AN ABSTRACT OF THE THESIS OF

LEONARD RAY HAFF (Name) for the DOCTOR OF PHILOSOPHY (Degree)

in STATISTICS (Major) presented on May 30, 1973 (Date)

Title: BAYESIAN REGRESSION WITH AUTOREGRESSIVE PRIORS

Let $\tilde{y} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)'$ be a multivariate normal random vector with mean θ and covariance matrix I. Denote by $f_{\tilde{y}|\theta}(y)$ the density of $\tilde{y}|\theta$. Then,

$$E_\theta(\tilde{y}) = (E_\theta\tilde{y}_1, E_\theta\tilde{y}_2, \ldots, E_\theta\tilde{y}_n)' = (\theta_1, \theta_2, \ldots, \theta_n)' = \theta'.$$

An outcome y (a value of $\tilde{y}$) is a vector of data resulting from an ordinary regression experiment. A Bayesian approach is developed for estimating θ. It is motivated by knowledge a priori that the θ's fall on a response curve which is reasonably smooth. To formulate this knowledge in a precise way, priors on θ are taken which characteristically imply that for a given θ_{i_o}, nearby θ_i have higher correlations with θ_{i_o} than is the case with remote θ_i. The priors studied are called "autoregressive priors" (AR) because of their resemblence to autoregressive models of time series analysis. The general AR prior of order p is determined by

$$\tilde{\epsilon}_t = [\sum_{r=0}^{p} a_t 1_N(t-r) \mathcal{P}^{p-r}](\tilde{\theta}_{t-p} - \mu_{t-p}) ,$$

$\tilde{\epsilon}_t$ i.i.d., $E\tilde{\epsilon}_t = 0$, $\theta_o = 0$, $a_o = 1$, the a_r are unknown, $r = 2, \ldots,$ and $\mathcal{P}$ is the forward operator; i.e., $\mathcal{P}\, u_t = u_{t+1}$.

Let $f_{\tilde{\theta}}(\theta)$ be an AR prior on θ. The marginal distribution of $\tilde{y}$,

$$f_{\tilde{y}}(y) = \int f_{\tilde{y}|\theta}(y) f_{\tilde{\theta}}(\theta) d\theta,$$

depends upon a vector of hyperparameters, P_o, which were used to define the AR prior. A noninformative prior or a vague prior is taken as the distribution of $\tilde{P}_o$ (except in the general AR situation where priors are put on the a's to help determine the order of the process.) This prior is combined with $f_{\tilde{y}}(y)$ to obtain a posterior density whose mode, $\hat{P}_o$, is taken to estimate P_o. Denote $E(\tilde{\theta}|y)$ by $\hat{\theta}$. In general, $\hat{\theta}$ also depends upon P_o. The vector $\hat{\theta}(\hat{P}_o)$ is taken to estimate θ.

A sequence of AR priors evolves with increasing generality. The celebrated James-Stein estimate is interpreted in terms of an AR prior of order zero and the estimation procedure sketched above. Some success is achieved in understanding the risk function

$$E_{\tilde{\theta}}(\hat{\theta}(\hat{P}_o) - \theta)'(\hat{\theta}(\hat{P}_o) - \theta)$$

as the AR priors increase in generality. The Bayesian estimates are seen to be superior to the maximum likelihood estimates over a large class of response curves (perhaps the only curves of practical interest).

Also discussed is the situation where $\text{cov}(\tilde{y}|\theta) = \sigma^2 I$, σ^2 unknown.

Bayesian Regression With Autoregressive Priors

by

Leonard Ray Haff

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Completed May 1973

Commencement 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 May 30, 1973

Typed by Clover Redfern for Leonard Ray Haff

ACKNOWLEDGMENT

The personal example and technical direction provided by Dr. H.D. Brunk (the author's Major Professor) has influenced the author in many subtle ways--far more than he can comprehend. To Dr. Brunk the author expresses his wholehearted gratitude.

This thesis could not have been written without the attitudes and techniques acquired from courses taught by Dr. David Thomas and Dr. Justus Seely. Dr. Thomas and Dr. Seely serve the department well with fine courses in mathematical statistics and linear model theory. In particular, the author expresses his appreciation of these gentlemen.

What stands out as a highlight in the author's training is the numerous informal, reading and conference situations directed by Dr. Charles Land and Dr. Kenneth Rowe. Being influenced by these teachers, the author began to understand that statistical thinking is useful, aesthetically pleasing and so on.

The author is very grateful for financial support provided by an N.I.H. Biometry traineeship and by the G.I. Bill. Thanks, also, to the O.S.U. Computer Center for a generous allocation of unsponsored research funds. The author is also indebted to Jo An Barnes whose expert fortran programming helped illustrate an important part of this thesis.

Actually, it was the encouragement given by the author's wife, Joanne, that prompted him to attempt the doctoral program. To her and to our children--Andrew, John, and Rebecca--the author expresses his deepest love and appreciation.

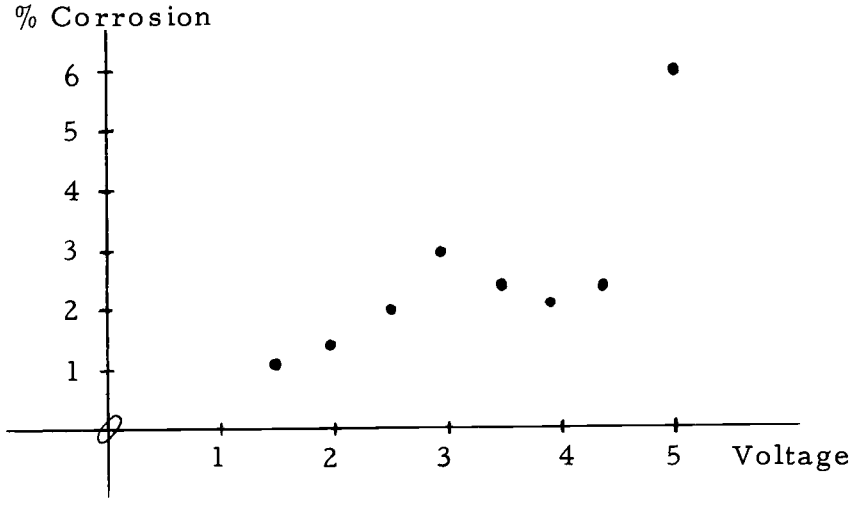
TABLE OF CONTENTS

BAYESIAN REGRESSION WITH AUTOREGRESSIVE PRIORS

I. INTRODUCTION

The Estimation Problem

Consider the common stress-response type of an experiment where at each level of stress X_i, we observe exactly one experimental outcome $\tilde{y}_i = y_i$, $i = 1, \ldots, n$. The experimental situation is typified by Figure 1.1 where we have eight equal masses of metal assumed identical in shape and composition; each mass is subjected to one level of stress (voltage) and the response (percentage corrosion) is recorded.



Voltage	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
% Corrosion	1.1	1.4	2.1	3.1	2.5	2.2	2.5	5.9

Figure 1.1. (Data taken from Graybill (1961).)

More generally, where m independent observations y_{ij}; $i = 1, \ldots, n$; $j = 1, \ldots, m$; are available at each X_i, denote $(1/m) \sum_{j=1}^{m} y_{ij}$ by $y_{i\cdot}$. All experimental observations (mn in number) are assumed to be values taken on by independent random variables.

To initiate the notational convention regarding random vectors, random variables, and their respective values; let us write

$$\tilde{y} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)'$$

for the random vector $\tilde{y}$ (column vector) and use subscripts only when designating components of a vector. Likewise for the values of random vectors we have

$$\tilde{y}(\omega) = (\tilde{y}_1(\omega), \tilde{y}_2(\omega), \ldots, \tilde{y}_n(\omega))' = (y_1, y_2, \ldots, y_n)' = y$$

where ω is some point in the underlying probability space.

Assuming that $E\tilde{y} = (E\tilde{y}_1, E\tilde{y}_2, \ldots, E\tilde{y}_n)' = (\theta_1, \theta_2, \ldots, \theta_n)' = \theta$ is meaningful, the broad objective of this thesis is the estimation of θ with particular objectives summarized as follows:

i. Assuming, further, that $\tilde{y} \mid \theta \sim N_n(\theta, I)$, we desire an improvement over the maximum likelihood estimator ($\bar{\theta} = \text{m.l.e.}(\theta) = y$). The James-Stein estimator $\bar{\bar{\theta}} = (1-(n-2)/y'y)y$ (James, W. and Stein, C., 1960) is of considerable interest since it dominates $\bar{\theta}$ for $n \geq 3$ in

the sense of having strictly smaller, expected squared error loss $\forall\ \theta \in \mathbb{R}^n$. In Chapter 2, $\bar{\bar{\theta}}$ is derived from Bayesian assumptions which seem inappropriate for many experimental situations. Accordingly, when these assumptions are modified to accomodate a given experimental situation (for example, the situation of Figure 1.1) we argue that the resulting estimate of θ should be seriously regarded as a competitor of $\bar{\bar{\theta}}$. In particular, we might have strong motivation for preferring this new estimate to $\bar{\theta}$.

ii. We seek the improvement referred to in i. by considering possible distributions of θ which effectively convey the attitude that the components of θ have a structure inferred from basic physics or by some other reasoning, e.g.,

$$\tilde{\theta}_t = a\tilde{\theta}_{t-1} + \tilde{\epsilon}_t\ ;\quad \tilde{\epsilon}_t \text{ i.i.d.} \tag{1.1}$$

iii. In Chapter 4, the objective is focused upon a preliminary investigation of the extent to which autoregressive priors (1.1 for example) can be used to advantage in various, ordinary regression situations. (The general autoregressive prior is stated in Equation (1.3) below.)

iv. The approach taken is nonparametric in a certain sense. We do not assume that the parameters (components of θ) fall on a response curve which is adequately described by an

expression in some closed form, say, a polynomial. None the less, we seek distributions of $\tilde{\theta}$ which enforce the condition that--the response curve is smooth; it is continuous and does not deviate in a completely wild fashion from a polynomial of low degree.

Thus, the approach is Bayesian and with the above remarks in mind we proceed to find meaningful ways of writing

$$E(\tilde{y}|\theta) = \theta$$

$$\text{cov}(\tilde{y}|\theta) = (1/\rho_1)I$$

$$E\tilde{\theta} = (E\tilde{\theta}_1, E\tilde{\theta}_2, \ldots, E\tilde{\theta}_n)' = (\mu_1, \mu_2, \ldots, \mu_n)' = \mu$$

and

$$\text{cov}(\tilde{\theta}) = E(\tilde{\theta}-\mu)(\tilde{\theta}-\mu)' = E(\tilde{\theta}\tilde{\theta}') - \mu\mu' = (1/\rho_2)V,$$

where V is positive definite.

Distributional Assumptions

The basic framework is given by assuming--

$$\tilde{y}|\theta \sim N_n(\theta, (1/\rho_1)I), \quad \tilde{\theta} \sim N_n(\mu, (1/\rho_2)V) \tag{1.2}$$

where $\theta, \mu \in \mathbb{R}^n$; $\rho_i > 0$ $i = 1, 2$, and V is positive definite. Or, in terms of density functions we have

$$f_{\tilde{y}|\theta}(y) \propto \exp[(-\rho_1/2)(y-\theta)'(y-\theta)]$$

$$f_{\tilde{\theta}}(\theta) \propto [\det(V/\rho_2)]^{-1/2} \exp[(-\rho_2/2)(\theta-\mu)'V^{-1}(\theta-\mu)] .$$

In Chapter 2 (1.2) is stated with the understanding that μ and V are fixed but unspecified; $\rho_1 = 1$; and ρ_2 is variable. Additionally, the structure of Chapter 3 is defined by letting μ vary. Chapter 5 treats the most general assumptions considered in this thesis. There (1.2) is stated to mean that ρ_1, ρ_2 and μ are unknown and V is written in terms of one or more unknown parameters--i.e., the a_i where

$$\tilde{\theta}_t - \mu_t = \sum_{i=1}^{p} a_i(\tilde{\theta}_{t-i} - \mu_{t-i}) + \tilde{\epsilon}_t \qquad (1.3)$$

$i = 1, \ldots, p$; $\tilde{\theta}_{t-i} = \mu_{t-i} = 0$ for $t-i \leq 0$; $\tilde{\epsilon} \sim N_n(0, (1/\rho_2)I)$.

The process of Equation (1.3) is designated as "the general autoregressive process of order p." This departs somewhat from standard usage in time series analysis. In Chapter 5 the difference is stated and discussed further.

The following are special cases of well known results in Bayesian inference. (See Lindley and Smith (1972) for a convenient summary of the general statements.)

Result 1.1. The posterior density

$$f_{\tilde{\theta}|y}(\theta) \propto f_{\tilde{y}|\theta}(y) f_{\tilde{\theta}}(\theta)$$

is multinormal $N_n(A^{-1}a, A^{-1})$ where

$$A = \rho_2 V^{-1} + \rho_1 I, \quad a = \rho_2 V^{-1}\mu + \rho_1 y.$$

Result 1.2. The marginal density

$$f_{\tilde{y}}(y) = \int f_{\tilde{y}|\theta}(y) f_{\tilde{\theta}}(\theta) d\theta$$

is multinormal $N_n(\mu, V/\rho_2 + I/\rho_1)$.

In the sequel the parameters $\mu_1, \mu_2, \ldots, \mu_n$, ρ_1 and ρ_2 are referred to as "hyperparameters." This designation, also, applies to the a_i $i = 1, \ldots, p$ of Equation (1.3).

Let

$$P = (a_1, \ldots, a_p, \mu_1, \ldots, \mu_n, \rho_1, \rho_2)'$$

$$P^* = (a_1, \ldots, a_p)'$$

$$P^{**} = (\mu_1, \ldots, \mu_n, \rho_1, \rho_2)'.$$

Questions relevant to the specification of distributions for $\tilde{P}^{**}|P^*$ and $\tilde{P}$ are dealt with in Chapters 3 and 5 respectively.

Previous Related Research

In a Biometrics article by Dr. F.L. Ramsey (1972) entitled "A Bayesian Approach to Bioassay," the situation depicted in Figure 1.2 was examined.

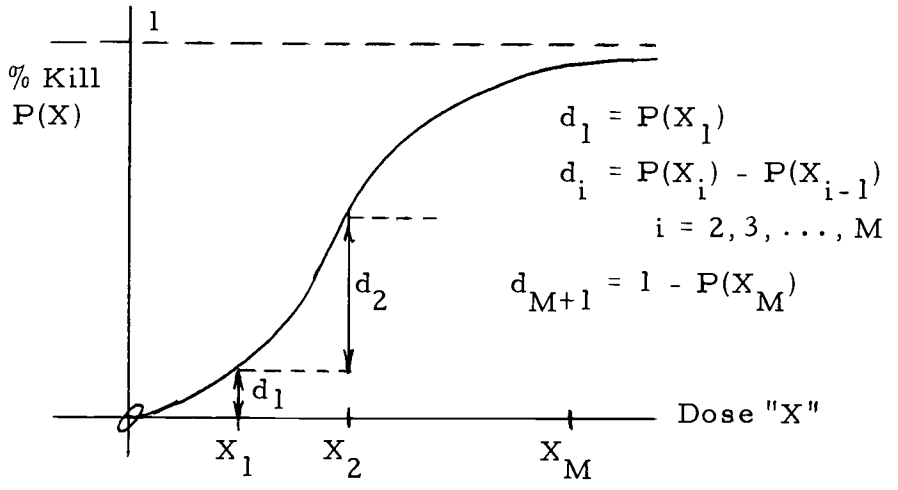


Figure 1.2.

(See, also, Antoniak (1969).)

For the experiment of Figure 1.2,

n_i = number of creatures subjected to dose X_i,

s_i = number of survivors among those at dose X_i.

The likelihood function is

$$L \propto \prod_{i=1}^{M} [P(X_i)]^{s_i} [(1-P(X_i))]^{n_i - s_i} .$$

To enforce the assumption that $P(X)$ is monotone increasing, Ramsey imposed a Dirichlet prior on the increments d_i, $i = 1, \ldots, M+1$. Specifically, where $\{a_i : i = 1, \ldots, M+1\}$ are non-negative constants and $\sum_{i=1}^{M+1} a_i = 1$, let $\beta \geq 0$ and take

$$\pi \propto \prod_{i=1}^{M+1} (d_i)^{\beta\alpha_i - 1} .$$

His stated objectives were:

i. For a fixed X, estimate $P(X)$.

ii. For a fixed γ, find that does X^* which has $P(X^*) = \gamma$.

The success Ramsey enjoyed in realizing i. and ii. above was brought to my attention by Professor H. D. Brunk sometime in the spring of 1972. Professor Brunk made the observation that for a large number of X's Ramsey's prior led to increments with small dependencies. He then surmised that one should be successful attacking a more general regression problem (e.g., the kind depicted in Figure 1.1) by pursuing this technique. That is, perhaps one should proceed by writing a conjugate prior for the parameters of interest in a manner that will render the increments independent.

For the corrosion data of Figure 1.1 our initial investigation considered

$$\tilde{y} | \theta \sim N_8(\theta, I), \quad \tilde{\theta} = \mu + T\tilde{\epsilon}$$

where

$$T = \begin{pmatrix} 1&1&1&0&0&0&0&0&0&0 \\ 1&1&1&1&0&0&0&0&0&0 \\ \vdots & & & & & & & & & \vdots \\ 1&1&1&1&1&1&1&1&1&1 \end{pmatrix}_{8 \times 10}, \quad \tilde{\epsilon} = \begin{pmatrix} \tilde{\epsilon}_1 \\ \tilde{\epsilon}_2 \\ \vdots \\ \tilde{\epsilon}_{10} \end{pmatrix} \sim N_{10}(0, (1/\rho)I) ,$$

$\mu = X\tau$; $\tau \in \mathbb{R}^1$, $\rho \geq 0$ are unknown and

$X = (1.5\ 2.0\ 2.5\ 3.0\ 3.5\ 4.0\ 4.5\ 5.0)'$. This implies

$$\Delta\tilde{\theta}_{t-1} = \tilde{\theta}_t - \tilde{\theta}_{t-1} = \tau + \tilde{\epsilon}_t, \quad t = 1, \ldots, 8, \quad \tilde{\theta}_o = 0;$$

i.e., a process with independent increments.

In accord with the basic framework stated on page 3 we have

$$\tilde{\theta} \sim N_8(\mu, (1/\rho)TT').$$

A discussion of the appropriateness of this prior along with a detailed analysis of the ensuing estimation problem is given in Chapter 3.

The corrosion data is analyzed again in Chapter 5. There, more general assumptions are made concerning dependencies among the components of θ. As indicated earlier, we assume (in Chapter 5) that the parameters minus their prior means are generated by the general autoregressive process of order p (Equation (1.3)). For the corrosion application we take $p = 3$ and discuss a solution of the resulting estimation problem.

II. THE BAYES ESTIMATE $\hat{\theta} = E(\tilde{\theta} | \tilde{y} = y)$

Throughout this chapter we write

$$\tilde{y} | \theta \sim N_n(\theta, I) \quad \text{and} \quad \tilde{\theta} \sim N_n(\mu, (1/\rho)V) \tag{2.1}$$

with the understanding that μ and V are fixed; V is assumed positive definite and $\rho > 0$ is variable. The sequence proceeds roughly as follows. First, $\hat{\theta} = E(\tilde{\theta} | y)$ is introduced as the estimate to be taken for θ (see Result 1.1, p. 5) and alternate expressions for θ are derived. Then, an investigation of some of the basic geometry relating the prior mean μ, the Bayes estimate $\hat{\theta}$, and $\bar{\theta}$ (the m.l.e.) is outlined. We next set $\mu = 0$ and $V = I$ $(\hat{\theta} = (1 - \rho/(1+\rho))y)$ and take the mode of

$$f_{\tilde{\pi} | y}(\pi) \propto f_{\tilde{y} | \pi}(y) f_{\tilde{\pi}}(\pi) , \tag{2.2}$$

where $\pi = \frac{\rho}{1+\rho}$, as an estimate of π. Here

$$f_{\tilde{y} | \pi}(y) = \int f_{\tilde{y} | \theta}(y) f_{\tilde{\theta}}(\theta) d\theta \propto \pi^{(n/2)} \exp((-1/2)\pi y'y)$$

and $f_{\tilde{\pi}}(\pi)$ is some prior distribution of π. When $f_{\tilde{\pi}}(\pi)$ is taken to be an (appropriately defined) improper gamma density, the modal calculation results in an estimator θ^* (indexed by the prior) having the property that $E_\theta(\theta^* - \theta)'(\theta^* - \theta) < E_\theta(\bar{\theta} - \theta)'(\bar{\theta} - \theta)$ $\forall\, \theta \in \mathbb{R}^n$. It is shown that the positive part version of the James-Stein rule corresponds to

the "flat" gamma prior. More precisely, in taking $f_{\tilde{\pi}}(\pi) \propto (1/\pi) 1_{(0,\infty)}(\pi)$, the mode of the posterior density (2.2) results in a θ^*.

$$\theta^* = \bar{\theta}^+ = (1-(n-2)/y'y)^+ y \ .$$

This motivates (to a certain extent) the work in succeeding chapters since it is proposed that a serious competitor of the James-Stein estimator (and related estimators) should result by following the strategy sketched above--after making more realistic assumptions regarding μ, ρ, and V.

If we consider V matrices which are not proportional to the identity, how can we write such a matrix in an advantageous manner? Note the covariance structure generated by the following considerations: To accommodate the notion that the response curve is "smooth," let us write--

$$\tilde{y} | \theta \sim N_n(\theta, I) \quad \text{and} \quad \tilde{\phi} \sim N_n(0, (1/\rho)I) \tag{2.3}$$

where $\tilde{\phi}_1 = \tilde{\theta}_1$, $\tilde{\phi}_k = \tilde{\theta}_k - \tilde{\theta}_{k-1}$, $k = 2, 3, \ldots, n$. Observe that $\rho\tilde{\phi}'\tilde{\phi} \sim \chi^2_n$; consequently

$$E\tilde{\phi}'\tilde{\phi} = E \sum_{i=1}^{n} (\tilde{\theta}_i - \tilde{\theta}_{i-1})^2 = n/\rho$$

or

$$\frac{\sum_{i=1}^{n} E(\tilde{\theta}_i - \tilde{\theta}_{i-1})^2}{n} = \frac{1}{\rho}.$$

The prior precision ρ quantifies the belief that "on the average, we do not expect differences between adjacent θ's to be exceedingly large." In the next chapter, a procedure for estimating ρ is given.

The present chapter is concluded by deducing consequences of the assumption (2.3). That assumption is--in a sense--one step removed from the assumption,

$$\tilde{y} | \theta \sim N_n(\theta, I), \quad \tilde{\theta} \sim N_n(0, (1/\rho)I) \tag{2.4}$$

where estimators of known quality are derived (e.g. $\bar{\tilde{\theta}}^{+}$).

As the chapter concludes, it becomes clear that good estimates of θ do result from the intermediate model (2.3). In Chapter 3 we keep the covariance structure on $\tilde{\theta}$ implied by (2.3) but we take an additional step and assume that $E\tilde{\theta} = \mu \neq 0$. An estimate for θ is then written in terms of estimates for μ and ρ. The intermediate model provides another service. It suggests the possibility of filling out the V matrix by imposing autoregressive priors on the θ's of a more general nature (Equation (1.3) for example).

Alternate Expressions for $\hat{\theta}$

In this section and in that which follows, the general distributional situation of (2.1) is assumed.

Since $f_{\tilde{\theta}|y}(\theta) \propto \exp((-1/2)(\theta-A^{-1}a)'A(\theta-A^{-1}a))$ where $A = \rho V^{-1} + I$, $a = \rho V^{-1}\mu + y$ (Result 1.1); as an estimate of θ we will take the posterior mean

$$\hat{\theta} = E(\tilde{\theta}|y) = A^{-1}a . \tag{2.5}$$

Equation (2.5) can be expressed in a more informative manner if we use the standard matrix result

$$(P+Q)^{-1} = P^{-1}(P^{-1}+Q^{-1})^{-1}Q^{-1}$$

and write

$$\begin{aligned} \hat{\theta} = A^{-1}a &= (\rho V^{-1}+I)^{-1}(\rho V^{-1}\mu+y) \\ &= ((1/\rho)V+I)^{-1}(1/\rho)V(\rho V^{-1}\mu+y) \\ &= ((1/\rho)V+I)^{-1}\mu + ((1/\rho)V+I)^{-1}(1/\rho)Vy . \end{aligned} \tag{2.6}$$

Equation (2.6) expresses $\hat{\theta}$ as a weighted sum of the prior mean and the data in an interesting way; i.e., the sum of the "weights" is the identity matrix.

Another expression for $\hat{\theta}$ is achieved by writing the spectral decomposition of V. We are guaranteed an orthogonal matrix $R = (R_1, R_2, \ldots, R_n)$, $RR' = R'R = \sum_{i=1}^{n} R_i R_i' = I$, where

$R'VR = \Lambda = \delta(\lambda_i)$, $i = 1, \ldots, n$. In general $\delta(u_i)$ will designate that diagonal matrix whose ith diagonal element is u_i. Define $R_i R_i' = E_i$. Then

$$\begin{aligned}\hat{\theta} &= (1/\rho R\Lambda R' + I)^{-1}\mu + [(1/\rho R\Lambda R' + I)^{-1}(1/\rho R\Lambda R')]y \\ &= R\delta\left(\frac{\rho}{\lambda_i+\rho}\right)R'\mu + R\delta\left(\frac{\lambda_i}{\lambda_i+\rho}\right)R'y \qquad (2.7) \\ &= \left(\sum_{i=1}^{n}\left(\frac{\rho}{\lambda_i+\rho}\right)E_i\right)\mu + \left(\sum_{i=1}^{n}\left(\frac{\lambda_i}{\lambda_i+\rho}\right)E_i\right)y, \quad i = 1, 2, \ldots, n.\end{aligned}$$

The spectral decomposition result for positive semidefinite matrices is used in several places throughout this paper. In each instance, the notation is identical to that introduced above. Accordingly, the symbols for eigenvalues, eigenvectors, and projections (the E_i) have meaning only within the immediate context of their appearance.

Results Regarding the Nature of $\hat{\theta}$--Some Inequalities

From (2.6) or (2.7), $\hat{\theta}$ is seen to be a fairly complicated mixture of μ and $\bar{\theta} = y$ ($\hat{\theta}$ is a convex sum of μ and $\bar{\theta}$ only when V is proportional to the identity). In this section the concern is with the effect that ρ has on the basic features of the geometry relating μ, $\hat{\theta}$, and $\bar{\theta}$.

First, from (2.7) we easily have

Lemma 2.1. (i) $\lim_{\rho \to 0} \hat{\theta} = \bar{\theta} = y$ (data)

(ii) $\lim_{\rho \to \infty} \hat{\theta} = \mu$ (prior mean).

In what follows, $\|v\| = \sqrt{v'v}$, $v \in \mathbb{R}^n$.

Lemma 2.2. (i) $\|\hat{\theta}-y\|^2 = (\mu-y)'R\delta[(\frac{\rho}{\lambda_i+\rho})^2]R'(\mu-y)$

(ii) $\|\hat{\theta}-\mu\|^2 = (\mu-y)'R\delta[(\frac{\lambda_i}{\lambda_i+\rho})^2]R'(\mu-y)$.

Proof of (i). Again, from (2.7)

$$\|\hat{\theta}-y\|^2 = [R\delta(\frac{\rho}{\lambda_i+\rho})R'\mu+R\delta(\frac{\lambda_i}{\lambda_i+\rho})R'y-y]'$$

$$\times [R\delta(\frac{\rho}{\lambda_i+\rho})R'\mu+ R\delta(\frac{\lambda_i}{\lambda_i+\rho})R'y-y]$$

$$= \|\delta(\frac{\rho}{\lambda_i+\rho})R'\mu + \delta(\frac{\lambda_i}{\lambda_i+\rho})R'y - R'y\|^2$$

$$= \|\delta(\frac{\rho}{\lambda_i+\rho})R'\mu + [\delta(\frac{\lambda_i}{\lambda_i+\rho})-I]R'y\|^2$$

$$= \|\delta(\frac{\rho}{\lambda_i+\rho})R'\mu - \delta(\frac{\rho}{\lambda_i+\rho})R'y\|^2$$

$$= (\mu-y)'R\delta[(\frac{\rho}{\lambda_i+\rho})^2]R'(\mu-y).$$

The proof of (ii) is similar.

Lemma 2.3. Let W be an $n \times n$ symmetric matrix, $\lambda_1 \geq \ldots \geq \lambda_n$ be its eigenvalues. Then

(i) $\sup_x (x'Wx/x'x) = \lambda_1$

and

(ii) $\inf_x (x'Wx/x'x) = \lambda_n$.

For a complete proof of Lemma 2.3 followed by related results see Rao (1968).

The limiting process of Lemma 2.1 can be understood in view of the following theorem and its corollaries.

Theorem 2.1. Let $\lambda_1 > \ldots > \lambda_n$ be the eigenvalues of V where $\mathrm{cov}(\theta) = (1/\rho)V$; then

(i) $(\frac{\rho}{\lambda_1+\rho})\|\mu-y\| \leq \|\hat{\theta}-y\| \leq (\frac{\rho}{\lambda_n+\rho})\|\mu-y\|$.

(ii) $(\frac{\lambda_n}{\lambda_n+\rho})\|\mu-y\| \leq \|\hat{\theta}-\mu\| \leq (\frac{\lambda_1}{\lambda_1+\rho})\|\mu-y\|$.

Proof of (i). From Lemmas 2.2, 2.3

$$\sup_y \frac{\|\hat{\theta}-y\|^2}{\|\mu-y\|^2} = \sup_y \frac{(\mu-y)'R\delta[(\frac{\rho}{\lambda_i+\rho})^2]R'(\mu-y)}{(\mu-y)'(\mu-y)} = (\frac{\rho}{\lambda_n+\rho})^2 ,$$

$$\therefore \quad \|\hat{\theta}-y\| \leq (\frac{\rho}{\lambda_n+\rho})\|\mu-y\| \quad \forall\ y \in \mathbb{R}^n.$$

Also, $\inf_{y} \frac{\|\hat{\theta}-y\|^2}{\|\mu-y\|^2} = (\frac{\rho}{\lambda_1+\rho})^2$. (Lemma 2.3(ii).) Therefore

$$\|\hat{\theta}-y\| \geq (\frac{\rho}{\lambda_1+\rho})\|\mu-y\| \qquad \forall\ y \in \mathbb{R}^n.$$

Again, the proof of (ii) is similar.

The following

Corollary 2.1. (i) $\|\hat{\theta}-y\| \leq \|\mu-y\|$

(ii) $\|\hat{\theta}-\mu\| \leq \|\mu-y\|$

is obvious from Theorem 2.1.

Corollary 2.2. (i) If $\rho \leq \lambda_n$, then $\|\hat{\theta}-y\| \leq \|\hat{\theta}-\mu\|$.

(ii) If $\rho \geq \lambda_1$, then $\|\hat{\theta}-y\| \geq \|\hat{\theta}-\mu\|$.

Proof of (i). From Theorem 2.1 we have $\|\hat{\theta}-y\| \leq (\frac{\rho}{\lambda_n+\rho})\|\mu-y\|$. Suppose that $\rho \leq \lambda_n$; then $\lambda_n+\rho \geq 2\rho$ or $\frac{\rho}{\lambda_n+\rho} \leq 1/2$ so that

$$\|\hat{\theta}-y\| \leq (1/2)\|\mu-y\| \leq (1/2)\|(\mu-\hat{\theta}) + (\hat{\theta}-y)\|$$

$$\leq (1/2)\|\mu-\hat{\theta}\| + (1/2)\|\hat{\theta}-y\|\ ,$$

which implies $\|\hat{\theta}-y\| \leq \|\hat{\theta}-\mu\|$.

The proof of (ii) is similar to that of (i).

The James-Stein Estimate $\bar{\bar{\theta}}$ and Related Results

Within the framework

$$\tilde{y} \mid \theta \sim N_n(\theta, I), \quad \tilde{\theta} \sim N_n(0, (1/\rho)I), \quad \rho > 0, \tag{2.8}$$

Efron and Morris (1973) provide an elementary investigation of the James-Stein estimator

$$\bar{\bar{\theta}} = (1-(n-2)/y'y)y$$

and discuss such modifications of $\bar{\bar{\theta}}$ as

$$\bar{\bar{\theta}}^+ = (1-(n-2)/y'y)^+ y .$$

The "plus" notation is intended as usual; i.e., for real a

$$a^+ = \max(0, a).$$

Charles Stein's original paper (1956) demonstrated the inadmissibility of $\bar{\theta}$ (the m.l.e.) in the sense that

$$(1/n)E_\theta(\theta-\bar{\bar{\theta}})'(\theta-\bar{\bar{\theta}}) < (1/n)E_\theta(\theta-\bar{\theta})'(\theta-\bar{\theta}) = 1, \quad \forall\, \theta \in \mathbb{R}^n.$$

In what follows, whenever reference is made to the "loss function" $L(\theta, \theta^*)$ or to the "risk function" $R(\theta, \theta^*)$ where θ^* is some estimator of θ, the intention is that

$$L(\theta, \theta^*) = (1/n)(\theta-\theta^*)'(\theta-\theta^*) = (1/n)\sum_{i=1}^{n}(\theta_i-\theta_i^*)^2$$

and $R(\theta, \theta^*) = E_\theta L(\theta, \theta^*)$. Regarding the notation for expectations, $E_\theta(\cdot)$, $E_y(\cdot)$, and $E(\cdot)$ are used to denote expectations with respect to the distributions of $\tilde{y}|\theta$, $\tilde{\theta}|y$, and $(\tilde{y},\tilde{\theta})$ respectively. "An estimator θ^{**} dominates an estimator θ^*" will be taken to mean that $R(\theta, \theta^{**}) \leq R(\theta, \theta^*)$ $\forall\ \theta \in \mathbb{R}^n$ and $R(\theta_1, \theta^{**}) < R(\theta_1, \theta^*)$ for some $\theta_1 \in \mathbb{R}^n$. "Uniform dominance" will be used to describe the situation where strict inequality holds $\forall\ \theta \in \mathbb{R}^n$.

A. Baranchik in his unpublished thesis (Stanford) proved the uniform dominance of $\bar{\bar{\theta}}^+$ over $\bar{\bar{\theta}}$.

The estimator $\bar{\bar{\theta}}$ can be obtained from (2.8) by replacing $\frac{\rho}{1+\rho}$ in the posterior mean

$$\hat{\theta} = E(\tilde{\theta}|y) = (1-\rho/(1+\rho))y \tag{2.9}$$

(recall the assumptions in (2.8)) with an estimate based upon information provided by the marginal distribution

$$\tilde{y} \sim N_n(0, (1+\rho)/\rho I)\ . \tag{2.10}$$

Within (2.10), $\tilde{y}'\tilde{y}$ is a complete sufficient statistic and a standard calculation shows that

$$E\left(\frac{n-2}{\tilde{y}'\tilde{y}}\right) = \frac{\rho}{(1+\rho)} \cdot$$

(Let $\tilde{u} = \left(\frac{\rho}{(1+\rho)}\right)\tilde{y}'\tilde{y}$; then $\tilde{u} \sim \chi^2_n$ and

$$E\left(\frac{n-2}{\tilde{y}'\tilde{y}}\right) = \left(\frac{\rho}{(1+\rho)}\right)E\left(\frac{n-2}{\tilde{u}}\right)$$

$$= \left(\frac{\rho}{(1+\rho)}\right)(n-2)\frac{1}{\Gamma(n/2)2^{n/2}}\int_0^\infty \left(\frac{1}{u}\right)u^{\frac{n}{2}-1}e^{-u/2}du$$

$$= \left(\frac{\rho}{(1+\rho)}\right)(n-2)\frac{\Gamma\left(\frac{n}{2}-1\right)2^{\frac{n}{2}-1}}{\Gamma(n/2)2^{n/2}} = \frac{\rho}{1+\rho} \quad .\ )$$

Hence the criterion of m.v.u.e. (as applied to the marginal distribution of $\tilde{y}$) is seen to result in the James-Stein estimator.

Instead of imposing unbiasedness (within the framework of (2.10)), we might have a suitable prior density $f_{\tilde{\rho}}(\rho)$ in mind defining $\rho \geq 0$; and write $f_{\tilde{\rho}|y}(\rho) \propto f_{\tilde{y}|\rho}(y)f_{\tilde{\rho}}(\rho)$ with the intention of using the mode of this posterior density to estimate ρ. Here, $f_{\tilde{y}|\rho}(y)$ is the density defined by (2.10); i.e.,

$$f_{\tilde{y}|\rho}(y) \propto \pi^{(n/2)}\exp((-1/2)\pi y'y), \tag{2.11}$$

where $\pi = \frac{\rho}{(1+\rho)}$ and $\rho \geq 0$. The natural (conjugate) prior for π -- considering (2.11)--is $f_{\tilde{\pi}}(\pi) \propto \pi^{(\alpha-1)}\exp(-\pi/\beta)1_{(0,1)}(\pi)$, $\alpha > 0$, $\beta > 0$. Combining this prior with the likelihood function of (2.11) results in the posterior

$$f_{\tilde{\pi}|y}(\pi) \propto \pi^{(n/2)+(a-1)} \exp((-1/2)\pi y'y - \pi/\beta) 1_{(0,1)}(\pi) . \quad (2.12)$$

Calculations for the mode of (2.12) proceed, as usual,

$$L = \ln f_{\tilde{\pi}|y}(\pi) = (\text{const.}) + (n/2 + (a-1)) \ln \pi - (1/2)\pi y'y - \pi/\beta$$

$$\frac{\partial L}{\partial \pi}\Big|_{\hat{\pi}} = \frac{(n/2 + a - 1)}{\pi} - (1/2)y'y - 1/\beta = 0$$

which implies

$$(n/2 + a - 1) - \hat{\pi}(y'y + 2/\beta) = 0$$

or

$$\hat{\pi} = \frac{(n-2) + 2a}{y'y + 2/\beta} = \frac{n+2(a-1)}{y'y + 2/\beta}$$

provided that $\hat{\pi} < 1$. Assume that $n \geq 2$. What is the mode of (2.12) if a given $y'y$ renders the stationary point larger than 1? In that even the mode is at $\pi = 1$ since

$$\frac{\partial L}{\partial \pi} = (1/2)(y'y + 2/\beta)[(1/\pi)\frac{(n+2(a-1))}{y'y + 2/\beta} - 1] = (1/2)(y'y + 2/\beta)[(1/\pi)\hat{\pi} - 1]$$

implies that the slope (as a function of π) is positive to the left of the stationary point. The modal estimate is therefore

$$\hat{\pi} = \min\{1, \frac{n+2(a-1)}{y'y + 2/\beta}\} .$$

Substituting for $\pi = \frac{\rho}{1+\rho}$ in (2.9) we have

$$\hat{\theta} = (1 - \frac{n+2(a-1)}{y'y+2/\beta})^{+} y \ . \tag{2.13}$$

From (2.13) the positive part version of the James-Stein estimator is the consequence of taking the "flat" gamma prior $a = 0$ and $\beta = \infty$.

The following theorem concerns the naive estimator ($\beta = \infty$) $\hat{\theta}_o$,

$$\hat{\theta}_o = (1 - \frac{(n-2)+2a}{y'y})y \ . \tag{2.14}$$

Theorem 2.1. If $n > 2(a+1)$, then the estimator $\hat{\theta}_o$ uniformly dominates $\bar{\theta}$. $R(\theta, \hat{\theta}_o)$ is minimized for the case where $a = 0$; i.e., where $\hat{\theta}_o$ is equal to the James-Stein estimator.

Proof. The proof will follow the strategy Efron and Morris (1973) have used to establish the uniform dominance of $\bar{\bar{\theta}}$ over $\bar{\theta}$. The key idea is that, where $\tilde{\theta} \sim N_n(0, (1/\rho)I)$, $\rho > 0$, we have $\|\theta\|^2 \sim (\frac{1}{\rho})\chi^2_n$; and the family of distributions of $\|\theta\|^2$ being complete as a function of ρ is also complete as a function of $\frac{\rho}{1+\rho}$.

Next, let

$$R_1 = 1 - (2-c/(n-2))(c/n)E_\theta(\frac{n-2}{\tilde{y}'\tilde{y}}) \tag{2.15}$$

where $c = (n-2) + 2a$ and

$$R_2 = (1/n)E_\theta(\hat{\theta}_o - \theta)'(\hat{\theta}_o - \theta) \ . \tag{2.16}$$

Since $\tilde{y}|\theta \sim N_n(\theta, I)$, $\tilde{y}'\tilde{y}$ is distributed as a noncentral χ^2

on n degrees of freedom with noncentrality parameter $\theta'\theta$. R_1 is therefore a function of $\theta'\theta$. To see that R_2 is also a function of $\theta'\theta$, consider

$$R_2 = (1/n)E_\theta((1-c/\tilde{y}'\tilde{y})\tilde{y}-\theta)'((1-c/\tilde{y}'\tilde{y})y-\theta)$$
$$= (1/n)E_\theta((1-c/\tilde{y}'\tilde{y})^2\tilde{y}'\tilde{y} - 2(1-c/\tilde{y}'\tilde{y})\tilde{y}'\theta+\theta'\theta) .$$

It is sufficient to argue that $E_\theta(\tilde{y}'\theta/\tilde{y}'\tilde{y})$ is a function of $\theta'\theta$. To see why this expectation is a function of $\theta'\theta$, let $v = \theta/\|\theta\|$ and $u^{(i)}$ be the ith unit vector of $\mathbb{R}^n$; i.e.,

$$u_k^{(i)} = \begin{cases} 1 & i = k \\ 0 & i \neq k \end{cases} \qquad i = 1, 2, \ldots, n.$$

Let $U_1, U_2, \ldots, U_n$ be the orthonormal basis resulting from a Gram-Schmidt application on v, $u^{(i)}$ $i = 2, 3, \ldots, n$ (assuming, of course, that $v_1 \neq 0$ --otherwise--more care is required in selecting the $u^{(i)}$'s). $U = (U_1, U_2, \ldots, U_n)$ is an orthogonal matrix such that $U'\theta = \|\theta\|u^{(1)}$. Set $U'\tilde{y} = \tilde{z}$, then $\tilde{z}|\theta \sim N_n(\|\theta\|u^{(1)}, I)$. Finally,

$$E_\theta(\tilde{y}'\theta/\tilde{y}'\tilde{y}) = E_\theta(\frac{(U'\tilde{y})'(U'\theta)}{\tilde{y}'\tilde{y}}) = E_\theta(\frac{\|\theta\|\tilde{z}_1}{\tilde{y}'\tilde{y}})$$
$$= \|\theta\|E_\theta((1/\tilde{y}'\tilde{y})E_\theta(\tilde{z}_1|y'y)) .$$

The function $E_\theta(\tilde{z}_1|y'y)$ answers the following question. What is the center of mass of a "thin" spherical shell $\tilde{y}'\tilde{y} = S$ when the mass

has a normal (spherical) distribution about $\|\theta\| u^{(1)}$? The expectation clearly depends only upon the values of $y'y$ and $\theta'\theta$ so we write

$$E_\theta(\tilde{y}'\theta \,|\, \tilde{y}'\tilde{y}) = f(\tilde{y}'\tilde{y}, \theta'\theta)$$

where f is some function measurable with respect to the sigma field generated by $y'y$ for a given $\theta'\theta$.

Next, we show that $ER_1 = ER_2 \quad \forall\ \rho > 0$ ($\forall\ \frac{\rho}{1+\rho}$ such that $0 < \frac{\rho}{1+\rho} < 1$). With this done, the completeness of $\theta'\theta$ implies $R_1 = R_2$ (a.e.).

Proof that $ER_1 = ER_2$.

(i) Calculation of ER_1:

$$EE_\theta(\frac{n-2}{\tilde{y}'\tilde{y}}) = EE_y(\frac{n-2}{\tilde{y}'\tilde{y}}) = E(\frac{n-2}{\tilde{y}'\tilde{y}}) = (\frac{\rho}{1+\rho}) \qquad \text{(see p. 20)}$$

$$\therefore\ ER_1 = 1 - (2 - \frac{c}{n-2})(\frac{c}{n})(\frac{\rho}{1+\rho})\ .$$

(ii) Calculation of ER_2:

$$\frac{1}{n} EE_\theta(\hat{\theta}_o - \theta)'(\hat{\theta}_o - \theta)$$

$$= \frac{1}{n} EE_y(\hat{\theta}_o - \theta)'(\hat{\theta}_o - \theta)$$

$$= \frac{1}{n} EE_y[(1 - \frac{c}{\tilde{y}'\tilde{y}})\tilde{y} - \theta]'[(1 - \frac{c}{\tilde{y}'\tilde{y}})\tilde{y} - \theta]$$

$$= \frac{1}{n} E[(1 - \frac{c}{\tilde{y}'\tilde{y}})\tilde{y} - (\frac{1}{1+\rho})\tilde{y}]'[(1 - \frac{c}{\tilde{y}'\tilde{y}})\tilde{y} - (\frac{1}{1+\rho})\tilde{y}] \quad +$$

$$+ \frac{1}{1+\rho} \quad (\text{since } \tilde{\theta} \mid y \sim N_n((\frac{1}{1+\rho})y,\ (\frac{1}{1+\rho})I))$$

$$= \frac{1}{n} E[(\frac{\rho}{1+\rho} - \frac{c}{\tilde{y}'\tilde{y}})^2 \tilde{y}'\tilde{y}] + \frac{1}{1+\rho}$$

$$= \frac{1}{n} E[\tilde{y}'\tilde{y}(\frac{\rho}{1+\rho})^2 - 2(\frac{\rho c}{1+\rho}) + \frac{c^2}{\tilde{y}'\tilde{y}}] + \frac{1}{1+\rho}$$

$$= \frac{1}{n} [(\frac{\rho}{1+\rho})^2 n(\frac{1+\rho}{\rho}) - \frac{2\rho c}{1+\rho} + \frac{c^2}{n-2} E(\frac{n-2}{\tilde{y}'\tilde{y}})] + \frac{1}{1+\rho}$$

$$(\text{since } \tilde{y} \sim N_n(0, (\frac{1+\rho}{\rho})I) \text{ and } (\frac{\rho}{1+\rho})\tilde{y}'\tilde{y} \sim \chi^2_n)$$

$$= \frac{1}{n}[n(\frac{\rho}{1+\rho}) - \frac{2\rho c}{1+\rho} + \frac{c^2}{n-2} E(\frac{n-2}{\tilde{y}'\tilde{y}})] + \frac{1}{1+\rho}$$

$$= 1 - \frac{2c}{n}(\frac{\rho}{1+\rho}) + \frac{c^2}{n(n-2)} E(\frac{n-2}{\tilde{y}'\tilde{y}})$$

$$= 1 - (2-\frac{c}{n-2})(\frac{c}{n})(\frac{\rho}{1+\rho}) . \qquad (2.17)$$

The conclusion is that $R_1 = R_2$ (a.e.).

Now assume that $n > 2(a+1)$. Since

$$R_1 = (1/n)E_\theta(\hat{\theta}_o - \theta)'(\hat{\theta}_o - \theta) = 1 - (2 - c/(n-2))(c/n)E_\theta((n-2)/\tilde{y}'\tilde{y}) ,$$

and since $(1/n)E_\theta(\bar{\theta}-\theta)'(\bar{\theta}-\theta) = 1$, the uniform dominance of $\hat{\theta}_o$ over $\bar{\theta} = y$ is established by showing that

$$0 < (2 - c/(n-2))(c/n) . \qquad (2.18)$$

Now, $c/n > 0$ since $c > 0$. The truth of (2.18) easily follows since

$$2 - \frac{c}{n-2} = 2 - \frac{(n-2)+2a}{n-2} = 1 - \frac{2a}{n-2}$$

and we have assumed that $n > 2(a+1)$ or $n - 2 > 2a$.

Finally, the fact that the James-Stein estimator uniformly dominates all estimators described by Equation (2.14) (under the conditions of the above theorem) is transparent from the identity

$$(2 - \frac{c}{n-2})(c/n) = \frac{1}{n(n-2)}((n-2)^2 - 4a^2) , \qquad (2.19)$$

(2.19) being maximized where $a = 0$. This concludes the proof.

Noninformative Priors

In the last section, the marginal distribution of $\tilde{\tilde{y}}$ (see (2.10)) was

$$f_{\tilde{\tilde{y}}|\pi}(y) \propto \pi^{(n/2)} \exp((-1/2)\pi y'y) \qquad (2.20)$$

where $\pi = \frac{\rho}{1+\rho}$. It was found that

$$f_{\tilde{\pi}}(\pi) \propto (1/\pi) 1_{(0,1)}(\pi)$$

combined with (2.20) to give a posterior distribution $g_{\tilde{\pi}|y}(\pi)$ whose mode, $\hat{\pi}$, had the property

$$\hat{\theta}(\hat{\pi}, y) = E(\tilde{\theta}|y, \hat{\pi}) = \bar{\bar{\theta}}^{+} = (1 - \frac{n-2}{y'y})^{+} y .$$

For convenience, let us state

$$g_{\tilde{\pi}|y}(\pi) \propto \pi^{(n/2)-1}\exp((-1/2)\pi y'y)1_{(0,1)}(\pi) \cdot \qquad (2.21)$$

The more complicated model of the next chapter will necessitate thinking in terms of distributions of $\tilde{\rho}$. In this regard, $\pi = \frac{\rho}{1+\rho}$ and $\frac{d\pi}{d\rho} = \frac{1}{(1+\rho)^2}$; so that, from (2.21),

$$g_{\tilde{\rho}|y}(\rho) \propto (\frac{\rho}{1+\rho})^{(n/2)-1}\exp((-1/2)(\frac{\rho}{1+\rho})y'y)\frac{1}{(1+\rho)^2}1_{(0,\infty)}(\rho)$$

$$= (\frac{\rho}{1+\rho})^{(n/2)}\exp((-1/2)(\frac{\rho}{1+\rho})y'y)(\frac{1}{\rho(1+\rho)})1_{(0,\infty)}(\rho). \qquad (2.22)$$

Figure 2.1 shows densities $g_{\tilde{\rho}|y}$ for $n = 8$ and values of $y'y$ as indicated. The posterior (2.22) can be viewed as the result of combining $f_{\tilde{y}|\rho}(= f_{\tilde{y}|\pi})$ of (2.20) with the prior

$$f_{\tilde{\rho}}(\rho) \propto \frac{1}{\rho(1+\rho)}1_{(0,\infty)}(\rho) \cdot \qquad (2.23)$$

The interpretation of Figure 2.1 is clear. A large $y'y$ renders the mode, $\hat{\rho}$, of (2.22) close to zero. Consequently, the estimate

$$\hat{\theta}(\hat{\rho}) = (1 - \frac{\hat{\rho}}{1+\hat{\rho}})y$$

is practically indistinguishable from the m.l.e. . This behavior (given large $y'y$) is obvious, of course, in view of the formula

$$\bar{\bar{\theta}} = (1 - \frac{n-2}{y'y})y \ .$$

Note that for the data set of Figure 1.1, $y'y \doteq 69.3$.

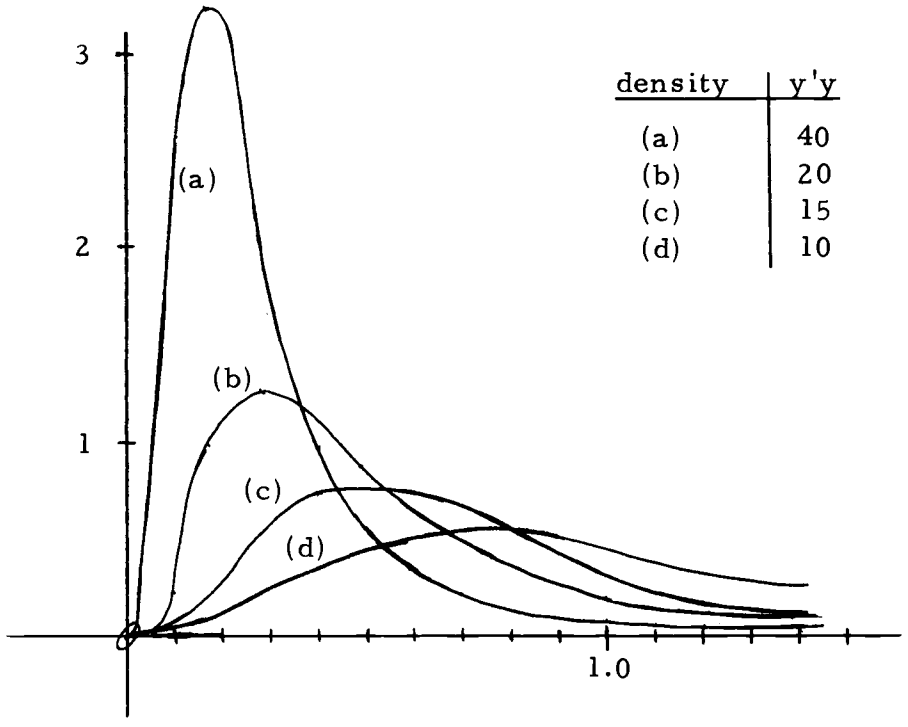


Figure 2.1. Posterior densities $g_{\tilde{\rho}|y}$ for $n = 8$ and $y'y$ as shown.

The prior density of (2.23) is a "flat" prior in a very interesting sense. The following definition of "noninformative priors" is stated in Box and Taio (1972).

Definition of Noninformative Priors

If $\phi(\rho)$ is a one-to-one transformation of ρ, we shall say that a prior distribution of ρ which is locally proportional to $|\frac{d\phi}{d\rho}|$ is noninformative for the parameter ρ if, in terms of ϕ, the likelihood curve is data translated, that is, the data only serve to change the location of the likelihood

$$\ell_{\tilde{\phi}|y}(\phi) = f_{\tilde{y}|\phi}(y) .$$

Mathematically, a data translated likelihood must be expressible in the form

$$\ell_{\tilde{\phi}|y}(\phi) = r[\phi(\rho)-t(y)]$$

where r is a known function independent of the data and $t(y)$ is a function of y.

Assume that $\ell_{\tilde{\phi}|y}$ is continuous, has a unique maximum, and is data translated (again, see Box and Taio (1972)). An important idea, here, is that if we have little knowledge a priori about the parameter ρ, we might be almost equally willing to accept one value of a one-to-one function $\phi(\rho)$ as another. This state of indifference may be expressed by taking $\phi(\tilde{\rho})$ to be locally uniform. If the prior $f_{\phi(\tilde{\rho})}(\rho)$ is locally uniform, then $f_{\tilde{\phi}|y}(\phi) \propto \ell_{\tilde{\phi}|y}(\phi)$ (locally). A consequence (of data translated likelihood and the above mentioned

regularity conditions of $\ell_{\tilde{\phi}|y}$) is that high posterior density intervals for $\tilde{\phi}$, of a given percent, have constant length (locally).

Let us return to the likelihood function of (2.20). Thinking in terms of priors of ρ rather than on $\pi = \frac{1}{1+\rho}$, let us write (2.20) as $\ell_{\tilde{\rho}|y}(\rho) = f_{\tilde{y}|\rho}(y) \propto (\frac{\rho}{1+\rho})^{(n/2)} \exp((-1/2)(\frac{\rho}{1+\rho})y'y)$. Multiplication by $(y'y)^{(n/2)}$ does not change the likelihood (as a function of ρ) so we have

$$\ell_{\tilde{\rho}|y}(\rho) \propto (y'y\,\frac{\rho}{1+\rho})^{(n/2)} \exp((-1/2)(\frac{\rho}{1+\rho})y'y)$$

$$= \exp((n/2)(\ln(y'y) + \ln(\frac{\rho}{1+\rho})))$$

$$\times \exp((-1/2)\exp(\ln(y'y) + \ln(\frac{\rho}{1+\rho})))$$

$$\propto \ell_{\ln\tilde{\pi}|y}(\ln\pi) .$$

In the logarithmic metric, the data acting through $y'y$ only change the location of the likelihood. What does this say about a prior on ρ? We want $\phi(\tilde{\rho}) = \ln(\frac{\tilde{\rho}}{1+\tilde{\rho}})$ to be locally uniform so that $f_{\phi(\tilde{\rho})|y}(\phi) \propto \ell_{\phi(\tilde{\rho})|y}(\phi)$ (locally). Consequently,

$$f_{\tilde{\rho}}(\rho) \propto \left|\frac{d\,\ln(\frac{\rho}{1+\rho})}{d\rho}\right| 1_{(0,\infty)}(\rho) = \frac{1}{\rho(1+\rho)}\, 1_{(0,\infty)}(\rho)$$

is by definition our noninformative prior on ρ. Note that we have obtained the prior density of (2.23). Conversely, if one is motivated

at the start to impose the noninformative prior $f_{\tilde{\rho}}(\rho) = \frac{1}{\rho(1+\rho)} 1_{(0,\infty)}(\rho)$, he would arrive at the posterior (2.22). Since $\hat{\theta}(\pi) = E(\tilde{\theta} \mid y,\pi) = (1-\pi)y$, a change of variables (from ρ to π) might seem reasonable. In making this change, the posterior density of (2.22) is transformed to that of (2.21). The mode of (2.21) resulted in the final output $\tilde{\tilde{\theta}}^{+}$ as an estimate of θ.

The Regression Problem--An Intermediate Model

We have seen that from the Bayesian assumptions

$$\tilde{y} \mid \theta \sim N_n(\theta, I), \quad \tilde{\theta} \sim N_n(0, (1/\rho)I) ,$$

estimates of θ are obtained which uniformly dominate $\bar{\theta}$. In certain practical situations it might be desirable to sacrifice uniform dominance for sharp performance over specified subsets of $\mathbb{R}^n$. For example, the investigator might prescribe a distribution for $\tilde{\theta}_t - \tilde{\theta}_{t-1}$ consistent with his belief that the response curve is "smooth." The performance (risk) of his estimator over such subsets of $\mathbb{R}^n$ as

$$\mathcal{S}_t = \{\theta \in R^n: \sum_{i=1}^{n} (\theta_i - \theta_{i-1})^2 \leq t, \quad \theta_o = 0\}$$

might then be of interest.

Instead of having $\tilde{\theta} \sim N_n(0, (1/\rho)I)$, which can be viewed as an

autoregressive process of order zero:

$$\tilde{\theta}_o = 0, \quad \tilde{\theta}_t = \tilde{\epsilon}_t \quad t = 1, 2, \ldots, n$$

where $\tilde{\epsilon} \sim N_n(0, (1/\rho)I)$, let us consider an autoregressive process of order 1:

$$\tilde{\theta}_o = 0, \quad \tilde{\phi}_t = \tilde{\theta}_t - \tilde{\theta}_{t-1} = \tilde{\epsilon}_t, \quad t = 1, 2, \ldots, n \tag{2.24}$$

where, again, $\tilde{\epsilon} \sim N_N(0, (1/\rho)I)$. Note that $E(\tilde{\theta}_t | \tilde{\theta}_{t-1} = \theta_{t-1}) = \theta_{t-1}$.

With the latter assumption (2.24), we have

$$\tilde{y} | \theta \sim N_n(\theta, I), \quad T_o\tilde{\theta} \sim N_n(0, (1/\rho)I)$$

where

$$T_o = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & & 0 \\ \vdots & & & & \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix}, \quad W_o = T_o^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \ddots & \\ 1 & 1 & 1 & \cdots & & 1 \end{pmatrix}.$$

Therefore $\tilde{\theta} \sim N_n(0, (1/\rho)V_o)$ where

$$(T_o^{-1})(T_o^{-1})' = V_o = \begin{pmatrix} 1 & 1 & \cdots & & 1 \\ & 2 & 2 & \cdots & 2 \\ & & 3 & 3 & 3 \\ & & & & \\ 1 & 2 & 3 & 4 & n \end{pmatrix} \tag{2.25}$$

To accommodate the regression situation of Figure 1.1,

$(n = 10)$, reindex the parameters so that θ_k is the mean response at $X_k = 1.5 + (k-1)(.5)$, $k = 1, 2, \ldots, 8$ (now $\theta \in \mathbb{R}^8$). $\mathrm{Cov}(\tilde{\theta}) = (1/\rho)V$ where V is the lower right submatrix of (2.25) whose first row has 3 in each position. Let T be equal to T_o with the exception that the element in row 1, column 1 (of T) is equal to $1/\sqrt{3}$. Now, we have

$$\tilde{\phi} = T\tilde{\theta} \sim N_8(0, (1/\rho)I) \quad \text{(since } TVT' = I) \qquad (2.26)$$

and

$$\tilde{z} = T\tilde{y} \sim N_8(\phi, TT').$$

Within the framework of (2.26), $z = \text{m.l.e.}(\phi) = \bar{\phi}$ and

$$R'(\bar{\phi}, \phi) = (1/8)E_\phi(\tilde{z}-\phi)'(TT')^{-1}(\tilde{z}-\phi)$$

$$= (1/8)E_\phi[T^{-1}(\tilde{z}-\phi)]'[T^{-1}(\tilde{z}-\phi)] = (1/8)E_\theta(\tilde{y}-\theta)'(\tilde{y}-\theta).$$

The corresponding Bayesian estimate of ϕ is

$$\hat{\phi}(\rho) = E(\tilde{\phi}\,|\,z, \rho) = (1/\rho)((1/\rho)I+TT')^{-1}z.$$

Since we are primarily interested in estimating θ, observe that the risk in estimating θ with $T^{-1}\hat{\phi}(\rho)$ is the same as the risk in estimating ϕ with $\hat{\phi}(\rho)$ since

$$R(\theta, T^{-1}\hat{\phi}) = (1/n)E_{\theta}(T^{-1}\hat{\phi}-\theta)'(T^{-1}\hat{\phi}-\theta)$$

$$= E_{\theta}(T^{-1}\hat{\phi}-T^{-1}(T\theta))'(T^{-1}\hat{\phi}-T^{-1}(T\theta))$$

$$= E_{\phi}(\hat{\phi}-\phi)'(TT')^{-1}(\hat{\phi}-\phi).$$

Define $R'(\phi,\hat{\phi}) = (1/n)E_{\phi}(\hat{\phi}-\phi)(TT')^{-1}(\hat{\phi}-\phi)$. To understand $R(\theta,\hat{\theta})$ (where $\hat{\theta} = T^{-1}\hat{\phi}(\rho) = T^{-1}(1/\rho)((1/\rho)I+TT')^{-1}Ty$) is it therefore sufficient to study $R'(\phi,\hat{\phi})$. Let us record

$$R'(\phi,\hat{\phi}) = E_{\phi}[1/\rho((1/\rho)I+TT')^{-1}\tilde{z}-\phi]'(TT')^{-1}[1/\rho((1/\rho)I+TT')^{-1}\tilde{z}-\phi]$$

$$= [1/\rho((1/\rho)I+TT')^{-1}\phi-\phi]'(TT')^{-1}[1/\rho((1/\rho)I+TT')^{-1}\phi-\phi]$$

$$+ \operatorname{tr}(TT')^{-1}(1/\rho)((1/\rho)I+TT')^{-1}(TT')(1/\rho)((1/\rho)I+TT')^{-1}. \tag{2.27}$$

Remark. By writing the spectral decomposition of TT',

$$TT' = R\Lambda R' = \sum_{i=1}^{8}\lambda_i(R_iR_i') = \sum_{i=1}^{8}\lambda_iE_i$$

(regarding notational convention, see page 14), Equation (2.27) can be expressed in a more suitable form for computer calculations. Specifically,

$$R'(\phi,\hat{\phi}) = \rho^2\sum_{i=1}^{8}\frac{\lambda_i}{(1+\rho\lambda_i)^2}\phi'E_i\phi + \sum_{i=1}^{8}\left(\frac{1}{1+\rho\lambda_i}\right)^2. \tag{2.28}$$

To derive (2.28), first observe that

$$\text{(i)}\quad (TT')^{-1} = \sum_{i=1}^{8} (1/\lambda_i)E_i \ ;$$

$$\text{(ii)}\quad ((1/\rho)I+TT')^{-1} = [(1/\rho)I + \sum_{i=1}^{8} \lambda_i E_i]^{-1}$$

$$= [\sum_{i=1}^{8} ((1/\rho)+\lambda_i)E_i]^{-1} = \sum_{i=1}^{8} (\frac{\rho}{1+\rho\lambda_i})E_i \ ;$$

$$\text{(iii)}\quad (1/\rho)((1/\rho)I+TT')^{-1} - I = -\sum_{i=1}^{8} (\frac{\rho\lambda_i}{1+\rho\lambda_i})E_i \ . \qquad (2.29)$$

The first expression on the right side of (2.27) becomes

$$\phi'[\sum_{i=1}^{8} (\frac{-\rho\lambda_i}{1+\rho\lambda_i})E_i][\sum_{i=1}^{8} (1/\lambda_i)E_i][\sum_{i=1}^{8} (\frac{-\rho\lambda_i}{1+\rho\lambda_i})E_i]\phi$$

$$= \phi'[\sum_{i=1}^{8} (\frac{\rho}{1+\rho\lambda_i})E_i][\sum_{i=1}^{8} (\frac{\rho\lambda_i}{1+\rho\lambda_i})E_i]\phi$$

$$= \phi'[\sum_{i=1}^{8} \frac{\rho^2\lambda_i}{(1+\rho\lambda_i)^2}E_i]\phi = \rho^2 \sum_{i=1}^{8} \frac{\lambda_i}{(1+\rho\lambda_i)^2}\ \phi'E_i\phi \ .$$

And the second term on the right side of (2.27) is

$$\mathrm{tr}(TT')^{-1}(1/\rho)[(TT')((1/\rho)I+TT')^{-1}(TT')^{-1}](TT')(1/\rho)((1/\rho)I+TT')^{-1}$$

$$= \mathrm{tr}(1/\rho)^2((1/\rho)I+TT')^{-2} = \mathrm{tr}(1/\rho)^2 \sum_{i=1}^{8} \left(\frac{\rho}{1+\rho\lambda_i}\right)^2 E_i = \sum_{i=1}^{8} \left(\frac{\rho}{1+\rho\lambda_i}\right)^2$$

after noting that

$$((1/\rho)I+TT')^{-1} = \rho(\rho I+(TT')^{-1})^{-1}(TT')^{-1} = (TT')(TT'+(1/\rho)I)^{-1}(TT')^{-1}$$

and that trace $(E_i) = 1$ (trace (E_i) = trace $(R_i R_i') = R_i' R_i = 1.$)

The expression for $R'(\phi, \hat{\phi})$ of (2.28) is still fairly complicated. The following approximation seems convenient. Define

$$\xi_\rho = \max\{\lambda_i \left(\frac{\rho}{1+\rho\lambda_i}\right)^2\}_{i=1}^{8} .$$

Since

$$\rho^2 \sum_{i=1}^{8} \frac{\lambda_i}{(1+\rho\lambda_i)^2} \phi' E_i \phi = \phi' R \delta[\lambda_i \left(\frac{\rho}{1+\rho\lambda_i}\right)^2] R' \phi ,$$

$$\sup_{\phi} \frac{\phi' R \delta[\lambda_i \left(\frac{\rho}{1+\rho\lambda_i}\right)^2] R' \phi}{\phi'\phi} = \xi_\rho$$

(see Lemma 2.3) and we have $R'(\phi, \phi) \leq \phi'\phi\xi_\rho + \sum_{i=1}^{8} \left(\frac{1}{1+\rho\lambda_i}\right)^2$. Define

$$B(\phi'\phi, \rho) = \phi'\phi\xi_\rho + \sum_{i=1}^{8} \left(\frac{1}{1+\rho\lambda_i}\right)^2 .$$

Some calculations of $B(\phi'\phi, \rho)$ are included in Table 2.1. Recall that $R(\theta, \bar{\theta}) = 1$. Observe the first column of Table 2.1 in connection with Lemma 2.1. Note that for $\phi_1'\phi_1 < \phi_2'\phi_2$, $B(\phi_1'\phi_1, \rho) \leq B(\phi_2'\phi_2, \rho)$. Table 2.1 reflects the previously stated willingness to sacrifice uniform dominance for sharp performance over ϕ contained in certain intervals. Accordingly, the calculations indicate good performance (for a given ρ) when $\phi'\phi$ is small and degradation of performance as $\phi'\phi$ increases. A $\phi'\phi$ which is exceedingly large for a given precision, ρ, indicates that our prior is inappropriate. Recall that (see page 11)

$$\rho\tilde{\phi}'\tilde{\phi} = \rho \sum_{i=1}^{n} (\tilde{\theta}_i - \tilde{\theta}_{i-1})^2 \sim \chi_n^2$$

and

$$(1/n)E\tilde{\phi}'\tilde{\phi} = 1/\rho \ .$$

For the sake of illustration, let the θ's be exactly equal to the data points observed in Figure 1.1. This being the case, $\phi'\phi = 19.72$. It appears that for moderate choices of ρ, the estimator $\hat{\theta} = T^{-1}\hat{\phi}(\rho)$ performs well (relative to the m.l.e.) over a very large class of response curves.

How does one interpret Table 2.1 in terms of the unknown values of $\phi'\phi$ and ρ ? A point of view taken by Raiffa and Schlaifer (1961)

Table 2.1. Values of $B(\phi'\phi, \rho)$.

	Prior Precision ρ									
$\phi'\phi$	0.00	0.20	0.40	0.60	0.80	1.00	1.50	2.00	3.00	5.00
10	1.00	0.67	0.59	0.58	0.59	0.61	0.72	0.83	1.08	1.70
20	1.00	0.74	0.72	0.76	0.84	0.92	1.19	1.44	1.98	3.25
30	1.00	0.80	0.84	0.95	1.09	1.23	1.66	2.06	2.88	4.81
40	1.00	0.86	0.97	1.13	1.34	1.53	2.13	2.67	3.79	6.37
50	1.00	0.92	1.09	1.32	1.59	1.84	2.59	3.28	4.69	7.93
60	1.00	0.98	1.22	1.50	1.84	2.15	3.06	3.89	5.59	9.48
70	1.00	1.04	1.34	1.69	2.09	2.46	3.53	4.51	6.49	11.0
80	1.00	1.10	1.46	1.87	2.34	2.76	4.00	5.12	7.39	12.6
90	1.00	1.17	1.59	2.06	2.59	3.07	4.47	5.73	8.30	14.2
100	1.00	1.23	1.71	2.24	2.84	3.38	4.94	6.34	9.20	15.7

is that prior probabilities should not be assigned (this corresponds to specifying ρ in our problem) without due consideration of posterior consequences. The statistician might be aided in arriving at his "betting odds" by examining the observed data and the entries of Table 2.1 for a variety of subjective choices on ρ. In this way he might gain access to his true prior opinions.

The author's inclination is more in the direction of an Empirical Bayes approach. Fundamentally, he desires to avoid (when possible) making subjective numerical assignments for hyperparameters. Accordingly, the information contained in Table 2.1 admits the following interpretation. From (2.26) the marginal distribution of $\tilde{z}|\rho$ is $N_n(0, (1/\rho)I+TT')$ or

$$f_{\tilde{z}|\rho}(z) \propto [\det((1/\rho)I+TT')]^{(-1/2)}\exp((-1/2)z'((1/\rho)I+TT')^{-1}z)$$

$$= [\prod_{i=1}^{8}(\frac{1+\rho\lambda_i}{\rho})]^{(-1/2)}\exp((-1/2)\sum_{i=1}^{8}(\frac{\rho}{1+\rho\lambda_i})z'E_iz)$$

(after writing the spectral decomposition for TT'). Let $z'R_i = w_i$, then $z'E_iz = z'R_iR_i'z = w_i^2$ and $w = (w_1, w_2, \ldots, w_8)'$ is a sufficient statistic. Empirical Bayesian considerations suggest that ρ should be estimated with some measurable function of w, say $\hat{\rho} = \hat{\rho}(w)$. Recalling our earlier success with a less complicated model, $\hat{\rho}$ might be the posterior mode obtained when (2.30) is combined with a

noninformative prior on ρ. It is seen in the next chapter that a non-informative prior on ρ is not readily available (if it in fact exists). An improper gamma prior is used in its place. Assume that $\hat{\rho}$ is a measurable function of $\tilde{z}$. Assume, further, that the loss function $L'(\phi, \hat{\phi}) = (1/8)(\hat{\phi}-\phi)'(TT')^{-1}(\hat{\phi}-\phi)$ is measurable with respect to the sigma field generated by $\hat{\rho}$. Then $E_\phi[(\hat{\phi}(\hat{\rho})-\phi)'(TT')^{-1}(\hat{\phi}(\hat{\rho})-\phi)\,|\,\hat{\rho}]$ is a measurable function of $\tilde{z}$ and we can meaningfully write

$$E_\phi[(\hat{\phi}(\hat{\rho})-\phi)'(TT')^{-1}(\hat{\phi}(\hat{\rho})-\phi)] = E_\phi[E_\phi\{(\hat{\phi}(\hat{\rho})-\phi)'(TT')^{-1}(\hat{\phi}(\hat{\rho})-\phi)\}\,|\,\hat{\rho}\}] .$$

In Table 2.1, given $\phi'\phi \leq C_o$, the row corresponding to C_o contains upper bounds for $E_\phi[(\hat{\phi}-\phi)'(TT')^{-1}(\hat{\phi}-\phi)\,|\,\hat{\rho}](1/8)$. The columns serve as guidelines for truncating $\hat{\rho}$ if we insist that $E_\phi(\hat{\phi}-\phi)'(TT')^{-1}(\hat{\phi}-\phi)(1/8) < 1$. For $\phi'\phi \leq C_o$, define

$$\rho^* = \begin{cases} k_1 & \hat{\rho} < k_1 \\ \hat{\rho} & k_1 \leq \hat{\rho} \leq k_2 \\ k_2 & k_2 \leq \hat{\rho} \end{cases} .$$

Note that $\hat{\rho}^*$ is measurable. Suppose that

$$(1/8)E_\phi[(\hat{\phi}(\hat{\rho}^*)-\phi)'(TT')^{-1}(\hat{\phi}(\hat{\rho}^*)-\phi)\,|\,\hat{\rho}^*] < 1,$$

then

$$(1/8)E_\phi(\hat{\phi}-\phi)'(TT')^{-1}(\hat{\phi}-\phi) = (1/8)E_\phi[E_\phi\{(\hat{\phi}-\phi)'(TT')^{-1}(\hat{\phi}'-\phi)\,|\,\hat{\rho}^*\}]$$

$$< 1.$$

III. AN EXERCISE IN ESTIMATION

A First Order Autoregressive Model

Figure 1.1 depicts a regression situation where we assume that the response curve passes through the natural origin $(\theta_o = 0)$. In general, suppose that we have such a natural origin and that the "levels of stress" are equally spaced. The attitude a priori that the response curve is "reasonably smooth" can be made precise by starting with a simple model of the following kind:

$$\Delta\tilde{\theta}_{k-1} = \tilde{\theta}_k - \tilde{\theta}_{k-1} = \tau + \tilde{\epsilon}_k \qquad (3.1)$$

$$k = 1, \ldots, n; \quad \tilde{\theta}_o = 0$$

$$\tilde{\epsilon} \sim N_n(0, (1/\rho)I) .$$

This implies

$$\tilde{\theta}_k = k\tau + \sum_{i=1}^{k} \tilde{\epsilon}_i \qquad (3.2)$$

or

$$(\tilde{\theta}_k - \mu_k) = (\tilde{\theta}_{k-1} - \mu_{k-1}) + \tilde{\epsilon}_k$$

where $\mu_k = k\tau \quad k = 1, \ldots, n$. Expression (3.2) denotes a first order autoregressive process in terms of the components of $\tilde{\theta} - \mu$ where

$$\mu = (1, 2, \ldots, n)'\tau .$$

In stating (3.1), the investigator is expressing his prior opinion that the response curve is roughly linear in nature, i.e.,

$$E[\Delta \tilde{\theta}_{k-1}] = \tau$$

and that it deviates from a linear function in a manner that is not too wild--this variation being expressed by the stationary process--

$$\tilde{\epsilon}_t \text{ i.i.d.}, \quad E\tilde{\epsilon}_t = 0, \quad \text{and} \quad \text{var}\,\tilde{\epsilon}_t = (1/\rho).$$

From (3.2)

$$\tilde{\theta} = \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n \end{pmatrix} \tau + \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \tilde{\epsilon}_1 \\ \tilde{\epsilon}_2 \\ \vdots \\ \tilde{\epsilon}_n \end{pmatrix}.$$

Or, with obvious notation, $\tilde{\theta} = X_o\tau + W_o\tilde{\epsilon}$. Since $\tilde{\epsilon} \sim N_n(0, (1/\rho)I)$, $\tilde{\theta} \sim N_n(X_o\tau, (1/\rho)W_oW_o')$. If, indeed, the response curve is approximately linear, the preceding prior on θ is more realistic than the prior,

$$\tilde{\theta} \sim N_n(0, (1/\rho)W_oW_o'),$$

of the intermediate model studied in Chapter 2. Recall that the intermediate model resulted in sharp estimates of θ (relative to the m.l.e.) over a large class of response curves. In what follows,

$$V_o = W_o W_o' = \begin{pmatrix} 1 & & & & 1 \\ & 2 & & & 2 \\ & & 3 & & 3 \\ & & & & \\ 1 & 2 & 3 & & n \end{pmatrix} . \qquad (3.3)$$

The experimental design and the structure of V_o impose the following "smoothness condition" on the response curve:

$$\text{Corr}(\tilde{\theta}_j, \tilde{\theta}_k) = \sqrt{j/k}, \quad k \geq j .$$

Note, also, that $\text{Var}(\tilde{\theta}_k) = k/\rho$, $k = 1, 2, \ldots, n$. This introduces prior knowledge of the following form: Our knowledge of a natural origin $(\theta_o = 0)$ is accompanied by increasing lack of knowledge regarding the θ's as we successively increase the level of stress.

As in Chapter 2, the following notational adjustments are made to accommodate the particular experimental situation of Figure 1.1. Reindex the parameters so that θ_k is the mean response at $X_k = 1.5 + (k-1)(.5)$, $k = 1, \ldots, 8$. Then $\text{cov}(\tilde{\theta}) = (1/\rho)V$ where V is the lower right submatrix of V_o (3.3). With these adjustments,

$$\tilde{\theta} = \begin{pmatrix} 1.5 \\ 2.0 \\ 2.5 \\ \\ \\ 5.0 \end{pmatrix} \tau + \underset{8 \times 10}{\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & & & & \vdots \\ \vdots & & & & & & & & & \vdots \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}} \begin{pmatrix} \tilde{\epsilon}_1 \\ \tilde{\epsilon}_2 \\ \\ \vdots \\ \\ \tilde{\epsilon}_{10} \end{pmatrix}$$

or $\tilde{\theta} = X\tau + W\tilde{\epsilon}$. Assuming that $\tilde{\epsilon} \sim N_{10}(0, (1/\rho)I)$, then

$\tilde{\theta} \sim N_8(\mu, (1/\rho)WW')$ where $\mu = X\tau$ and

$$V = WW' = \begin{pmatrix} 3 & 3 & \text{———} & 3 \\ | & 4 & \text{———} & 4 \\ | & | & 5 & \text{———} & 5 \\ | & | & | & & \\ 3 & 4 & 5 & 6 & 10 \end{pmatrix} . \qquad (3.4)$$

From Result 1.2, page 6, the marginal distribution of $\tilde{y}$ is $N_8(\mu, (1/\rho)V+I)$; i.e.,

$$f_{\tilde{y}|\rho,\mu}(y) = \int f_{\tilde{y}|\theta}(y) f_{\tilde{\theta}}(\theta) d\theta$$

$$\propto (\det((1/\rho)V+I))^{-1/2} \exp((-1/2)(y-\mu)'((1/\rho)V+I)^{-1}(y-\mu)) .$$

Since

$$\hat{\theta}(\rho,\tau) = E(\tilde{\theta}|y,\rho,\tau) = ((1/\rho)V+I)^{-1}(X\tau) + ((1/\rho)V+I)^{-1}(1/\rho)Vy$$

(see (2.7)), our immediate aim is to estimate ρ and τ. Having done so, $\hat{\theta}(\hat{\rho},\hat{\tau})$ is stated as the estimate to be taken for θ.

Distributional Assumptions on Hyperparameters

Two approaches to the problem of estimating (ρ,τ) are discussed below. In each case, the strategy suggested in Chapter 2 is followed to the extent possible. We will combine the marginal distribution of $\tilde{y}$ (which depends upon a vector of hyperparameters, P) with a prior $f_{\tilde{P}}(P)$ to obtain a posterior

$$f_{\tilde{P}|y}(P) \propto f_{\tilde{y}|P}(y) f_{\tilde{P}}(P) .$$

The mode of $f_{\tilde{P}|y}(P)$ is then taken as the estimate of P. In Chapter 2 ($P = \rho$) it was seen that excellent results were obtained by writing a noninformative prior for ρ. The likelihood function was

$$f_{\tilde{y}|\rho}(y) \propto \left(\frac{\rho}{1+\rho}\right)^{n/2} \exp\left((-1/2)\left(\frac{\rho}{1+\rho}\right) y'y\right) . \tag{3.5}$$

Compare (3.5) with a likelihood function of the present chapter (derived below):

$$f_{\tilde{w}|\rho}(w) \propto \prod_{i=1}^{n-1} \{(\rho/(\lambda_i+\rho))^{1/2} \exp((-1/2)(\rho/(\lambda_i+\rho)) w_i^2)\} \tag{3.6}$$

Although certain approximations are possible (see Box and Tiao (1972)), it is not clear to the author how one would arrive at a metric $\phi(\rho)$ in which (3.6) is data translatable. However, within the context of the model where ρ was defined--i.e.,

$$\tilde{\theta}_k - \tilde{\theta}_{k-1} = \tau + \tilde{\epsilon}_k; \quad k = 1, \ldots, n, \quad \tilde{\theta}_o = 0,$$

$$\tilde{\epsilon}|\rho \sim N_n(0, (1/\rho)I) ,$$

a prior distribution $f_{\tilde{\rho}}(\rho)$ can be assigned which is noninformative in a sense. Suppose that Nature generates θ vectors according to the above process. The statistician--having little knowledge of how smooth the response curve is--may feel that

$$f_{\tilde{\rho}}(\rho) \propto (1/\rho)1_{(0,\infty)}(\rho) \tag{3.7}$$

is an appropriate prior for ρ. This prior distribution, (3.7), is "noninformative" with respect to the distribution of $\tilde{\epsilon}|\rho$. The author is persuaded that noninformative priors (in the sense of Box and Tiao) or the more usual vague priors ((3.7) for example) should be taken on hyperparameters such as ρ. However, in a full Bayesian analysis

$$f_{\tilde{\rho}}(\rho) \propto \rho^{(\alpha-1)}\exp(-\rho/\beta)1_{(0,\infty)}(\rho)$$

is probably the first prior on ρ that one would consider. The Bayesian might then examine the experimental data together with estimates of θ resulting from a variety of choices of α and β. Both points of view (regarding ρ) are illustrated in this chapter.

The improper uniform prior on the real line is henceforth assigned as the distribution of $\tilde{\tau}$; i.e., $f_{\tilde{\tau}}(\tau) \propto 1_{\mathbb{R}^1}(\tau)$. We shall also assume that $\tilde{\rho}$ and $\tilde{\tau}$ are independent.

The Estimation Procedure

Returning to the regression problem--the likelihood function (the marginal density of $\tilde{y}$) is

$$\ell_{\tilde{\rho},\tilde{\tau}}(\rho,\tau|y) = f_{\tilde{y}|\rho,\tau}(y) \propto (\det((1/\rho)V+I))^{-1/2}$$
$$\times \exp((-1/2)(y-\mu)'((1/\rho)V+I)^{-1}(y-\mu)) ;$$

where $\mu = X\tau$; X and V are defined on page 43.

We first address the problem of estimating ρ. In this regard τ is a nuisance parameter and will be eliminated (from present consideration) by a data reduction based on location invariance. To facilitate discussion, let us state--

Proposition 3.1. Let $\mathcal{F}$ be the family of $N_n(X\underline{\beta}, \underline{\Sigma})$ distributions, $\underline{\Sigma} > 0$, where X is a known $n \times k$ matrix of rank $r < n$ and $\underline{\beta}$ is an unknown $k \times 1$ vector of parameters. Further let

(i) $S_{(n-r) \times n}$ be any matrix such that $\underline{R}(S') = \underline{R}(X)^{\perp}$; i.e., the range space of X' is equal to the orthogonal subspace of $\underline{R}(X)$;

(ii) $\mathcal{G} = \{g_{\mu^*} : g_{\mu^*}(y) = y+\mu^*, \ \mu^* \in \underline{R}(X)\}$. (It is easily seen that $\mathcal{G}$ is a translation group on the sample space.)

Then

(i) The family $\mathcal{F}$ is invariant under the group $\mathcal{G}$.

(ii) The statistic $\tilde{z} = S\tilde{y}$ is a maximal invariant under $\mathcal{G}$.

For a proof of the above proposition, consult Seely (1972). Dr. Seely used this idea to find estimators for two variance components (Model II ANOVA) which are location invariant. This general method was introduced by W.A. Thompson (1963). The above proposition, however, was formulated by Seely who approached the problem

(designated as restricted maximum likelihood estimation) with a quite general set of assumptions (compared with Thompson's).

In our example where $\mu = X\tau = (1.5, 2.0, 2.5, \ldots, 5.0)'\tau$, let $u^{(i)}$ $i = 1, \ldots, 8$, be the ith unit vector in $\mathbb{R}^8$. After applying the Gram-Schmidt process to the basis vectors

$$X,\ u^{(i)},\quad i = 2, \ldots, 8;$$

we have an orthonormal basis--

$$v_1, v_2, \ldots, v_8;\quad (v_1 = X/\|X\|).$$

Now, let $(v_2, v_3, \ldots, v_8)' = S'_{7 \times 8}$ and consider $\tilde{z} = S\tilde{y} \sim N_7(0, S((1/\rho)V+I)S')$ or

$$\tilde{z} \sim N_7(0, (1/\rho)SVS'+I) \quad (\text{since } SS' = I_{7 \times 7}). \tag{3.8}$$

In our Bayesian situation, a partial motivation for basing the estimate of ρ upon $\tilde{z}$ might be as follows. The assumed independence of $\tilde{\rho}$ and $\tilde{\tau}$ might suggest that our estimator of ρ should be based upon a statistic which is location invariant (the sufficient statistic for the family of (3.8) for example).

To expedite what follows, let n be an arbitrary positive integer; decompose the product-- $SVS' = R\Lambda R' = \sum_{i=1}^{n-1} \lambda_i E_i$ (where $R = (R_1, \ldots, R_{n-1})$, $RR' = R'R = I$, $E_i = R_i R_i'$, $\Lambda = \delta(\lambda_i)$, $S_{(n-1) \times n}$

has rank $n-1$, and S is constructed as above where

$SX = 0_{(n-1)\times 1}$) and record the following:

Proposition 3.2.

(i) $(\mathrm{cov}(\tilde{z}))^{-1} = ((1/\rho)SVS'+I)^{-1} = \sum_{i=1}^{n-1} (\frac{\rho}{\lambda_i+\rho})E_i$

(ii) $\det(\mathrm{cov}(\tilde{z})) = (1/\rho)^{n-1} \prod_{i=1}^{n-1} (\lambda_i+\rho)$.

Proof. Statement (i) was essentially shown in (2.23). Statement (ii) easily follows since

$$\det(\mathrm{cov}(\tilde{z})) = \det[\sum_{i=1}^{n-1} (\frac{\lambda_i+\rho}{\rho}) E_i] = \det[R\,\delta(\frac{\lambda_i+\rho}{\rho})\,R']$$

$$= \det\,\delta(\frac{\lambda_i+\rho}{\rho}) = \prod_{i=1}^{n-1} (\frac{\lambda_i+\rho}{\rho}).$$

Recalling that $\tilde{z} \sim N_7(0, (1/\rho)SVS'+I)$, the likelihood is

$$f_{\tilde{z}|\rho}(z) \propto [\prod_{i=1}^{7} (\frac{\rho}{\lambda_i+\rho})]^{1/2} \exp[(-\frac{1}{2})\rho \sum_{i=1}^{7} (\frac{1}{\lambda_i+\rho}) z'E_i z]$$

or letting $t_i^2 = z'E_i z = (z'R_i)(z'R_i)$ we have

$$f_{\tilde{t}|\rho}(t) \propto \prod_{i=1}^{7} \{(\frac{\rho}{\lambda_i+\rho})^{1/2} \exp[(-\frac{1}{2})(\frac{\rho}{\lambda_i+\rho})t_i^2]\}. \qquad (3.9)$$

We will estimate ρ with the mode of the posterior distribution

$$f_{\tilde{\rho}|t}(\rho) \propto \prod_{i=1}^{7} \{(\frac{\rho}{\lambda_i+\rho})^{1/2} \exp\{(-\frac{1}{2})(\frac{\rho}{\rho+\lambda_i})t_i^2\}\}(1/\rho)1_{(0,\infty)}(\rho) .$$

Let $L(\rho) = \ln f_{\tilde{\rho}|t}(\rho)$. Then

$$L(\rho) = (\text{const.}) + (5/2)\ln\rho - (1/2)\sum_{i=1}^{7} \ln(\lambda_i+\rho) - (1/2)\sum_{i=1}^{7} (\frac{\rho}{\lambda_i+\rho})t_i^2 .$$

$$dL/d\rho = (5/2)(1/\rho) - (1/2)\sum_{i=1}^{7} (\frac{1}{\lambda_i+\rho}) - (1/2)\sum_{i=1}^{7} (\frac{\lambda_i}{(\rho+\lambda_i)^2})t_i^2$$

$$= (5/2)(1/\rho) - \sum_{i=1}^{7} \frac{\lambda_i(1+t_i^2)+\rho}{2(\lambda_i+\rho)^2} .$$

Our estimate (denoted $\hat{\rho}$) is found among the positive zeros of $\frac{dL}{d\rho} = 0$ which implies

$$5 - \rho\sum_{i=1}^{7} \frac{\lambda_i(1+t_i^2)+\rho}{(\lambda_i+\rho)^2} = 0.$$

The posterior density $f_{\tilde{\rho}|t}(\rho)$ (corresponding to the data set of Figure 1.1) is proportional to the curve shown in Figure 3.1.

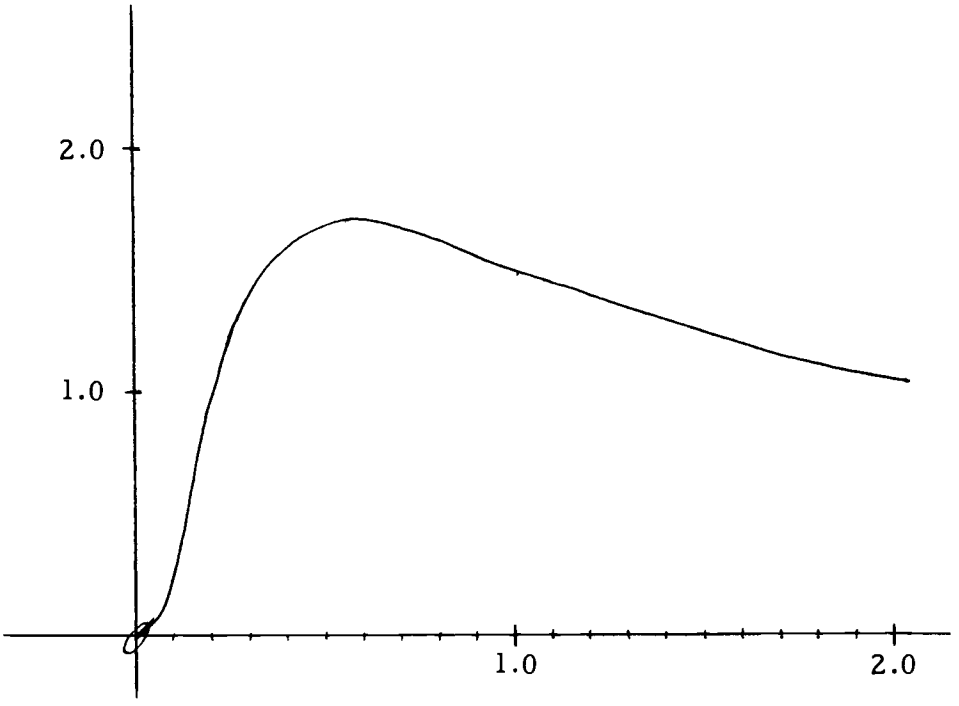


Figure 3.1.

After solving (3.12) for $\hat{\rho}$, we write

$$\hat{\tau} = \frac{X'[(1/\hat{\rho})V+I]^{-1}y}{X'[(1/\hat{\rho})V+I]^{-1}X} \tag{3.10}$$

as the estimate taken for τ. Note that $\hat{\tau}$ is an approximation to the Gauss-Markov estimate for τ within the framework

$$\tilde{y}\Big|_{\rho,\tau} \sim N_8(X\tau, (1/\rho)V+I) \ . \tag{3.11}$$

Figure 3.2 shows $\hat{\theta}(\hat{\rho},\hat{\tau}) = ((1/\hat{\rho})V+I)^{-1}(X\hat{\tau}) + ((1/\hat{\rho})V+I)^{-1}(1/\hat{\rho})Vy$

where the components of y are the experimental observations of

Figure 1.1. Also shown in Figure 3.2 is an estimate for θ (denoted $\hat{\theta}(\rho^*, \tau^*)$) arrived at by another means. The calculation for (ρ^*, τ^*) is outlined in the next paragraph.

Another procedure for estimating (ρ, τ) involves a two dimensional gradient search. Again, consider the marginal density of $\tilde{y}$; i.e.,

$$f_{\tilde{y}|\rho,\tau}(y) \propto (\det((1/\rho)V+I))^{-1/2}\exp((-1/2)(y-X\tau)'((1/\rho)V+I)^{-1}(y-X\tau)) .$$

Let R be the orthogonal matrix specified by $R'VR = \Lambda = \delta(\lambda_i)$ $i = 1, \ldots, 8$ and $\lambda_1 > \lambda_2 > \ldots > \lambda_8$. The density of $\tilde{u} = R'\tilde{y}$ can be written as

$$f_{\tilde{u}|\rho,\tau}(u) \propto \left[\prod_{i=1}^{8}\left(\frac{\lambda_i+\rho}{\rho}\right)\right]^{-1/2}\exp\left[\left(-\frac{1}{2}\right)(u-W\tau)'\delta\left(\frac{\rho}{\lambda_i+\rho}\right)(u-W\tau)\right] \quad (3.12)$$

where $W = R'X$. The prior $f_{\tilde{\rho},\tilde{\tau}}(\rho, \tau) \propto (1/\rho)1_{[\rho \geq 0]}(\rho, \tau)$ combines with (3.12) to yield the posterior--

$$f_{\tilde{\rho},\tilde{\tau}|u}(\rho, \tau) \propto (1/\rho)\left[\prod_{i=1}^{8}\left(\frac{\lambda_i+\rho}{\rho}\right)\right]^{-1/2}\exp\left[\left(-\frac{1}{2}\right)(u-W\tau)'\delta\left(\frac{\rho}{\lambda_i+\rho}\right)(u-W\tau)\right]$$

$$\times 1_{[\rho \geq 0]}(\rho, \tau) .$$

Let $L(\rho, \tau) = \ln f_{\tilde{\rho},\tilde{\tau}|u}(\rho, \tau)$. The calculations for the gradient of L, $\nabla L = (L_\rho, L_\tau)'$, are routine and we merely state the partial derivatives:

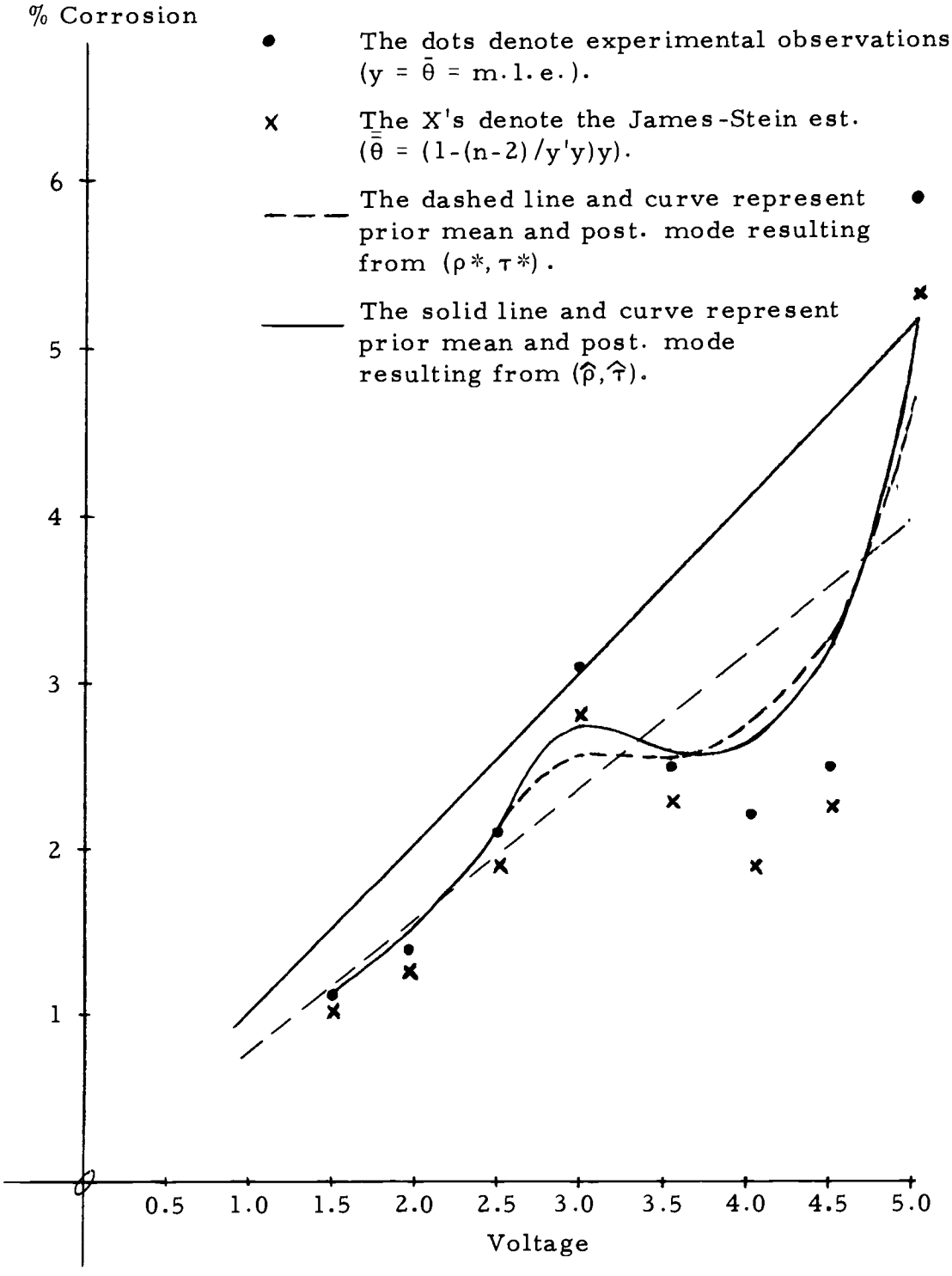

% Corrosion
The dots denote experimental observations (y = θ̄ = m.l.e.).
The X's denote the James-Stein est. (θ̄ = (1-(n-2)/y'y)y).
The dashed line and curve represent prior mean and post. mode resulting from (ρ*, τ*).
The solid line and curve represent prior mean and post. mode resulting from (ρ̂, τ̂).
6
5
4
3
2
1
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Voltage

Figure. 3.2.

$$\partial L/\partial \rho = 3/\rho - (1/2) \sum_{i=1}^{8} \{ \frac{(\lambda_i+\rho)+\lambda_i(u_i - W_i\tau)^2}{(\lambda_i+\rho)^2} \}$$

$$\partial L/\partial \tau = \sum_{i=1}^{8} W_i (\frac{\rho}{\lambda_i+\rho})(u_i - \tau W_i) \ .$$

Specific details pertinent to the gradient--search are outlined in Chapter 5 where the same technique is used in a more general setting.

Let us discuss the disparity between the two Bayesian estimates depicted in Figure 3.2. Where location invariance was used to estimate ρ ($\hat{\rho} = .625$, $\hat{\tau} = 1.01$), the resulting estimate of θ (solid curve) more closely resembles the data than does the dashed curve. The dashed curve depicts the estimate of θ which resulted from a gradient--search done to estimate (ρ, τ) $((\rho^*, \tau^*) = (1.2, .89))$. The estimated precision ρ^* is nearly double the estimated precision $\hat{\rho}$. To calculate $\hat{\rho}$ we used a distribution

$$\tilde{z} | \rho \sim N_7(0, (1/\rho)SVS'+I) \ ,$$

which expressed our willingness to surrender sufficient information about the original multinormal mean $X\tau$. (We started with $\tilde{y}\big|_{\rho, \tau} \sim N_8(X\tau, (1/\rho)V+I)$.) For example, we sacrificed that portion of the data which gave information about τ being positive. This is reminiscent of the James-Stein estimate (also shown in Figure 3.2)

where the likelihood defined by

$$\tilde{y} \mid \pi \sim N_8(0, \pi I), \quad \pi = \frac{\rho}{1+\rho} ,$$

was used to estimate π. The James-Stein estimate can be interpreted in terms of a posterior estimate of π when a noninformative prior on ρ is applied to the likelihood function. The two Bayesian estimates (discussed above) are written in terms of posterior estimates of hyperparameters obtained by applying vague priors (in the usual sense) to appropriate likelihood functions.

From a more subjective point of view, let us examine the situation when the likelihood function of (3.9),

$$f_{\tilde{t} \mid \rho}(t) \propto \prod_{i=1}^{7} \{(\frac{\rho}{\lambda_i+\rho})^{(1/2)} \exp((-1/2)(\frac{\rho}{\lambda_i+\rho})t_i^2)\},$$

is combined with the proper gamma prior

$$f_{\tilde{\rho}}(\rho) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \rho^{(\alpha-1)} \exp(-\rho/\beta) 1_{(0,\infty)}(\rho)$$

to give the posterior

$$f_{\tilde{\rho} \mid t}(\rho) \propto \rho^{7/2+(\alpha-1)} \Big[\prod_{i=1}^{7} (\lambda_i+\rho)\Big]^{-1/2} \exp\Big[(-\frac{1}{2})\rho \sum_{i=1}^{7} (\frac{1}{\lambda_i+\rho})t_i^2 - \rho/\beta\Big] 1_{(0,\infty)}(\rho) .$$

Let $\hat{\rho} = \hat{\rho}(\alpha, \beta, t)$ (the estimate taken for ρ) be the mode of $f_{\tilde{\rho} \mid t}$.

As before, let us use the marginal distribution of $\tilde{y}$,

$$\tilde{y}\Big|_{\rho,\tau} \sim N_8(X\tau, (1/\rho)V+I) ,$$

to estimate τ. After calculating $\hat{\rho}(\alpha, \beta, t)$, estimate τ with

$$\hat{\tau} = \frac{X'((1/\hat{\rho})V+I)^{-1}y}{X'((1/\hat{\rho})V+I)^{-1}X} \qquad \text{(see (3.10))}.$$

Figure 3.3 shows estimates of θ,

$$\hat{\theta}(\alpha, \beta) = ((1/\hat{\rho})V+I)^{-1}X\hat{\tau} + ((1/\hat{\rho})V+I)^{-1}(1/\hat{\rho})Vy \quad \text{(see (2.6))},$$

for several pairs (α, β). In each case we have taken $E\tilde{\rho} = \alpha\beta = 2$ (recall that $\operatorname{cov} \tilde{y}|\theta = I$).

It is seen that this procedure allows the experimenter much flexibility if he has definite prior opinions regarding the smoothness of the response curve.

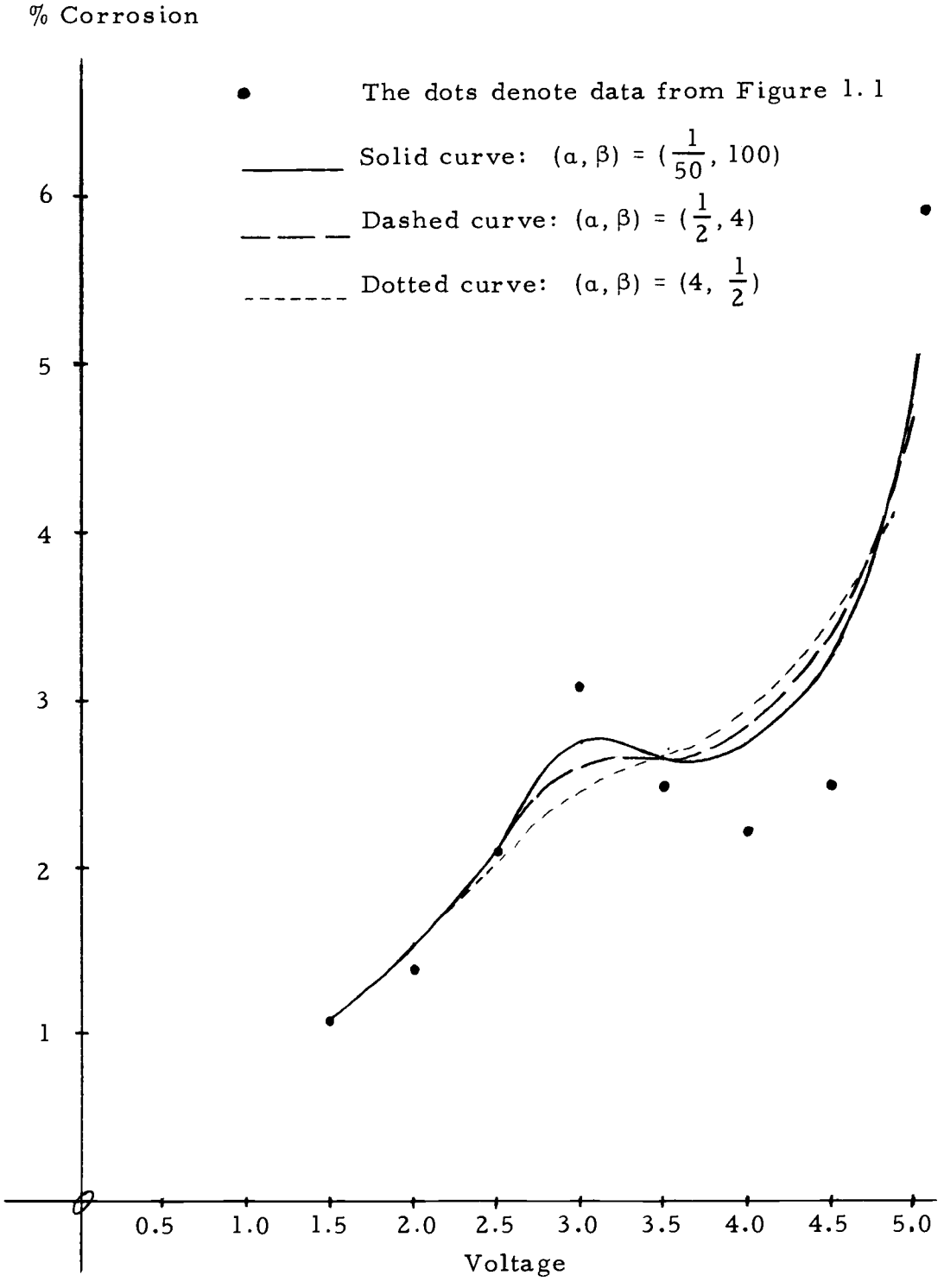

% Corrosion
The dots denote data from Figure 1.1
Solid curve: $(\alpha, \beta) = (\frac{1}{50}, 100)$
Dashed curve: $(\alpha, \beta) = (\frac{1}{2}, 4)$
Dotted curve: $(\alpha, \beta) = (4, \frac{1}{2})$
6
5
4
3
2
1
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Voltage

Figure 3.3.

IV. GENERAL AUTOREGRESSIVE PRIORS

Autoregressive and Moving Average Models

Included among the more important models in the standard theory of time series is

$$\tilde{y}_t + \alpha_1 \tilde{y}_{t-1} + \ldots + \alpha_p \tilde{y}_{t-p} = \tilde{\epsilon}_t \qquad (4.1)$$

$t = p+1, \ldots,$ and

$$\tilde{y}_t = \beta_o \tilde{\epsilon}_t + \beta_1 \tilde{\epsilon}_{t-1} + \ldots + \beta_q \tilde{\epsilon}_{t-q}, \qquad (4.2)$$

where, in each case, the $\tilde{\epsilon}_t$ are i.i.d. .

Equation (4.1) describes the autoregressive (AR) process or the stochastic difference equation of order p. The general moving average (MA) model of order q is stated in (4.2). The following notation will be helpful; let us define--

(i) The forward operator φ :

$$\varphi \tilde{y}_t = \tilde{y}_{t+1}$$

(ii) The difference operator Δ:

$$\Delta \tilde{y}_t = \tilde{y}_{t+1} - \tilde{y}_t .$$

($\Delta = \varphi - 1$ since $\varphi \tilde{y}_t - \tilde{y}_t = \tilde{y}_{t+1} - \tilde{y}_t = \Delta \tilde{y}_t$.)

(iii) The backwards operator $\mathcal{B}$:

$$\mathcal{B}\tilde{y}_t = \tilde{y}_{t-1} .$$

(iv) Indicator function on the set N of nonnegative integers:

$$1_N(j) = \begin{cases} 1 & j \in N \\ 0 & j \notin N \end{cases} .$$

In this notation, (4.1) becomes

$$\left(\sum_{r=0}^{p} a_r \mathcal{P}^{p-r} \right) \tilde{y}_{t-p} = \tilde{\epsilon}_t \qquad (4.3)$$

with associated polynomial equation

$$\sum_{r=0}^{p} a_r X^{p-r} = 0 , \qquad (4.3')$$

and (4.2) becomes

$$\left(\sum_{r=0}^{q} \beta_r \mathcal{B}^r \right) \tilde{\epsilon}_t = \tilde{y}_t \qquad (4.4)$$

with associated polynomial equation

$$\sum_{r=0}^{q} \beta_r X^r = 0 . \qquad (4.4')$$

The autoregressive model is commonly used to predict an event y_t given past events $y_{t-1}, \cdots$. An important idea in this regard is that when--

(i) the stochastic process satisfies (4.3), and

(ii) the roots of (4.3') are strictly less than one in absolute value, then $E(\tilde{y}_t \mid y_{t-1}, y_{t-2}, \cdots) = -\beta_1 y_{t-1} - \cdots - \beta_p y_{t-p}$ predicts y_t with minimum mean square error. (See Anderson (1970).)

Throughout this paper we have been abusing standard terminology a bit by referring to

$$\left[\sum_{r=0}^{p} (a_r 1_N(t-r))\rho^{p-r}\right]\tilde{\theta}_{t-p} = \tilde{\epsilon}_t, \quad a_o = 1, \ \tilde{\epsilon}_t \text{ i.i.d.}, \tag{4.5}$$

$\tilde{\theta}_o = 0$, $t = 1, 2, \ldots,$ as "the general autoregressive model of order p." Equation (4.5) states:

$$\begin{aligned}
\tilde{\theta}_1 &= \tilde{\theta}_o + \tilde{\epsilon}_1, \quad \tilde{\theta}_o = 0 \\
\tilde{\theta}_2 &= a_1\tilde{\theta}_1 + \tilde{\epsilon}_2 \\
&\vdots \\
\tilde{\theta}_k &= a_1\tilde{\theta}_{k-1} + \ldots + a_p\tilde{\theta}_{k-p} + \tilde{\epsilon}_k, \quad k \geq p .
\end{aligned}$$

In the example of Chapter 3, our prior on the θ's was a special case of (4.5); there we had $p = 1$ and $a_1 = 1$.

The model of (4.5) relates to a "moving averages" model of the

form

$$\tilde{\theta}_t = \sum_{r=0}^{q} (\beta_r 1_N(t-q))\tilde{\epsilon}_{t-r}, \quad \tilde{\epsilon}_k \text{ i.i.d.}, \tag{4.6}$$

in a manner analogous to the way that the AR model of (4.3) relates to an appropriate MA model of the form (4.4). To be precise, let us state a standard result from time series analysis and look at an analogue to the standard result.

Standard Result From Time Series Analysis

Theorem 4.1. Assume that (4.3) holds and that the roots of (4.3') are strictly less than one in absolute value, then there exists β_r, $r = 0, 1, \ldots$ such that $\tilde{y}_t = \sum_{r=0}^{\infty} \beta_r \tilde{\epsilon}_{t-r}$ with equality holding in the sense

$$E(\tilde{y}_t - \sum_{r=0}^{s} \beta_r \tilde{\epsilon}_{t-r})^2 \to 0 \quad \text{as} \quad s \to \infty.$$

Proof. See Anderson (1972) for a proof of this theorem along with the statement and proof of a similar result which holds when one starts with a moving averages model (of finite order) and wishes to write an autoregressive model.

Theorem 4.2. Let $\{a_n\}_{n=1}^{\infty}$ be an arbitrary sequence of real

numbers except that $a_1 \neq 0$. Define

$$T_n = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ a_2 & a_1 & & 0 \\ a_3 & a_2 & a_1 & 0 \\ \vdots & \ddots & \ddots & \ddots \\ a_n & a_{n-1} & \cdots & a_1 \end{pmatrix} .$$

Then there exists a sequence of real numbers $\{b_n\}_{n=1}^{\infty}$ such that

$$T_n^{-1} = \begin{pmatrix} b_1 & 0 & & 0 \\ b_2 & b_1 & 0 & 0 \\ b_3 & b_2 & b_1 & 0 \\ \vdots & \ddots & \ddots & \ddots \\ b_n & b_{n-1} & \cdots & b_1 \end{pmatrix} .$$

Proof. First, T_n^{-1} is lower triangular since it is well known that the inverse of a lower triangular matrix is a lower triangular matrix. The fact that T_n^{-1} has the special structure of T_n follows by induction on n.

(i) $T_1^{-1} = (1/a_1)$

$T_2^{-1} = (1/a_1)^2 \begin{pmatrix} a_1 & 0 \\ -a_2 & a_1 \end{pmatrix}$. Note that $b_1 = 1/a_1$.

(ii) Suppose that T_{n-1}^{-1} is a lower triangular matrix having the property that all elements on any given minor diagonal are equal; i.e.,

$$T_{n-1}^{-1} = \begin{pmatrix} b_1 & 0 & & 0 \\ b_2 & b_1 & 0 & 0 \\ & \ddots & \ddots & \\ b_{n-1} & & & b_1 \end{pmatrix}$$

(iii) Consider

$$T_n = \left(\begin{array}{cccc|c} a_1 & 0 & & 0 & 0 \\ a_2 & a_1 & & 0 & 0 \\ & & & & \\ a_{n-1} & & & a_1 & 0 \\ \hline a_n & a_{n-1} & \cdots & a_2 & a_1 \end{array}\right) = \begin{pmatrix} T_{n-1} & O_{(n-1)\times 1} \\ P_{1\times(n-1)} & a_1 \end{pmatrix}$$

Define $Q = (q_1, q_2, \ldots, q_{n-1})$ such that

$$q_1 = -(1/a_1)(a_n b_1 + \ldots + a_2 b_{n-1})$$

$$q_k = b_{n-(k-1)}, \quad k = 2, \ldots, n-1.$$

The claim is that

$$\begin{pmatrix} T_{n-1}^{-1} & 0 \\ 0 & 1/a_1 \end{pmatrix} = T_n^{-1}.$$

Since

$$\begin{pmatrix} T_{n-1} & 0 \\ P & a_1 \end{pmatrix} \begin{pmatrix} T_{n-1}^{-1} & 0 \\ Q & 1/a_1 \end{pmatrix} = \begin{pmatrix} I & 0 \\ PT_{n-1}^{-1} + a_1 Q & 1 \end{pmatrix},$$

the proof is completed by showing that $PT_{n-1}^{-1} + a_1 Q = 0$. The first entry of $PT_{n-1}^{-1} + a_1 Q$ is seen to be zero by the definition of q_1.

All other components are seen to be zero since we have

$$a_{n-(k-1)}b_1 + a_{n-k}b_2 + \ldots + a_2 b_{n-k} + a_1 b_{n-(k-1)} = 0,$$

$$k = 2, \ldots, n-1,$$

by inductive hypothesis and $q_k = b_{n-(k-1)}$ be definition. This implies $a_{n-(k-1)}b_1 + a_{n-k}b_2 + \ldots + a_2 b_{n-k} = -a_1 q_k$. Therefore we have $PT_{n-1}^{-1} + a_1 Q = 0$.

An Analogue to the Standard Result (Theorem (4.1))

Corollary 4.1. Let $\mu_1, \ldots, \mu_n$ be arbitrary real numbers, also $\beta_1, \ldots, \beta_q$, such that

$$\begin{aligned}
\tilde{\theta}_1 &= \mu_1 + \tilde{\epsilon}_1 \\
\tilde{\theta}_2 &= \mu_2 + \beta_1 \tilde{\epsilon}_1 + \tilde{\epsilon}_2 \\
&\vdots \\
\tilde{\theta}_n &= \mu_n + \sum_{i=0}^{q} \beta_i \tilde{\epsilon}_{n-i}
\end{aligned}$$

$$\tilde{\epsilon}_t \text{ i.i.d.}, \quad E\tilde{\epsilon}_t = 0; \quad \beta_o = 1,$$

or in matrix notation, $\tilde{\theta} - \mu = T\tilde{\epsilon}$. Then there exists $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that

$$\tilde{\epsilon}_1 = \tilde{\theta}_1 - \mu_1$$
$$\tilde{\epsilon}_2 = \alpha_1(\tilde{\theta}_1 - \mu_1) + (\tilde{\theta}_2 - \mu_2)$$
$$\tilde{\epsilon}_3 = \alpha_2(\tilde{\theta}_1 - \mu_1) + \alpha_2(\tilde{\theta}_2 - \mu_2) + (\tilde{\theta}_3 - \mu_3)$$
$$\vdots$$

i.e., the parameters minus theor prior means are exactly represented as an autoregressive model.

Proof. The proof is immediate from the fact that T^{-1} has the same structure as T.

If we start with an AR model of the form (4.5), a converse statement to Corollary 4.1 is obvious.

Autoregressive Priors in Regression Situations

The chapter is concluded with a discussion of the applicability of autoregressive priors in ordinary regression situations.

Let us start by assuming that the components of our multinormal mean satisfy the AR process of (4.3); i.e.,

$$\left(\sum_{r=0}^{p} \alpha_i \rho^{p-r}\right)\tilde{\theta}_{t-p} = \tilde{\epsilon}_t, \quad \tilde{\epsilon}_t \text{ i.i.d.}, \quad t = p+1, \ldots . \tag{4.9}$$

Then, let us specify the set $\{\alpha_i\}_{i=1}^{p}$ in such a way that the α_i have the fortunate property of defining an equation--

$$\sum_{r=0}^{p} a_r X^{p-r} = 0$$

whose roots are strictly less than one in absolute value. It can be seen that the V matrix (where $\mathrm{cov}(\tilde{\theta}) = (1/\rho)V$) is specified by solutions of the Yule-Walker equations. Box and Jenkins (1971) present a readable account of this technique within the usual context of time series. That is, where observable random variables $\tilde{y}_{t-p}$ appear in (4.9) in place of the unobservable $\tilde{\theta}_{t-p}$.

The Bayesian might be hard pressed to specify a set $\{a_i\}_{i=1}^{p}$ having the property referred to above in any meaningful way. He would probably try to estimate the a's. In the standard discussion of time series, the a_i are estimated either by variations on the Yule-Walker procedure or by maximum likelihood calculations (where other special assumptions are required). Because of the unique features of our Bayesian approach, neither of these methods is applicable. The first approach is impossible since the θ's are not observed. We have major technical problems imitating the standard maximum likelihood approach. This is due to the relatively complicated error structure,

$$(1/\rho)V+I \quad (\text{where} \quad \mathrm{cov}(\tilde{\theta}) = (1/\rho)V) ,$$

in the marginal distribution of $\tilde{y}$. Specifically, it is not clear at all

how one might write the V matrix in manageable terms. See Anderson (1972), page 183, for a discussion of maximum likelihood estimation within the usual context of time series.

With the general moving average formulation

$$(\sum_{r=0}^{q} \beta_r \mathcal{B}^r)\tilde{\epsilon}_t = \tilde{\theta}_t, \quad \tilde{\epsilon}_t \text{ i.i.d.}, \quad E\tilde{\epsilon}_t = 0, \quad \text{var}\,\tilde{\epsilon}_t = 1/\rho \tag{4.10}$$

there are no technical problems in computing the covariance structure. From

$$\tilde{\theta}_r = \tilde{\epsilon}_r + \beta_1 \tilde{\epsilon}_{r-1} + \ldots + \beta_{r-s}\tilde{\epsilon}_s + \ldots + \beta_q \tilde{\epsilon}_{r-q}$$

$$\tilde{\theta}_s = \tilde{\epsilon}_s + \beta_1 \tilde{\epsilon}_{s-1} + \ldots + \beta_q \tilde{\epsilon}_{s-q}$$

$$\text{with} \quad r > s > q$$

we have

$$\text{var}(\tilde{\theta}_t) = (1 + \sum_{i=1}^{q} \beta_i^2)(1/\rho), \quad t > q,$$

$$\text{cov}(\tilde{\theta}_r, \tilde{\theta}_s) = \beta_{r-s}\beta_o + \beta_{r-s+1}\beta_1 + \ldots + \beta_q \beta_{q-(r-s)} .$$

In a given application, the V matrix depends upon the unknown β's. These hyperparameters can be estimated using the apparatus developed in the next chapter.

We next discuss an example where a moving average model of

the form (4.10) is used. Beyond this, the concept of MA priors is not pursued. The example is cited because the covariance structure involved (on the θ's) would seem appropriate in many regression situations. Also, the particular MA prior is easily handled within our Bayesian structure,

$$\tilde{y} \mid \theta \sim N_n(\theta, I), \quad \tilde{\theta} \sim N_n(\mu, (1/\rho)V).$$

To introduce the example, let us consider the following covariance structure:

$$\operatorname{cov}(\tilde{\theta}) = (1/\rho)\begin{pmatrix} 1 & 1-\gamma & 1-2\gamma & & 1-(n-1)\gamma \\ 1-\gamma & 1 & 1-\gamma & & \\ & & & & 1-\gamma \\ 1-(n-1)\gamma & \cdots & & 1-\gamma & 1 \end{pmatrix} \tag{4.11}$$

where $0 < (n-1)\gamma \leq 2$. For convenience, assume that $E\tilde{\theta} = 0$. We have $\operatorname{corr}(\tilde{\theta}_j, \tilde{\theta}_k) = 1 - (k-j)$ for $k \geqslant j$. The structure of (4.11) would seem appropriate in many regression situations. It can be generated by the MA model

$$\tilde{\theta}_t = \sum_{j=0}^{q} \frac{1}{\sqrt{q+1}}\, \tilde{\epsilon}_{t-j}, \quad \tilde{\epsilon}_t \text{ i.i.d.}, \quad t = 1, \ldots, ;$$

i.e.,

$$\theta = \sqrt{\frac{1}{q+1}} \begin{pmatrix} \epsilon_{1-q} & & & \epsilon_0 & \epsilon_1 & & \epsilon_n \\ 1 & & 1 & 1 & 1 & & \\ & 1 & 1 & & 1 & 1 & \\ 0 & & \searrow & 1 & & \searrow & 0 \\ & & & & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \epsilon_{1-q} \\ \epsilon_{2-q} \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{4.12}$$

Suppose that $E\tilde{\epsilon}_t = 0$ and $\text{var}\,\tilde{\epsilon}_t = 1/\rho \quad \forall\, t$. In view of (4.11) we easily have

$$\begin{aligned} &\text{var}(\tilde{\theta}_t) = (1/\rho) \quad \forall\, t \\ &\text{cov}(\tilde{\theta}_r, \tilde{\theta}_s) = \begin{cases} 1-(r-s)\gamma & r-s < q+1 \\ 0 & r-s \geq q+1 \end{cases} \end{aligned} \tag{4.13}$$

where $r \geq s$, and $\gamma = 1/(q+1)$.

In this example we might think of each θ_t as being the sum of $q+1$ independent and identically distributed "jolts"--the earliest contribution being ϵ_{t-q}; the latest, ϵ_t. Let us indicate a procedure for estimating q and the other hyperparameter ρ. Again suppose that

$$\tilde{y} \mid \theta \sim N_n(\theta, I) \quad \text{and} \quad \tilde{\theta} \sim N_n(0, (1/\rho)V)$$

with V specified according to (4.11). The marginal distribution of $\tilde{y}$ is-- $\tilde{y}\big|_{\rho, q} \sim N_n(0, (1/\rho)V+I)$. As a prior on the hyperparameters, take

$$f_{\tilde{\rho},\tilde{q}}(\rho, q) \propto (1/\rho)1_{(0,\infty)}(\rho)1_N(q)$$

where N is the set of nonegative integers. Then

$$f_{\tilde{\rho},\tilde{q}|y}(\rho, q) \propto [\det((1/\rho)V+I)]^{-1/2} \times \exp[(-1/2)y'((1/\rho)V+I)^{-1}y](1/\rho)1_{(0,\infty)}(\rho)1_N(q) .$$

We can maximize the posterior (over ρ) for a given nonnegative integer q according to methods discussed in the last chapter. If we do so for each $q = 0, 1, 2, \ldots,$ let $\hat{\rho}_o, \hat{\rho}_1, \ldots, \hat{\rho}_k, \ldots$ be the sequence of critical points. Assuming the posterior density is maximized at the point $\hat{\rho}_{n*}$, we write the modal estimate for (ρ, q) as $(\hat{\rho}, n*)$.

Earlier we noted some of the technical problems which arise when we assume that the components of θ satisfy the stochastic difference equation

$$(\sum_{r=0}^{p} a_r \mathcal{P}^{p-r})\tilde{\theta}_{t-p} = \tilde{\epsilon}_t, \quad \tilde{\epsilon}_t \text{ i.i.d.} .$$

Accordingly, for the remainder of this paper we restrict attention to autoregressive priors of the form--

$$(\sum_{r=0}^{p} (a_r 1_N(t-r))\mathcal{P}^{p-r})\tilde{\theta}_{t-p} = \tilde{\epsilon}_t, \quad \tilde{\epsilon}_t \text{ i.i.d.} \tag{4.14}$$

$\theta_o = 0, \quad t = 1, 2, \ldots, n.$

Recall Corollary 4.1 with regards to the model (4.14) and the computation of $\mathrm{cov}(\tilde{\theta}) = (1/\rho)V$. Also, it is seen in the next chapter that a model of the form (4.14) yields an expression for $((1/\rho)V+I)^{-1}$ which is easy to write down.

As a further example, let us consider the special case of (4.14) where $p = 1$, $t = 1, \ldots, n$. We have--

$$\begin{aligned} \tilde{\theta}_1 &= \tilde{\epsilon}_1 \\ \tilde{\theta}_2 &= a\tilde{\theta}_1 + \tilde{\epsilon}_2 \\ &\vdots \\ \tilde{\theta}_n &= a\tilde{\theta}_{n-1} + \tilde{\epsilon}_n \end{aligned} \quad \text{or} \quad \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -a & 1 & 0 & \cdots & 0 \\ 0 & -a & 1 & \cdots & 0 \\ & & & \ddots & \\ 0 & & & -a & 1 \end{pmatrix} \begin{pmatrix} \tilde{\theta}_1 \\ \tilde{\theta}_2 \\ \vdots \\ \tilde{\theta}_n \end{pmatrix} = \begin{pmatrix} \tilde{\epsilon}_1 \\ \tilde{\epsilon}_2 \\ \vdots \\ \tilde{\epsilon}_n \end{pmatrix}$$

Let us write this as $W\tilde{\theta} = \tilde{\epsilon}$. Then where $\tilde{y}|\theta \sim N_n(\theta, I)$ and $\tilde{\epsilon} \sim N_n(0, (1/\rho)I)$, the marginal distribution of $\tilde{y}$ is

$$\tilde{y}\Big|_{a,\rho} \sim N_n(0, (1/\rho)(W'W)^{-1}+I) .$$

In Chapter 5 we consider general procedures for estimating the autoregressive coefficients. However, simple estimates are available in the present situation which do not require the apparatus developed for the more general model. For example, suppose that n is an even integer and define

$$W_e = \begin{pmatrix} -a & 1 & 0 & \cdots & 0 \\ & -a & 1 & 0 & \cdots & 0 \\ & & -a & 1 & 0 & \cdots & 0 \\ & 0 & & & -a & 1 \end{pmatrix} (n/2) \times n$$

Then $\tilde{u} = W_e \tilde{y} \sim N_{(n/2)}(0, (1/\rho)I+K)$ where

$$K = \begin{pmatrix} a^2+1 & & 0 \\ & a^2+1 & \\ 0 & & \ddots \quad a^2+1 \end{pmatrix} \quad (n/2) \times (n/2)$$

We have $\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_{n/2}$ which are i.i.d. from a $N(0, 1/\rho+a^2+1)$ population. A "method of moments" estimator is found by setting

$$0 = \frac{\sum_{i=1}^{n/2} (\hat{a}y_{2i-1} - y_{2i})}{n/2} = \bar{u} \; ;$$

which implies

$$\hat{a} = \frac{\sum_{i=1}^{n/2} y_{2i}}{\sum_{i=1}^{n/2} y_{2i-1}} \; . \qquad (4.15)$$

Let us consider, yet, another estimate for "a." Define

$$t_1' = (-a, 1, -a, 1, \ldots, -a, 1)$$
$$t_2' = (0, -a, 1, -a, 1, \ldots, -a, 1, 0). \qquad (t_1, t_2 \in \mathbb{R}^n)$$

Note that (4.13) was found by solving $t_1'y = 0$. We might consider

$$v = \gamma t_1'y + (1-\gamma)t_2'y, \quad \gamma \in [0, 1], \quad E\tilde{v} = 0.$$

If we ask for that value of γ which minimizes $\mathrm{var}(\tilde{v})$, an easy calculation shows that the optimal γ has $1/2$ as a limit as $n \to \infty$.

After setting $\gamma = 1/2$ and $v = 0$, we obtain an estimate

$$\hat{a} = 1 + \frac{y_n - y_1}{y_1 + y_2 + \ldots + y_{n-1}}.$$

We now turn to the more general considerations of Chapter 5. There, the concern is focused upon two fairly difficult technical problems. First, how does one arrive at the order of the autoregressive prior? In answering this question one confronts another problem. What can we do to obtain estimates for the various hyperparameters we have introduced?

V. THE AUTOREGRESSIVE PRIOR

$$\tilde{\epsilon}_t = [\sum_{r=0}^{p} a_r 1_N(t-r)\, \rho^{p-r}](\tilde{\theta}_{t-p} - \mu_{t-p})$$

ESTIMATION CONSIDERATIONS

An Outline of the Technical Problem and the Approach Taken

In the notation of Corollary 4.1,

$$\tilde{\epsilon}_t = [\sum_{r=0}^{p} a_r 1_N(t-r)\, \rho^{p-r}](\tilde{\theta}_{t-1} - \mu_{t-p}), \quad \tilde{\epsilon}_t \text{ i.i.d.}, \quad t = 1, \ldots, n \tag{5.1}$$

can be written as

$$\tilde{\epsilon} = T^{-1}(\tilde{\theta} - \mu) . \tag{5.2}$$

Recall that

$$T^{-1} = \begin{pmatrix} 1 & 0 & \text{———} & 0 \\ a_1 & 1 & 0 & \text{———} & 0 \\ a_2 & a_1 & 1 & 0 & \text{———} & 0 \\ & & & & & \\ 0 & & & a_p & a_1 & 1 \end{pmatrix}_{n \times n}, \quad n > p.$$

From (5.2) we have

$$\tilde{\theta} = \mu + T\tilde{\epsilon}. \tag{5.3}$$

As before, we assume

$$\tilde{y} | \theta \sim N_n(\theta, I) \quad \text{and} \quad \tilde{\epsilon} \sim N_n(0, (1/\rho)I) .$$

This--combined with (5.3)--implies that

$$\tilde{\theta} \sim N_n(\mu, (1/\rho)TT') .$$

The marginal distribution of $\tilde{y}$ now depends upon a vector

$$P_o' = (a_1, \ldots, a_p, \rho, \mu_1, \ldots, \mu_n) .$$

We have

$$\begin{aligned} f_{\tilde{y}|P_o}(y) &= \int f_{\tilde{y}|\theta}(y) f_{\tilde{\theta}}(\theta) d\theta \\ &\propto [\det(I+(1/\rho)TT')]^{-1/2} \\ &\quad \times \exp[(-1/2)(y-\mu)'(I+(1/\rho)TT')^{-1}(y-\mu)] . \end{aligned} \tag{5.4}$$

The calculation for T (in terms of the a's) is routine but its elements become complicated. Therefore, it is convenient to work with expressions involving T^{-1}. Denote T^{-1} by W and express $(I+(1/\rho)TT')^{-1}$ as

$$\begin{aligned} (I+(1/\rho)TT')^{-1} &= (I+\rho(TT')^{-1})^{-1}\rho(TT')^{-1} \\ &= (I+\rho W'W)^{-1}\rho W'W \\ &= \rho W'W(I+\rho W'W)^{-1} . \end{aligned} \tag{5.5}$$

Note that

$$\det[\rho W'W(I+\rho W'W)^{-1}] = \rho^n \det(I+\rho W'W)^{-1} \tag{5.6}$$

(since $\det(W'W) = 1$).

Set

$$P_1' = (\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_p)$$

and

$$P_2' = (\tilde{\rho}, \tilde{\mu}_1, \ldots, \tilde{\mu}_n), \quad (\tilde{P}_o' = (\tilde{P}_1', \tilde{P}_2')').$$

Assume that $\tilde{P}_1$ and $\tilde{P}_2$ are independent. The distribution of $\tilde{P}_o$ is specified by taking

$$f_{\tilde{P}_1}(P_1) \propto \exp((-1/2) \sum_{i=1}^{p} (\alpha_i/\xi_i)^2), \quad \xi_1 \geq \xi_2 \geq \cdots \geq \xi_p,$$

and

$$f_{\tilde{P}_2}(P_2) \propto (1/\rho) 1_{(0,\infty)}(\rho), \quad \mu_i \in \mathbb{R}^1, \quad i = 1, 2, \ldots, n.$$

We have

$$f_{\tilde{P}_o}(P_o) = f_{\tilde{P}_1}(P_1) f_{\tilde{P}_2}(P_2)$$

$$\propto \exp((-1/2) \sum_{i=1}^{p} (\alpha_i/\xi_i)^2)(1/\rho) 1_{(0,\infty)}(\rho), \qquad (5.7)$$

$$\xi_1 \geq \xi_2 \geq \cdots \geq \xi_p, \quad \mu_i \in \mathbb{R}^1, \quad i = 1, 2, \ldots, n.$$

Let ξ_i, $i = 1, 2, \ldots, p$, be specified subject to the above inequalities. The prior on P_1 indicates a belief that for a given θ_i, nearby θ's are likely to have more effect on θ_i than remote ones.

Assume, temporarily, that P_2 is known. The marginal density (5.4) is now written as $f_{\tilde{y}|P_1}(y)$. This marginal density in

combination with the prior $f_{\underset{\sim}{P}_1}(P_1)$ yields a posterior

$$
\begin{aligned}
f_{\underset{\sim}{P}_1|\,y}(P_1) &\propto f_{\underset{\sim}{y}|\,P_1}(y) f_{\underset{\sim}{P}_1}(P_1) \\
&\propto [\rho^n \det(I+\rho W'W)^{-1}]^{1/2} \\
&\quad \times \exp[(-\rho/2)(y-\mu)'W'W(I+\rho W'W)^{-1}(y-\mu)] \\
&\quad \times \exp\Big[(-1/2)\sum_{i=1}^{p}(\alpha_i/\xi_i)^2\Big], \qquad \xi_1 \geq \xi_2 \geq \cdots \geq \xi_p .
\end{aligned}
\tag{5.8}
$$

Suppose that a "smooth" curve passes through the points (X_i, μ_i), $i = 1, \ldots, n$, where X_i is the ith level of stress, $\mu_i = E(\underset{\sim}{\theta}_i)$. (In Chapter 3 we took $\mu = X\tau$, $X' = (1.5, 2.0, \ldots, 5.0)$, τ unknown.) Also, suppose that the precision ρ is not small to the extent that the prior on θ has negligible effect (see Lemma 2.1). With P_2 given, the mode of (5.8) is taken as the estimate for P_1. If the α's are estimated near zero, the overall effect is too "smooth" the estimate of θ. This smoothing is a combined effect of prior and sample. The smoothing effect of the prior alone depends on taking the ξ's small. Thus, prior choices of the ξ_i, $i = 1,2,\ldots,p$, combine with information given by the data (in (5.8)) to say something about the related questions regarding "the smoothness of the response curve" and a property that might be called the effective order of the autoregressive prior.

Now let us assume, temporarily, that P_1 is known. Denote the marginal density (5.4) as $f_{\tilde{y}|P_2}(y)$. The posterior density of present interest is

$$\begin{aligned} f_{\tilde{P}_2|y}(P_2) &\propto f_{\tilde{y}|P_2}(y) f_{\tilde{P}_2}(P_2) \\ &\propto [\rho^n \det(I+\rho W'W)^{-1}]^{1/2} \\ &\quad \times \exp[(-\rho/2)(y-\mu)'W'W(I+\rho W'W)^{-1}(y-\mu)] \\ &\quad \times (1/\rho) 1_{(0,\infty)}(\rho), \quad \mu_i \in \mathbb{R}^1, \ i = 1, 2, \ldots, n. \end{aligned} \tag{5.9}$$

The mode of (5.9) (denoted by $\hat{P}_2$) is taken as the estimate of P_2. It can be calculated using the methods discussed in Chapter 3.

Actually, we treat the vector $P_o = (P_1', P_2')'$ as completely unknown. Since

$$\hat{\theta}(P_o) = \rho W'W(I+\rho W'W)^{-1}\mu + W'W(I+\rho W'W)^{-1}(W'W)^{-1}y$$

(see (2.7)), it is necessary to estimate P_o before we can estimate θ. The mode, $\hat{P}_o$, of

$$f_{\tilde{P}_o|y}(P_o) \propto f_{\tilde{y}|P_o}(y) f_{\tilde{P}_o}(P_o) \tag{5.10}$$

(see (5.4) and (5.7)) is assigned as the estimate of P_o. After calculating $\hat{P}_o$, θ is estimated with $\hat{\theta}(\hat{P}_o)$.

Although a standard gradient-search procedure might be used to

find the stationary points of (5.10), an algorithm is given below for exploiting the special structure of $f_{\tilde{P}_o|y}(y)$. It is felt that iterations on P_o might become complicated or inefficient due to

(i) its dimension (in a given application),

(ii) the possibility that mild perturbations of P_1 (with P_2 fixed) might have an effect on $f_{\tilde{P}_o|y}$ which is drastic in comparison with the effect of mild perturbations of P_2 (with P_1 fixed).

We acknowledge the fact that within P_2, changes in ρ have an effect on $f_{\tilde{P}_o|y}$ which are distinctly different than the effect of changes in μ_i, $i = 1, \ldots, n$. However, we will take $(P_1', P_2')'$ as an "approximately natural" partition of P_o and sketch our algorithm as follows.

(a) Let $P_2^{(1)}$ be an initial guess for P_2. Setting $P_2 = P_2^{(1)}$, calculate the mode of (5.8).

(b) Use the mode calculated in (a) as a starting value for P_1. With this given P_1, calculate the mode of (5.9).

(c) Repeat the process initiated as (a) by using the mode found in (b) as the new starting value $P_2^{(2)}$ in (a).

The essential features of this procedure are discussed by Lindley and Smith (1972). They remark that this sequence of iterations "typically converges." The author has not addressed the question of convergence in a mathematical manner. However, in the

present situation, the iterations have converged nicely (without failure) in a large number of examples.

The gradient-search technique used to carry out the calculations in (a) and (b) above is devised to adapt itself to the "pace of convergence" in each case. The technique is outlined on page 87. Let us now illustrate the Lindley and Smith procedure and our "tailor made" gradient-search with a final application to the corrosion data of Figure 1.1.

The Corrosion Data Revisited

Consistent with the notation and distributional assumptions made regarding the example of Chapter 3, let us take

$$\tilde{y} \mid \theta \sim N_8(\theta, I) \quad \text{and} \quad \tilde{\theta} \sim N_8(X\tau, (1/\rho)V)$$

where $X = (1.5, 2.0, \ldots, 5.0)'$, $p > 0$, and V is positive definite. Now, let us assume that V is specified by an autoregressive prior of order 3 at most on the components of $\theta - X\tau$. Writing (5.1) with $\mu = X\tau$, $p = 3$, and $n = 8$ we have

$$
\begin{aligned}
\tilde{\theta}_1 - X_1\tau &= \tilde{\epsilon}_1 \\
(\tilde{\theta}_2 - X_2\tau) + a_1(\tilde{\theta}_1 - X_1\tau) &= \tilde{\epsilon}_2 \\
(\tilde{\theta}_3 - X_3\tau) + a_1(\tilde{\theta}_2 - X_2\tau) + a_2(\tilde{\theta}_1 - X_1\tau) &= \tilde{\epsilon}_3 \\
&\vdots \\
(\tilde{\theta}_k - X_k\tau) + a_1(\tilde{\theta}_{k-1} - X_{k-1}\tau) + a_2(\tilde{\theta}_{k-2} - X_{k-2}\tau) + a_3(\tilde{\theta}_{k-3} - X_{k-3}\tau) &= \epsilon_k
\end{aligned}
\tag{5.11}
$$

$$k = 4, 5, \ldots, 8.$$

A more convenient form of (5.11) is

$$\tilde{\epsilon} = W(\tilde{\theta} - X\tau)$$

where

$$
W = T^{-1} = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
a_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
a_2 & a_1 & 1 & 0 & 0 & 0 & 0 & 0 \\
a_3 & a_2 & a_1 & 1 & 0 & 0 & 0 & 0 \\
 & & & \vdots & & & & \\
0 & 0 & 0 & 0 & a_3 & a_2 & a_1 & 1
\end{pmatrix}_{(8 \times 8)}
\tag{5.12}
$$

Our distributional assumptions are completely described by

$$\tilde{y} \mid \theta \sim N_8(\theta, I) \quad \text{and} \quad \tilde{\theta} \mid P_o \sim N_8(X\tau, (\rho W'W)^{-1})$$

where $P_o = (P_1', P_2')'$ with

$$P_1' = (a_1, a_2, a_3)$$

$$P_2' = (\rho, \tau) .$$

For this example, (5.8) becomes

$$f_{\tilde{P}_1|y}(P_1) \propto [\rho^8 \det(I+\rho W'W)^{-1}]^{1/2} \times \exp[(-\rho/2)(y-X\tau)'W'W(I+\rho W'W)^{-1} \times (y-X\tau) - (1/2)\sum_{i=1}^{3}(\alpha_i/\xi_i)^2] . \quad (5.13)$$

Two special cases are illustrated in the following:

(i) $\xi_i^2 = \infty, \quad i = 1, 2, 3,$

(ii) $\xi_1^2 = 10.0, \quad \xi_2^2 = 0.1, \quad \xi_3^2 = 0.01.$

The distribution of $\tilde{P}_2|y$ is (see (5.9))

$$f_{\tilde{P}_2|y}(P_2) \propto [\rho^8 \det(I+\rho W'W)^{-1}]^{1/2} \times \exp[(-\rho/2)(y-X\tau)'W'W(I+\rho W'W)^{-1}(y-X\tau)] \times (1/\rho)1_{(0,\infty)}(\rho) . \quad (5.14)$$

Denote $\ln f_{\tilde{P}_1|y}(P_1)$ by $F(\alpha_1, \alpha_2, \alpha_3)$. Then

$$F(\alpha_1, \alpha_2, \alpha_3) = (\text{const.}) + (1/2)(8 \ln \rho + \ln \det \Sigma^{-1}) - (1/2)(y-X\tau)'((1/\rho)V+I)^{-1}(y-X\tau) - (1/2)\sum_{i=1}^{3}(\alpha_i/\xi_i)^2 , \quad (5.15)$$

where $\not{\Sigma}^{-1} = (\rho W'W+I)^{-1}$ and $V = TT' = (W'W)^{-1}$. Also, let

$$G(\rho,\tau) = \ln f_{\underset{\sim}{P}_2|y}(P_2) = (\text{const.}) + (1/2)\ln\det((1/\rho)V+I)^{-1} - (1/2)(y-X\tau)'((1/\rho)V+I)^{-1}(y-X\tau) - \ln\rho. \tag{5.16}$$

Let

$$\nabla F = \left(\frac{\partial F}{\partial a_1}, \frac{\partial F}{\partial a_2}, \frac{\partial F}{\partial a_3}\right)' \quad \text{and} \quad \nabla G = \left(\frac{\partial G}{\partial \rho}, \frac{\partial G}{\partial \tau}\right)'$$

denote the gradients of F and G respectively. Certain standard results from the differential calculus in matrix notation are needed to carry out the gradient calculations. In the Appendix we have definitions, standard results, and derivations of all differentiation formulae which appear in the following.

Remark 5.1. Let $W_i = \frac{\partial W}{\partial a_i}$ (W is defined by (5.12)) $i = 1, 2, 3$. Then W_i is that lower triangular matrix whose elements are all zero except those on the ith minor diagonal--all of which are 1. For example,

$$W_1 = \frac{\partial W}{\partial a_1} = \begin{pmatrix} 0 & & & 0 \\ 1 & \ddots & 0 & \\ & 1 & \ddots & \\ 0 & 0 & 1 & 0 \end{pmatrix}_{(8 \times 8)}.$$

Remark 5.2. Denote $\frac{\partial}{\partial \alpha_i} W'W$ by U_i, $i = 1, 2, 3$.

Then $U_i = W_i'W + W'W_i$.

In the following, let $\not\Sigma = \rho W'W + I$.

Formula 5.1. $\frac{\partial}{\partial \alpha_i} \ln \det \not\Sigma^{-1} = -\rho \operatorname{trace}(U_i \not\Sigma^{-1})$, $i = 1, 2, 3$.

Formula 5.2. $\frac{\partial}{\partial \alpha_i} ((1/\rho)V+I)^{-1} = \rho M U_i \not\Sigma^{-1}$ where $M = I - \rho(W'W)\not\Sigma^{-1}$.

Formula 5.3. $\frac{\partial}{\partial \rho} \ln \det ((1/\rho)V+I)^{-1} = 8/\rho - \operatorname{trace}(W'W \not\Sigma^{-1})$.

Formula 5.4. $\frac{\partial}{\partial \rho} ((1/\rho)V+I)^{-1} = W'W \not\Sigma^{-2}$.

Formula 5.5. $\frac{\partial}{\partial \tau} z'((1/\rho)V+I)^{-1} z = -2X'((1/\rho)V+I)^{-1} z$ where $z = y - X\tau$.

Proceeding with the gradient calculations we have

$$\frac{\partial F}{\partial \alpha_i} = (1/2) \frac{\partial}{\partial \alpha_i} \ln \det \not\Sigma^{-1} - (1/2)(y-X\tau)' \frac{\partial}{\partial \alpha_i} ((1/\rho)V+I)^{-1}(y-X\tau) - \alpha_i/\xi_i^2,$$

$$i = 1, 2, 3.$$

Applying the above differentiation results,

$$\frac{\partial F}{\partial \alpha_i} = (-\rho/2) \operatorname{trace}(U_i \Sigma^{-1}) - (\rho/2)(y-X\tau)'MU_i \not\Sigma^{-1}(y-X\tau) - \alpha_i/\xi_i^2 .$$

The gradient of F is

$$\nabla F = (-\rho/2)\begin{pmatrix} \text{trace } U_1\Sigma^{-1} + (y-X\tau)'MU_1\Sigma^{-1}(y-X\tau) + 2a_1/\rho\xi_1^2 \\ \text{trace } U_2\Sigma^{-1} + (y-X\tau)'MU_2\Sigma^{-1}(y-X\tau) + 2a_2/\rho\xi_2^2 \\ \text{trace } U_3\Sigma^{-1} + (y-X\tau)'MU_3\Sigma^{-1}(y-X\tau) + 2a_3/\rho\xi_3^2 \end{pmatrix}.$$

For the gradient of G we have:

$$\frac{\partial G}{\partial \rho} = (1/2)\frac{\partial}{\partial \rho}\ln\det((1/\rho)V+I)^{-1}$$

$$- (1/2)(y-X\tau)'\frac{\partial}{\partial \rho}((1/\rho)V+I)^{-1}(y-X\tau) - 1/\rho$$

$$= (1/2)(8/\rho - \text{trace}(W'W\Sigma^{-1})) - (1/2)(y-X\tau)'W'W\Sigma^{-2}(y-X\tau) - 1/\rho,$$

$$\frac{\partial G}{\partial \tau} = (-1/2)\frac{\partial}{\partial \tau}(y-X\tau)'((1/\rho)V+I)^{-1}(y-X\tau)$$

$$= X'((1/\rho)V+I)^{-1}(y-X\tau).$$

Therefore,

$$\nabla G = \begin{pmatrix} 3/\rho - (1/2)\text{trace}(W'W\Sigma^{-1}) - (1/2)(y-X\tau)'W'W\Sigma^{-2}(y-X\tau) \\ X'((1/\rho)V+I)^{-1}(y-X\tau) \end{pmatrix}.$$

The flowchart in Figure 5.1 illustrates the above indicated procedure (see page 86) for obtaining $\hat{P}_o$. For the corrosion application, let

$$P_1^{(1)} = (a_1^{(1)}, a_2^{(1)}, a_3^{(1)})' = (1.0, 0.5, 0.0)'$$

and

$$P_2^{(1)} = (\rho^{(1)}, \tau^{(1)}) = (0.5, 1.0)'$$

denote starting values assigned to P_1 and P_2 respectively. In addition, let

$$K^{(1)} = (K_1^{(1)}, K_2^{(1)}, \ldots, K_{11}^{(1)})'$$

$$= (0.0, 0.1, \ldots, 1.0)'$$

be an auxiliary vector used to define the search points. The gradient-search starts by taking $P_2^{(1)}$ as given. The gradient of F is evaluated at $P_1^{(1)}$. Then, 11 search points are defined on line with the gradient at $P_1^{(1)}$,

$$Q_i^{(1)} = P_1^{(1)} + K_i^{(1)} \nabla F]_{P_1^{(1)}}, \quad i = 1, 2, \ldots, 11 . \tag{5.17}$$

(Acutally, the $Q_i^{(1)}$ are calculated only if the gradient at $P_1^{(1)}$ has magnitude $> .01$. See the diagram on the next page.) Suppose $Q_{j_o}^{(1)}$ is that search point among those at (5.17) which maximized F. Define $P_1^{(2)} = Q_{j_o}^{(1)}$, $K^{(2)} = (j_o/11)K^{(1)}$ (an adjustment for the pace of convergence), and calculate $\nabla F]_{P_1^{(2)}}$ (etc.) .

In Figure 5.2, estimates of θ (based on the data from Figure 1.1) are shown corresponding to the priors,

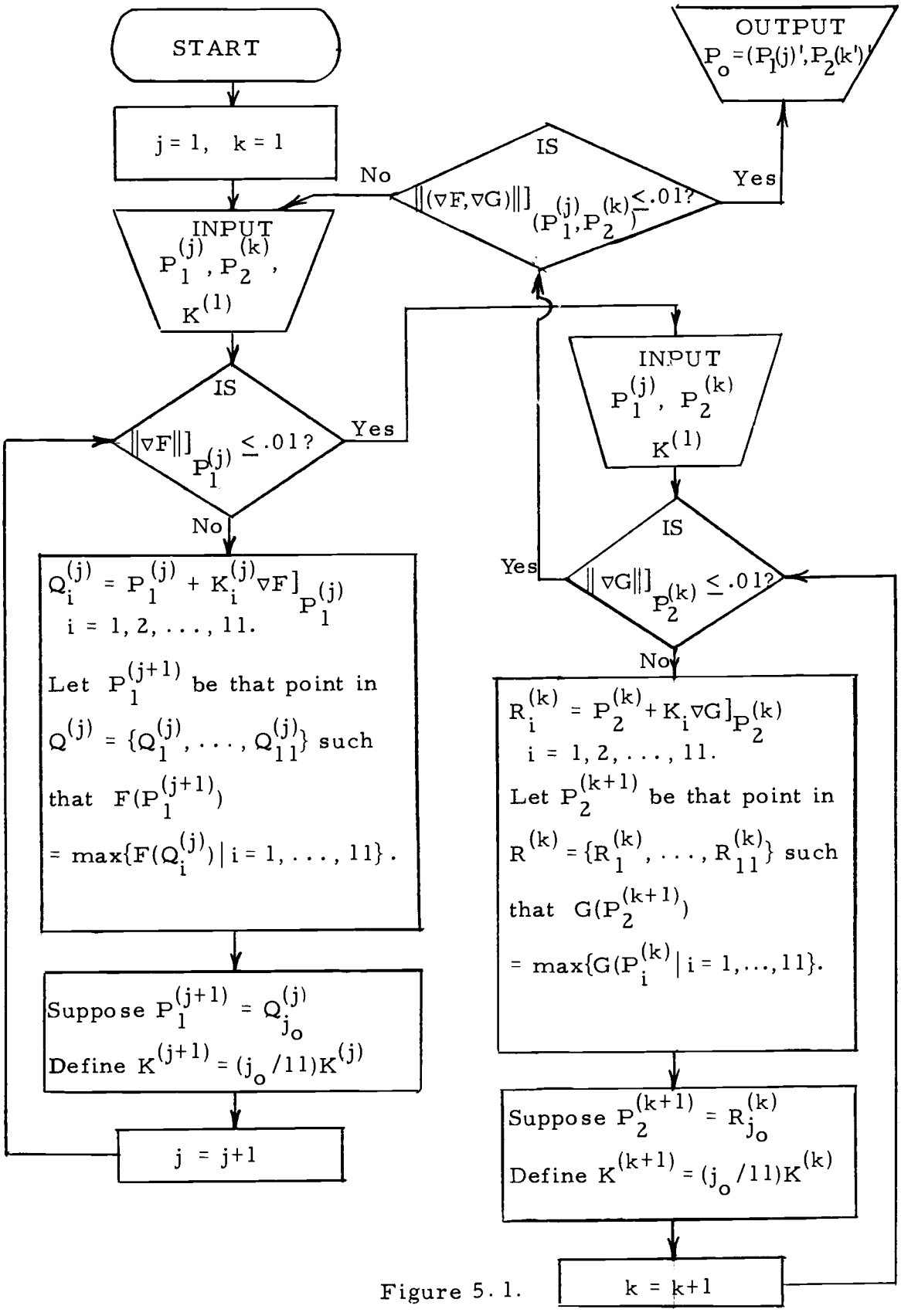

START
j = 1, k = 1
INPUT $P_1^{(j)}, P_2^{(k)}, K^{(1)}$
IS $\|\nabla F\|]_{P_1^{(j)}} \leq .01$?
Yes
No
$Q_i^{(j)} = P_1^{(j)} + K_i^{(j)} \nabla F]_{P_1^{(j)}}$
$i = 1, 2, \ldots, 11.$
Let $P_1^{(j+1)}$ be that point in
$Q^{(j)} = \{Q_1^{(j)}, \ldots, Q_{11}^{(j)}\}$ such
that $F(P_1^{(j+1)})$
$= \max\{F(Q_i^{(j)}) \mid i = 1, \ldots, 11\}$.
Suppose $P_1^{(j+1)} = Q_{j_o}^{(j)}$
Define $K^{(j+1)} = (j_o/11)K^{(j)}$
j = j+1
IS $\|(\nabla F, \nabla G)\|]_{(P_1^{(j)}, P_2^{(k)})} \leq .01$?
No
Yes
OUTPUT
$P_o = (P_1(j)', P_2(k')'$
INPUT $P_1^{(j)}, P_2^{(k)}, K^{(1)}$
IS $\|\nabla G\|]_{P_2^{(k)}} \leq .01$?
Yes
No
$R_i^{(k)} = P_2^{(k)} + K_i \nabla G]_{P_2^{(k)}}$
$i = 1, 2, \ldots, 11.$
Let $P_2^{(k+1)}$ be that point in
$R^{(k)} = \{R_1^{(k)}, \ldots, R_{11}^{(k)}\}$ such
that $G(P_2^{(k+1)})$
$= \max\{G(P_i^{(k)}) \mid i = 1, \ldots, 11\}$.
Suppose $P_2^{(k+1)} = R_{j_o}^{(k)}$
Define $K^{(k+1)} = (j_o/11)K^{(k)}$
k = k+1

Figure 5.1.

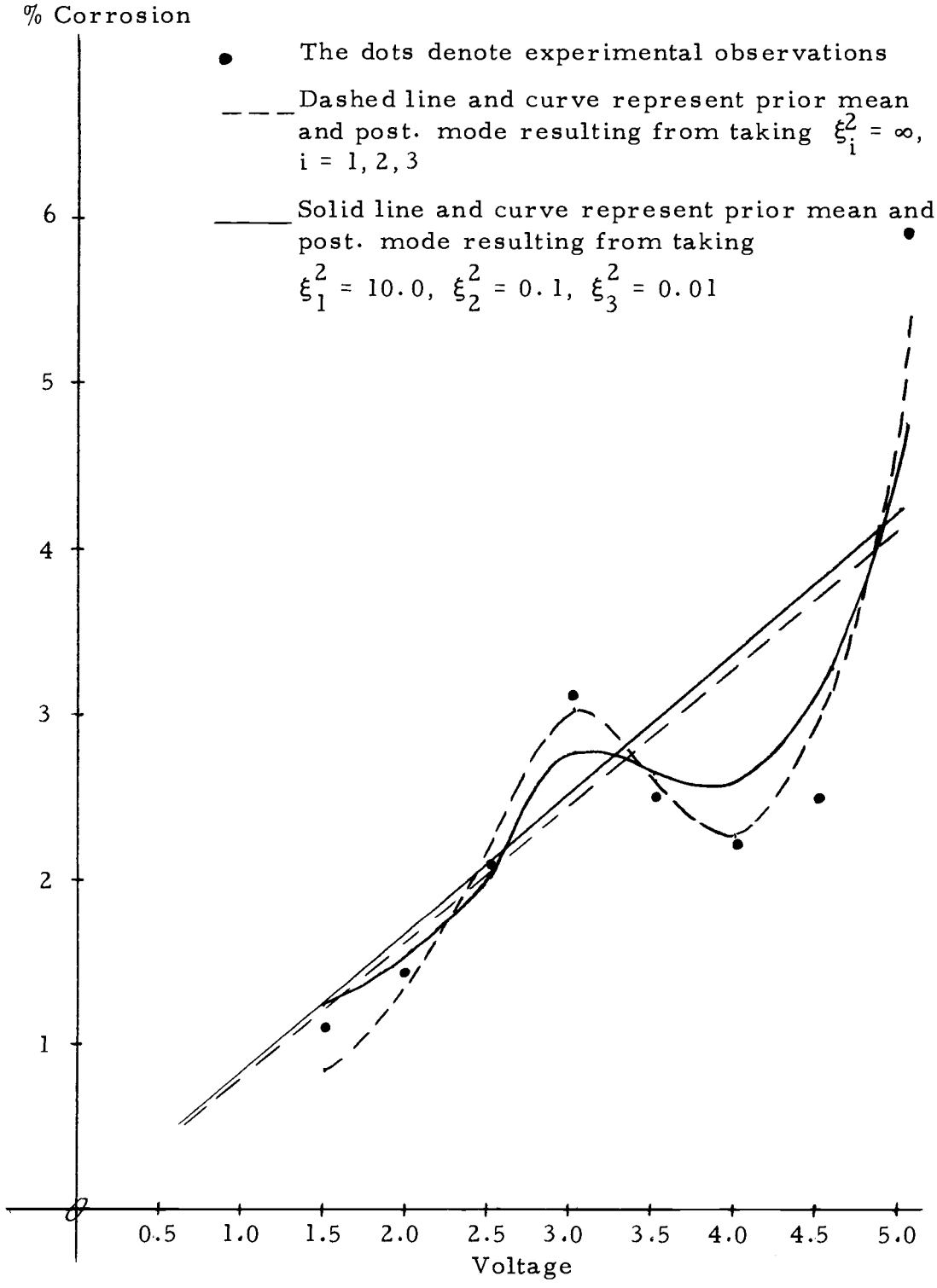

% Corrosion
The dots denote experimental observations
Dashed line and curve represent prior mean and post. mode resulting from taking $\xi_i^2 = \infty$, i = 1, 2, 3
Solid line and curve represent prior mean and post. mode resulting from taking
$\xi_1^2 = 10.0$, $\xi_2^2 = 0.1$, $\xi_3^2 = 0.01$
6
5
4
3
2
1
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Voltage

Figure 5.2.

(i) $\xi_i^2 = \infty, \quad i = 1, 2, 3$

and (ii) $\xi_1^2 = 10, \quad \xi_2^2 = 0.1, \quad \xi_3^2 = 0.01$.

The dashed curve denotes the estimate of θ resulting from (i). In this situation we might be tending to overfit the data for the following reasons. We have vague priors on all hyperparameters (five in number). Under this condition, the number of hyperparameters is probably excessive in comparison with the number of data points (eight). In effect, the data was underconstrained. This results in a quite vague prior on θ. Corresponding to (ii) is the solid curve in Figure 5.2. Prior (ii) expresses the idea that appreciable correlations exist only between adjacent θ's. Consequently, we observe more smoothing than was the case with (i).

The Case of Unknown $\sigma^2 = 1/\rho_1$, $\underset{\sim}{y} | \theta \sim N_n(\theta, (1/\rho_1)I)$

We conclude the chapter by indicating the estimation procedure taken when $\underset{\sim}{y} | \theta \sim N_n(\theta, (1/\rho_1)I)$ and $\underset{\sim}{\theta} \sim N_n(X\underline{\beta}, (1/\rho_2)V)$, $\rho_1 > 0$, $\rho_2 > 0$, $\underline{\beta} \in R^k$, and X is a known $n \times k$ matrix of rank $r > k$. Again, assume that V is a positive definite matrix specified by the autoregressive process of (5.1) (an AR process of order p with the α's unknown). From Result 1.2, the marginal distribution of $\underset{\sim}{y}$ is $N_n(X\underline{\beta}, I/\rho_1 + V/\rho_2)$. In this context, let

$$P_1 = (a_1, a_2, \ldots, a_p)'$$

$$P_2 = (\rho_1, \rho_2, \underline{\beta})'$$

and $$P_o = (P_1', P_2')' .$$

Again, suppose that P_1 and P_2 are independent and take

$$f_{\underset{\sim}{P}_1}(P_1) \propto \exp[(-1/2) \sum_{i=1}^{p} (a_i/\xi_i)^2] ,$$

$$f_{\underset{\sim}{P}_2}(P_2) \propto 1/\rho_1\rho_2 \qquad \rho_i > 0, \quad i = 1, 2 .$$

Analogous to the posterior densities of (5.13) and (5.14) we have

$$g_{\underset{\sim}{P}_1|y}(P_1) \propto [(\rho_1\rho_2)^n \det(\rho_1 I + \rho_2 W'W)^{-1}]^{1/2}$$
$$\times \exp[(-\rho_1\rho_2/2)(y-X\beta)'W'W(\rho_1 I+\rho_2 W'W)^{-1}(y-X\beta)]$$
$$\times \exp((-1/2) \sum_{i=1}^{p} (a_i/\xi_i)^2 ,$$

$$g_{\underset{\sim}{P}_2|y}(P_2) \propto [(\rho_1\rho_2)^n \det(\rho_1 I + \rho_2 W'W)^{-1}]^{1/2}$$
$$\times \exp[(-\rho_1\rho_2/2)(y-X\beta)'W'W(\rho_1 I+\rho_2 W'W)^{-1}(y-X\beta)$$
$$\times (1/\rho_1\rho_2), \quad \rho_i > 0, \quad i = 1, 2.$$

The formulas needed to carry out a gradient-search for P_o (as conducted in the last section) are derived in Appendix II.

BIBLIOGRAPHY

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

Antoniak, C.E. 1969. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Unpublished Ph.D. thesis, UCLA, 1969.

Baranchik, A.J. 1964. Multiple regression and estimation of the mean of a multivariate normal distribution. Stanford Univ. Technical Report No. 51.

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.

Box, G.E.P. and Gwilym M. Jenkins. 1971. Time series analysis forcasting and control. San Francisco, Holden-Day. 553 p.

Box, G.E.P. and George C. Tiao. 1973. Bayesian inference in statistical analysis. Reading, Mass., Addison-Wesley Publishing Company. 588 p.

Burrill, Claude W. 1972. Measure, integration, and probability. New York, McGraw-Hill Inc. 464 p.

Efron, B. and Carl Morris. 1972. Empirical Bayes on vector observations: An extension of Stein's method. Biometrika 59:335-347.

Efron, B. and Carl Morris. 1973. Stein's estimation rule and its competitors - an Empirical Bayes approach. Journal of the American Statistical Association 68:117-131.

Ferguson, T.S. 1967. Mathematical Statistics a decision theoretic approach. New York, McGraw-Hill, Inc. 396 p.

Goldberger . 1964. Econometric theory. New York, Wiley. 350 p.

Graybill, F.A. 1961. An introduction to linear statistical models. Vol. 1. New York, McGraw-Hill, Inc. 463 p.

James W. and Charles Stein. 1961. Estimation with quadratic loss. Proceedings of the Fourth Berkeley Symposium, Berkeley, Univ. of California Press. Vol. 1:361-379.

Lindley, D.V. and A.F.M. Smith. 1971. Bayes estimates for the linear model. Journal of the Royal Statistical Society, ser. B. 34:1-42.

Luenberger David G. 1969. Optimization by vector space methods. New York, Wiley. 319 p.

Raiffa, Howard and Robert Schlaifer. 1960. Applied statistical decision theory. Cambridge, Mass., The M.I.T. Press. 356 p.

Rao, C.R. 1965. Linear statistical inference and its applications. New York, Wiley. 522 p.

Seely, Justus. 1972. Restricted maximum likelihood estimation for two variance components. Oregon State Univ. Dept. of Statistics, Technical Report No. 26.

Stein, Charles. 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium, Vol. 1:197-206.

Thompson, W.A. Jr. 1963. The problem of negative estimates of variance components. The Annals of Mathematical Statistics 33:273-289.

Wilks, Samuel S. 1962. Mathematical statistics. New York, Wiley. 644 p.

Zacks, S. 1972. The theory of statistical inference. New York, Wiley. 609 p.

APPENDIX

APPENDIX

Differentiation Results Involving Matrices

The presentation in Chapter 5 depended upon certain differentiation results involving matrices. The necessary calculations (shown below) are based on definitions and formulas which are found in Goldberger (1964), the CRC publication Standard Mathematical Tables (1969), or in various books on nonlinear programming.

Definition A.1. If the elements of a matrix $M = (m_{ij})_{p \times q}$ and the elements of a vector $m = (m_1, m_2, \ldots, m_r)'$ are functions of a scalar x, then

(i) $\frac{\partial M}{\partial x}$ denotes a matrix of order $p \times q$ with elements $\frac{\partial m_{ij}}{\partial x}$

(ii) $\frac{\partial m}{\partial x}$ denotes a vector in $\mathbb{R}^r$ with elements $\frac{\partial m_i}{\partial x}$.

Definition A.2. If y is a scalar function of $p \times q$ variables m_{ij} which are the elements of a matrix $M = (m_{ij})$, the expression $\frac{\partial y}{\partial M}$ denotes a matrix with elements $\frac{\partial y}{\partial m_{ij}}$.

Result A. If M and N are matrices whose elements are functions of a scalar x, B and C are matrices whose elements are not functions of x, then

(i) $\frac{\partial}{\partial x}(M+N) = \frac{\partial M}{\partial x} + \frac{\partial N}{\partial x}$

(ii) $\frac{\partial MN}{\partial x} = (\frac{\partial M}{\partial x})N + M(\frac{\partial N}{\partial x})$

(iii) $\frac{\partial C'MC}{\partial x} = C'(\frac{\partial M}{\partial x})C$

(iv) If C and M are in $\mathbb{R}^n$ then, $\frac{\partial C'M}{\partial M} = C$

(v) If M is in $\mathbb{R}^n$ and C is in $\mathbb{R}^{n \times n}$ then,

$$\frac{\partial M'CM}{\partial M} = CM + C'M$$

(vi) $\frac{\partial\, \text{trace}(M)}{\partial M} = I$

(vii) $\frac{\partial \ln \det(M)}{\partial M} = (M')^{-1}$

(viii) $\frac{\partial M^{-1}}{\partial x} = -M^{-1}(\frac{\partial M}{\partial x})M^{-1}$

(ix) $\frac{\partial y}{\partial x} = \text{trace}(\frac{\partial y}{\partial M})(\frac{\partial M'}{\partial x})$.

In the most general framework of the last chapter we had

$$\tilde{y} | \theta \sim N_n(\theta, (1/\rho_1)I), \quad \tilde{\theta} \sim N_n(X\underline{\beta}, (1/\rho_2)V),$$

$\rho_1 > 0$, $\rho_2 > 0$, $\underline{\beta} \in \mathbb{R}^k$, X a known $n \times k$ matrix of rank $r \leq k$, and $V = TT'$ where

$$W = T^{-1} = \begin{pmatrix} 1 & 0 & & & 0 \\ a_1 & 1 & 0 & & 0 \\ a_2 & a_1 & 1 & & 0 \\ & & & & \\ 0 & & & a_p \cdots a_1 & 1 \end{pmatrix}$$

The marginal distribution of $\tilde{y}|P_o$ was

$$N_n(X\underline{\beta}, I/\rho_1+V/\rho_2) \quad \text{(see (5.21))},$$

where $P_o = (a_1, \ldots, a_p, \rho_1, \rho_2, \beta)'$. We were to estimate P_o with the mode of

$$f_{\tilde{P}_o|y}(P_o) \propto f_{\tilde{y}|P_o}(y) f_{\tilde{P}_o}(P_o) .$$

The gradient of $\ln f_{\tilde{P}_o|y}$, $\nabla \ln f_{\tilde{P}_o|y}$, can be constructed from the following:

Formula A.1. Let

$$w = \ln \det(1/\rho_1 I + 1/\rho_2 V)^{-1} \qquad \text{(A.1)}$$

$$= n \ln \rho_1 + n \ln \rho_2 + \ln \det(\rho_1 I + \rho_2 W'W)^{-1} .$$

(Since $\ln \det((1/\rho_1)I+(1/\rho_2)V)^{-1} = \ln \det \rho_1(\rho_1 I+\rho_2 W'W)^{-1}\rho_2 W'W$

$$= \ln \det(\rho_1\rho_2)W'W(\rho_1 I+\rho_2 W'W)^{-1} =$$

$$= \ln(\rho_1\rho_2)^n \det W'W(\rho_1 I+\rho_2 W'W)^{-1}$$

$$= n \ln \rho_1 + n \ln \rho_2 + \ln \det(\rho_1 I+\rho_2 W'W)^{-1} . \quad)$$

Then

$$\frac{\partial w}{\partial a_i} = \frac{\partial}{\partial a_i} \ln \det(\rho_1 I+\rho_2 W'W)^{-1}$$

$$= -\rho_2 \operatorname{trace}[U_i(\rho_2 W'W+\rho_1 I)^{-1}]$$

where

$$U_i = \frac{\partial}{\partial a_i} W'W = W_i'W + W'W_i, \quad i = 1, 2, \ldots, p .$$

Proof. The equation for U_i follows from R A (ii). Now,

$$\frac{\partial w}{\partial a_i} = \frac{\partial}{\partial a_i} \ln \det(\rho_1 I+\rho_2 W'W)^{-1}$$

$$= \operatorname{trace} \frac{\partial \ln \det[\rho_1 I+\rho_2 W'W]^{-1}}{\partial[(\rho_1 I+\rho_2 W'W)^{-1}]} \frac{\partial[(\rho_1 I+\rho_2 W'W)^{-1}}{\partial a_i} \qquad \text{(from F A (ix))}$$

$$= -\operatorname{trace}(\rho_1 I+\rho_2 W'W)(\rho_1 I+\rho_2 W'W)^{-1}[\frac{\partial}{\partial a_i}(\rho_1 I+\rho_2 W'W)](\rho_1 I+\rho_2 W'W)^{-1}$$

(from R A (vii) and R A (viii))

$$= -\operatorname{trace} \rho_2 [\frac{\partial}{\partial a_i}(W'W)](\rho_1 I+\rho_2 W'W)^{-1} \qquad \text{(from R A (i))}$$

$$= -\rho_2 \operatorname{trace}[U_i(\rho_1 I+\rho_2 W'W)^{-1}] .$$

Note that Formula 5.1 is verified by setting $\rho_1 = 1$ in the above Formula A.1 and its proof. In the same way, Formulas 5.2,

5.3, 5.4, and 5.5 are verified by the following Formulas A.2, A.3, A.4, and A.5 respectively.

Formula A.2.

$$\frac{\partial}{\partial a_i}(1/\rho_2 V+1/\rho_1 I)^{-1} = \rho_1\rho_2 MU_i(\rho_2 W'W+\rho_1 I)^{-1}$$

where

$$M = I - \rho_2 W'W(\rho_2 W'W+\rho_1 I)^{-1} .$$

Proof.

$$\frac{\partial}{\partial a_i}(1/\rho_2 V+1/\rho_1 I)^{-1} = \frac{\partial}{\partial a_i}(\rho_1\rho_2)W'W(\rho_2 W'W+\rho_1 I)^{-1}$$

$$= (\rho_1\rho_2)\frac{\partial}{\partial a_i}(W'W)(\rho_2 W'W+\rho_1 I)^{-1} .$$

$$= (\rho_1\rho_2)[U_i(\rho_2 W'W+\rho_1 I)^{-1} - (W'W)(\rho_2 W'W+\rho_1 I)^{-1}$$

$$\times \{\frac{\partial}{\partial a_i}(\rho_2 W'W+\rho_1 I)\}(\rho W'W+\rho_1 I)^{-1}]$$

(from R A (ii) and F A (viii))

$$= (\rho_1\rho_2)[U_i(\rho_2 W'W+\rho_1 I)^{-1} - (W'W)(\rho_2 W'W+\rho_1 I)^{-1}$$

$$\times (\rho_2 U_i)(\rho_2 W'W+\rho_1 I)^{-1}] .$$

$$= (\rho_1\rho_2)[I-\rho_2(W'W)(\rho_2 W'W+\rho_1 I)^{-1}]U_i(\rho_2 W'W+\rho_1 I)^{-1}$$

$$= (\rho_1\rho_2)MU_i(\rho_2 W'W+\rho_1 I)^{-1} .$$

Formula A.3.

(i) $\frac{\partial}{\partial \rho_2} \ln \det(1/\rho_2 V + 1/\rho_1 I)^{-1} = n/\rho_2 - \text{trace}[W'W(\rho_2 W'W + \rho_1 I)^{-1}]$

(ii) $\frac{\partial}{\partial \rho_1} \ln \det(1/\rho_2 V + 1/\rho_1 I)^{-1} = n/\rho_1 - \text{trace}(\rho_2 W'W + \rho_1 I)^{-1}$.

Proof of (i).

$$\frac{\partial}{\partial \rho_2} \ln \det(1/\rho_2 V + 1/\rho_1 I)^{-1}$$

$$= \frac{\partial}{\partial \rho_2} [n \ln \rho_1 + n \ln \rho_2 + \ln \det(\rho_1 I + \rho_2 W'W)^{-1}]$$

$$= n/\rho_2 + \frac{\partial}{\partial \rho_2} \ln \det(\rho_1 I + \rho_2 W'W)^{-1}$$

$$= n/\rho_2 + \text{trace} \frac{\partial \ln \det(\rho_1 I + \rho_2 W'W)^{-1}}{\partial [(\rho_1 I + \rho_2 W'W)^{-1}]} \frac{\partial [(\rho_1 I + \rho_2 W'W)^{-1}]}{\partial \rho_2}$$

(from R A (ix))

$$= n/\rho_2 - \text{trace}(\rho_1 I + \rho_2 W'W)(\rho_1 I + \rho_2 W'W)^{-1}$$

$$\times \{\frac{\partial}{\partial \rho_2} (\rho_1 I + \rho_2 W'W)\}(\rho_1 I + \rho_2 W'W)^{-1}$$

(from R A (vi) and R A (viii))

$$= n/\rho_2 - \text{trace}[W'W(\rho_1 I + \rho_2 W'W)^{-1}]$$

The proof of (ii) is transparent from the proof of (i).

Formula A.4.

(i) $\frac{\partial}{\partial \rho_2}(1/\rho_2 V+1/\rho_1 I)^{-1} = \rho_1^2(W'W)(\rho_2 W'W+\rho_1 I)^{-2}$

(ii) $\frac{\partial}{\partial \rho_1}(1/\rho_2 V+1/\rho_1 I)^{-1} = (1/\rho_1)^2(1/\rho_2 V+1/\rho_1 I)^{-2}$.

Proof of (i).

$$\frac{\partial}{\partial \rho_2}(1/\rho_2 V+1/\rho_1 I)^{-1}$$

$$= -(1/\rho_2 V+1/\rho_1 I)^{-1}\{\frac{\partial}{\partial \rho_2}(1/\rho_2 V)\}(1/\rho_2 V+1/\rho_1 I)^{-1} \qquad \text{(R A (viii))}$$

$$= (1/\rho_2)^2(1/\rho_2 V+1/\rho_1 I)^{-1} V(1/\rho_2 V+1/\rho_1 I)^{-1}$$

$$= (1/\rho_2)^2[\rho_2 V^{-1}(\rho_2 V^{-1}+\rho_1 I)^{-1}\rho_1]V[\rho_2 V^{-1}(\rho_2 V^{-1}+\rho_1 I)^{-1}\rho_1]$$

$$= \rho_1^2 V^{-1}(\rho_2 V^{-1}+\rho_1 I)^{-2}$$

$$= \rho_1^2 W'W(\rho_2 W'W+\rho_1 I)^{-2} .$$

Proof of (ii).

$$\frac{\partial}{\partial \rho_1}(1/\rho_2 V+1/\rho_1 I)^{-1}$$

$$= -(1/\rho_2 V+1/\rho_1 I)^{-1}\{\frac{\partial}{\partial \rho_1}(1/\rho_1 I)\}(1/\rho_2 V+1/\rho_1 I)^{-1} \qquad \text{(R A (viii))}$$

$$= (1/\rho_1)^2(1/\rho_2 V+1/\rho_1 I)^{-2} .$$

Formula A.5.

$$\frac{\partial}{\partial \underline{\beta}} (y-X\underline{\beta})'(1/\rho_2 V+1/\rho_1 I)^{-1}(y-X\underline{\beta}) = -2X'(1/\rho_2 V+1/\rho_1 I)^{-1}(y-X\underline{\beta})$$

Proof.

$$\frac{\partial}{\partial \underline{\beta}} (y-X\underline{\beta})'(1/\rho_2 V+1/\rho_1 I)^{-1}(y-X\underline{\beta})$$

$$= \frac{\partial}{\partial \beta} \{\underline{\beta}'X'(1/\rho_2 V+1/\rho_1 I)^{-1}X\underline{\beta} - 2y'(1/\rho_2 V+1/\rho_1 I)^{-1}X\underline{\beta}$$

$$+ y'(1/\rho_2 V+1/\rho_1 I)^{-1}y\}$$

$$= 2X'(1/\rho_2 V+1/\rho_1 I)^{-1}X\underline{\beta} - 2y'(1/\rho_2 V+1/\rho_1 I)^{-1}X$$

(from R A (iv) and R A (v)).