ is the unique

minimizing function of the quantity in (20), where $\bar{v}$ is the bounded isotonic regression of g with respect to the weights w and the bounds 0 and v_{m+1}.

Observe that if $\gamma_i = 1$ for $i = k,k+1,\ldots,m$, then the prior distribution of the v_i's, given in Assumption 7', will become "flat" as β_i becomes large for $i = k,k+1,\ldots,m$. That such a prior is appropriate to represent vague knowledge (apart from Assumption 6') is substantiated not only by its shape, but by the fact that the g_i's and the w_i's of this section converge to the g_i's and w_i's of the previous section as $\beta_i \to \infty$ and $\gamma_i = 1$ for $i = k,k+1,\ldots,m$. Hence, by Definition A4 and Theorem A3 of the appendix, the $\bar{v}$ of this section converges to the $\bar{v}$ of the previous section when $\gamma_i = 1$ and $\beta_i \to \infty$ for $i = k,k+1,\ldots,m$. In the previous section we had no prior distribution on the v_i's.

As was discussed in 3.2.2, the problem of expressing prior knowledge which leads to a negative value of some g_i can often be overcome by assuming that z_i is known.

We now present the example mentioned in the introductory chapter of this thesis as well as in 3.3.0.

Example

We make Assumptions 1 through 5 as given in 3.2.2 except that we assume σ^2 is known. We also make Assumptions 6' and 7' as given in 3.3.1 and 3.3.2 respectively. We make the following additional assumptions:

(i) $k = 0$.

(ii) $V = I$.

(iii) $\sigma_0^2 = \sigma_1^2 = \cdots = \sigma_m^2$ (unknown).

(iv) $\mu = 0$.

(v) $\sigma^2 = 1$.

(vi) $m \geq 2$.

(vii) $a > 0$, where $a = \sum_{i=0}^{m} (2\gamma_i - 1)$.

Assumptions (iii) and (iv) imply that

$$B = \{v_i; \quad 0 < v_0 = v_1 = \cdots = v_m < 1\}$$

The quasi-order $\bar{\prec}$ on $X = \{0,1,\ldots,m\}$ is defined by

$$i \bar{\prec} j$$

for all $i,j \in X$.

Assumptions (iii) and (iv) imply that the basic rule becomes

$$\bar{\mu} = (1 - \bar{z}_o)\bar{\theta}$$

where $\bar{\theta}$ is the least squares estimate of θ.

To obtain $\bar{z}_o$, we must minimize the quantity given in (20), which is

$$\sum_{i=0}^{m} [v_i - g_i \log(v_i)] w_i ,$$

subject to v being isotonic and $0 < v \leq 1$.

For this particular example, an expression for $\bar{z}_o$ is easier to obtain by reformulating the problem somewhat. Using the fact that

$$v_o = v_1 = \cdots = v_m,$$

the quantity to be minimized is

$$v_o(\sum_{i=0}^{m} w_i) - [\log(v_o)](\sum_{i=0}^{m} g_i w_i)$$

Define $\bar{w}_o$ and $\bar{g}_o$ by

$$\bar{w}_o = \sum_{i=0}^{m} w_i \text{ and } \bar{g}_o \bar{w}_o = \sum_{i=0}^{m} g_i w_i = a .$$

Then we wish to minimize

$$[v_o - \bar{g}_o \log(v_o)] \bar{w}_o \qquad (21)$$

subject to $0 < v_o \leq 1$.

Thus, the set B, defined above, may be

replaced with the set $\bar{B}$, where $\bar{B} = (0,1]$. The set X is likewise replaced with the set $\bar{X} = \{0\}$. The quasi-order $\bar{\leq}$ is replaced with the usual total-order, $\leq$, on $\bar{X}$.

Hence, our problem is now to minimize (21) subject to v_o being isotonic on $\bar{X}$ and $0 < v_o \leq 1$. Thus, by Theorem A4 of the appendix to this thesis,

$$\bar{v}_o = \min \{1, \bar{g}_o\} .$$

Therefore

$$\bar{z}_o = \bar{v}_o / v_{m+1} = \bar{v}_o = \min \{1, \bar{g}_o\} .$$

For the remainder of this example, we shall assume a loss function L defined by

$$L(\theta,\delta) = (\theta-\delta)'(\theta-\delta).$$

for all parameters θ and estimates δ.

We now define S by

$$S = \bar{\theta}'\bar{\theta}.$$

Then Baranchik's Theorem [Efron and Morris, 1973] states that the rule, δ_1, defined by

$$\delta_1 = [1 - \frac{m-1}{S} \tau(S)]\bar{\theta}$$

will dominate $\bar{\theta}$, if

(i) $\tau(S)$ is nondecreasing in S, $\tau(S) \geq 0$ and

(ii) $\tau(S) \to t$ as $S \to \infty$, and $0 < t \leq 2$.

Since $\mu = 0$, w_i, as defined in (18), becomes

$$w_i = \bar{\theta}_i^2 + 2/\beta_i$$

for $i = 0,1,\ldots,m$. If we define b by

$$b = 2\sum_{i=0}^{m} 1/\beta_i,$$

then

$$\bar{w}_o = \sum_{i=0}^{m} w_i = S+b .$$

Thus

$$\bar{g}_o = a/\bar{w}_o = a/(S+b)$$

implies that

$$\bar{z}_o = \min\left\{1 , \frac{a}{S+b}\right\} .$$

We define $\tau(S)$ by

$$\tau(S) = \frac{S}{m-1}\,\bar{z}_o = \begin{cases} \dfrac{S}{m-1} & \text{if } S \leq a-b \\ \dfrac{S}{m-1}\cdot\dfrac{a}{S+b} & \text{otherwise .} \end{cases}$$

It is easy to see that condition (i) of Baranchik's Theorem is satisfied. Since $\tau(S) \to \frac{a}{m-1}$ as $S \to \infty$, the rule

$$\bar{\mu} = (1-\bar{z}_o)\bar{\theta}$$

will dominate $\bar{\theta}$ when $a \leq 2(m-1)$ (See Assumption vii).

In particular, the methods of 3.3.1, which would correspond to $b = 0$ and $a = m+1$, yield a rule which

dominates $\bar{\theta}$ when $m \geq 3$.

Now, the James-Stein rule is

$$\delta_2 = [1 - (m-1)/S]\bar{\theta} .$$

It is well known that δ_2 dominates $\bar{\theta}$. But the plus-rule version of the James-Stein rule, which is

$$\delta_2^+ = [1 - \min \{1,(m-1)/S\}]\bar{\theta},$$

uniformly improves on δ_2 [Efron and Morris, 1973].

Clearly, if $a = m-1$ and $b = 0$, we have

$$\tilde{\mu} = \delta_2^+ .$$

3.3.3 Estimating the hyperparameters: Bernoulli prior, σ^2 Known .

As in section 3.2.3, we shall restrict the possible values of the parameters z_i, for $i = 0,1,\ldots,m$ to zero and one in order to obtain "classical-like" estimates from the basic rule.

We make Assumptions 1 through 5 of 3.2.2 except that we assume σ^2 is known.

We shall also assume $k = 0$ and $\mu = 0$.

The likelihood function which is proportional to the quantity given in (14) may be written so as to be

$$\propto \prod_{i=0}^{m} z_i^{\frac{1}{2}} e^{-\frac{1}{2} z_i w_i} \qquad (22)$$

where $w_i = \frac{1}{\sigma^2}\bar{\theta}_i^2$ for $i = 0,1,\ldots,m$.

Since the result in (22) was obtained under the assumption that $0 < z_i < 1$ for $i = 0,1,\ldots,m$, we use the following prior distribution:

$$\prod_{i=0}^{m} P[\tilde{z}_i = z_i] \qquad (23)$$

where for $0 < \gamma < \frac{1}{2}$ and $0 < \epsilon < \frac{1}{2}$

$$P[\tilde{z}_i = z_i] = \begin{cases} p_i & \text{if } z_i = 1-\gamma \\ q_i = 1-p_i & \text{if } z_i = \epsilon \end{cases}$$

for $i = 0,1,\ldots,m$.

Thus, the posterior is

$$\propto \prod_{i=0}^{m} z_i^{\frac{1}{2}} P[\tilde{z}_i = z_i] e^{-(\frac{1}{2}) z_i w_i} . \qquad (24)$$

For ϵ and γ sufficiently small,

$$\sqrt{\epsilon}\, q_i e^{-\frac{1}{2}\epsilon w_i} < (1-\gamma)^{\frac{1}{2}} p_i e^{-\frac{1}{2}(1-\gamma) w_i}$$

for $i = 0,1,\ldots,m$. Thus, the value of z_i which maximizes the posterior when ϵ and γ are sufficiently small is $z_i = 1-\gamma$ for $i = 0,1,\ldots,m$.

As in section 3.2.3, we let

$$p_i = C_i\sqrt{\epsilon} \quad , \quad q_i = 1-p_i$$

for $i = 0,1,\ldots,m$. Then

$$z_i^{\frac{1}{2}}\, P[\tilde{z}_i=z_i] = \sqrt{\epsilon}\, C_i^{1_{\{1\}}(z_i)}(1-\sqrt{\epsilon}\, C_i)^{1_{\{\epsilon\}}(z_i)} 1_{\{\epsilon,1\}}(z_i)$$

for $i = 0,1,\ldots,m$. Thus, the posterior distribution converges as $\epsilon \to 0$ and $\gamma \to 0$ to a distribution which is

$$\propto \prod_{i=0}^{m} [C_i e^{-(\frac{1}{2})w_i}]^{z_i} .$$

That is, the $\tilde{z}_i$'s are independent binary variables leading to a posterior distribution which is a product of Bernoulli distributions with parameters given by

$$\frac{C_i}{e^{\frac{1}{2}w_i} + C_i}$$

for the probability that $\tilde{z}_i = 1$ for $i = 0,1,\ldots,m$.

The mode is obtained by taking $\bar{z}_i = 1$ if and only if

$$w_i < 2\log(C_i)$$

for $i = 0,1,\ldots,m$. That is, we perform the classical test of the hypothesis

$$H_o^{(i)}: \theta_i=0 \quad \text{against} \quad H_1^{(i)}: \theta_i\neq 0$$

with critical region defined by

$$w_i \geq 2\log C_i$$

(w_i is the square of the usual test statistic) and take

$$\bar{\mu}_i = \begin{cases} \bar{\theta}_i & \text{if } H_o^{(i)} \text{ is rejected} \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 0,1,\ldots,m$.

Since the tests are independent, there is no difficulty in obtaining the overall level of significance for the test procedure.

4. MONTE CARLO COMPARISONS: σ^2 UNKNOWN

4.0 Introduction

Fifteen different polynomials were considered; five were quadratics; five were cubics; and five were quartics. For each polynomial, say $R(x)$, a random sample of size two was taken from a population which was normal with a mean equal to $R(x)$, and a variance of one. At each of the seven points $x = -3,-2,-1,0,1,2,3$. Thus, a total of $N = 14$ observations was obtained. Based on the 14 observations, five estimators of R were evaluated. The estimators used were

1. GM = Gauss Markov = Least squares
2. GMT = T. W. Anderson's rule (see 3.2.3 or [Anderson, 1971]
3. GMD = method recommended in [Draper and Smith, 1966]
4. GMP = the rule derived in 3.2.3 (the mode of a posterior distribution)
5. IR = the basic rule using Isotonic Regression derived in 3.2.2.

The loss was computed for each of the five rules. The loss attained in using the estimate $\bar{R}$ to estimate

the polynomial R was defined as "the average squared error loss over the interval $I = [-3, 3]$". That is

$$L(R,\bar{R}) = (1/6)\int_I [R(x)-\bar{R}(x)]^2 dx.$$

This loss function is discussed in 3.1.

The process was repeated twenty times for each polynomial $R(x)$. That is, in each of the twenty times, fourteen observations were taken. From these fourteen observations a loss for each of the five estimates was computed. Thus, for the polynomial $R(x)$, we obtained twenty losses using the GM rule, twenty losses using the GMT rule, etc. Section 4.1 contains tables in which the average and the variance of the twenty losses for each estimator is given. The tables also include the number of times the estimated polynomial was of the same degree as the actual polynomial for each estimator.

The average loss and the variance of the losses are denoted by AL and VL respectively, and defined by

$$AL = (1/20)\sum_{i=1}^{20} L(R,\bar{R}_i) \qquad \text{and}$$

$$VL = (1/19)\sum_{i=1}^{20} [L(R,\bar{R}_i) - AL]^2$$

where $\bar{R}_i$ denotes the estimate of R obtained on the

i^{th} repitition of the experiment for $i = 1,2,\ldots,20$.

The polynomials, R, were expressed as the sum of orthonormal polynomials, ψ_i (where ψ_i is of degree i) for $i = 0,1,\ldots,6$. That is

$$R(x) = \sum_{i=0}^{6} \theta_i \psi_i(x)$$

where

$$\sum_{x=-3}^{3} \psi_i(x)\psi_j(x) = \delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

The tables in Section 4.1 identify R by specifying the values of θ_i for $i = 0,1,\ldots,6$.

The design matrix used was

$$Q = [\psi_j(x_i)] \quad \begin{matrix} i = 1,2,\ldots,14 \\ j = 0,1,\ldots,6 \end{matrix} .$$

Thus, the observed vector y of the random vector $\tilde{y}$, where

$$\tilde{y} \sim N_{14}(Q\theta, I)$$

led to the GM estimate $\bar{\theta}$, where $\bar{\theta} = Q'y$. Thus, the **GM rule** produced a six degree polynomial estimate in each case. This is to be expected since

$$P[\bar{\theta}_6 = 0] = 0.$$

In each of the other four estimation procedures, it was

assumed that the coefficient of the constant term, θ_0, was known to be nonzero.

Thus, in the GMD procedure, the hypothesis $H_o^{(i)}: \theta_i=0$ against the alternative $H_1^{(i)}: \theta_i\neq 0$ was tested for $i = 1,2,\ldots,6$. In each test, a standard F-test was performed using a significance level of 5%. The denominator of the test statistic was the same in each of the six tests. In fact, it was $y'y - \bar{\theta}'\bar{\theta}$.

The GMT rule was obtained as described in 3.2.3 with the C_i's chosen so that each individual test would have a significance level of 5%.

For the GMP rule, the C_i's of the GMT rule were used. The domain of the posterior distribution was unrestricted (except that each z_i was either zero or one for $i = 1,2,\ldots,6$) so that it was possible for the GMP rule to produce estimates, $\underline{\theta}_i$, of θ_i in which

$$\underline{\theta}_1\neq 0,\ \underline{\theta}_2=0,\ \underline{\theta}_3=0,\ \underline{\theta}_4=0,\ \underline{\theta}_5=0,\ \underline{\theta}_6\neq 0.$$

Such a result could not be obtained from the GMT rule since if the hypothesis that $\theta_6=0$ is rejected, the parameter elimination is stopped and a polynomial of degree six is fitted.

The IR rule that was used is described in 3.2.2. A "truncated beta" prior distribution was put on the z_i's. Vague knowledge was expressed regarding

$$\frac{1}{\sigma^2} = v_{m+1}$$

by taking $\gamma_{m+1} = 1$ and $\beta_{m+1} = \infty$. In this case, $m=6$, and $N=14$. Then, in order to insure that $n > 0$, where

$$n = N-(m+1-k) - 2\sum_{i=k}^{m}(\gamma_i-1),$$

it was necessary to choose γ_i for $i = 1,2,\ldots,6$ so that

$$2\sum_{i=1}^{6}(\gamma_i-1) < 8.$$

It was also desired to choose the γ_i's to be consistent with the restriction $z_1 \leq z_2 \leq \cdots \leq z_6$. Thus γ_i was taken to be nondecreasing in i. The values used were

$$\gamma_1 = \gamma_2 = 1, \quad \gamma_3 = 1.5, \quad \gamma_4 = \gamma_5 = 2.0, \quad \gamma_6 = 2,49999.$$

These were the only values for the γ_i's that were tried. It would be interesting to see the change in performance of the IR rule with other choices of the γ_i's.

The known value of z_0 was taken as zero in the

in the sense of the procedure described in 3.2.2. The order on the set $X = \{1,2,\ldots,6\}$ used was the usual order, $\leq$. This means that the set B was defined by

$$B = \{(v_1,v_2,\ldots,v_7);\ 0 < v_1 \leq v_2 \leq \ldots \leq v_7\}$$

where $v_7 = 1/\sigma^2$ (presumed unknown).

There was also some Monte Carlo work done by taking ten observations at each of the points $-3,-2,-1,0,1,2,3$. So that in that case $N = 70$ observations on which to base each estimate. In that study, the same loss function was used, but the average loss was based on ten observations of the loss. Also, in that study the IR was based on a flat prior as described in section 3.2.1. In that study the IR rule was compared to the GM and the GMD rules. While this IR rule had no difficulty outperforming the GM rule, especially when the actual degree of the polynomial sampled from was less than five, it did not perform as well as the GMD rule, especially when the actual degree of the polynomial sampled from was less than three.

Based on these observations it seems that in order to have the IR rule compete with the GMD or GMT rules one should express a fairly strong opinion that the z_i's are near 1, unless he has good reason to do

otherwise. It may be that this would not hold if one expressed an opinion other than that of vagueness about $1/\sigma^2$.

4.1 Tabulated Results

Quadratic

(1) $\theta_0 = 30$, $\theta_1 = 20$, $\theta_2 = 10$, $\theta_i = 0$ $\quad i = 3,4,5,6$

	AL	VL	# times correct degree
GM	.78096	.36885	0
GMT	.37350	.24063	17
GMD	.36320	.22308	17
GMP	.36320	.22308	17
IR	.21138	.02536	16

(2) $\theta_0 = 0$, $\theta_1 = 10$, $\theta_2 = 4$, $\theta_i = 0$ $\quad i = 3,4,5,6$

	AL	VL	# times correct degree
GM	1.09831	1.11057	0
GMT	.50750	1.06915	17
GMD	.47061	.94630	18
GMP	.47061	.94630	18
IR	.20241	.07655	16

(3) $\theta_0 = 15$, $\theta_1 = .5$, $\theta_2 = .05$, $\theta_i = 0$ $\quad i = 3,4,5,6$

	AL	VL	# times correct degree
GM	.93745	.66490	0
GMT	.29541	.43898	2
GMD	.29582	.29558	2
GMP	.26715	.29961	2
IR	.14092	.02168	3

(4) $\theta_0 = 4$, $\theta_1 = 10$, $\theta_2 = 20$, $\theta_i = 0$ $\quad i = 3,4,5,6$

	AL	VL	# times correct degree
GM	.78068	.16709	0
GMT	.24000	.04793	17
GMD	.22417	.03364	17
GMP	.22417	.03364	17
IR	.17824	.00978	17

(5) $\theta_0 = 105.8299$, $\theta_1 = 0$, $\theta_2 = 91.6516$, $\theta_i = 0$ $\quad i = 3,4,5,6$

	AL	VL	# times correct degree
GM	.76443	.20960	0
GMT	.28164	.20511	17
GMD	.22023	.11066	17
GMP	.16997	.06971	18
IR	.14183	.01395	16

Cubic

(6) $\theta_0 = 10$, $\theta_1 = 6$, $\theta_2 = -2$, $\theta_3 = 2_1$ $\theta_i = 0$ $\quad i = 4,5,6$

	AL	VL	# times correct degree
GM	.86323	.20929	0
GMT	.39045	.11722	14
GMD	.46637	.15661	13
GMP	.41361	.12101	13
IR	.27024	.04816	15

(7) $\theta_0 = 10,\ \theta_1 = 5,\ \theta_2 = 3,\ \theta_3 = 1,\ \theta_i = 0 \quad i = 4,5,6$

	AL	VL	# times correct degree
GM	.86210	.36926	0
GMT	.44486	.18113	2
GMD	.46246	.20863	2
GMP	.35094	.09404	2
IR	.28643	.02703	5

(8) $\theta_0 = 10,\ \theta_1 = 10,\ \theta_2 = -20,\ \theta_3 = 30,\ \theta_i = 0 \quad i = 4,5,6$

	AL	VL	# times correct degree
GM	.65137	.49892	0
GMT	.31807	.44597	18
GMD	.31806	.44592	18
GMP	.31806	.44592	18
IR	.17346	.01960	19

(9) $\theta_0 = 10,\ \theta_1 = 0,\ \theta_2 = 0,\ \theta_3 = 3,\ \theta_i = 0 \quad i = 4,5,6$

	AL	VL	# times correct degree
GM	.95592	.71419	0
GMT	.52958	.67972	16
GMD	.50934	.70962	15
GMP	.40302	.61922	17
IR	.33509	.10354	18

(10) $\theta_0 = 10,\ \theta_1 = 1,\ \theta_2 = 2,\ \theta_3 = -2,\ \theta_i = 0 \quad i = 4,5,6$

	AL	VL	# times correct degree
GM	.78607	.27760	0
GMT	.53724	.13719	10
GMD	.69695	.14145	12
GMP	.68629	.17239	13
IR	.38964	.06184	14

Quartic

(11) $\theta_0 = 1,\ \theta_1 = 2,\ \theta_2 = 2,\ \theta_3 = 2,\ \theta_4 = 2,\ \theta_i = 0 \quad i = 5,6$

	AL	VL	# times correct degree
GM	.73146	.14500	0
GMT	.73621	.17419	10
GMD	.82296	.03301	8
GMP	.78936	.04893	6
IR	.73354	.10198	9

(12) $\theta_0 = 1,\ \theta_1 = 5,\ \theta_2 = 4,\ \theta_3 = 3,\ \theta_4 = 2,\ \theta_i = 0 \quad i = 5,6$

	AL	VL	# times correct degree
GM	.76137	.23555	0
GMT	.60645	.14911	10
GMD	.67361	.17767	9
GMP	.61509	.17082	10
IR	.58568	.09763	13

(13) $\theta_0 = 1$, $\theta_1 = 10$, $\theta_2 = 0$, $\theta_3 = 2$, $\theta_4 = 10$, $\theta_i = 0$ $i = 5,6$

	AL	VL	# times correct degree
GM	.86864	.38225	0
GMT	.38346	.26674	18
GMD	.49036	.29075	18
GMP	.46181	.28761	18
IR	.23218	.02550	19

(14) $\theta_0 = 1$, $\theta_1 = 15$, $\theta_2 = 10$, $\theta_3 = 5$, $\theta_4 = 1$, $\theta_i = 0$ $i = 5,6$

	AL	VL	# times correct degree
GM	.93403	.51401	0
GMT	.38883	.03788	4
GMD	.36887	.03908	4
GMP	.36708	.03997	4
IR	.33466	.04143	6

(15) $\theta_0 = 1$, $\theta_1 = 50$, $\theta_2 = 25$, $\theta_3 = 50$, $\theta_4 = 10$, $\theta_i = 0$ $i = 5,6$

	AL	VL	# times correct degree
GM	.67665	.11935	0
GMT	.39573	.07236	18
GMD	.39175	.06698	18
GMP	.39175	.06698	18
IR	.36158	.05240	19

4.2 Graphical Results

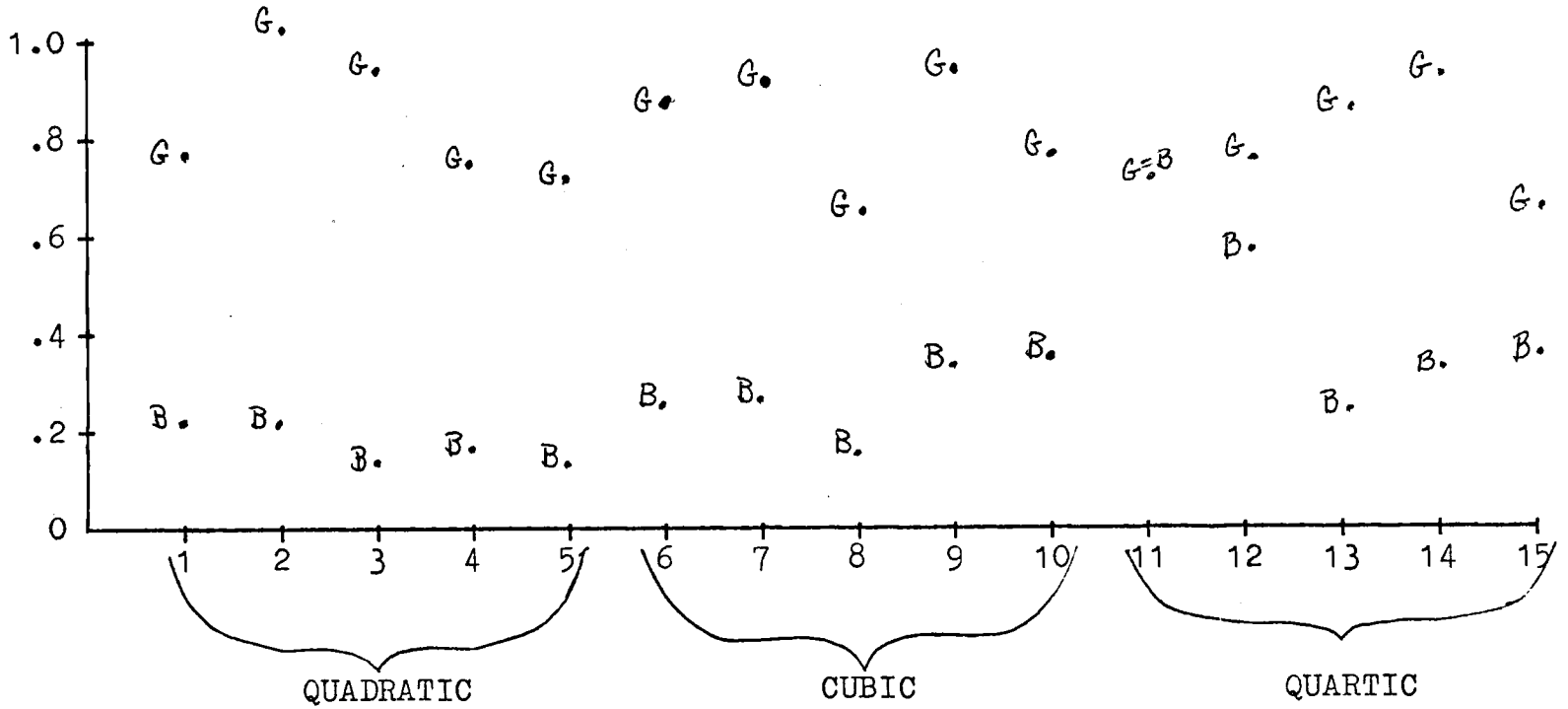
In the graphs which follow, the vertical scale gives the values of the average loss as taken from the preceding tables. The integers on the horizontal scale correspond to the situation described in the like numbered table.

"Best" is defined as the minimum value of the average loss (for each particular situation) attained by the five rules, GM, GMT, GMD, GMP, and IR. The plotted points are labeled with the final letter of the designations for each rule except for the IR rule. For example, GMT is labeled as T and IR is labeled as I .

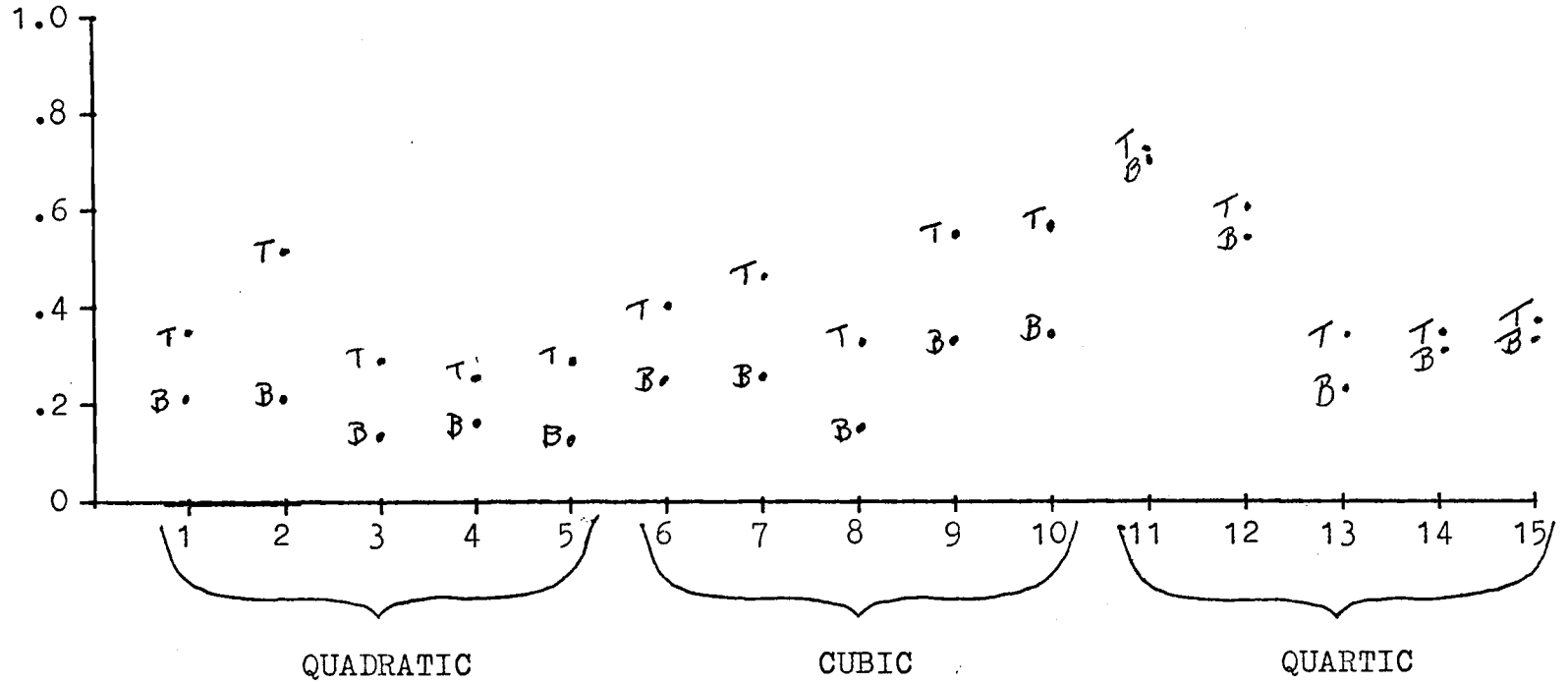
Notice that the IR rule does uniformly well. In fact it is "Best" except for case (11); and then it is within .0021 of the "Best". The IR rule performed well even when the assumption $\sigma_6^2 \leq \sigma_5^2 \leq \ldots \leq \sigma_1^2$ was radically violated.

We also see that the GMP rule (the generalization of GMT) outperforms GMT ten of the 15 times. The GMP rule does as well or outperforms GMD every time.

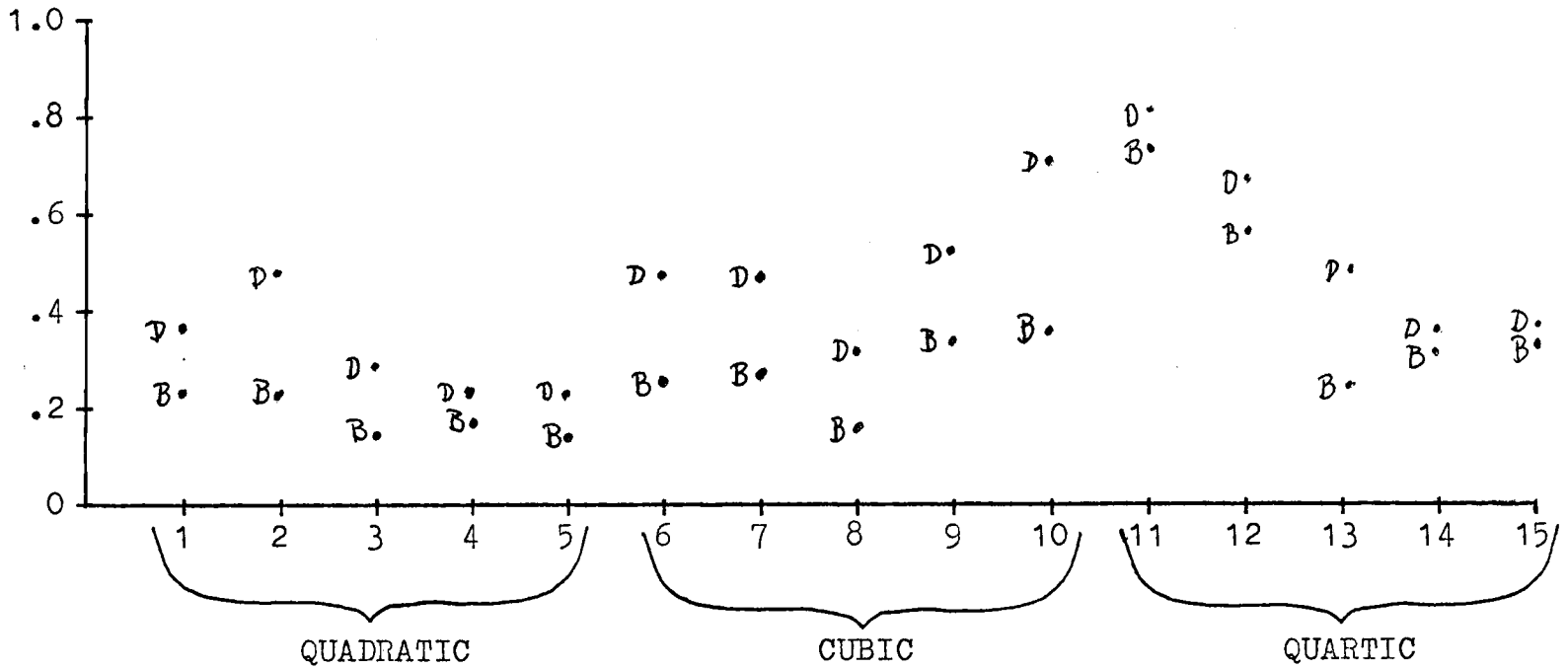
GM v.s. Best


1.0
.8
.6
.4
.2
0
G.
B.
G=B
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
QUADRATIC
CUBIC
QUARTIC

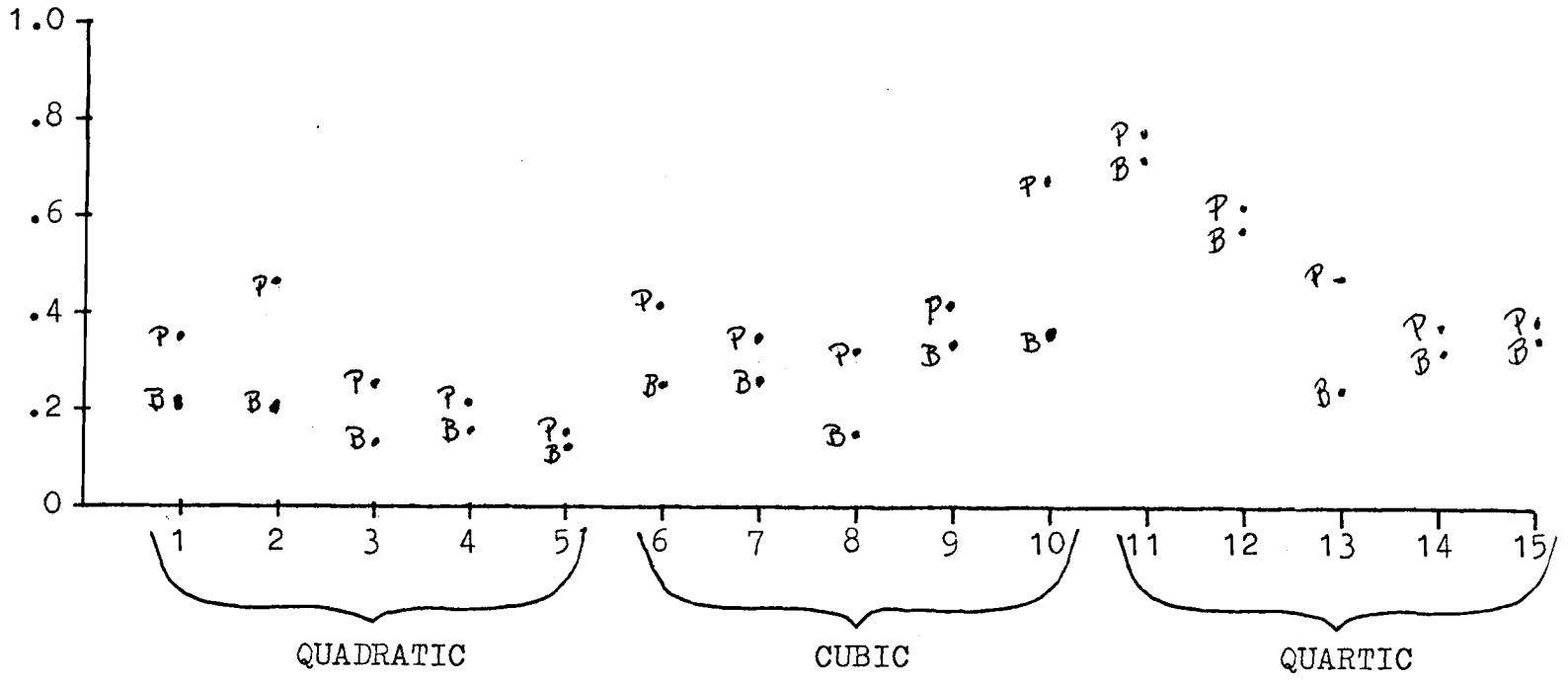
GMT v.s. Best



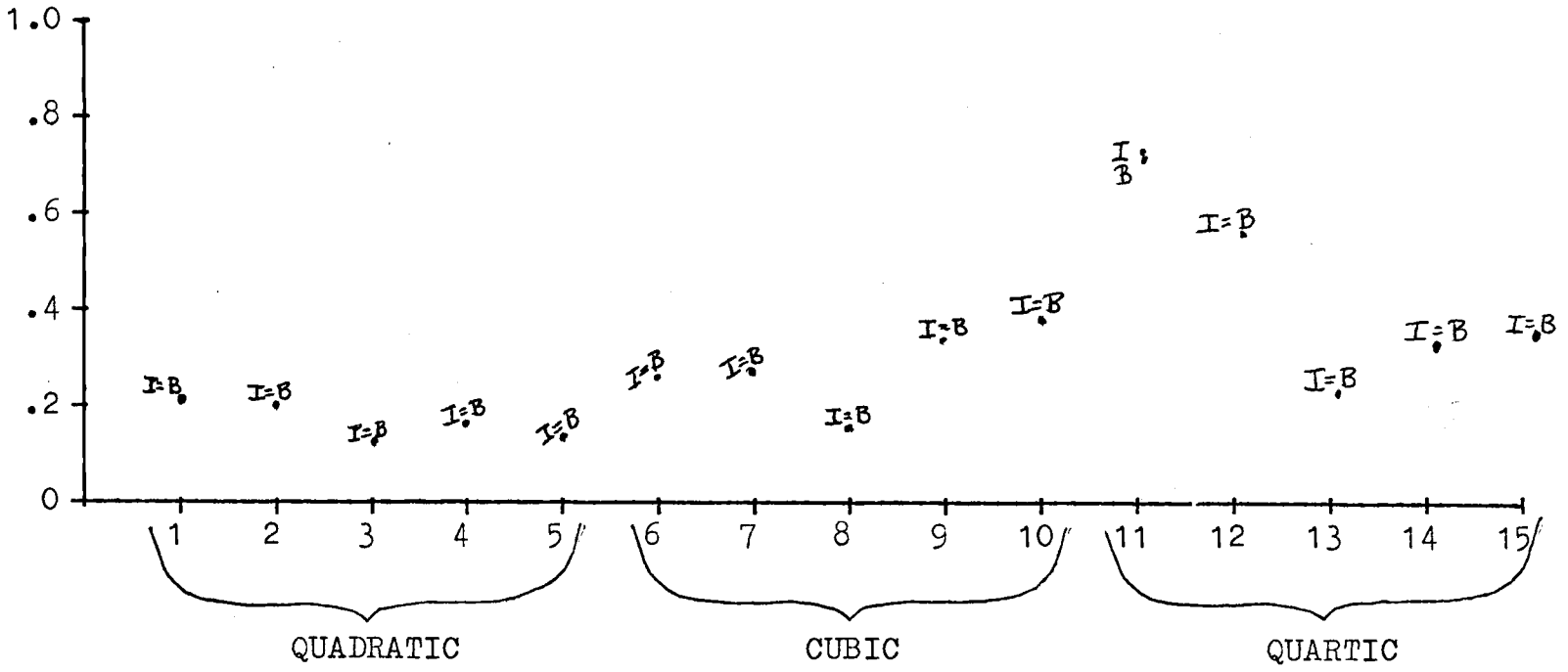
GMD v.s. Best



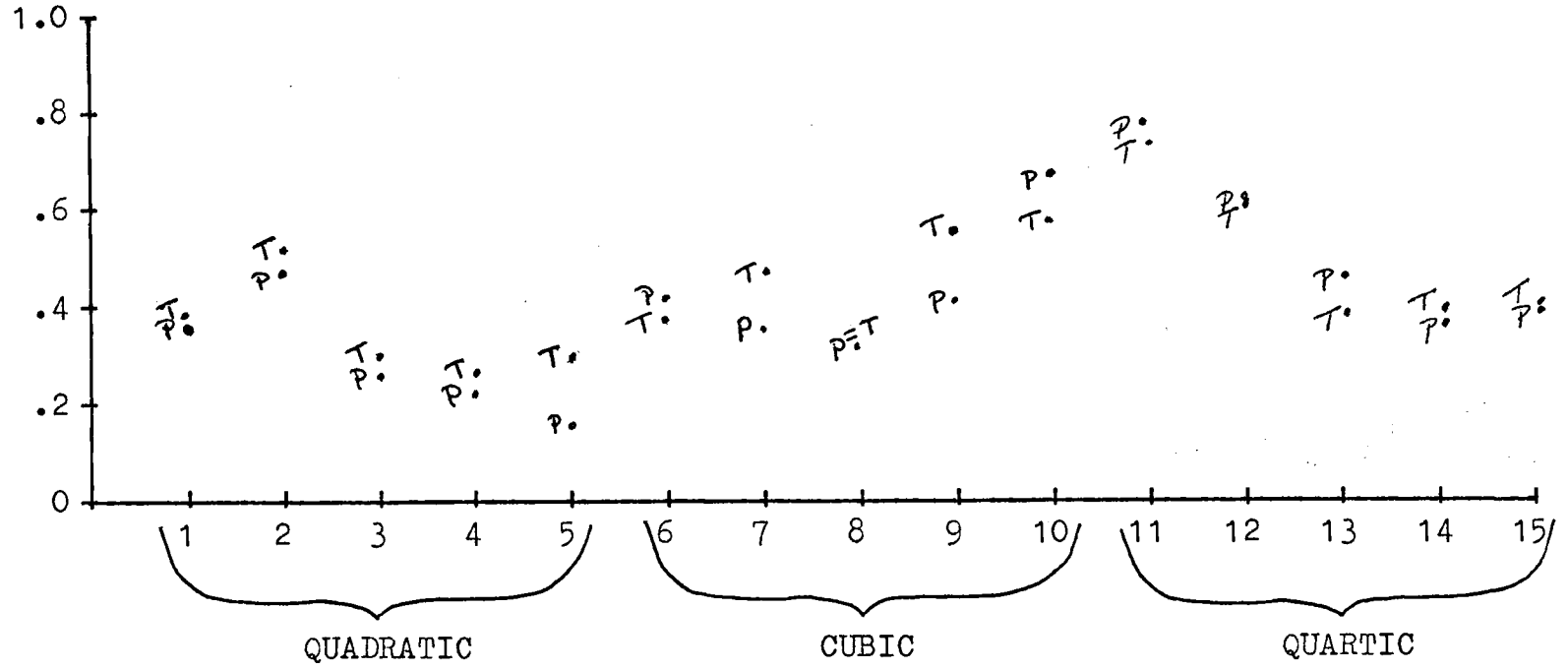
GMP v.s. Best



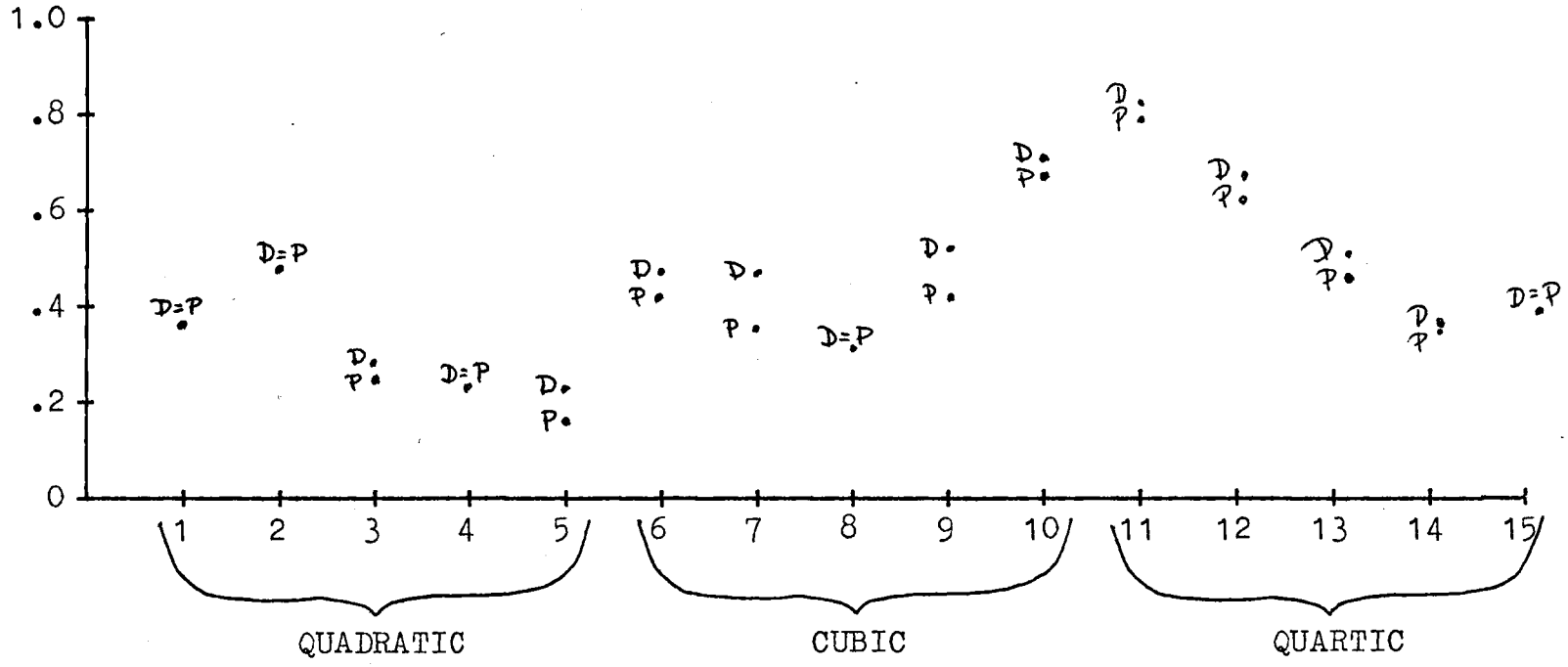
IR v.s. Best


1.0
.8
.6
.4
.2
0
I=B
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
QUADRATIC
CUBIC
QUARTIC

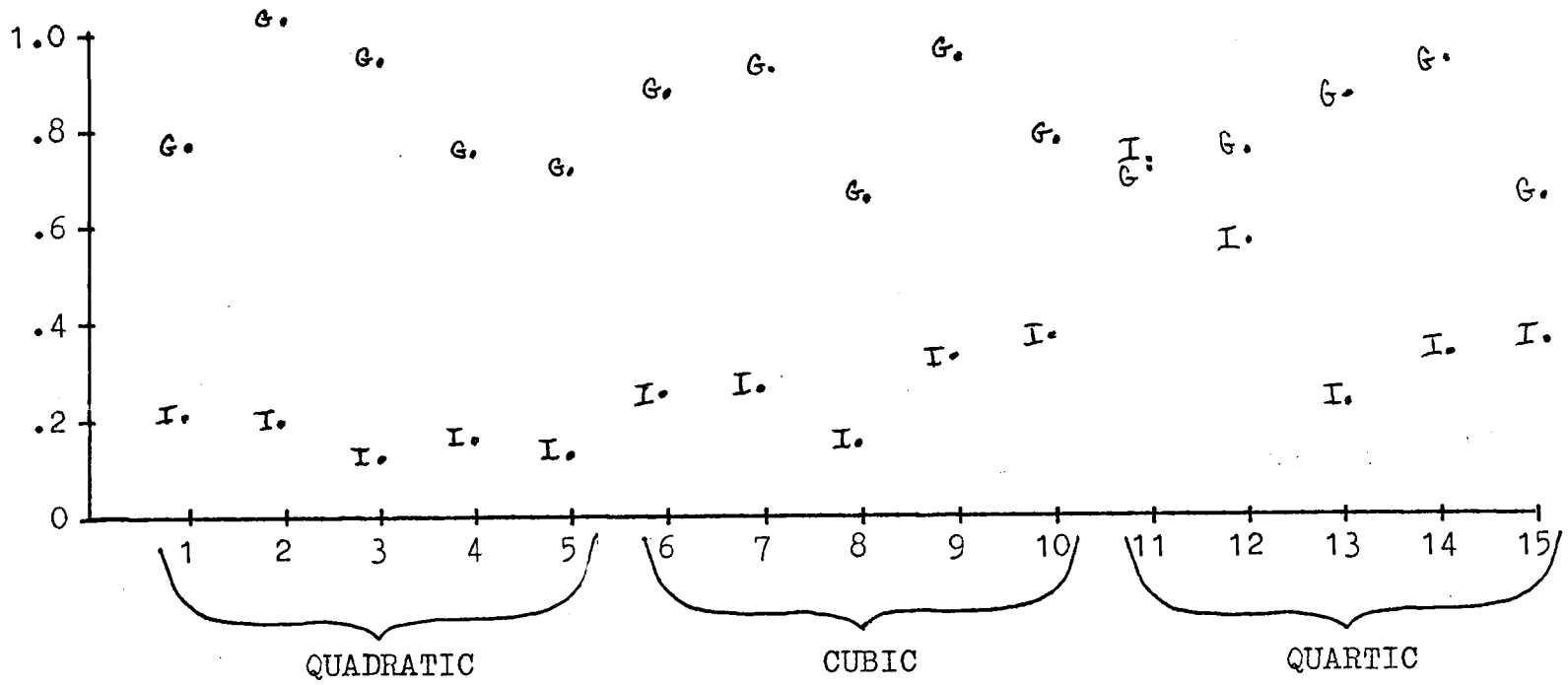
GMP v.s. GMT



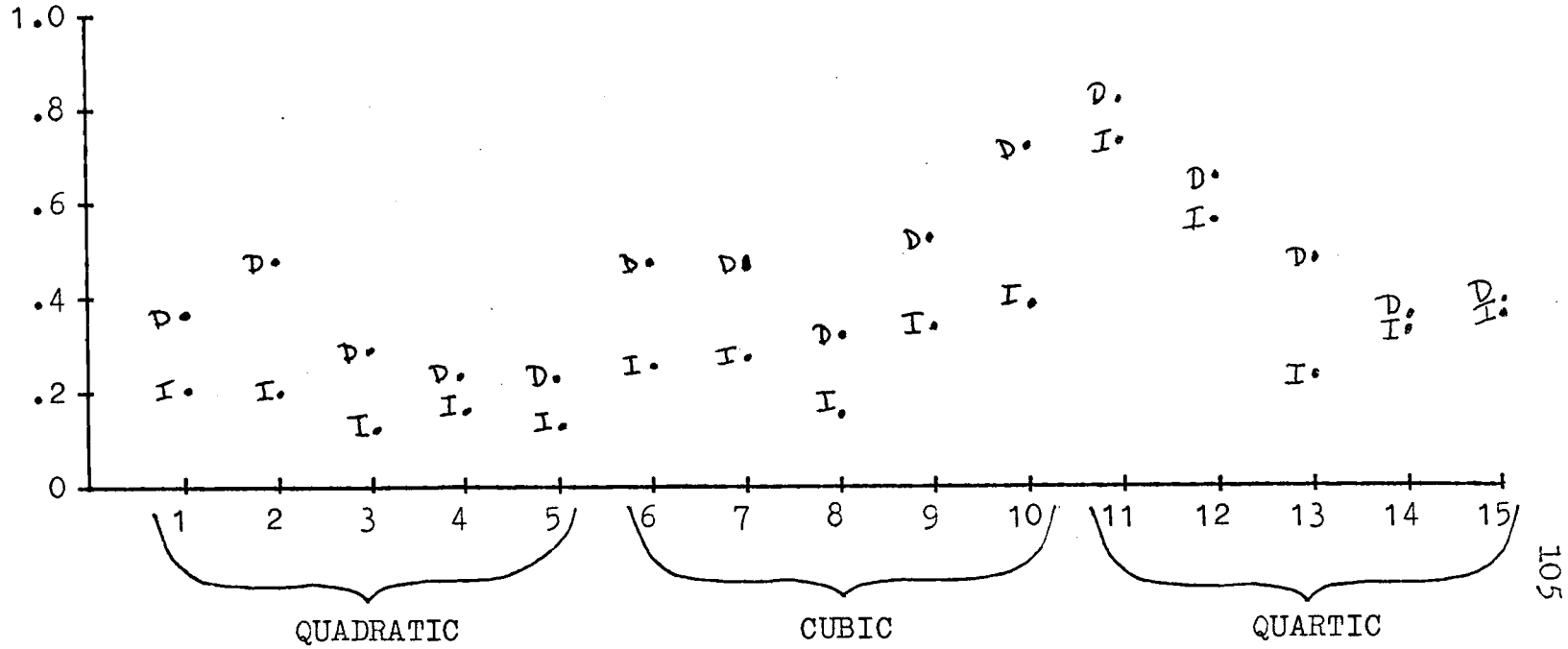
GMP v.s. GMD


1.0
.8
.6
.4
.2
0
D=P
D
P
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
QUADRATIC
CUBIC
QUARTIC

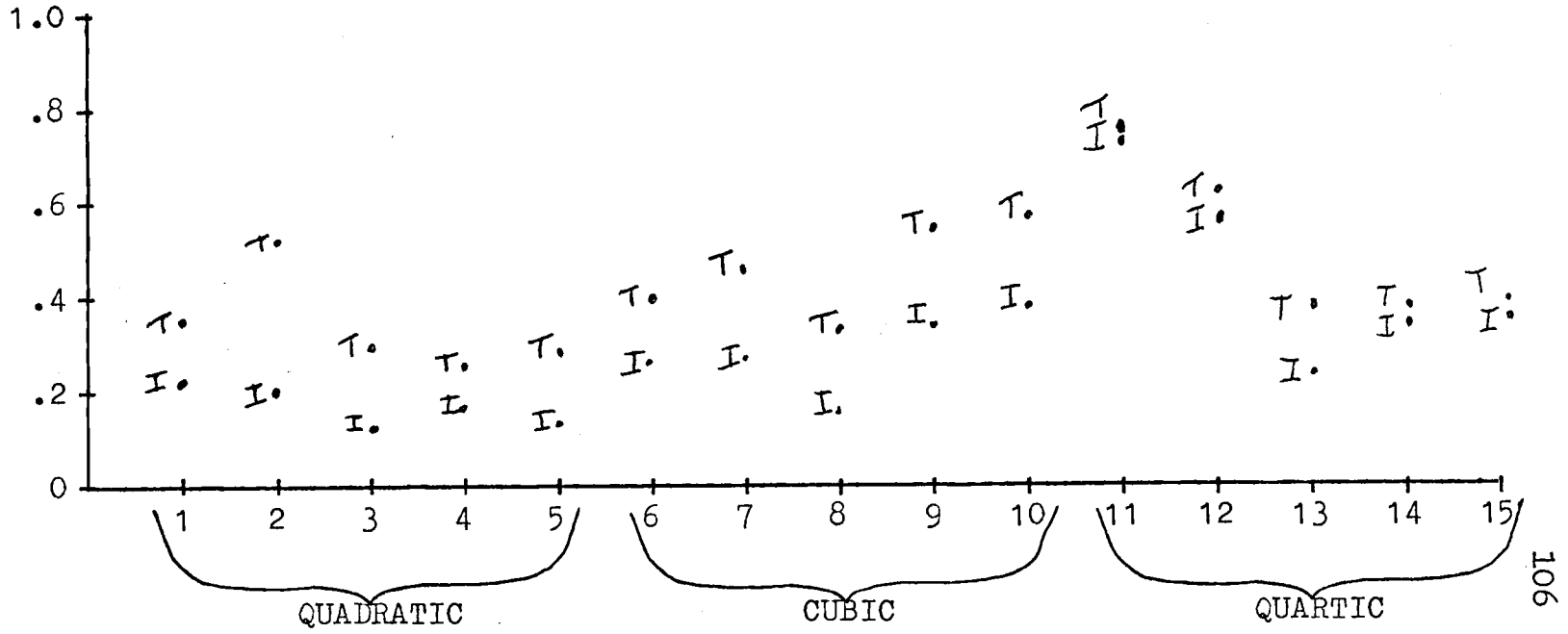
IR v.s. GM



IR v.s. GMD


1.0
.8
.6
.4
.2
0
D
I
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
QUADRATIC
CUBIC
QUARTIC

IR v.s. GMT



BIBLIOGRAPHY

Anderson, T.W. 1971. The statistical analysis of time series. New York, Wiley. 704 p.

Baranchik, A.J. 1970. A family of minimax estimators of the mean of a multivariate normal distribution. The Annals of Mathematical Statistics 41:642-645.

Bargmann, R.E. 1969. Standard mathematical tables. Ohio, Chemical Rubber. 118-140 pp.

Barlow, R.E. and D.J. Bartholomew and J.M. Bremner and H.D. Brunk. Statistical inference under order restrictions. 1972. London, Wiley. 388 p.

Bartle, Robert G. 1966. The elements of integration. New York, Wiley. 129 p.

Brunk, H.D. 1965. Conditional expectation given a -lattice and applications. The Annals of Mathematical Statistics 36:1339-1350.

Brunk, H.D. and Donald A. Pierce. 1972. Estimation of multivariate densities for computer aided differential diagnosis of disease. Oregon State University Department of Statistics Technical Report No. 31

DeGroot, Morris H. 1970. Optimal statistical decisions. New York, McGraw-Hill. 489 p.

Draper, N.R. and H. Smith. 1966. Applied regression analysis. New York, Wiley. 407 p.

Efron, B. and C. Morris. 1972. Empirical bayes on vector observations: An extension of Steins method. Biometrica 59: 335-347.

Efron, B. and C. Morris. 1973. Stein's estimation rule and its competitors--an empirical bayes approach. Journal of The American Statistical Association 68: 117-130.

Ferguson, Thomas S. 1967. Mathematical statistics. New York, Academic Press. 396 p.

Halpern, E.F. 1973. Polynomial regression from a bayesian approach. Journal of The American Statistical Association 68: 137-143.

Lindley, P.V. 1971. Bayesian statistics, a review. Regional conference series in applied mathematics. Pennsylvania, Society for Industrial and Applied Mathematics.

Lindley, D.V. and A.F.M. Smith. 1972. Bayes estimates for the linear model. Journal of the Royal Statistical Society 34: 1-41.

Luenberger, David G. 1969. Optimization by vector space methods. New York, Wiley. 326 p.

Rudin, Walter. 1964. Principles of mathematical analysis. New York, McGraw-Hill. 270 p.

Searle, S.R. 1971. Linear models. New York, Wiley. 532 p.

APPENDICES

APPENDIX

The material in the appendix is included for the purpose of aquainting the reader (or reminding the already informed reader) of those aspects of isotonic regression and bounded isotonic regression alluded to in the text of this thesis. Most of the facts on isotonic regression were taken from [Barlow, Bartholomew, Bremner and Brunk, 1972]. Most of the facts on bounded isotonic regression were mentioned in that reference as being obtainable by generalizing the theorems on isotonic regression, but proofs were omitted. Thus, the thesis author has provided proofs of these facts which were deemed most important in the thesis.

Definition A1

A binary relation "$\preceq$" on a set X establishes a simple order on X if

i. it is reflexive: $x \preceq x$ for all x in X ;

ii. it is transitive: x, y, z in X , $x \preceq y$, $y \preceq z$ imply $x \preceq z$;

iii. it is antisymmetric: x,y in X , $x \preceq y$, $y \preceq x$ imply $x = y$;

iv. every two elements are comparable: x,y in X implies either $x \preceq y$ or $y \preceq x$.

A partial order is reflexive, transitive and anti-symmetric.

A quasi-order is reflexive and transitive.

Definition A2

A real valued function f on X is isotonic with respect to a quasi-ordering " $\preceq$ " on X if x,y in X , $x \preceq y$ imply $f(x) \leq f(y)$.

Note: For the remainder of the appendix, we shall assume X is a nonempty, finite set. We shall also assume the function w defined on X is strictly positive.

Definition A3

Let g be a given function on X and w a given positive function on X . An isotonic function g^* on X is an isotonic regression of g with weights w , if and only if it minimizes in the class of isotonic functions f on X the sum

$$\sum_{x \in X} [g(x)-f(x)]^2 w(x) \ . \qquad (a_1)$$

Definition A4

Let a and b be real numbers with $a \leq b$. An isotonic function g^* on X with $a \leq g^* \leq b$ is called a bounded isotonic regression of g with weights w and bounds a and b if and only if it minimizes the

sum in (a1) in the class of isotonic functions f for which $a \leq f(x) \leq b$ for all x in X.

Note: In [Barlow, Bartholomew, Bremner and Brunk, 1972], the a and b of Definition A4 are allowed to be functions on X. Since such generality was not needed in this thesis, we will use Definition A4 as it is stated.

Notation: Let $X = \{x_1, x_2, \ldots, x_k\}$, $\preceq$ be a quasi-order on X.

$$K = \{f;\ f \text{ is (bounded) isotonic on } X\},$$

$$C = \{(y_1, y_2, \ldots, y_k)' \in R^k;\ y_i = f(x_i),\ i = 1,2,\ldots,k,$$

$$\text{for some } f \in K\},$$

where y' denotes the transpose of the vector y and R^k Euclidian k-space. Also, let

$$W = (w(x_i)\delta_{ij}) \quad i,j = 1,2,\ldots,k.$$

If g is a real valued function on X, let

$$\bar{g} = (g(x_1), \ldots, g(x_k))'.$$

That is, the bar over a function on X is the vector of the function's values.

Remark: With the above notation we may define an inner product " $\langle \cdot, \cdot \rangle$ " on R^k by

$$\langle u,v \rangle = u'Wv$$

for all u,v in R^k and a norm "$\| \cdot \|$" on R^k by

$$\|y\|^2 = y'Wy$$

for all y in R^k. (We are now assuming $w > 0$.)

Then the sum in (a1) may be written as

$$\|\bar{g} - \bar{f}\|^2 .$$

Hence, an (a bounded) isotonic regression g^* is that element of C which is closest to $\bar{g}$ in this metric. It is easily seen that C is a closed convex subset of R^k . Therefore, a well known theorem guarantees the existence and uniqueness of g^* . One may consult [Luenberger, 1969] for example. A necessary and sufficient condition that g^* be the (bounded) isotonic regression of g is that

$$\langle \bar{g}-\bar{g}^* , \bar{g}^*-\bar{f} \rangle \geq 0 \qquad \text{(a2)}$$

for all f in K .

Theorem A1: If g_1 and g_2 are (bounded) isotonic functions on X such that $g_1(x) \leq g(x) \leq g_2(x)$ for all x in X , and if g^* is the (bounded) isotonic regression of g , then also $g_1(x) \leq g^*(x) \leq g_2(x)$ for all x in X .

Proof: We repeat the method of proof given on page 29 of [Barlow, Bartholomew, Bremner, and Brunk, 1972].

Define the function h on X by

$$h(x) = \max[g^*(x), g_1(x)] .$$

It is easily seen that h is (bounded) isotonic. If $g^*(x) \geq g_1(x)$ for a particular x in X, then $h(x) = g^*(x)$ so that $g(x) - g^*(x) = g(x) - h(x)$; while if $g^*(x) < g_1(x)$ then $0 \leq g(x) - h(x) = g(x) - g_1(x) < g(x) - g^*(x)$. Thus, for all x in X,

$$[g(x) - h(x)]^2 \leq [g(x) - g^*(x)]^2$$

implies

$$\sum_{x \in X} [g(x)-h(x)]^2 w(x) \leq \sum_{x \in X} [g(x)-g^*(x)]^2 w(x)$$

with strict inequality if $g^*(x) < g_1(x)$ for some x in X . The proof that $g^*(x) \leq g_2(x)$ is similar. []

Notation: Let Φ be a convex function which is finite on an interval I containing the range of the function g and infinite elsewhere. Let Φ' be an arbitrary determination of its derivative (any value between or equal to the left and right derivatives), defined and finite on I . For real numbers u and v set

$$\Delta(u,v) = \Delta_{\Phi}(u,v) = \begin{cases} \Phi(u)-\Phi(v)-(u-v)\Phi'(v) & \text{if } u,v \in I \\ \infty & \text{if } u \notin I,\ v \in I . \end{cases}$$

Results: (Ai) Φ' is nondecreasing.

(Aii) $\Delta(u,v) \geq 0$, with strict inequality if $u \neq v$ and Φ is strictly convex.

(Aiii) $\Delta(r,t) = \Delta(r,s) + \Delta(s,t) + (r-s)[\Phi'(s)-\Phi'(t)]$ if $s,t \in I$.

Theorem A2: The sum

$$\sum_{x \in X} \Delta(g(x),f(x))w(x) ,$$

is minimized over the class of (bounded) isotonic f by taking $f = g^*$, where g^* is the (bounded) isotonic regression of g with weights w_i. The minimizing function is unique if Φ is strictly convex.

Proof. The proof given is for the bounded case with bounds 0 and 1 . This proof is easily adapted to the unbounded case and is shorter than that given in [Barlow, Bartholomew, Bremner and Brunk, 1972]. Using (Aiii) with $r = g(x)$, $t = f(x)$, and $s = g^*(x)$ we have

$$\sum_{x \in X} \Delta(g(x),f(x))w(x) = \sum_{x \in X} \Delta(g(x),g^*(x))w(x) + \sum_{x \in X} \Delta(g^*(x),f(x))w(x) + H(f) \tag{a3}$$

where

$$H(f) = \sum_{x \in X} [g(x)-g^*(x)][\Phi'(g^*(x))-\Phi'(f(x))]w(x) .$$

From (a3), we see that it suffices to show that $H(f) \geq 0$ for all bounded isotonic f. Let f be a fixed isotonic function on X such that $0 \leq f(x) \leq 1$ for all x in X. Define A, A_0, and A_1 by

$$A = \{x \in X;\ 0 < g^*(x) < 1\}$$
$$A_0 = \{x \in X;\ g^*(x) = 0\}$$
$$A_1 = \{x \in X;\ g^*(x) = 1\}.$$

Since X is finite there exists a real number c, $c > 0$, small enough, so that

(i) $0 \leq g^*(x) - c[\bar{\phi}'(g^*(x)) - \underline{\phi}'(f(x))] \leq 1$ for all x in A,

(ii) $c[\bar{\phi}'(f(x)) - \underline{\phi}'(0)] \leq 1$ for all x in A.

(iii) $c[\bar{\phi}'(1) - \underline{\phi}'(f(x))] \leq 1$ for all x in A, and

(iv) $g^*(x) - g^*(y) \geq c[\bar{\phi}'(g^*(x)) - \underline{\phi}'(g^*(y))]$ for all x, y in X such that $y \geq x$.

Define f' by

$$f'(x) = g^*(x) - c[\bar{\phi}'(g^*(x)) - \underline{\phi}'(f(x))]$$

for all x in X.

Since $\underline{\phi}'$ is nondecreasing, (i), (ii) and (iii) imply $0 \leq f'(x) \leq 1$ for all x in X.

Since (iv) implies that f' is isotonic, f' is bounded isotonic.

Now,

$$H(f) = \sum_{x\in X}[g(x)-g^*(x)][(1/c)\{g^*(x)-f'(x)\}]w(x)$$

$$= (1/c) < \bar{g} - \bar{g}^* , \bar{g}^* - \bar{f} > .$$

Since f' is bounded isotonic, (a2) implies $H(f) \geq 0$.

To prove uniqueness, (a3) and the result just obtained imply

$$\sum_{x\in X}\Delta(g(x),f(x))w(x) \geq \sum_{x\in X}\Delta(g(x),g^*(x))w(x) + \sum_{x\in X}\Delta(g^*(x),f(x))w(x) \qquad (a4)$$

for all bounded isotonic f. When Φ is strictly convex (Aii) implies that last sum in (a4) is strictly positive when $f \neq g^*$. Thus

$$\sum_{x\in X}\Delta(g(x),f(x))w(x) > \sum_{x\in X}\Delta(g(x),g^*(x))w(x)$$

when $f \neq g^*$. []

Theorem A3: Let C be a closed convex subset of a finite dimensional Hilbert space H with inner products $<\cdot,\cdot>_\alpha$ for $0 \leq \alpha \leq a$. Let $\|\cdot\|_\alpha$ denote the norm induced by $<\cdot,\cdot>_\alpha$ for $0 \leq \alpha \leq a$. Assume

(i) $\|x\|_{\alpha_1} \leq \|x\|_{\alpha_2}$ for all x in H when $\alpha_1 < \alpha_2$, α_1 , α_2 in $[0,a]$,

(ii) $\|x\|_{\alpha} \to \|x\|_0$ as $\alpha \to 0$ for each x in H , and

(iii) $\bar{g}_{\alpha} \to \bar{g}_0$ as $\alpha \to 0$, where $\bar{g}_{\alpha}$ in H for $0 \leq \alpha \leq a$.

(Note: Since H is finite dimensional, all norms on H are equivalent. So, by assuming convergence in any particular norm, we have convergence in all norms on H .)

Let $P_{\alpha}\bar{g}_{\alpha}$ denote the point in C which minimizes $\|\bar{g} - \bar{f}\|_{\alpha}$ over $\bar{f}$ in C . Then $P_{\alpha}\bar{g}_{\alpha} \to P_0\bar{g}_0$ as $\alpha \to 0$.

Proof: First we show $P_{\alpha}\bar{g}_0 \to P_0\bar{g}_0$ as $\alpha \to 0$. Let $\{\alpha_n\}_{n=1}^{\infty}$ be a sequence such that $0 < \alpha_n < a$ for all n and $\alpha_n \to 0$ as $n \to \infty$.

Then by (i) and the definition of $P_{\alpha}\bar{g}_{\alpha}$ we have

$$\|\bar{g}_0 - P_{\alpha_n}\bar{g}_0\|_0 \leq \|\bar{g}_0 - P_{\alpha_n}\bar{g}_0\|_{\alpha_n} \leq \|\bar{g}_0 - P_0\bar{g}_0\|_{\alpha_n}$$

for all n . By taking the limit superior ($\overline{\lim}$) as $n \to \infty$ and using (ii), we have

$$\overline{\lim_{n\to\infty}} \|\bar{g}_0 - P_{\alpha_n}\bar{g}_0\|_0 \leq \|\bar{g}_0 - P_0\bar{g}_0\|_0 \ .$$

But by the definition of $P_0\bar{g}_0$ we have

$$\|\bar{g}_0 - P_{\alpha_n}\bar{g}_0\|_0 \geq \|\bar{g}_0 - P_0\bar{g}_0\|_0$$

for all n. Thus

$$\lim_{n\to\infty}\|\bar{g}_0 - P_{\alpha_n}\bar{g}_0\|_0 = \|\bar{g}_0 - P_0\bar{g}_0\|_0 .$$

By the parallelogram law we have

$$\|(P_{\alpha_m}\bar{g}_0-\bar{g}_0) + (\bar{g}_0-P_{\alpha_n}\bar{g}_0)\|_0^2 =$$

$$2\|P_{\alpha_m}\bar{g}_0 - \bar{g}_0\|_0^2 + 2\|\bar{g}_0 - P_{\alpha_n}\bar{g}_0\|_0^2$$

$$- 4\|\bar{g}_0 - (\tfrac{1}{2}P_{\alpha_m}\bar{g}_0 + \tfrac{1}{2}P_{\alpha_n}\bar{g}_0)\|_0^2$$

$$\leq 2\|P_{\alpha_m}\bar{g}_0 - \bar{g}_0\|_0^2 + 2\|\bar{g}_0 - P_{\alpha_n}\bar{g}_0\|_0^2 - 4\|\bar{g}_0 - P_0\bar{g}_0\|_0^2$$

since C is convex and $P_\alpha\bar{g}_0$ in C.

The right hand side of the inequality converges to zero as $n, m \to \infty$. Therefore, the sequence $\{P_{\alpha_n}\bar{g}_0\}_{n=1}^{\infty}$ is Cauchy in $\|\cdot\|_0$ and has a limit in C. Thus,

$$\|\bar{g}_0 - P_0\bar{g}_0\|_0 = \lim_{n\to\infty}\|\bar{g}_0 - P_{\alpha_n}\bar{g}_0\|_0 = \|\bar{g}_0 - \lim_{n\to\infty}P_{\alpha_n}\bar{g}_0\|_0 .$$

Hence, by the definition and uniqueness of $P_0\bar{g}_0$ we have

$$\lim_{\alpha\to 0} P_\alpha\bar{g}_0 = P_0\bar{g}_0 .$$

Now let $\epsilon > 0$ be given. Then (iii) and the result just obtained imply that there exists a α_0, $0 < \alpha_0 < a$

such that

$$\|\bar{g} - \bar{g}_\alpha\|_a < \epsilon/2 \quad \text{and} \quad \|P_0\bar{g}_0 - P_\alpha\bar{g}_0\|_0 < \epsilon/2$$

for all α in $[0,\alpha_0]$. But from (i) we have

$$\|\bar{g} - \bar{g}_\alpha\|_\alpha < \epsilon/2 \quad \text{and} \quad \|P_0\bar{g}_0 - P_\alpha\bar{g}_0\|_0 < \epsilon/2$$

for all α in $[0,\alpha_0]$. Using the triangle inequality, (i), and the fact that P is a projection and norm reducing (see [Brunk, 1965]), we have

$$\begin{aligned}\|P_0\bar{g}_0 - P_\alpha\bar{g}_\alpha\|_0 &\le \|P_0\bar{g}_0 - P_\alpha\bar{g}_0\|_0 + \|P_\alpha\bar{g}_0 - P_\alpha\bar{g}_\alpha\|_\alpha \\ &\le \|P_0\bar{g}_0 - P_\alpha\bar{g}_0\|_0 + \|\bar{g}_0 - \bar{g}_\alpha\|_\alpha \\ &< \epsilon/2 + \epsilon/2 = \epsilon\end{aligned}$$

for all α in $[0,\alpha_0]$. []

The remainder of the appendix is devoted to computation of isotonic regression and bounded isotonic regression. We shall begin with the Minimum Violator algorithm. The Maximum Violator algorithm is analogous and will not be stated. First we give a definition.

Definition A4

An element x in the partially order set X is an immediate predecessor of an element y if $x \precsim y$ but

there is no z in X distinct from x and y such that $x \precsim z \precsim y$.

Minimum Violator algorithm (taken from [Barlow, Bartholomew, Bremner, and Brunk, 1972]).

This algorithm applies when the partial order is such that each element has exactly one immediate predecessor, except for one element, called the root, which has no predecessor.

> "The algorithm starts with the finest possible partition into blocks, the individual points of X. We look for violators: y is a violator if $g(y) < g(x)$ where x is the immediate predecessor of y. A minimum violator is a violator y for which $g(y)$ attains its minimum value among the values of g at violators. The algorithm begins by selecting a minimum violator and pooling it with its immediate predecessor to form a block. At an arbitrary stage of the algorithm we have a partition into blocks. Each block has a weight, the sum of the weights of its individual elements; a value, the weighted average of the values of g at its individual points; and a root, that one of its points whose immediate predecessor is not in the block. The immediate predecessor of any block is the block containing the immediate predecessor of its root. When a block and its immediate predecessor are pooled, the root of the new block is the root of the immediate predecessor block.
>
> A block is a violator if its value is smaller than that of its immediate predecessor block. A minimum violator block is a violator whose value is at least as small as that of any other violator. Each step of the algorithm consists in pooling a minimum violator with its immediate predecessor block. This is continued until there are no violators. At this point the blocks are sets of constancy for g^* and the value of g^* at each point of any block is just the value of the block.
>
> If the partial order is such that each element has exactly one immediate sucessor, except

for one with no successor, an analogous maximum violator algorithm can be used. Of course, these algorithms apply in particular to the special case of a complete order."

Henceforth we shall assume that $X = \{x_1,\ldots,x_k\}$ is equipped with a simple order $\preceq$. That is we assume $x_1 \preceq x_2 \preceq \ldots \preceq x_k$.

Max-Min formula (simple order)

Let g^* be the isotonic regression of g with weights w. Then

$$g^*(x_i) = \max_{s\le i} \min_{t\ge i} Av(s,t) \quad i = 1,2,\ldots,k , \tag{a5}$$

where

$$Av(s,t) = \frac{\sum_{r=s}^{t} g(x_r)w(x_r)}{\sum_{r=s}^{t} w(x_r)} . \tag{a6}$$

(This formula and other equivalent formulas are found in [Barlow, Bartholomew, Bremner, and Brunk, 1972].)

We now present a formula for the case of bounded isotonic regression with a simple order.

Max-Min formula (simple order)

Let g^* be the bounded isotonic regression of g with weights w and bounds 0 and 1. Then if $g(x) > 0$ for all x in X,

$$g^*(x_1) = \min\{1,\ Av(1,t),\ t \geq 1\} \qquad (a7)$$

$$g^*(x_i) = \max\{g^*(x_{i-1}),\ \min\{1,\ Av(i,t),\ t \geq i\}\} \qquad (a8)$$

for $i = 2,3,\ldots,k$.

(Again [Barlow, Bartholomew, Bremner, and Brunk, 1972] contains more general formulas, but no proof.)

We shall prove the formulas in (a7) and (a8). First we establish some notation and then give three lemmas.

Notation: Let g^* be defined by (a7) and (a8),

$$K = \{f;\ f:X\to[0,1],\ f(x_i)\leq f(x_{i+1}),\ i = 1,2,\ldots,k-1\} ,$$

$$G_i = \sum_{j=1}^{i} g(x_j)w(x_j), \qquad G_i^* = \sum_{j=1}^{i} g^*(x_j)w(x_j) , \quad \text{and}$$

$$W_n = \sum_{i=1}^{n} w(x_i) .$$

Lemma A1

Let $g \geq 0$, then $G_i^* \leq G_i$, $i = 1,2,\ldots,k$.

Proof: Let $\bar{g}$ denote the isotonic regression (unbounded) of g with weights w . By comparing the formula of (a5) with those of (a7) and (a8), it is clear that $g^* \leq \bar{g}$. Thus, $G_n^* \leq \bar{G}_n$, where $\bar{G}_n = \sum_{i=1}^{n} \bar{g}(x_i)w(x_i)$. But it is established in [Barlow, Bartholomew, Bremner, and Brunk, 1972] that $\bar{G}_n \leq G_n$ $n = 1,2,\ldots,k$. []

Lemma A2

Let $g \geq 0$, then for $i = 1,2,\ldots,k-1$, $G_i^* < G_i$ implies $g^*(x_{i+1}) - g^*(x_i) = 0$.

Proof: Let $B = \{1,2,\ldots,k\}$ and $A = \{n \text{ in } B \ ; \ g^*(x_n) = 1\}$. If A is empty, then $g^* < 1$ and is therefore the isotonic (unbounded) regression of g with weights w (i.e. g^* is given by formula (a5)). It is shown that Lemma A2 holds when g^* is the isotonic regression in [Barlow, Bartholomew, Bremner, and Brunk, 1972].

Thus, assume A is not empty. Let T be the least element of A . Then

$$g^*(x_i) = 1 \quad \text{for} \quad i \geq T \quad \text{and}$$

$$g^*(x_i) < 1 \quad \text{for} \quad i < T \ .$$

The lemma holds for $i \geq T$. If $T = 1$, we are done. So assume $1 < T \leq k - 1$.

Since $T > 1$, $g^*(x_1) < 1$. Hence

$$g^*(x_1) = \min\{Av(1,t), t \geq 1\} \ .$$

Define t_1 as the largest member of B for which

$$g^*(x_1) = Av(1,t_1) \ .$$

Thus, we have

$$g^*(x_1) = \frac{G_{t_1}}{W_{t_1}} < \frac{G_t}{W_t} \tag{a9}$$

for all $t > t_1$.

Claim 1: $g^*(x_n) = g^*(x_1)$ for $n = 1,2,\ldots,t_1$. Hence Lemma A2 holds for $i < t_1$.

Proof of Claim 1: If $t_1 = 1$ we are done. So assume $t_1 > 1$. The claim is trivial for $n = 1$.

Now

$$G^*_{t_1} = \sum_{n=1}^{t_1} g^*(x_n)w(x_n) \geq \sum_{n=1}^{t_1} g^*(x_1)w(x_n)$$

since g^* in K . But

$$\sum_{n=1}^{t_1} g^*(x_1)w(x_n) = g^*(x_1)w_{t_1} = G_{t_1}$$

by (a9). Thus, we have

$$G^*_{t_1} \geq G_{t_1} \tag{a10}$$

with strict inequality if the claim fails to hold for some n , $1 < n \leq t_1$. But a strict inequality in (a10) would contradict Lemma A1, hence the claim is true.

As a result of Lemma A1 and (a10) we have

$$G^*_{t_1} = G_{t_1} \ . \tag{a11}$$

Thus Lemma A2 holds for $i \leq t_1$.

Since $g^*(x_1) < 1$, Claim 1 implies that $t_1 < T$. If $t_1 + 1 = T$, we are done. So assume $t_1 + 1 < T$. Define t_2 to be the largest member of $\{n\ ;\ n \geq t_1 + 1\}$ such that

$$\min\{Av(t_1+1,t),\ t \geq t_1+1\} = Av(t_1+1,t_2) .$$

Claim 2:

$$g^*(x_{t_1+1}) = Av(t_1 + 1,\ t_2) .$$

Proof of Claim 2: We have $Av(t_1 + 1,\ t_2) < 1$, since otherwise, $g^*(x_{t_1+1}) = 1$ which implies $t_1 + 1 = T$. To complete the proof of Claim 2, it suffices to show that

$$Av(t_1+1,\ t_2) > g^*(x_{t_1}) . \tag{a12}$$

Suppose not, then

$$G_{t_2} - G_{t_1} \leq g^*(x_{t_1})[W_{t_2} - W_{t_1}] = g^*(x_1)[W_{t_2} - W_{t_1}] \tag{a13}$$

by Claim 1. By (a9), $G_{t_1} = g^*(x_1)W_{t_1}$, so (a13) implies

$$\frac{G_{t_2}}{W_{t_2}} \leq g^*(x_1) .$$

But this contradicts (a9) since $t_2 > t_1$. Hence Claim 2 is true.

As a result of Claim 2 and the definition of t_2 we have

$$g^*(x_{t_1+1}) = \frac{G_{t_2}-G_{t_1}}{W_{t_2}-W_{t_1}} < \frac{G_t-G_{t_1}}{W_t-W_{t_1}} \qquad (a14)$$

for $t > t_2$. We may also note that Claim 2 and (a12) imply

$$g^*(x_{t_1+1}) > g^*(x_{t_1}) .$$

Claim 3: $g^*(x_n) = g^*(x_{t_1+1})$ for $n = t_1 + 1,\ldots,t_2$.

Hence, Lemma A2 holds for i t_2 .

Proof of Claim 3: The claim is trivial for $t_2 = t_1+1$. So assume $t_2 \geq t_1 + 1$. The claim is trivial for $n = t_1 + 1$. Now

$$G^*_{t_2} = G^*_{t_1} + \sum_{i=t_1+1}^{t_i} g^*(x_i)w(x_i) \geq G^*_{t_1} + g^*(x_{t_1+1})[W_{t_2} - W_{t_1}]$$

since g^* in K . But (a11) and (a14) imply

$$G^*_{t_1} + g^*(x_{t_1+1})[W_{t_2}-W_{t_1}] = G_{t_1} + G_{t_2} - G_{t_1} .$$

Hence

$$G^*_{t_2} \geq G_{t_2}$$

with strict inequality if Claim 3 fails to hold for some n

such that $t_1 + 1 < n \leq t_2$. Since a strict inequality would contradict Lemma A1, Claim 3 is true and

$$G^*_{t_2} = G_{t_2} \tag{a15}$$

Thus, Lemma A2 holds for $i \leq t_2$. Since $t_1 + 1 < T$, $g^*(x_{t_1+1})$ 1 . From Claim 3, $g^*(x_{t_2}) < 1$. Therefore $t_2 < T$. If $t_2 + 1 = T$, the proof of Lemma A2 is complete. Otherwise, we may define t_3 (and t_4 , t_5 , etc. if necessary) in the obvious way. The process must terminate within T steps. This completes the proof of Lemma A2. []

Lemma A3

Let $g \geq 0$. Then

$$[g^*(x_k)-f(x_k)][G_k-G^*_k] \geq 0 \quad \text{for all} \quad f \text{ in } K .$$

Proof: From Lemma A1, $G_k - G^*_k \geq 0$. If $g^*(x_k) = 1$, then $g^*(x_k) - f(x_k) \geq 0$, since f in K . Thus Lemma A3 is true if $g^*(x_k) = 1$.

If $g^*(x_k) < 1$, then $g^* < 1$ implies g^* is the isotonic regression of g . Lemma A3 is proved in [Barlow, Bartholomew, Bremner, and Brunk, 1972] for the case g^* is the isotonic regression of g . []

Theorem A4

Let $g \geq 0$ and g^* be defined by (a7) and (a8). Then g^* is the bounded isotonic regression of g. (Hence, the Max-Min formula is valad.)

Proof: From (a2), it suffices to show that for each f in K

$$\sum_{i=1}^{k} [g(x_i)-g^*(x_i)][g^*(x_i)-f(x_i)]w(x_i) \geq 0 .$$

Abel's partial summation formula yields

$$\sum_{i=1}^{k} [g(x_i)-g^*(x_i)][g^*(x_i)-f(x_i)]w(x_i) =$$

$$\sum_{i=1}^{k} \{[f(x_i)-f(x_{i-1})] - [g^*(x_i)-g^*(x_{i-1})]\}[G_{i-1}-G^*_{i-1}]$$

$$+ [g^*(x_k)-f(x_k)][G_k-G^*_k] ,$$

where $x_0 = f(x_0) = g^*(x_0) = G_0 = G^*_0 = 0$.

The non-negativeness is now clear from the fact that f in K and Lemmas A1, A2, and A3.