DEPARTMENT OF STATISTICS

University of Wisconsin
1210 W. Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 726

September 1983

THE ANALYSIS OF SOME PENALIZED LIKELIHOOD SCHEMES

by

Finbarr O'Sullivan

This research was partially supported by a University of Wisconsin-Madison Dean's Fellowship, by ONR under Contract No. N00014-77-C-0675 and by NASA under Grant NAG 5-128.

ABSTRACT

There are many areas in applied science where the non-parametric estimation of regression functions is important.

In this thesis a general penalized likelihood method for non-parametrically estimating regression functions under a variety of observational models is developed. The existence and numerical approximation of the estimators is studied and a cross-validatory method for estimating the smoothing parameter is presented. Implementation of the method is algorithmically straightforward.

The procedures developed are applied to the estimation of atmospheric temperature profiles from satellite radiance data and are found to compare favorably with the currently used methodology.

A thesis under the supervision of Grace Wahba.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 THE BASIC PRINCIPLE

The method of *Penalized Likelihood* estimation was introduced by Good and Gaskins in 1971. They argued, from a Bayesian perspective, that in order to estimate a probability density function given a random sample, $a \le X_1, X_2 \ldots X_n \le b$, the *penalized log likelihood functional*

$$\omega(f) = 2 \sum_{i=1}^{n} \log f(X_i) - \lambda J(f) \qquad \lambda>0$$

should be considered. $J(f)$ is a "flamboyancy" or penalty functional such as $\int_a^b [f''(t)]^2 dt$, designed to represent prior notions about the behaviour of the density f. The estimation of the density is done by maximizing ω over a suitably chosen space of alternatives. The parameter λ controls the amount by which the data are smoothed to give the estimate.

The Good and Gaskins idea is not new to statistics, nor is it restricted to density estimation (see Kimeldorf and Wahba 1970, Leonard 1978, Silverman 1979, Tapia and Thompson 1978, Wegman 1982 and, indeed, Whittaker 1923). In this thesis some generalized versions of the penalized likelihood estimation are studied. The following specification of the method will be used: x is some function/

vector of parameters of interest. Conditional upon a data vector z, the estimation of x is done by minimizing the quantity:

$$I_\lambda(x) = Q(x|z) + \lambda J(x) \qquad \lambda>0$$

over some set C of plausible alternatives. $Q(x|z)$ could represent the -2 log [*sampling distribution* of $z|x$] but other choices are feasible. We give a few concrete examples.

1.2 SOME EXAMPLES

(i) Density Estimation

In 1982 Silverman introduced an interesting penalized likelihood estimator of a log probability density function. There are many density estimators which are similar in spirit (see Cox 1982 and Leonard 1978).

Given observations $a \le X_1 \le X_2 \le \cdots \le X_n \le b$ Silverman's penalized likelihood estimate for the log density of the X_i's is prescribed to be the minimizer of

$$A_0(g) = -n^{-1} \sum_{i=1}^{n} g(X_i) + \lambda \int_a^b [g^{(m)}(t)]^2 dt$$

over $K_1 = \{g \in W_2^m[a,b]\colon \int_a^b e^{g(t)} dt = 1\}$ $m \ge 1$. It can be shown (see Silverman 1982) that minimizing A_0 over K_1 is equivalent to minimizing

$$I_\lambda(g) = -n^{-1} \sum_{i=1}^{n} g(X_i) + \int_a^b e^{g(t)} dt + \lambda \int_a^b [g^{(m)}(t)]^2 dt$$

over $K = \{g \in W_2^m[a,b]: \int_a^b e^{g(t)}\,dt < \infty\}$.

(ii) Generalized Linear Regression Models

Here we have independent observations $y_1, y_2, \ldots, y_n$ where each observation y_i has a one-parameter exponential density, p, of the form:

$$p(y_i) = \exp\{[y_i\theta_i - b(\theta_i)]/a_i(\phi) + c(y_i,\phi)\} \qquad (*)$$

for suitable choices of a_i, b and c (note that ϕ, termed the scale parameter, is constant for all i). The mean and the variance of y_i can be expressed in terms of θ_i and ϕ as

$$E(y_i) = \mu_i = b'(\theta_i) \text{ and } \operatorname{var}(y_i) = b''(\theta_i)a_i(\phi)$$

The functions $a_i(\phi)$ are usually of the form ϕ/w_i (where the w_i are known). Specification of the model is completed by supplying a linearizing transformation, $g(\mu_i)$, of the mean.

$$g(\mu_i) = L_i x = \eta_i$$

The function g, referred to as the *link function*, is assumed to be monotonic and differentiable. The L_i's are linear functionals and x is the unknown function/vector of parameters to be estimated. Penalized Likelihood estimates of x are obtained by considering I_λ given by:

$$I_\lambda(x) = \sum_{i=1}^{n} [b(\theta_i) - y_i\theta_i]/a_i(\phi) + \lambda J(x)$$

over an appropriate space of functions.

Notice that a wide variety of distributions, including the normal, Poisson, binomial and gamma, can be written in the form of equation (*). Consequently, a great deal of useful statistical modeling can be accomplished within the Generalized Linear Interactive Modeling (GLIM) framework (see Baker and Nelder 1978). Unfortunately however, good graphical procedures for exploring the data $\{y_i\}$ do not exist (see Raynor and Bates 1983) so that the model building process can be rather slow. The primary motivation for obtaining a non-parametric determination of the regression function, x, is that it can be used as an Exploratory Data Analysis tool by the GLIM modeler.

(iii) Cox's Proportional Hazard Regression Model

Following Miller (1981), let $T_1, T_2, \ldots, T_n; C_1, C_2, \ldots, C_n$ be independent random variables. C_i is the censoring time associated with the survival time T_i. We observe (Y_1, δ_1), $(Y_2, \delta_2), \ldots, (Y_n, \delta_n)$ where

$$Y_i = \min(T_i, C_i), \quad \delta_i = I\{T_i \le C_i\}.$$

Covariates x_i corresponding to the survival times T_i are also available. In Cox's Proportional Hazards Model the hazard rate $\mu(t;x)$ for the distribution of T given x

is written as

$$\mu(t;x) = \mu_0(t)e^{\theta(x)}$$

where θ is some regression function on the covariates. Let the ordered observed times be written as

$$y_{(1)}, y_{(2)}, \ldots, y_{(n)},$$

and $R_{(i)}$ be the risk set, i.e. the set of individuals surviving up to $y_{(i)}$. The Penalized (Partial) Likelihood estimator of θ is the minimizer of the functional

$$I_\lambda(\theta) = \sum_{i\epsilon U} \{\log [\sum_{j\epsilon R_{(i)}} e^{\theta(x_j)}] - \theta(x_i)\} + \lambda J(\theta)$$

where U is the set of uncensored observations. The practical use of this model is currently being evaluated by Crowley and O'Sullivan (1983), a report should be forthcoming.

(iv) Normal Regression Model

The observed data $\{z_i\}_{i=1}^n$ are functionals of x contaminated by white noise.

$$z_i = \eta_i(x) + \varepsilon_i \quad \varepsilon \overset{iid}{\sim} N(0,\sigma)$$

where η_i's are functionals of x. x is estimated by minimizing:

$$I_\lambda(x) = \frac{1}{n} \sum_{i=1}^{n} [z_i - \eta_i(x)]^2 + \lambda J(x), \quad \lambda>0 .$$

Apart from the data smoothing applications of this model (see Kimeldorf and Wahba 1970), there are a number of inversion problems in engineering where the normal regression assumption becomes a useful approximation (see Westwater 1979).

1.3 CONTENTS OF THESIS

There are a number of questions which arise in connection with the actual implementation of the penalized likelihood method. The penalized likelihood estimate is the minimizer of the penalized likelihood functional, I_λ, over an appropriate space of functions. The technical issues addressed in the following chapters include:

- Existence of the penalized likelihood estimate.
- Numerical computation of the estimates for fixed λ.
- Choice of the smoothing parameter λ.

The analysis is carriedout first for the general case and then applied to various particular models. Existence questions are dealt with in Chapter 2; the approach is similar to that taken by Wegman (1982). Numerical methods are presented in Chapter 3 - versions of the Newton-Raphson and Gauss-Newton algorithms are analyzed. The material in these latter two chapters involves a fair bit of mathematical machinery, the elements of which are contained in an appendix at the end of this chapter. Generalized Cross

Validation-type estimators of the smoothing parameter λ are proposed and partially justified in Chapter 4, while the final chapter features an application of the methodology to the remote sensing of atmospheric temperature profiles. An asymptotic analysis of penalized likelihood estimators is currently under study (see Cox and O'Sullivan 1983).

APPENDIX A. BACKGROUND MATHEMATICS

NOTATION

H is a real Hilbert space with inner product $<\cdot,\cdot>$ and norm $\|\cdot\|$ (so $\|x\|^2 = <x,x>$ $x \in H$). The dual space of continuous linear functionals on H is denoted by H^*. $\|\cdot\|$ induces a norm $\|\cdot\|_*$ on H^* given by

$$\|\ell\|_* = \sup_{x \in H} \frac{|\ell(x)|}{\|x\|} \quad \ell \in H^*.$$

For every $\ell \in H^*$ $\exists y_\ell \in H$ such that

$$\ell(x) = <y_\ell, x> \quad \forall\, x \in H$$

The relation $\ell \Leftrightarrow y_\ell$ establishes an isometric isomorphism between the spaces H and H^*. We also have that H^* is a Hilbert space with inner product:

$$<\ell,k>_* = <y_\ell, y_k> \quad \ell,k \in H^*$$

Finally if H_1 and H_2 are two Hilbert spaces then the space of continuous linear operators from H_1 into H_2 will be denoted by $L(H_1,H_2)$.

THE WEAK TOPOLOGY

In the usual or norm topology on H, an open ball, B_r, of radius r about the origin is defined by:

$$B_r = \{x \in H : \|x\| < r\}$$

However the space H^* generates another topology on H called the *weak* topology: this is the weakest topology on H with respect to which the elements of H^* are continuous on H.

The usual topological notions of convergence, compactness, closedness etc. will be used in many places (see Simmons 1963). Whenever we are talking about these notions, unless explicit reference is made to the weak topology, the norm topology will be what we have in mind. Thus when a sequence $\{x_n\}$ converges to a point x in H we will write:

$$x_n \to x \quad \text{as} \quad n \to \infty$$

if norm convergence is intended $(\lim_{n\to\infty} \|x_n - x\| = 0)$.

Whereas if $\{x_n\}$ converges weakly to x we will write:

$$x_n \overset{w}{\to} x \quad \text{as} \quad n \to \infty$$

meaning $\lim_{n\to\infty} |\ell(x_n) - \ell(x)| \to 0 \quad \forall\, \ell \in H^*$.

PROPERTIES OF NON-LINEAR FUNCTIONALS ON H

We will be interested in various properties of non-linear functionals defined on H, the most important of these being convexity, continuity and differentiability.

CONVEXITY

Definition 1.A.1. A set C in H is *convex* if $\forall\ x,y \in C$

$$\lambda x + (1-\lambda)y \in C \quad 0<\lambda<1.$$

Definition 1.A.2. A functional f on H is *convex* if $\forall\ x,y \in H$ and $0<\lambda<1$

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y).$$

f is strictly convex if the above inequality is strict whenever $x \neq y$.

Definition 1.A.3. The *epigraph* of a functional f on H is the set:

$$\text{epi } f = \{(u,a) \in H \times R\colon f(u) \leq a\}\ .$$

It is the set of points in $H \times R$ which lie above the graph of f. The notions of convex sets and convex functions are intimately related; a function is convex if and only if it's epigraph is convex. One important property of convex functions on finite-dimensional spaces is the following:

Lemma 1.A.1. Suppose H is finite dimensional. If f is strictly convex and bounded below and has a minimum in H then $f(x) \to \infty$ as $\|x\| \to \infty$ (i.e. f is coercive on H).

Proof: See Rockafellar (1970).

Definition 1.A.4. A function f is *quasi-convex* if $\forall\ x,y \in H$ and $0 < \lambda < 1$

$$f(\lambda x + (1-\lambda)y) \le \max\ (f(x), f(y))$$

f is strongly quasi-convex if the above inequality is strict whenever $x \neq y$.

It is easy to see that if f is convex or quasi-convec then a unique local minimizer of f is the unique global minimizer of f. Also if f is strictly convex or strongly quasi-convex any minimizer of f is the unique global minimizer.

Definition 1.A.5. A function f is *uniformly convex* on H if $\forall\ x \neq y$

$$f(\frac{x+y}{2}) \le \tfrac{1}{2}f(x) + \tfrac{1}{2}f(y) - \delta(\|x-y\|)$$

for some real-valued, continuous, monotone function δ; $\delta(t) \ge 0$ for $t \ge 0$ with $\delta(t) = 0$ if and only if $t = 0$.

Local versions of convexity, quasi-convexity, uniform convexity etc. can also be defined. For instance f is locally convex at $x \in H$ if f is convex in some neighborhood of x. We get the following lemma.

Lemma 1.A.2. If f is a convex function and x^* is a minimizer of f then if f is locally uniformly convex (or locally strictly convex) at x^* then x^* is the unique global minimizer of f .

CONTINUITY AND LOWER SEMI-CONTINUITY

Definition 1.A.6. A function f is (weakly) *continuous* at $\bar{u}$ if for all sequences, $\{u_n\}$, (weakly) converging to $\bar{u}$, $\lim_{n \to \infty} f(u_n) = f(\bar{u})$.

Definition 1.A.7. A function f is (weakly) *lower semi-continuous* on H if it satisfies one of the two equivalent conditions:

$\forall\, a \in R \quad \{u \in H : f(u) \leq a\}$ *is (weakly) closed.*

or

$$\bar{u} \in H, \quad \liminf_{n} f(u_n) \geq f(\bar{u})$$

for all sequences (weakly) converging to $\bar{u}$. A further characterization of lower semi-continuity is also available.

Theorem 1.A.3. A function $f : H \to R$ is *(weakly)* lower semi-continuous if and only if it's epigraph is *(weakly)* closed.

The concept of lower semi-continuity plays an important role in minimization problems because of the following result.

Theorem 1.A.4. Let f be a weakly lower semi-continuous functional defined on a weakly compact set C; then there exists x_0 in C such that $f(x_0) = \inf_{x \in C} f(x)$.

Proof:

Let $m = \inf_{x \in C} f(x)$; then there exists a sequence $\{x_n\}$ in C such that $\lim_{n \to \infty} f(x_n) = m$. Since C is weakly compact, there is a weakly convergent subsequence $\{x_{n_i}\}$ of $\{x_n\}$: $x_{n_i} \overset{w}{\to} x_0$ for some x_0 in C. But by weak lower semi-continuity and the fact that $\lim_{n_i \to \infty} f(x_{n_i}) = m$ we have

$$f(x_0) \le \liminf_{n_i \to \infty} f(x_{n_i}) = m \le f(x_0)$$

Thus $f(x_0) = m$. □

The above theorem has two important corollaries.

Corollary 1.A.5. A weakly lower semi-continuous functional achieves its infimum on every closed and bounded convex subset of H.

Proof:

In a real Hilbert space, every closed convex set is weakly closed (see Ekeland and Teman (1973) pp. 3-7) while a weakly closed and bounded set is weakly compact. Therefore Theorem 1.A.4 applies. □

Corollary 1.A.6. Let f be a weakly lower semi-continuous functional on an unbounded closed convex set C.

Suppose $\lim_{\|x\|\to\infty} f(x) = \infty$ in C. Then f achieves its infimum on C.

Proof:

If $f(x) = \infty \ \forall \ x \in C$ then we are done. Otherwise for M large enough $\inf_{C} f = \inf_{C\cap MB} f$, where B is the unit ball. The result now follows by corollary 1.A.5 since $C\cap MB$ is a closed and bounded convex set. □

FRECHET DIFFERENTIABILITY

Let f be a mapping from a Hilbert space H_1 (norm $\|\cdot\|_1$) into a Hilbert space H_2 (norm $\|\cdot\|_2$).

Definition 1.A.8. f is Frechet *differentiable* at $x \in H_1$ if there exists a bounded linear operator $f'(x) \in L(H_1,H_2)$ satisfying:

$$\lim_{\|y\|_2\to 0} \frac{\|f(x+y) - f(x) - f'(x)y\|_1}{\|y\|_2} = 0$$

$f'(x)$ is called the Frechet derivative of f at x. Higher order Frechet derivatives are analogously defined; the second Frechet derivative of f at x, $f''(x)$, is an element of $L(H_1,L(H_1,H_2))$ etc.

Versions of the usual mean value theorems apply to Frechet differentiable mappings. A couple of these results, which we will use later on, are worth noting.

Theorem 1.A.7. Let $f : D \subset H \to R$, and assume that f has a second Frechet derivative at each point of a convex set $D_0 \subset D$. Then, for any $x,y \in D_0$, there is a $t \in [0,1]$ such that

$$f(y) - f(x) - f'(x)(y-x) = \tfrac{1}{2}f''(x+t(y-x))(y-x)(y-x) .$$

Theorem 1.A.8. Assume that $f : D \subset H \to R$, has a second Frechet derivative at $x \in D$. Then

$$\lim_{\|h\| \to 0} (1/\|h\|^2)[f(x+h) - f(x) - f'(x)h - \tfrac{1}{2}f''(x)hh] = 0$$

for any $h \in H$.

Definition 1.A.9. A function f on H is *uniformly positive definite* if for some positive constant M we have:

$$f''(x)hh \geq M\|h\|^2$$

for all $x,h \in H$.

Functions which are uniformly positive definite on H clearly satisfy: $f(x) \to \infty$ as $\|x\| \to \infty$.

THE GRADIENT AND HESSIAN OF A FUNCTIONAL

The usual notions of gradients and hessians of functionals on R^n generalize to real Hilbert spaces. Suppose $g : H \to R$ is a functional on H. The first and second Frechet derivatives of g at a point x in H are denoted by $g'(x)$ and $g''(x)$ respectively. Now since $g'(x) \in H^*$ and $g''(x) \in L(H,H^*)$, by the isometric isomorphism between H and H^*, $\exists \cdot \nabla g(x) \in H$ and $H_g(x) \in L(H,H)$ with the properties that:

$$<\nabla g(x),y> = g'(x)y \ \forall \ y \in H$$

$$\|\nabla g(x)\| = \|g'(x)\|_*$$

and

$$<H_g(x)y,z> = g''(x)yz \ \forall \ y,z \in H$$

$$\|H_g(x)y\| = \|g'(x)y\|_* \ \forall \ y \in H$$

In R^n $\nabla g(x)$ and $H_g(x)$ correspond to the gradient and hessian, respectively.

CHAPTER 2

EXISTENCE THEORY

2.1 INTRODUCTION

In this chapter we give conditions under which the Penalized Likelihood estimators will exist. Attention will be restricted to the case where the penalty functional J is, at least, convex (indeed, more often than not, $J(x) = \|Px\|^2$ where P is a projection operator with finite dimensional null space). One of the main results, Theorem 2.2.5 and its corollary, gives conditions when the existence of a minimizer of $Q(x|z)$ guarantees the existence of the Penalized Likelihood estimator; this related to Silverman's 1982 Annals of Statistics conjecture. We begin with some general considerations and go on to discuss some applications in section 3. The major references for this chapter are Daniel (1971), Ekeland and Teman (1973) and Ortega and Rheinbold (1970). The notation introduced in Chapter 1 will be employed throughout.

2.2 SOME GENERAL CONSIDERATIONS

Consider the functional:

$$I_\lambda(x) = Q(x) + \lambda J(x), \quad \lambda>0.$$

Our program is to find sufficient conditions for there to exist minimizers of I_λ in closed convex unbounded subsets C of H. The next theorem, a restatement of corollary 1.A.6., clarifies this problem somewhat.

Theorem 2.2.1.

Let C be a closed convex subset of a Hilbert space H. Suppose $I_\lambda : C \to R$ is coercive (i.e. $I_\lambda(x) \to \infty$ as $\|x\| \to \infty$, $x \in C$) and is weakly lower semi-continuous on C, then I_λ attains its infimum on C.

Proof: [See corollary 1.A.6]

In the context of this theorem we now ask ourselves what conditions must the functionals Q and J satisfy in order that

(i) I_λ is weakly lower semi-continuous on C?

(ii) I_λ is coercive on C?

The first of these can be answered as follows.

Theorem 2.2.2.

Suppose C is a closed convex subset of H. Let Q be weakly continuous on C, and J (bounded below) continuous and convex on C. Then I_λ is weakly lower semi-continuous.

Proof:

J is continuous and convex so it is weakly lower continuous (see the appendix of Chapter 1). Now since $\lambda > 0$

and Q is weakly lower semi-continuous (it is weakly continuous) it follows that $Q + \lambda J$ is weakly lower semi-continuous. □

Conditions guaranteeing the coerciveness of I_λ on general convex sets C are not as easy to identify. However, when Q and J are well behaved, and C satisfies some mild conditions, the problem becomes more tractable. Suppose H can be split into two subspaces, H_0 and H_1, with $H = H_0 \oplus H_1 : H_0$ is finite-dimensional. Let $J(x) = \|Px\|^2$ where P is a projection operator on H with null space H_0 and range H_1. We consider subsets C of H which satisfy the following property:

Definition 2.2.3.

A convex set C in $H_0 \oplus H_1$ satisfies *property 1* if for any sequence in C for which $\|(I-P)x_n\| \to \infty$ and $\{\|Px_n\|\}$ is bounded $\exists$ sequences $\{\mu_n\} \subset [0,1]$, $\{u_n\} \subset H_0$, $\{y_n\} \subset C$ and $z \in H$ for which

$$u_n + z = \mu_n x_n + (1-\mu_n)y_n$$

with $\|u_n\| \to \infty$ and $\{y_n\}$ bounded.

Definition 2.2.4.

A convex set C in $H_0 \oplus H_1$ satisfies *property 2* if it satisfies property 1 and the choice of z does not depend on the sequence $\{x_n\}$.

Convex sets satisfying property 1 are common in applications; for instance in $W_2^m[a,b]$ with $H_0 = \{x \in W_2^m[a,b]: x^{(m)} \equiv 0\}$ the sets $C_{q,\alpha} = \{x \in W_2^m[a,b]: x^{(q)} \geq \alpha\}$, $0 \leq q < m$ and α some continuous function. To see this let $\{x_n\} \subset C_{q,\alpha}$ be such that $\|Px_n\|^2 = \int_a^b [x_n^{(m)}(t)]^2 dt$ is bounded and $\|(I-P)x_n\| \equiv \|P_0 x_n\|^2 \to \infty$. Since $\|Px_n\|$ is bounded, $q < m$ and α is continuous, $\exists$ a polynomial, $M(t)$, such that $-M^{(q)}(t) < (Px_n)^{(q)}(t) < M^{(q)}(t)$ $t \in [a,b]$, and $\alpha < M$. Let $u_n = \frac{1}{2} P_0 x_n$, $y_n = 2M - Px_n$ and $z = 2M$.

By definition, $\{y_n\}$ is a bounded sequence in $C_{q,\alpha}$, and since

$$u_n + z = \frac{1}{2} x_n + \frac{1}{2} y_n$$

we have that $C_{q,\alpha}$ indeed satisfies property 1.

Theorem 2.2.5.

Let C be a convex subset of H satisfying property 1. Suppose Q satisfies the following conditions:

(i) coercive on $[x + H_0] \cap C$, for $x \in H$

(ii) quasi-convex on C

(iii) bounded on bounded subsets of C

(iv) bounded below on C

then I_λ is coercive on C.

Proof:

$$I_\lambda(x) = Q(x) + \lambda\|Px\|^2$$

If I_λ is not coercive on C then there is a sequence $\{x_n\}$ in C for which $\{I_\lambda(x_n)\}$ is bounded but $\{x_n\} \to \infty$ as $n \to \infty$.

Case 1: $\{x_n\}$ has a subsequence, $\{x_{n_k}\}$, for which $\|Px_{n_k}\| \to \infty$.

Here, since Q is bounded below on C, we have that $I_\lambda(x_{n_k}) = Q(x_{n_k}) + \lambda\|Px_{n_k}\|^2 \to \infty$, contradicting the definition of $\{x_n\}$.

Case 2: $\{\|Px_n\|\}$ is bounded.

C satisfies property 1, therefore $\exists$ sequences $\{\mu_n\} \subset [0,1]$, $\{u_n\} \subset H_0$, $\{y_n\} \subset C$ and $z \in H$ for which

$$u_n + z = \mu_n x_n + (1-\mu_n)y_n$$

with $\|u_n\| \to \infty$ and $\{y_n\}$ bounded. Since Q is quasi-convex we have that:

$$Q(u_n + z) \le \max(Q(x_n), Q(y_n))$$

Since Q is coercive on $[z + H_0] \cap C$, and bounded on bounded subsets of C it follows that $Q(x_n) \to \infty$ which of course contradicts the definition of $\{x_n\}$. □

Corollary 2.2.6.

Let C be a convex set satisfying property 1 and suppose Q is bounded below and bounded on bounded subsets of C. If Q is strictly convex on C and has minimizers on subsets of C of the form $[x+H_0] \cap C \quad x \in H$, then I_λ is coercive on C.

Proof:

Since Q is strictly convex and has a minimum on $[x + H_0] \cap C \quad Q$ is coercive on sets of this form (see Lemma 1.A.1). □

The above results obviously hold when the set C satisfies property 2: the coerciveness of Q need then only be checked on $[z + H_0] \cap C$ for a fixed z in H.

2.3 APPLICATIONS

Consider the following set up $H = H_0 \oplus H_1$ with H_0 finite-dimensional, P is a projection operator on H with range space H_1 and null space H_0. $J(x) = \|Px\|^2$, and L_i/N_i are continuous linear/non-linear functionals on H. Finally, C is a closed convex subset of H satisfying property 1 of section 2. Within this framework we investigate the existence of minimizers of the I_λ's arising in a few specific models.

(i) SILVERMAN'S PENALIZED LIKELIHOOD DENSITY ESTIMATES

Given observations $a \le X_1 \le X_2 \le \cdots \le X_n \le b$ Silverman's (1982) penalized likelihood estimate for the log density of the X_i's is prescribed to be the minimizer of

$$A_0(g) = -n^{-1} \sum_{i=1}^{n} g(X_i) + \lambda \int_a^b [g^{(m)}(t)]^2 dt$$

over $C_1 = \{g \in W_2^m[a,b]: \int_a^b e^{g(t)} dt = 1\}$ $m \ge 1$. It can be shown (see Silverman 1982) that minimizing A_0 over C_1 is equivalent to minimizing

$$A(g) = -n^{-1} \sum_{i=1}^{n} g(X_i) + \int_a^b e^{g(t)} dt + \lambda \int_a^b [g^{(m)}(t)]^2 dt$$

over $C = \{g \in W_2^m[a,b]: \int_a^b e^{g(t)} dt < \infty\}$.

Since $m \ge 1$, if $g \in W_2^m[a,b]$ then $|g|$ is bounded and so $\int_a^b e^{g(t)} dt$ is finite. Consequently $C = W_2^m[a,b]$. It follows trivally that C satisfies property 2 of section 2 with $z \equiv 0$.

A is strictly convex and bounded below by zero on C. Since evaluation is continuous on W_2^m for $m \ge 1$ and since $\int_a^b e^g dt$ is weakly continuous (by continuity of e^x), $-n^{-1} \sum_{i=1}^{n} g(X_i) + \int_a^b e^{g(t)} dt$ is weakly continuous and consequently bounded on bounded subsets of C. Therefore, by Theorem 2.2.2, A is weakly lower semi-continuous and so by the Corollary 2.2.6 A has a minimizer in C provided $-n^{-1} \sum_{i=1}^{n} g(X_i) + \int_a^b e^{g(t)} dt$ has a minimizer in

$\{1,t,t^2,\ldots,t^{m-1}\}$. This provides an alternative justification of Theorem 4.1 in Silverman's (1982) paper.

(ii) GENERALIZED LINEAR MODELS

Wedderburn (1976) discusses the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models over finite-dimensional parameter spaces. Here we use some of his results to discuss twO particular models in some detail.

Log-linear Poisson regression

This model is especially useful for studying systems in which the measurement process is subject to shot noise (see Nicot and Yandell (1983) for an interesting application involving the mapping of fungal activity in potato fields). Barndorff-Nielsen (1978) discusses a parametric regression form of the model (see also Kalbfleisch and Sprott 1974). The data consist of independent counts $\{y_i,\ i = 1,2,\ldots,n\}$. Each $y_i \sim P(\mu_i)$ and μ_i is modeled log-linearly as:

$$\mu_i = e^{L_i x} \quad i = 1,2,\ldots,n$$

where L_i's are continuous linear functionals of x. Thus the likelihood of x is given by:

$$\Phi(x) = \prod_{i=1}^{n} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

from which it follows that I_λ has the form:

$$I_\lambda(x) = \sum_{i=1}^{n} \{e^{L_i x} - y_i L_i x + \log(y_i!)\} + \lambda\|Px\|^2$$

Let $\dim(H_0) = m$; then $\exists\{\phi_j\}_{j=1}^{m} : H_0 = \text{span}\ \{\phi_1, \phi_2, \ldots, \phi_m\}$. Now given $x \in [z + H_0]$

$$x = \sum_{j=1}^{m} \beta_j \phi_j + z$$

Let $B = \{\beta \in R^m : \sum_{j=1}^{m} \beta_j \phi_j + z \in [z + H_0] \cap C\}$. B is a closed convex subset of R^m. Let T be a matrix with ij^{th} element $L_i \phi_j$ $i = 1,2,\ldots,n$ $j = 1,2,\ldots,m$, then

$$L_i x = T_i' \beta + L_i z$$

where $T_i = (L_i\phi_1, L_i\phi_2, \ldots, L_i\phi_m)'$. If rank $(T) \geq m$ then it follows easily that I_λ is strictly convex on H. Also, since the L_i's are continuous linear functionals, I_λ is bounded on bounded subsets of H and so, by Theorem 2.2.2., I_λ is weakly lower semi-continuous ($\|Px\|^2$ is continuous and convex and $\sum_{i=1}^{n} \{e^{L_i x} - y_i L_i x + \log(Y_i!)\}$ is weakly continuous and bounded below). Consequently, by the corollary to Theorem 2.2.5, I_λ will achieve its minimum on C if, dropping extraneous terms, $\sum_{i=1}^{n} \{e^{L_i x} - y_i L_i x\}$ has a minimum in $[z + H_0] \cap C$ for any z in C. Now $\sum_{i=1}^{n} \{e^{L_i x} - y_i L_i x\}$ has a minimum in $[z + H_0] \cap C$ if and only if $Q(\beta) = \sum_{i=1}^{n} \{e^{T_i'\beta} - y_i^* T_i' \beta\}$ has a minimum in B, where $y_i^* = y_i / e^{L_i z}$.

Therefore we need only give conditions under which Q has a minimum in B.

Let T^0 be the matrix obtained from T by removing the rows corresponding to zero y_i counts. Following Wedderburn (1976) we can deduce:

Lemma 2.3.1.

If rank $(T) \geq m$ and the row space of T^0 is the same as the row space of T, then $Q(\beta)$ has a minimum in B.

Proof:

Q is convex and continuous on B, so, we are done once we establish that Q is coercive on B. Suppose $\|\beta\| \to \infty$ in B. Then since T is of full rank $\|T\beta\| \to \infty$, but this in turn implies that $\|T^{0'}\beta\| \to \infty$, that is $\sum_{i:y_i \neq 0} \{e^{T_i'\beta} - y_i^* T_i'\beta\} \to \infty$. It follows, arguing by contradiction, that $Q(\beta) \to \infty$. □

Therefore, under the hypotheses of the above lemma, I_λ achieves its minimum in C. The uniqueness of this minimum follows by the strict convexity of I_λ on C.

The logistic regression model

Cox (1970) discussed this model as a special case of more general quantal response models. The data in these models (in the form of y_i individuals responding out of n_i) are considered as independent binomial variates whose means are functionally related via a common regression. In the logistic case we observe $\{y_i, i = 1,2,...,n\}$ where $y_i \sim B(n_i,p_i)$ and the logits are written as:

$$\log [\frac{p_i}{1-p_i}] = L_i x \quad i = 1,2,...,n$$

Silverman (1978), Raynor and Bates (1983) and Villalobos (1983) have considered this model useful in the analysis of discrimination problems. The likelihood of x given the observed data is:

$$\Phi(x) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{n_i-y_i}$$

so that I_λ has the form:

$$I_\lambda(x) = \sum_{i=1}^{n} [n_i \log (1+e^{L_i x}) - y_i L_i x] + \lambda\|Px\|^2$$

Differentiating w.r.t. x we get for any $h \in H$:

$$I_\lambda''(x)hh = \sum_{i=1}^{n} n_i \frac{e^{L_i x}}{(1+e^{L_i x})^2} (L_i h)^2 + 2\lambda\|Ph\|^2$$

$$= 0 \text{ iff } \sum_{i=1}^{n} (L_i h)^2 = 0 \text{ and } Ph = 0. \qquad (*)$$

Let T be defined as in the Poisson model and suppose we are interested in minimizing I_λ over the whole of H. Equations (*) imply that I_λ is strictly convex provided rank $(T) \geq m$. I_λ is obviously bounded below, the continuity of the L_i's implies that I_λ is weakly lower semi-continuous and hence bounded on bounded subsets of H. It follows from Corollary 2.2.6 that when rank (T) ≥ m I_λ has a minimizer in H whenever $\sum_{i=1}^{n} [n_i \log (1+e^{T_i'\beta}) - y_i T_i'\beta]$ has a minimum in R^m. If the rank (T) ≥ m then $\sum_{i=1}^{n} [n_i \log (1+e^{T_i'\beta}) - y_i T_i'\beta]$ on R^m constitutes a regular exponential family distribution of order m (see Barndorff-Nielsen 1978) and using the Barndorff-Nielsen theory of such families we obtain the following result.

Lemma 2.3.2.

Let $S = \{s = (s_1, \ldots, s_\nu, \ldots, s_m) : s_\nu = \sum_{i=1}^{n} y_i L_i \phi_\nu;$ $y_i = 0, 1, \ldots, n_i\}$ and C_s be the closure of the convex hull of S. If rank (T) ≥ m then $\sum_{i=1}^{n} [n_i \log (1+e^{T_i'\beta}) - y_i T_i'\beta]$ has a minimum in R^m if and only if the observed value of s lies in the interior of C_s.

Proof:

The condition on the rank of T guarantees that $\sum_{i=1}^{n} [n_i \log (1+e^{T_i'\beta}) - y_i T_i'\beta]$ constitutes a regular exponential family distribution of order m and the result follows by appealing to Corollary 0.6, p. 153 of Barndorff-Nielsen (1978). □

Therefore, under the hypotheses of Lemma 2.3.2, I_λ has a minimizer in H. The uniqueness of this minimizer follows from the strict convexity of I_λ on H. For the case in which L_i's are evaluations and the n_i's are all 1's, Silverman (1978) and Villalobos (1983) have obtained a more intuitive characterization of the result in Lemma 2.3.2.

(iii) Cox's Proportional Hazard Regression Model

Suppose the explanatory variables $x_1,x_2.,\ldots,x_n$ lie in [a,b]. Following section 2 of Chapter 1, let the Penalized (Partial) Likelihood estimator of the proportional hazard regression function, θ, be the minimizer of the functional

$$I_\lambda(\theta) = \sum_{i\in U} \{\log[\sum_{j\in R_{(i)}} e^{\theta(x_j)}] - \theta(x_i)\} + \lambda\int_a^b [\theta^{(m)}(x)]^2 dx$$

over $C = \{\theta - \theta(a),\ \theta \in W_2^m[a,b]\}$ for $m \geq 1$. (The -log {partial likelihood} evaluated at $\theta - \theta(a)$ is equal to the - log {partial likelihood} evaluated at θ. Hence, the motivation minimizing I_λ over C.) Clearly C satisfies property 2 of section 2 with "$z \equiv 0$". Also I_λ is bounded below by zero and since evaluation is continuous ($m \geq 1$) the -log {partial likelihood} is weakly continuous.

With $p_{ij}(\theta) = \dfrac{e^{\theta(x_j)}}{\sum_{k\in R_{(i)}} e^{\theta(x_k)}}$ we have for any $\theta,\phi \in C$:

$$I_\lambda''(\theta)\phi\phi = \sum_{i\in U} [\sum_{j\in R_{(i)}} p_{ji}(\theta)\phi^2(x_j)$$

$$- [\sum_{j\in R_{(i)}} p_{ji}(\theta)\phi(x_j)]^2]$$

$$+ 2\lambda\|P\phi\|^2 \geq 0$$

$$= 0 \quad \text{iff} \quad \phi(x_j) = 0 \quad \text{for} \quad j \in \bigcup_{i\in U} R_{(i)}$$

$$\text{and} \quad P\phi = 0.$$

It follows that I_λ is strictly convex on C provided there are at least m distinct x_j's in $\bigcup_{i\in U} R_{(i)}$. Hence, by the standard arguments, I_λ has a minimum in C provided there exist $\beta \in R^{m-1}$ to minimize the finite dimensional -log {partial likelihood}:

$$Q(\beta) = \sum_{i\in U} \{\log [\sum_{j\in R_{(i)}} e^{T_j\beta}] - T_i\beta\}$$

where $T_i = (x_i, x_i^2, \ldots, x_i^{m-1})$.

(iv) THE NON-LINEAR REGRESSION MODEL

The penalized likelihood functional, I_λ, is given by

$$I_\lambda(x) = \sum_{i=1}^{n} [z_i - N_i(x)]^2 + \lambda\|Px\|^2$$

where N_i's are non-linear functionals. If the N_i's are weakly continuous, then I_λ is weakly lower semi-continuous and bounded on bounded subsets of H.

We need to work harder here to develop meaningful conditions under which I_λ is coercive on H. The conditions which we come up with are strongly influenced by our expierence with the radiative transfer equations (see Chapter 5 and Appendix B). We begin by considering *comparison* functionals I_Φ defined on C in the following manner.

$$I_\Phi(x) = \Phi(N_1(x), N_2(x), \ldots, N_n(x)) + \|Px\|^2 \qquad (**)$$

where $\Phi : R^n \to R$ is bounded below and satisfies $\Phi(x_1, x_2, \ldots, x_n) \leq \Phi(x_1^*, x_2^*, \ldots, x_n^*)$ whenever $|x_i| \leq |x_i^*|$ $\forall\, i, i = 1, 2, \ldots, n$.

These comparison functionals will be used to establish the coerciveness of I_λ for certain classes of non-linear functionals N_i. Whenever I_Φ is coercive then so is the corresponding I_λ; the next lemma is needed to prove this assertion.

Lemma 2.3.3.

If $\{x_k\}$ is any sequence in R^n for which $\Phi(x_k) \to \infty$ as $k \to \infty$ then $\|x_k\|_{E_n}$ (the Euclidean norm of x_k) also tends to infinity as $k \to \infty$.

Proof:

Suppose $\exists$ a bounded subsequence of $\{x_k\}$, $\{x_{k_j}\}$ say, then by definition of Φ, $\{\Phi(x_{k_j})\}$ must also be bounded. However, this contradicts the fact that $\Phi(x_k) \to \infty$ as $k \to \infty$. Therefore, $\forall\ M\ \exists\ k_M$ such that $\forall\ k \geq k_M$ $\|x_k\| \geq M$ i.e. $\lim_{k\to\infty} \|x_k\| = \infty$. □

Theorem 2.3.4.

Let C be an unbounded convex set and suppose I_Φ given by (**) is coercive on C, then I_λ is also coercive on C.

Proof:

Suppose not, then for some $M > 0$ $\exists\,\{x_\ell\} \subseteq C$ with the property that

$$I_\lambda(x_\ell) \leq M \quad \text{for} \quad \|x_\ell\| > \ell \quad \ell = 1,2,\ldots$$

For this sequence we must have, by definition of I_λ, that

$$\|Px_\ell\| \leq \frac{1}{\lambda} M \ \forall\ \ell .$$

However, the coerciveness of I_Φ on C implies that

$$\Phi(N_1(x_\ell), N_2(x_\ell), \ldots, N_2(x_\ell)) \to \infty \quad \text{as} \quad \ell \to \infty$$

Hence by Lemma 2.3.3.

$$\sum_{i=1}^{n} [N_i(x_\ell)]^2 \to \infty \Rightarrow \sum_{i=1}^{n} [z_i - N_i(x_\ell)]^2 \to \infty \Rightarrow I_\lambda(x_\ell) \to \infty$$

This contradicts the definition of $\{x_\ell\}$. Therefore $\forall\ M\ \exists\ \ell$ such that $\forall\ x \in C$ with $\|x\| > \ell$ $I_\lambda(x) \geq M$ i.e. I_λ is coercive on C. □

We finish this section by presenting a theorem which provides a set of conditions under which the functionals N_i yield coercive I_λ's on C. This result will be used to study the existence of smooth solutions to the ill-posed problem associated with the radiative transfer equations (see Appendix B).

Theorem 2.3.5.

Let C be a convex subset of H satisfying property 1. Suppose $\exists$ $\phi : R \to R$ which is monotonic increasing (and bounded below) in the modulus of its argument. Suppose further that

(s.1) $\sum\limits_{i=1}^{n} \phi[N_i(x)]$ is strictly convex on C

(s.2) $\sum\limits_{i=1}^{n} \phi[N_i(x)]$ has a proper minimum on

$$[x + H_0] \cap C,\ x \in H.$$

Then I_λ is coercive on C.

Proof:

Let $\Phi(x) = \sum\limits_{i=1}^{n} \phi(x_i)$. Then Φ clearly satisfies properties (**) above. Next define I_Φ as follows:

$$I_{\Phi}(x) = \sum_{i=1}^{n} \phi[N_i(x)] + \|Px\|^2$$

By (s.1) I_{Φ} is convex on C and since the N_i's are weakly continuous, I_{Φ} is bounded on bounded subsets of H. Finally, since $\sum_{i=1}^{n} \phi[N_i(x)]$ is convex and has a proper minimum in $[x + H_0] \cap C\ x \in H$, it follows that I_{Φ}, is coercive on C and so by Theorem 2.3.4. that I_{λ} is coercive on C . □

CHAPTER 3

NUMERICAL ALGORITHMS

3.1 INTRODUCTION

This chapter discusses some numerical methods for minimizing the penalized likelihood functionals. Recall that penalized likelihood functionals have the form:

$$I_{\lambda}(x) = Q(x|z) + \lambda J(x), \quad \lambda>0 \text{ .}$$

The goal is to develop effective methods for computing minimizers of such functionals over classes of functions of practical interest. Now this task, in even *simple* cases, such as the linear regression model with a quadratic penalty, is not completely trivial. When $Q(x|z)$ is not quadratic these numerical difficulties can be greatly magnified. Our discussion will restrict itself to the cases where the penalty term, J, is quadratic in x. Extensions to uniformly positive definite penalties should be fairly obvious.

To begin with, many numerical minimization algorithms proceed by successively obtaining quadratic approximations to the objective function; the minima of these quadratics are used to define a sequence of iterates which, one hopes, will converge to a minimizer of the objective function.

Due care has to be taken to insure that the successive quadratic approximants are computationally tractable, i.e. their minimizers are easily obtainable. This consideration is especially important for infinite dimensional problems. We restrict ourselves to fixed-step size Newton-like schemes. Iterates, $\{x^k\}$, are defined by:

$$x^{k+1} = x^k - [A(x^k)]^{-1} \Delta I_\lambda(x^k) \qquad k=0,1,\ldots \qquad (*)$$

where $A(x^k) \in L(H,H)\ \forall\ k$. Obvious choices for A come to mind. The standard Newton choice is:

$$\begin{aligned} A(x^k) &= H_{I_\lambda}(x^k) \\ &= H_Q(x^k) + \lambda H_J(x^k) \end{aligned}$$

where H denotes the Hessian. With Fisher's Scoring Technique $A(x^k)$ becomes:

$$\begin{aligned} A(x^k) &= E\{H_{I_\lambda}(x^k)\} \\ &= E\{H_Q(x^k)\} + \lambda H_J(x^k)\ . \end{aligned}$$

Here expectation is taken assuming x^k is the true value of x.

Finally if $Q(x|z) = (F(x|z), F(x|z))$ where $(\cdot,\cdot)$ is some inner product on the range space of F, $A(x^k)$ might be chosen according to a Gauss-Newton formulation i.e.:

$$A(x^k) = (\nabla F(x^k|z), \quad \nabla F(x^k|z)) + H_J(x^k).$$

Our reason for favouring schemes of the above type is that these methods can often be easily meshed with existing finite-dimensional software. The final section of this chapter outlines the implementation of such algorithms in Generalized Linear Interactive Models and Non-linear Regression Model contexts.

As with any iterative algorithm one is naturally interested in knowing convergence properties. Does the process converge to a minimizer of I_λ? If so, what is the rate of convergence? Our goal in this chapter is to show that some basic results concerning the convergence of Newton-like schemes (*), which have already been established in finite-dimensional situations, can be generalized to handle the broad class of minimization problems associated with penalized likelihood methods. Results about convergence properties are presented in the next section. Some specifics on implementation of the techniques for Generalized Linear Models and Non-Linear Regression Models are outlined in section 3 of this chapter. In Chapter 5, a non-linear integral equation arising in Satellite Meteorology is numerically inverted using the algorithm discussed here - the method appears to work quite well in practice.

The major reference for this Chapter is Ortega and

Rheinbold (1970).

3.2 CONVERGENCE PROPERTIES OF THE ALGORITHMS

Asymptotic Root Convergence Factors

Before getting into the convergence analysis, we first need to introduce a way of measuring the asymptotic rate of convergence of a convergent sequence.

Definition 3.2.1.

Suppose $\{x_k\}$ is a sequence in H converging to a point x^*. The *asymptotic root convergence factors*, $R_p\{x^k\}$, of the sequence are given by:

$$R_p\{x^k\} = \begin{cases} \limsup\limits_{k \to \infty} \|x^k - x^*\|^{1/k} & \text{if } p=1 \\ \limsup\limits_{k \to \infty} \|x^k - x^*\|^{1/p^k} & \text{if } p>1 \end{cases}$$

Notice that whenever $R_p\{x^k\} < 1$, then, for any $\varepsilon > 0$ with $R_p\{x^k\} + \varepsilon < 1$, there is a $k_0 \geq 0$, such that $\forall\ k \geq k_0$ either

$$\|x^k - x^*\| \leq (R_p\{x^k\} + \varepsilon)^{p^k} \quad \text{if } p>1$$

or

$$\|x^k - x^*\| \leq (R_1\{x^k\} + \varepsilon)^k \quad \text{if } p=1 .$$

In the latter case, the convergence of the sequence is at least as rapid as a geometric progression with ratio

$R_1 + \varepsilon < 1$. If $0 < R_1\{x^k\} < 1$, the convergence is called R-linear, while for $R_1\{x^k\} = 1$ or $R_1\{x^k\} = 0$ convergence is R-sublinear and R-superlinear, respectively. Similarly, if

$$0 < R_2\{x^k\} < 1$$

we say that the convergence is R-quadratic.

More detailed information about this measure of asymptotic rate of convergence can be found in chapter 9 or Ortega and Rheinbold (1970).

Convergence Analysis

We now turn to the analysis of the algorithms. Throughout this discussion the penalty term $J(x) = \|Px\|^2$ where P is, as usual, a projection operator on H with finite-dimensional null space. The objective functionals therefore have the form:

$$I_\lambda(x) = Q(x) + \lambda\|Px\|^2 \quad \lambda>0 \quad \text{and} \quad x \in H \tag{3.2.1}$$

Q will be assumed to be bounded below. We wish to find a minimizer of I_λ in a closed convex subset C of H. Let x^k be the current estimate and x^{k+1} the update. We consider iterative schemes of the form:

$$x^{k+1} = x^k - [A(x^k)]^{-1}\nabla I_\lambda(x^k) \quad k=0,1,\ldots \tag{3.2.2}$$

where $A(x^k) \in L(H,H)\ \forall\ k$.

In this section we prove the following result about the convergence of iterations of the form (3.2.2.).

Theorem 3.2.2.

Suppose Q in equation (3.2.1) is such that Q'' is continuous and Q' is weakly continuous on the interior of C. Let $x_\lambda^0 \in \text{int } C$ be such that the level set

$$L^0 = \{x \mid I_\lambda(x) \le I_\lambda(x_\lambda^0)\}$$

is weakly compact and I_λ has only finitely many critical points in L^0. Assume that $A: L^0 \to L(H,H)$ is a continuous mapping such that $\exists$ an invertible map $B(x) \in L(H,H)$ with $A(x) = B^*(x)B(x)$ (* denotes the adjoint) $\forall\ x \in L^0$. Suppose also $\exists\ \mu_0, \mu_1$ and γ_1 all positive with $0 < \frac{1}{2}\gamma_1 < \mu_0$ satisfying

$$\mu_0\|h\|^2 \le \langle h, A(x)h\rangle \le \mu_1\|h\|^2$$

and

$$I_\lambda''(x)hh \le \gamma_1\|h\|^2$$

$\forall\ x \in L^0$ and $h \in H$. Consider the process (3.2.2) i.e.

$$x^{k+1} = x^k - [A(x^k)]^{-1}\nabla I_\lambda(x^k) \qquad k=0,1,\ldots$$

Then $\{x^k\} \in L^0$, $\lim_{k\to\infty} x^k = x^*$, where $\nabla I_\lambda(x^*) = 0$, and if

$H_{I_\lambda}(x^*)$ is nonsingular, then the convergence is at least R-linear.

Comment.

A finite-dimensional version of this Theorem appears in Chapter 14 of Ortega and Rheinbold (1970). The proof we give here parallels their argument. Care, however, is needed in dealing with the *weak* compactness of L^0. The convergence properties of the Newton method $(A(x) = H_Q(x) + \lambda H_J(x))$ in abstract Banach spaces have, of course, been known for some time. The famous Newton-Kantorovic Theorem (see Chapter 4 of Rall (1969)) gives conditions under which the Newton method converges R-quadratically. In view of this one might expect that the rate of convergence for Fisher's scoring technique is better than R-linear.

We begin the proof of Theorem 3.2.2 by establishing four lemmas which will be used in the proof.

Lemma 3.2.3.

For $\lambda > 0$ let $I_\lambda = Q(x) + \lambda\|Px\|^2$, $x \in D \subset H$ where $Q: D \to R^1$ has a weakly continuous first Frechet derivative on $D_0 - D_0$, a weakly compact subset of D. Suppose that $\{x^k\} \in D_0$ is any sequence which satisfies $\lim_{k\to\infty} \|\nabla I_\lambda(x^k)\| = 0$. Then the set $\Omega = \{x \in D_0 | \nabla I_\lambda(x) = 0\}$ of critical points of I_λ in D_0 is non-empty and

$$\liminf_{k\to\infty\, x\in\Omega} \|x^k - x\| = 0 \ .$$

In particular if Ω consists of a single point x^* Then $\lim_{k\to\infty} x^k = x^*$ and $\nabla I_\lambda(x^*) = 0$.

Proof:

Since D_0 is weakly compact $\{x^k\}$ has a weakly convergent subsequence $\{x^{k_i}\}$ with $x^{k_i} \overset{w}{\to} x^*$ for some $x^* \in D_0$. Now $\nabla I_\lambda(x^{k_i}) = \nabla Q(x^{k_i}) + 2\lambda P x^{k_i}$ and so by the weak continuity of Q' on D_0 $\nabla I_\lambda(x^{k_i}) \overset{w}{\to} \nabla I_\lambda(x^*)$. Therefore, $|<\nabla I_\lambda(x^*), \nabla I_\lambda(x^{k_i})>| \to \|\nabla I_\lambda(x^*)\|^2$. However $|<\nabla I_\lambda(x^*), \nabla I_\lambda(x^{k_i})>| \le \|\nabla I_\lambda(x^*)\| \ \|\nabla I_\lambda(x^{k_i})\|$ (*Cauchy-Schwartz Inequality*) and since $\|\nabla I_\lambda(x^{k_i})\| \to 0$ (by definition of $\{x^{k_i}\}$) we have $\nabla I_\lambda(x^*) = 0$. That is to say Ω is non-empty. But this also means that

$$\lim_{k_i\to\infty} \|\nabla I_\lambda(x^{k_i}) - \nabla I_\lambda(x^*)\| = 0 \ .$$

Now

$$\nabla I_\lambda(x^{k_i}) - \nabla I_\lambda(x^*) = \nabla Q(x^{k_i}) - \nabla Q(x^*) + 2\lambda(Px^{k_i} - Px^*)$$

which implies

$$2\lambda\|Px^{k_i} - Px^*\| = \|\nabla I_\lambda(x^{k_i}) - \nabla I_\lambda(x^*) - \nabla Q(x^{k_i}) + \nabla Q(x^*)\|$$

$$\le \|\nabla I_\lambda(x^{k_i}) - \nabla I_\lambda(x^*)\| + \|\nabla Q(x^{k_i}) - \nabla Q(x^*)\| \ .$$

Hence, recalling that Q' is weakly continuous, we have

$$Px^{k_i} \to Px^* \quad \text{as} \quad i \to \infty .$$

But since the null space of P is finite-dimensional and $x^{k_i} \overset{w}{\to} x^*$ we have that

$$(I - P)x^{k_i} \to (I - P)x^* \quad \text{as} \quad i \to \infty .$$

Hence $x^{k_i} \to x^*$ as $i \to \infty$. To finish off the proof let $\delta_k = \inf_{x \in \Omega} \|x^k - x\|$ and suppose $\exists$ a subsequence $\{\delta_{k_j}\}$ such that $\delta_{k_j} \geq \delta > 0 \ \forall j$. Then, by an argument similar to that used above, the corresponding sequence of x's, $\{x^{k_j}\}$ say, must have a convergent subsequence whose limit lies in Ω. But this contradicts the definition of δ_{k_j}. Therefore $\lim_{k \to \infty} \delta_k = 0$. □

The proof of the next two lemmas is practically the same as the proofs of (14.1.15) and (14.1.16) in Ortega and Rheinbold (1970). They are included for completeness.

Lemma 3.2.4.

Let I_λ, D_0 and D be as in Lemma 3.2.3. Suppose that the set Ω of critical points of I_λ is D_0 is finite. Let $\{x^k\} \subset D_0$ be any sequence for which $\lim_{k \to \infty} \|x^k - x^{k+1}\| = 0$ and $\lim_{k \to \infty} \|\nabla I_\lambda(x^k)\| = 0$. Then $\lim_{k \to \infty} x^k = x^*$ and $\nabla I_\lambda(x^*) = 0$.

Proof:

Let Γ be the set of limit points of $\{x^k\}$. As in

Lemma 3.2.3, any limit point is also a critical point of I_λ so that $\Gamma \subset \Omega$; that is, Γ is finite. Suppose that $\Gamma = \{z^1, z^2, \ldots, z^m\}$ with $m > 1$; then

$$\delta = \min \{\| z^i - z^j \| : i \neq j, \quad i,j = 1, \ldots, m\} > 0$$

and we can choose $k_0 \geq 0$ such that $x^k \in \bigcup_{i=1}^{m} S(z^i, \delta/4)$ and $\| x^k - x^{k+1} \| \leq \delta/4$ for all $k \geq k_0$. $x^k \in S(z^1, \delta/4)$ implies that

$$\| z^i - x^{k+1} \| \geq \| z^i - z^1 \| - (\| z^1 - x^k \| + \| x^k - x^{k+1} \|)$$

$$\geq \delta - 2\delta/4 = \delta/2, \quad i \geq 2$$

and hence, necessarily, that $x^{k+1} \in S(z^1, \delta/4)$. By induction, therefore, $x^k \in S(z^1, \delta/4)$ for all $k \geq k_0$, which contradicts the fact that $z^2, \ldots, z^m$ are limit points of $\{x^k\}$; therefore, $m = 1$. □

<u>Lemma 3.2.5.</u>

Let $g\colon D \subset R^n \to R^1$ be Frechet differentiable on an open set $D_0 \subset D$ and suppose that $\{x^k\} \subset D_0$ converges to $x^* \in D_0$. Assume that $\nabla g(x^*) = 0$, and g has a second Frechet derivative at x^* and the Hessian $H_g(x^*)$ is invertible, and that there is an $\eta > 0$ and a k_0 for which

$$g(x^k) - g(x^{k+1}) \geq \eta \| \nabla g(x^k) \|^2, \quad \forall\, k \geq k_0 . \qquad (3.2.3)$$

Then the root convergence factor $R_1\{x^k\} < 1$. That is, the sequence converges, at least, R-linearly.

Proof:

For $A \in L(H,H)$

$$\|A\| \equiv \sup_{y \in H} \frac{|<y,Ay>|}{\|y\|^2} .$$

Set $\alpha = \|H_g(x^*)^{-1}\|^{-1}$. Then for given $\varepsilon \in (0,\alpha)$ we may choose $\delta > 0$ so that $S \equiv S(x^*,\delta) \in D_0$ and

$$\|\nabla g(x) - H_g(x^*)(x - x^*)\| \le \varepsilon \|x - x^*\| , \quad \forall\ x \in S .$$

Hence

$$\|\nabla g(x)\| \ge \|H_g(x^*)(x - x^*)\| - \|\nabla g(x) - H_g(x^*)(x - x^*)\|$$
$$\ge (\alpha - \varepsilon)\|x - x^*\| , \quad \forall\ x \in S .$$

Consequently in view of equation (3.2.3), there is a k_0 so that with $\gamma_0 = [\eta(\alpha - \varepsilon)^2]^{-1}$

$$\|x^k - x^*\| \le (\alpha - \varepsilon)^{-2}\|\nabla g(x^k)\|^2$$
$$\le \gamma_0[g(x^k) - g(x^{k+1})] , \quad \forall\ k \ge k_0 . \quad (3.2.4)$$

Now note that the mean-value Theorem 1.A.8 ensures that

$$g(x^k) - g(x^*) \le \gamma_1\|x^k - x^*\|^2 , \quad \forall\ k \ge k_2 ,$$

where $\gamma_1 = \frac{1}{2}\|H_g(x^*)\| + \varepsilon$, and $k_2 \ge k_0$ is sufficiently large. Hence, we obtain, using 3.2.3 and 3.2.4,

$$\begin{aligned} 0 \le g(x^{k+1}) - g(x^*) &\le g(x^k) - g(x^*) - \eta\|\nabla g(x^k)\|^2 \\ &\le g(x^k) - g(x^*) - (1/\gamma_0)\|x^k - x^*\|^2 \\ &\le \mu[g(x^k) - g(x^*)], \qquad \forall\, k \ge k_2, \end{aligned} \tag{3.2.5}$$

where $\mu = 1 - (\gamma_0\gamma_1)^{-1}$. Clearly (by a suitable choice of η) $0 \le \mu < 1$. Now letting $e_k = [g(x^k) - g(x^*)]^{\frac{1}{2}}$ for $k = 1,2,\ldots$, we have for $k \ge k_2$

$$e_{k+1} \le \mu^{\frac{1}{2}} e_k$$

so by induction

$$e_k \le \mu^{(k-k_2)/2} e_{k_2} \le \mu^{(k-k_2)/2}[1+e_{k_2}] .$$

Therefore

$$(e_k)^{1/k} \le \mu^{(k-k_2)/2k}[1+e_{k_2}]^{1/k} ,$$

which implies

$$\begin{aligned} \limsup_k\, (e_k)^{1/k} &\le \limsup_k\, \mu^{(k-k_2)/2k}[1+e_{k_2}]^{1/k} \\ &\le \mu^{\frac{1}{2}} < 1 . \end{aligned}$$

That is $R_1\{[g(x^k) - g(x^*)]^{\frac{1}{2}}\} < 1$. Therefore, by (3.2.4) and remembering that $g(x^k) - g(x^{k+1}) \le g(x^k) - g(x^*)$,

$$R_1\{x^k\} \leq R_1\{[g(x^k) - g(x^*)]^{\frac{1}{2}}\} < 1 . \quad \square$$

Lemma 3.2.6.

Let $g: D \subset H \rightarrow R$ be continuous on an open set D and Frechet differentiable on

$$L^0 = L^0[g(x^0)] = \{x \in D | g(x) \leq g(x^0)\}$$

for some x^0 in D. Suppose L^0 is weakly compact. Then for any $x \in L^0$ and $p \in H$ with $<\nabla g(x),p> > 0$ there is an $\alpha^* > 0$ such that

$$g(x) = g(x - \alpha^* p) \quad \text{and} \quad [x,x - \alpha^* p] \subset L^0 .$$

In particular if $\eta > 0$ is any number with the property that

$$g(x - \alpha p) < g(x) \quad \forall\ x - \alpha p \in (x,x - \eta p] \cap L^0$$

then $[x,x - \eta p] \subset L^0$.

Proof:

Set $\alpha^* = \sup J$ where

$$J = \{\alpha > 0 | [x,x - \alpha p] \subset D \ \text{ and } \ g(x - \beta p) < g(x), \ \forall\ \beta \in (0,\alpha]\} .$$

The function $\zeta(\alpha) = g(x - \alpha p) - g(x)$, $\alpha \in R^+$ is continuous, $\zeta(0) = 0$ and $\zeta'(0) = -g'(x)p = -<\nabla g(x),p> < 0$. Hence

$\exists\ \alpha > 0$ s.t. $\forall\ \beta \in (0,\alpha]$, $\zeta(\beta) < 0$. It follows from this and the openness of D that J is non-empty and so α^* is well-defined. Consider the half line $\ell = \{x - \alpha p, \alpha > 0\}$. Since L^0 is weakly compact, $L^0 \cap \ell$ is compact. Therefore the set $\{\alpha > 0: x - \alpha p \in L^0\}$ must also be compact. Which of course means that $\alpha^* < \infty$ and $[x, x - \alpha^* p] \subset L^0$.

Now suppose that $g(x - \alpha^* p) < g(x)$. Here since D is open and g is continuous $\exists\ \delta > 0$ such that $x - \alpha p \in D$ and $g(x - \alpha p) < g(x)\ \forall\ \alpha \in [\alpha^*, \alpha^* + \delta]$. But this contradicts the definition of α^* and so $g(x - \alpha^* p) = g(x)$.

The last statement follows immediately since if $\eta \geq \alpha^*$ then $g(x - \alpha^* p) < g(x)$ which contradicts the definition of α^*. □

We are now ready to prove Theorem 3.2.2. Recall the statement.

Theorem 3.2.2.

Suppose Q in equation (3.2.1) is such that Q'' is continuous and Q' is weakly continuous on the interior of C. Let $x_\lambda^0 \in \text{int}\ C$ be such that the level set

$$L^0 = \{x \in C \mid I_\lambda(x) \leq I_\lambda(x_\lambda^0)\}$$

is weakly compact and I_λ has only finitely many critical points in L^0. Assume that $A: L^0 \to L(H,H)$ is a continuous mapping such that $\exists$ an invertible map $B(x) \in L(H,H)$ with $A(x) = B^*(x)B(x)$ (* denotes the adjoint) $\forall\, x \in L^0$. Suppose also μ_0, μ_1 and γ_1 all positive with $0 < \frac{1}{2}\gamma_1 < \mu_0$ satisfying

$$\mu_0\|h\|^2 \leq \langle h, A(x)h\rangle \leq \mu_1\|h\|^2$$

and

$$I''_\lambda(x)hh \leq \gamma_1\|h\|^2$$

$\forall\, x \in L^0$ and $h \in H$. Consider the process (3.2.2) i.e.

$$x^{k+1} = x^k - [A(x^k)]^{-1}\nabla I_\lambda(x^k) \qquad k = 0,1,\ldots$$

Then $\{x^k\} \in L^0$, $\lim_{k\to\infty} x^k = x^*$, where $\nabla I_\lambda(x^*) = 0$, and if $H_{I_\lambda}(x^*)$ is nonsingular, then the convergence is at least R-linear.

Proof:

For ease of notation let $g \equiv I_\lambda$ i.e.

$$g(x) = Q(x) = \lambda\|Px\|^2.$$

For any $x \in L^0$ and $h \in H$

$$\mu_0\|h\|^2 \leq \langle h, A(x)h\rangle \leq \mu_1\|h\|^2. \qquad (3.2.6)$$

If $x^k \in L^0$, $B(x^k)$ is invertible, so

$$p^k = A(x^k)^{-1}\nabla g(x^k) = B^*(x^k)^{-1}B(x^k)^{-1}\nabla g(x^k)$$

is well-defined. Letting $h = B(x^k)^{-1}\nabla g(x^k)$ we have from (3.2.6) that

$$<B(x^k)^{-1}\nabla g(x^k), A(x^k)B(x^k)\nabla g(x^k)> \le \mu_1<B(x^k)^{-1}\nabla g(x^k), B(x^k)^{-1}\nabla g(x^k)>$$

which by the Cauchy-Schwartz inequality implies that $\frac{1}{\mu_1}\|\nabla g(x^k)\| \le \|p^k\|$. Similarly letting $h = A(x^k)^{-1}\nabla g(x^k)$ in (3.2.1) we obtain that $\|p^k\| \le \frac{1}{\mu_0}\|\nabla g(x^k)\|$. Hence

$$(\mu_1)^{-1}\|\nabla g(x^k)\| \le \|p^k\| \le (\mu_0)^{-1}\|\nabla g(x^k)\| . \qquad (3.2.7)$$

It is also clear that $<p^k, \nabla g(x^k)> > 0$ unless $\|A(x^k)^{-1}\nabla g(x^k)\| = 0$ i.e. $\nabla g(x^k) = 0$. In other words, $<p^k, \nabla g(x^k)> > 0$ unless the process terminates at x^k. Now let $\alpha \in (0,1]$ be such that $[x^k, x^k - \alpha p^k] \subset L^0$. Then, by the mean-value Theorem 1.A.7, there is an $\hat{\alpha} \in (0,\alpha]$ so that

$$g(x^k) - g(x^k - \alpha p^k) = \alpha<\nabla g(x^k), p^k> - \tfrac{1}{2}\alpha^2<p^k, H_g(x^k - \hat{\alpha}p^k)p^k>$$

But since $<\nabla g(x^k), p^k> = <p^k, A(x^k)p^k>$ and by hypothesis, $<p^k, H_g(x^k - \hat{\alpha}p^k)p^k> \le \tfrac{1}{2}\gamma_1\|p^k\|^2$ since $x^k - \hat{\alpha}p^k \in L^0$.

$$\therefore g(x^k) - g(x^k - \alpha p^k) = \alpha[\mu_0 - \tfrac{1}{2}\gamma_1]\|p^k\|^2 > 0. \qquad (3.2.8)$$

It follows from Lemma 3.2.2 that $x^{k+1} = x^k - p^k \in L^0$ and, by induction, that $\{x^k\} \subset L^0$.

Combining (3.2.7) and (3.2.8) we get

$$g(x^k) - g(x^{k+1}) \ge (\mu_1)^{-2}(\mu_0 - \tfrac{1}{2}\gamma_1)\|\nabla g(x^k)\|^2 \qquad (3.2.9)$$

So, since $\{g(x^k)\}$ is a monotonic decreasing sequence which is bounded below, the left hand side converges to zero. Hence $\lim_{k\to\infty} \nabla g(x^k) = 0$. Consequently, since $p^k = x^k - x^{k+1}$, (3.2.7) shows that $\lim_{k\to\infty} (x^{k+1} - x^k) = 0$, and the result follows by Lemmas 3.2.4 and 3.2.5. □

3.3 SPECIFIC IMPLEMENTATIONS OF THE ALGORITHM

We now give some details on the implementation of iterative minimization methods of the form

$$x^{k+1} = x^k - [A(x^k)]^{-1}\nabla I_\lambda(x^k)$$

in Generalized Linear Interative Models and Non-linear Regression models. Let x^0 be the current iterate. The following describes how the next iterate, x^1, is obtained. We begin by describing a Newton-Raphson method for use with Generalized Linear Interactive Models.

GENERALIZED LINEAR INTERACTIVE MODELS

Recall from Chapter 1 that I_λ has the form:

$$I_\lambda(x) = \sum_{i=1}^{n} [b(\theta_i) - y_i\theta_i]/a_i(\phi) + \lambda J(x) = Q(x|y) + \lambda J(x)$$

We will show that minimizing the Newton-Raphson approximation (with Fisher's Scoring technique)

$$I_\lambda(x) \approx Q(x^0) + Q'(x^0)(x - x^0) + \tfrac{1}{2}E(Q''(x^0))(x - x^0)(x - x^0)$$
$$+ \lambda J(x) \qquad (*)$$

is equivalent to minimizing I_λ^0 given by:

$$I_\lambda^0(x) = \sum_{i=1}^{n} \frac{1}{u_i} [z_i - L_i x]^2 + \lambda J(x)$$

where

$$z_i = L_i x^0 + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu i} ; \ u_i = 2 \text{ var } (y_i)(\frac{\partial \eta_i}{\partial \mu_i})^2 .$$

One should note from this that the determination of the next iterate, x^1, reduces to a generalized spline smoothing problem (see Wendelberger 1982). The derivation essentially follows Nelder and Wedderburn (1972).

Let the contribution of y_i to I_λ be denoted ℓ_i

$$\ell_i = [b(\theta_i) - y_i\theta_i]/a_i(\phi) .$$

Recall the following relations:

$$E(y_i) = b'(\theta_i) = \mu_i, \text{var } (y_i) = b''(\theta_1)a_i(\theta) = \frac{\partial \mu_i}{\partial \theta_i} a_i(\theta)$$

and

$$\eta_i = L_i x = g(\mu_i) .$$

(See Chapter 1 for more details.) By the chain rule, the Frechet derivative of ℓ_i w.r.t. x, ℓ_i' is given by

$$\ell_i' = \frac{\partial \ell_i}{\partial \mu_i} \mu_i'$$

$$= \frac{[\mu_i - y_i]}{\text{var }(y_i)} \frac{\partial \mu_i}{\partial \eta_i} L_i .$$

Similarly the second Frechet derivative is

$$\ell_i'' = \frac{\partial \ell_i'}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} L_i$$

$$= \left[\frac{\partial^2 \ell_i}{\partial \mu_i^2} \frac{\partial \mu_i}{\partial \eta_i} L_i + \frac{\partial \ell_i}{\partial \mu_i} \frac{\partial^2 \mu_i}{\partial \eta_i \partial \mu_i} L_i \right] \frac{\partial \mu_i}{\partial \eta_i} L_i .$$

Using the fact that $E\left(\frac{\partial \ell_i}{\partial \mu_i}\right) = 0$ we obtain

$$E[\ell_i''] = E\left[\frac{\partial^2 \ell_i}{\partial \mu_i^2}\right]\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 L_i L_i .$$

and $E\left[\frac{\partial^2 \ell_i}{\partial \mu_i^2}\right] = E\left[\frac{\partial}{\partial \mu_i}\left[\frac{\mu_i - y_i}{\text{var }(y_i)}\right]\right]$ which in turn is equal to 1/var (y_i). By the preceding remarks, we have that

$$Q'(x^0)(x - x^0) = \sum_{i=1}^{n} \ell_i'(x - x^0)$$

$$= \sum_{i=1}^{n} \frac{\mu_i - y_i}{\text{var }(y_i)} \frac{\partial \mu_i}{\partial \eta_i} L_i(x - x^0)$$

and also that

$$E[Q''(x^0)](x-x^0)(x-x^0) = \sum_{i=1}^{n} E[\ell_i''(x^0)](x-x^0)(x-x^0)$$

$$= \sum_{i=1}^{n} \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \frac{[L_i(x-x^0)]^2}{\mathrm{var}\ (y_i)}$$

Combining terms it follows that minimizing (*) w.r.t. x is equivalent to minimizing I_λ^0 given by

$$I_\lambda^0(x) = \sum_{i=1}^{n} \frac{1}{u_i}[z_i - L_i x]^2 + \lambda J(x)$$

where z_i and u_i are as given. □

NON-LINEAR REGRESSION MODELS

In these models we have that

$$I_\lambda(x) = \sum_{i=1}^{n} [z_i - N_i(x)]^2 + \lambda J(x) .$$

N_i's are non-linear functionals of x contaminated by noise. In the Gauss-Newton algorithm N_i's are linearized about the current iterate as

$$N_i(x) \approx N_i(x^0) + N_i'(x^0)(x - x^0)$$

and the next iterate, x^1, is the minimizer of the functional I_λ^0 given by:

$$I_{\lambda}^{0}(x) = \sum_{i=1}^{n} [z_i - N_i(x^0) - N_i'(x^0)(x - x^0)]^2 + \lambda J(x)$$

Again one should note from this that the determination of the next iterate, x^1, reduces to a generalized spline smoothing problem.

CHAPTER 4

ESTIMATION OF THE SMOOTHING PARAMETER

4.1 INTRODUCTION

Recall that the penalized likelihood estimator, x_λ, is the minimizer of the functional

$$I_\lambda(x) = Q(x|z) + \lambda J(x) \quad \lambda>0 .$$

The *smoothing* parameter λ controls the relative weighting given to the penalty term, large values of λ giving the penalty more weight.

There is a growing literature on methods for estimating smoothing parameters (see Atilgan 1983, Anderssen and Bloomfield 1974, Leonard 1982, Silverman 1978, and Stone 1974), one of the most popular of these methods is cross-validation. The cross-validatory procedure tries to choose a value of the smoothing parameter which is optimal for some specified predictive criterion (see Stone 1974). Roughly, the method works as follows: Data are divided into subsamples; for a given value of the smoothing parameter the estimator is computed on each of the subsamples and, using the predictive criterion, its performance is evaluated on the remaining subsamples. The "cross-validatory" assessment of λ is the combined performance of the estimator over all the subsamples. Using this technique the value of the smoothing parameter with the best cross-validatory assessment can be isolated.

Since Stone's 1974 paper, cross-validation has been widely applied in a number of diverse statistical contexts. Wahba and Wold (1975) were the first to use the method in a penalized likelihood setting. However, the method they proposed was computationally intensive and in 1979 Craven and Wahba came up with a refined procedure, Generalized Cross Validation, which has decided computational advantages over the Wahba and Wold proposal.

In this chapter we extend the Generalized Cross Validation (GCV) estimator to general penalized likelihood functionals. The estimator is described in the next section and some favourable heuristics are given in section 3. Although the results are of a preliminary nature, they suggest that the technique may have promise.

4.2 THE GENERALIZED CROSS VALIDATION PROCEDURE

Generalized Cross-Validation (GCV) was developed by Craven and Wahba (1979) to estimate a smoothing parameter similar to λ in a linear regression model context. In their model, the observations z_i are noise contaminated linear functionals of an unknown function, x, of interest.

$$z_i = L_i x + \text{noise}, \qquad i = 1,2,\ldots,m$$

where the L_i's are linear functionals. The estimator, x_λ, of x is the minimizer of the functional

$$\frac{1}{n} \sum_{i=1}^{n} [z_i - L_i x]^2 + \lambda \int [x^m(t)]^2 dt \qquad \lambda>0 .$$

It turns out that the residuals, $z_i - \hat{z}_i$, can be written as a linear transform of the data.

$$z - \hat{z} = [I - A(\lambda)]z$$

where $A(\lambda)$ is an n n matrix defined in terms of the linear functionals. Formally $A(\lambda)_{ij} = \frac{\partial L_i}{\partial z_j}(x_\lambda)$. The GCV estimate of λ is the minimizer of the function

$$V(\lambda) = \frac{\frac{1}{n} \sum_{i=1}^{n} [z_i - \hat{z}_i]^2}{(\frac{1}{n} \text{tr} [I - A(\lambda)])^2}$$

$$V(\lambda) = \frac{\frac{1}{n} \text{RSS}(\lambda)}{[1-\mu_1(\lambda)]^2}$$

$V(\lambda)$ is an assessment of λ combining goodness of fit and model complexity. The goodness of fit is measured by the residual sum of squares, $\frac{1}{n}\text{RSS}(\lambda)$, while the complexity of x_λ is measured by the "effective" number of parameters describing x_λ relative to the sample size, i.e. $\mu_1(\lambda)$. (See Wahba (1979) for an equivalent degrees-of-freedom interpretation of $\mathrm{tr}[I - A(\lambda)]$.) As $\lambda \to \infty$, x_λ tends to an m-1'th degree polynomial regression making $\frac{1}{n}\mathrm{RSS}(\lambda)$ take on its maximum and $\mu_1(\lambda)$ be a minimum. On the other hand, as $\lambda \to 0$, x_λ tends to a data interpolant minimizing $\frac{1}{n}\mathrm{RSS}(\lambda)$ and maximizing $\mu_1(\lambda)$. It has been shown that the GCV estimate of λ, for large sample sizes, comes close to minimizing the integrated squared error, $R(\lambda)$, between x and x_λ:

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[L_i x - L_i x_\lambda]^2 .$$

For further discussion see articles by Craven and Wahba (1979) and Speckman (1982). Wahba (1982) proposed a GCV based estimator of λ in a constrained version of the linear regression model. In order to extend the GCV procedure to general penalized likelihoods, we need to introduce some additional notation.

Some Notation and Mild Assumptions

Let x_0 be the true value of the function, i.e. the value corresponding to the observed data z. Define the following information operators on H by:

$$\text{i.1} \quad I_{sp}(x_0) = EH_Q[x_0] + \lambda H_J[x_0]$$

$$\text{i.2} \quad I_s(x_0) = EH_Q[x_0]$$

H_Q is the Hessian of Q, H_J the Hessian of J and expectations are taken with respect to the probability distribution of z.

We assume that $I_s(x_0)$ and $I_{sp}(x_0)$ (and their inverses) have cholesky-like decompositions (i.e. $I_s(x_0) = I_s^{\frac{1}{2}}(x_0) I_s^{\frac{1}{2}}(x_0)$) and that Q satisfies:

$$\text{q.1} \quad E[\nabla Q(x_0|z)] \equiv 0$$

$$\text{q.2} \quad \mathrm{Var}[\nabla Q(x_0|z)] = I_s(x_0)$$

$$\text{q.3} \quad E\| I_s^{-\frac{1}{2}}(x_0) \nabla Q(x_0|z)\|^2 = \nu_0 < \infty$$

where the inverses are generalized inverses.

Finally let $\mu_1(\lambda) = \nu_1(\lambda)/\nu_0$ and $\mu_2(\lambda)/\nu_0$ where ν_1 and ν_2 are given by:

$$\begin{aligned}\nu_1(\lambda) &= E\| I_{sp}^{-\frac{1}{2}} \nabla Q(x_0|z)\|^2 \\ &= \mathrm{tr}\{I_s(x_0) I_{sp}^{-1}(x_0)\}\end{aligned}$$

$$\begin{aligned}\nu_2(\lambda) &= E\| I_s^{\frac{1}{2}}(x_0) I_{sp}^{-1}(x_0) \nabla Q(x_0|z)\|^2 \\ &= \mathrm{tr}\{I_s^2(x_0) I_{sp}^{-2}(x_0)\} .\end{aligned}$$

(We tacitly assume the commutivity of the I-operators.)

The Proposed GCV Estimator

In the above notation, Craven and Wahba's GCV estimator is the minimizer of:

$$V(\lambda) = \frac{Q(x_\lambda|z)}{[1-\mu_1(\lambda)]^2}$$

where x_λ is the minimizer of I_λ, and $Q(x_\lambda|z)$ is given by:

$$Q(x_\lambda|z) = \frac{1}{n}\sum_{i=1}^{n}[z_i - L_i x_\lambda]^2 .$$

This suggests that a plausible GCV estimator for general Penalized Likelihoods is

$$V(\lambda) = \frac{Q(x_\lambda|z)}{[1-\mu_1(\lambda)]^2}$$

where μ_1 is now evaluated at x_λ;

$$\mu_1(\lambda) = \mathrm{tr}[I_s(x_\lambda)I_{sp}^{-1}(x_\lambda)]/\nu_0 .$$

The matrices required to calculate μ_1 are usually at the final stage of the iterative procedure used to calculate x_λ (see Chapter 5 for an example).

4.3 SOME JUSTIFICATION FOR USING THE GCV ESTIMATOR

We now give some heuristic arguments in favour of the

GCV estimator of λ proposed above. The argument relies heavily on how well x_λ can be approximated by x_λ^0:

$$x_\lambda^0 = x_0 - I_{sp}^{-1}(x_0)\nabla I_\lambda(x_0)$$

x_λ^0 also comes up in the asymptotic analysis of x_λ. Loosely, one can show (under suitable regularity conditions) that if $\lambda \to 0$ at an appropriate rate, the distance between x_λ^0 and x_0 is o_p of distance between x_λ and x_λ^0 (see Cox and O'Sullivan 1983). By strengthening such results it may be possible to rigourize the heuristics which follow.

Suppose we say that:

$$EQ(x_\lambda|z) = EQ(x_0|z) + E <\nabla Q(x_0|z),\ x_\lambda^0>$$
$$+ \tfrac{1}{2}E <I_s(x_0)(x_\lambda - x_0),\ (x_\lambda - x_0)>$$

and

$$\mu_1(\lambda) = tr[I_s(x_0)I_{sp}^{-1}(x_0)]/\nu_0 .$$

Then $2EV(\lambda)$ can be written as:

$$2EV(\lambda) = \frac{2EQ(x_0|z) - 2\mu_1(\lambda)\nu_0 + ER(\lambda)\nu_0}{[1 - \mu_1(\lambda)]^2}$$

in which $ER(\lambda)$

$$\begin{aligned} ER(\lambda) &= E < I_s(x_0)(x_\lambda - x_0),\ (x_\lambda - x_0)>/\nu_0 \\ &= b^2(x_\lambda) + \nu_2(x_\lambda) \\ &= \textit{bias} + \textit{variability} \end{aligned}$$

where

$$b^2(x_\lambda) = < I_s(x_0)(Ex_\lambda - x_0),\ (Ex_\lambda - x_0)>/\nu_0$$

and

$$\nu_2(x_\lambda) = E < I_s(x_0)(x_\lambda - Ex_\lambda),\ (x_\lambda - Ex_\lambda)>/\nu_0 \ .$$

Notice how the information operator enters naturally into $ER(\lambda)$. For example, in the logistic regression model of section 3 of Chapter 2, $R(\lambda)$ takes the form:

$$R(\lambda) = \frac{1}{n} \sum_{i=1}^{n} n_i p_i (1 - p_i)[L_i x_\lambda - L_i x_0]^2 .$$

Assuming $\nu_2(x_\lambda) \approx \nu_2(x_\lambda^0) = \mu_2(\lambda)$ we get (using Craven and Wahba arguments) that:

$$\frac{ER(\lambda) + 2EQ(x_0|z) - 2EV(\lambda)}{ER(\lambda)} = \frac{-\mu_1(2 - \mu_1)}{(1 - \mu_1)^2} + \frac{1}{b^2 + \mu_2}\ \frac{\mu_1^2}{(1 - \mu_1)^2}$$

from which it follows that

$$\frac{ER(\lambda) + 2EQ(x_0|z) - 2EV(\lambda)|}{ER(\lambda)} < h(\lambda)$$

where

$$h(\lambda) = \left[2\mu_1(\lambda) + \frac{\mu_1^2(\lambda)}{\mu_2(\lambda)}\right] \cdot \frac{1}{(1-\mu_1)^2} \ .$$

These considerations lead us to the following proposition.

Proposition 4.3.1.

Letting λ^* and $\tilde{\lambda}$ be the minimizers of $ER(\lambda)$ and $EV(\lambda)$ respectively, the expected inefficiency I^* defined by:

$$I^* = \frac{ER(\tilde{\lambda})}{ER(\lambda^*)}$$

satisfies:

$$I^* \leq \frac{1 - h(\lambda^*)}{1 - h(\tilde{\lambda})}$$

and if μ_1 and μ_1^2/μ_2 tend to zero along the sequences $\lambda^*(\nu_0)$ and $\tilde{\lambda}(\nu_0)$ then

$$I^* \to 1 \quad \text{as} \quad \nu_0 \to \infty \ .$$

Proof:

In view of the approximating assumptions, the proof is identical to Theorem 4.2 of Craven and Wahba's paper. □

A NUMERICAL EXPERIMENT

The heuristics are now supplemented by some numerical examples. Two separate logistic regression models were simulated; independent binomial observations $y_i, y_i \sim B(n,p_i)$ were generated according to:

$$\log\left[\frac{p_i}{1-p_i}\right] = x(t_i) \qquad i = 1,2,\ldots,80 \ .$$

The test logit function, x, was:

$$x(t_i) = -3.0 + 4.386\ e^{-t_i^2}$$

where $t_i = -1 + \frac{2(i-1)}{79}$, $i = 1,2,\ldots,80$.

For the first experiment $n = 10$, for the second $n = 1$. In each case the logit was estimated as the minimizer in $W_2^2[-1.1]$ of:

$$I_\lambda(x) = \sum_{i=1}^{80} \{n \log[1+e^{x(t_i)}] - y_i x(t_i)\} + \lambda \int_{-1}^{1} [x''(t)]^2 dt$$

with λ chosen by the GCV method proposed in section 2. For the logistic model $V(\lambda)$ reduces to:

$$V(\lambda) = \frac{\sum_{i=1}^{80} \{n\log[1+e^{x(t_i)}] - y_i x(t_i)\}}{(\mathrm{tr}[I - A(\lambda)])^2}$$

where $A(\lambda)$ is the A-matrix (see section 5.2) arising in the final Newton-Raphson minimization problem (see section 3.3), i.e.:

$$\sum_{i=1}^{80} \frac{1}{u_i} [z_i - x(t_i)]^2 + \lambda \int_{-1}^{1} [x''(t)]^2 dt$$

where

$$z_i = x_\lambda(t_i) + \frac{(y_i - n\hat{p}_i)}{n\hat{p}_i(1-\hat{p}_i)}$$

$$u_i^{-1} = n\hat{p}_i(1-\hat{p}_i)/2$$

and

$$\hat{p}_i = \frac{e^{x_\lambda(t_i)}}{1+e^{x_\lambda(t_i)}}$$

Figure 4.3.1 and 4.3.2 summarize the results of the study. Figure 4.3.1 plots the true and estimated logits while Figure 4.3.2 gives scaled plots of the functions $V(\lambda)$ and $R(\lambda)$. $R(\lambda)$ is given by:

$$R(\lambda) = \sum_{i=1}^{80} np_i(1-p_i)(x(t_i) - x_\lambda(t_i))^2.$$

These plots show that the generalized cross-validation function, $V(\lambda)$, chooses the smoothing parameter quite well in both cases.

Figure 4.3.1

True (solid) and Estimated (dashed) logits.

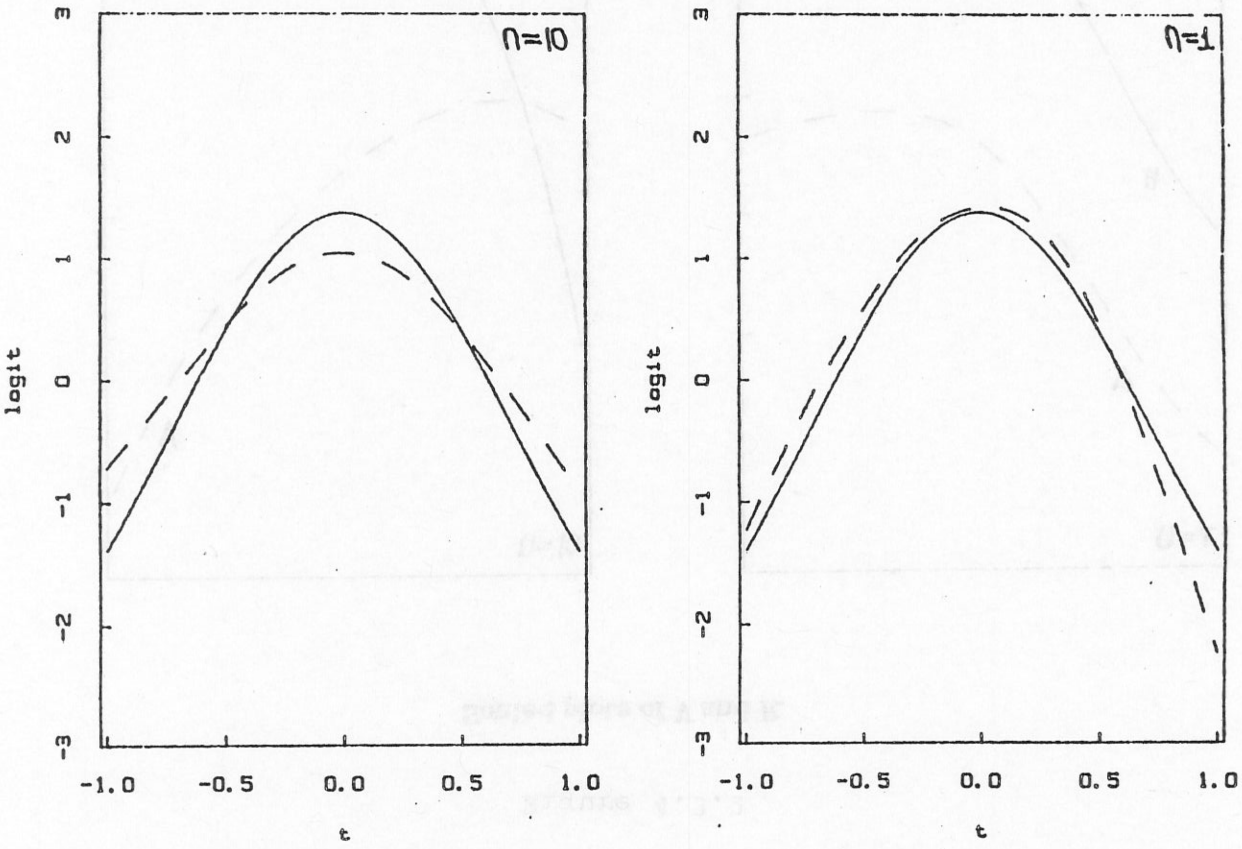
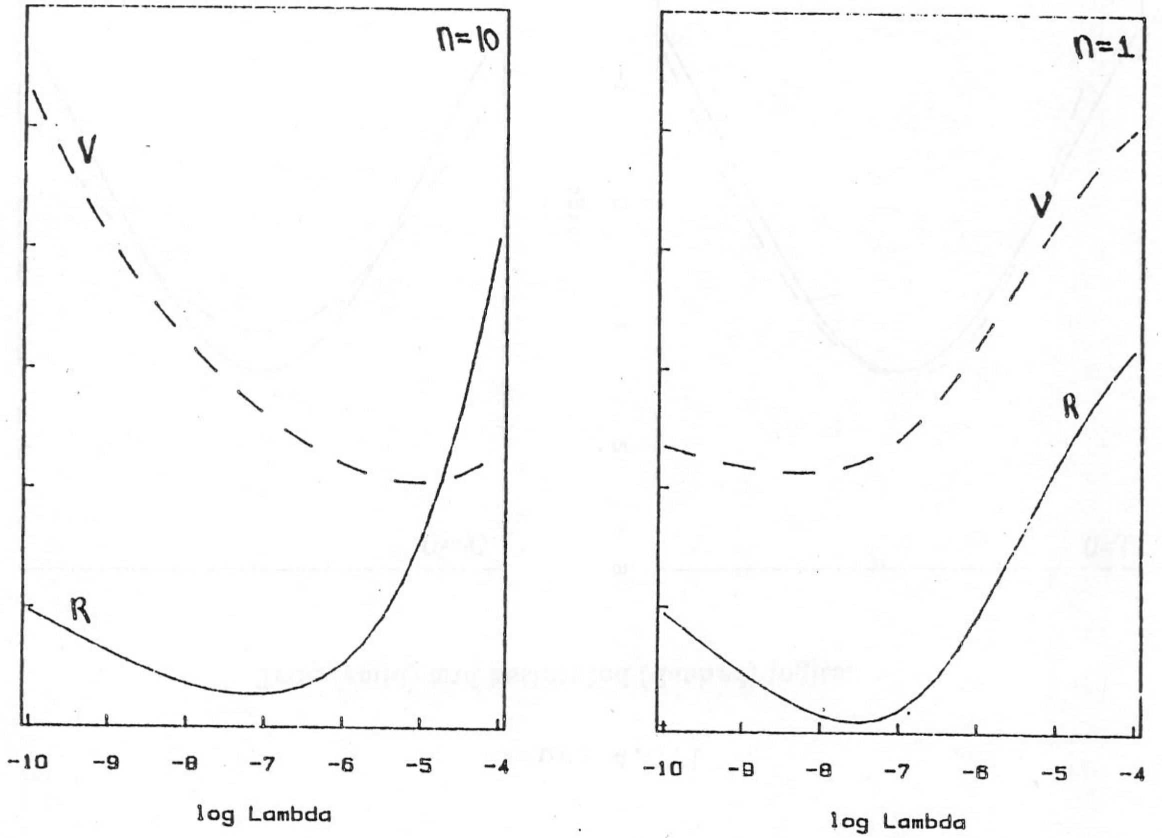


Figure 4.3.2

Scaled plots of V and R.



CHAPTER 5

A BAYESIAN METHOD FOR INVERTING SATELLITE RADIANCE DATA

5.1 INTRODUCTION

Remote sensing of the atmosphere is a rapidly developing science. Today's meteorological satellites, such as those in the TIROS-N series, have high resolution instruments on board, which measure the intensity of upwelling radiation at selected channel frequencies. An in-depth description of the data retrieved by the radiometers on the TIROS-N satellites can be found in Smith et. al. (1979). Satellite radiance data make it possible to obtain information on the atmosphere's temperature, moisture, and wind structure. One of the goals of the current Satellite Meteorology programme is to substantially upgrade the quality of the atmospheric information obtained from satellite soundings. A major challenge in this direction is to develop more refined numerical and statistical techniques for inverting the equations of radiative transfer.

For a non-scattering atmosphere in local thermodynamic equilibrium the radiative transfer equations (RTE's) (see Liou 1979) describe how the satellite up-welling radiance measurements relate to the underlying atmospheric temperature distribution T:

$$R_\nu(T) = B_\nu(T(p_0))\tau_\nu(p_0) - \int_0^{p_0} B_\nu[T(p)]\tau_\nu'(p)\,dp$$

where p_0 is the surface pressure, $\tau_\nu(p)$ is the transmittance of the atmosphere above pressure p at wavenumber ν, and B_ν is Plank's function given by:

$$B_\nu[T(p)] = c_1\nu^3/[\exp(c_2\nu/T(p))-1]$$

$$c_1 = 1.1906\times10^{-5}\text{erg} - \text{cm}^2 - \text{sec}^{-1}$$

$$c_2 = 1.43868\text{cm} - \text{deg}(\text{K})$$

The RTE's are, of course, an idealization. They describe the radiances which the satellite radiometer would record in the absence of such things as atmospheric attenuation due to clouds or instrument noise. Various clever techniques are used to correct for the cloud attenuation problem, after which it is reasonable to model the satellite data as:

$$z_{\nu_i} = R_{\nu_i}(T) + \varepsilon_i \qquad i = 1,2,\ldots,m$$

where z_{ν_i} is the satellite "cloud corrected" radiance measurement at wavenumber ν_i, and ε_i is the deviation of z_{ν_i} from the true radiance measurement $R_{\nu_i}(T)$. These deviations are probably best thought of as random variables having distributions which can be modeled from practical experience with the radiometer.

A number of procedures have been proposed to retrieve temperature profiles from satellite radiance data (see Fleming (1982) and references cited therein). The Statistical Eigenvector method of Smith and Woolf (1976) is currently favoured.[1] Recently Purser (1983) has pushed for a Bayesian synthesis of meteorological data. In this paper we desribe some Bayesian methods for estimating temperature profiles form the satellite radiance data. Prior information about atmospheric temperature profile structure, channel noise characteristics, and the non-linear dependence of the radiance measurements on temperature are explicitly taken into account. The discussion is restricted to the single profile case but extensions to the estimation of global temperature fields are quite feasible. The basic technique is presented in section 2; in section 3 we examine how a crude version of the Bayesian technique performs on real and simulated satellite radiance data.

5.2 THE BAYESIAN INVERSION TECHNIQUE

5.2.1 GENERAL FORMULATION

Let the temperature profile, T, be written as

$$T = T_0 + \delta$$

[1] Note added in proof: see also Smith(1983).

where T_0 is an initial guess for T (for example the regional climatology profile) and δ is the update to T_0 which is to be estimated from the data in hand. Our method of estimation consists of considering the functional:

$$I_\lambda(\delta) = \frac{1}{m}\sum_{i=1}^{m} w_i[z_i - R_{\nu_i}(T_0 + \delta)]^2 + \lambda J(\delta) \quad \lambda>0$$

and choosing the update, δ_λ, which minimizes I_λ over an appropriate space, C, of plausible candidates. The weighted residual sum of squares $\frac{1}{m}\sum_{i=1}^{m} w_i[z_i - R_{\nu_i}]^2$ is designed to model the information contained in the observed radiance data - there is an implicit assumption that the distribution of the ε_i in (5.1.2) is approximately Gaussian, with mean zero and variance w_i^{-1}. The second term, J, models past information or apriori beliefs about the behaviour of atmospheric temperature profile structure. In a strict Bayesian formulation, J behaves like the -2 log[*prior probability density* of δ] (see Box and Tiao 1973). The parameter λ specifies the relative weighting given to past and present information. A data-based method for choosing λ is given in section 3.2.

5.2.2 THE CHOICE OF PRIOR

Case A. Smoothness Prior

A general purpose smoothness prior for δ can be specified by setting

$$J(\delta) = \int_0^{p_0} [\delta''(p)]^2 dp .$$

In probabilistic terms (see Kimeldorf and Wahba 1970) this choice of prior says that apriori we believe:

$$\delta''(p) = \textit{white noise} .$$

The Bayes estimator of δ corresponding to this prior is the minimizer of I_λ given by:

$$I_\lambda(\delta) = \frac{1}{m} \sum_{i=1}^{m} w_i [z_i - R_{\nu_i} (T_0 + \delta)]^2 + \lambda \int_0^{p_0} [\delta''(p)]^2 dp .$$

From a practical point of view, however, the general purpose prior is suboptimal, since a more appropriate representation of prior information would be in terms of regional climatology.

Case B. Extracting a Prior from Regional Climatology

Suppose we represent profiles as a linear combination of basis functions $\{\phi_j\}$. The deviations of a profile from regional climatology can then be written as:

$$\delta = \sum_j \alpha_j \phi_j .$$

Regional climatology data bases can now be used to model the distribution of the basis coefficient vector, α. If the number of basis functions is finite, and the distribution of α was modelled by a multivariate normal distribution

$$\alpha = N(0,\Sigma_\alpha)$$

where Σ_α is the sample covariance matrix obtained from the regional climatology data base, the resulting Bayesian estimator of δ would be:

$$\delta_\lambda = \sum_j \hat{\alpha}_j \phi_j$$

where $\hat{\alpha}$ is the minimizer of:

$$I_\lambda(\alpha) = \frac{1}{m} \sum_{i=1}^{m} w_i [z_i - R_{\nu_i}(T_0 + \sum_j \alpha_j \phi_j)]^2 + \lambda \alpha^t \Sigma_\alpha^{-1} \alpha .$$

Under either of the above two priors, the iterative scheme used to approximate δ_λ is essentially the same. At each stage the non-linear radiance functionals, $R_{\nu_i}(T_0 + \delta)$, are linearized about the current iterate, δ^k, and the resulting quadratic minimization problem is solved.

5.3 COMPUTATIONAL METHODS

5.3.1 THE GAUSS-NEWTON ALGORITHM

An iterative Gauss-Newton procedure can be used to obtain the optimal correction defined by the minimizer of I_λ. Let T^k be the temperature profile estimate at the

k^{th} iteration.

$$T^k = T_0 + \delta^k \quad .$$

The iterative algorithm for obtaining the Bayes estimate can be stated as follows:

```
            Iterative Gauss-Newton Procedure
    while [convergence criteria fail to be met] do
            {
            Linearize the radiance functionals
            Solve for $\delta^{k+1}$
            Update convergence indicators
            }
    end
```

A detailed description of the iterative steps follows.

STEP 1: LINEARIZATION OF THE RADIANCE FUNCTIONALS

Linearizing $R_{\nu_i}(T)$ about T^k we obtain:

$$R_{\nu_i}(T) \approx R_{\nu_i}(T^k) + R'_{\nu_i}(T_k)(\delta - \delta^k)$$

where the action of the linear functional $R'_{\nu_i}(T^k)$ on a function h is given by:

$$R'_{\nu_i}(T^k)h = K_s(\nu_i,p_0)h(p_0) - \int_0^{p_0} K(\nu_i,p)h(p)dp$$

The kernels K_s and K are:

$$K_s(\nu_i, p_0) = B'_{\nu_i}[T^k(p)]\tau_{\nu_i}(p_0)$$

$$K(\nu_i, p) = B'_{\nu_i}[T^k(p)]\tau_{\nu_i}(p_0) \; .$$

STEP 2: SOLVING FOR THE NEXT ITERATE

After the linearization we get the next iterate, δ^{k+1}, by minimizing the functional

$$I_\lambda^{(k)}(\delta) = \sum_{i=1}^{m} w_i[z_i^k - R'_{\nu_i}(T^k)\delta]^2 + \lambda J(\delta)$$

$$z_i^k = z_i - R_{\nu_i}(T^k) + R'_{\nu_i}(T^k)\delta^k \quad i = 1,2,\ldots,m$$

Two cases corresponding to the smoothnesss and regional climatology priors will be considered.

<u>Solution for the Smoothness Prior</u>:

Here δ^{k+1} is the minimizer in $C = \{\delta: \int_0^{p_0} [\delta''(p)]^2 dp < \infty\}$ of the functional:

$$I_\lambda^{(k)}(\delta) = \frac{1}{m} \sum_{i=1}^{m} w_i[z_i^k - R'_{\nu_i}(T^k)\delta]^2 + \lambda \int_0^{p_0} [\delta''(p)]^2 dp \; .$$

It follows by straightforward calculus that δ^{k+1} can be written as:

$$\delta^{k+1}(p) = \hat{d}_1 + \hat{d}_2 p + \sum_{j=1}^{m} \hat{c}_j \xi_j(p)$$

where $\xi_j(p) = R'_{\nu_i}(T_k)Q_p(\cdot)$ and

$$Q_p(r) = \begin{cases} p^2r/2 - r^3/6, & p \geq r \\ pr^2/2 - p^3/6, & p \leq r \end{cases}$$

Letting $K_{ij} = <\xi_i, \xi_j> = \int_0^{p_0} \xi_i'' \xi_j'' dp$ for $i, j = 1, 2, \ldots, m$, and $T_{1i} = 1$, $T_{2i} = R'_{\nu_i}(T^k)p$ for $i = 1, 2, \ldots, m$, the coefficients $\hat{d}$ and $\hat{c}$ are the solution to the linear system:

$$[K + m\lambda I]c + Td = z^k$$

$$T^t c = 0 \quad .$$

Taking a QR-decomposition of the T matrix ($T = [Q_1 : Q_2]$ R with $R^t = [R_1 : 0]$ see Bunch et. al (1979)) $\hat{d}$ and $\hat{c}$ are given by the formulae:

$$\hat{c} = Q_2^t [Q_2^t K Q_2 + m\lambda I]^{-1} Q_2 z^k$$

$$\hat{d} = R_1^{-1} Q_1^t [z^k - K\hat{c}]$$

Note that the calculation of K_{ij} can be done by first fitting cubic spline interpolants, $\tilde{\xi}_j$, to ξ_j evaluated on a fine grid and proceeding to approximate K_{ij} as:

$$K_{ij} \approx <\tilde{\xi}_i, \tilde{\xi}_j> \; .$$

This method of approximation is used by Nychka (1983).

Solution for the Regional Climatology Prior:

δ^{k+1} is now given by

$$\delta^{k+1} = \sum_j \alpha_j^{k+1} \phi_j$$

where α^{k+1} is the minimizer of:

$$I_\lambda^k(\alpha) = \frac{1}{m} \sum_{i=1}^{m} w_i [z_i^i - X_i \alpha]^2 + \lambda \, \alpha^t \, \Sigma_\alpha^{-1} \, \alpha \, .$$

So $\alpha^{k+1} = [X^t D_\omega X + m\lambda \, \Sigma_\alpha^{-1}]^{-1} X^t D_\omega \, z^k$ where X and D_ω are:

$$X_{ij} = K_s(\nu_i,p_0)\phi_j(p_0) - \int_0^{p_0} K(\nu_i,p)\phi_j(p)\,dp$$

$$D_\omega = \mathrm{diag}(w_1,w_2,\ldots,w_m) \, .$$

The kernels K_s and K are given above.

STEP 3: CONVERGENCE CRITERIA

Standard convergence criteria can be used in the above iterations. We have used:

$$e_1^{k+1} = |I_\lambda(\delta^{k+1}) - I_\lambda(\delta^k)|/I_\lambda(\delta^k)$$

$$d_2^{k+1} = \|\delta^{k+1} - \delta^k\|/\|\delta^k\|$$

where $\|\cdot\|$ is an appropriate norm.

5.3.2 CHOICE OF THE SMOOTHING PARAMETER

An extension of Generalized Cross Validation (GCV) can be used to estimate the smoothing parameter, λ, empirically from the data. GCV was developed by Craven and Wahba (1979) to estimate a smoothing parameter similar to λ in a linear functional model context. In their model, the observations z_i are noise contaminated linear functionals of an unknown function, x, of interest.

$$z_i = L_i x + \text{noise}, \quad i = 1,2,\ldots,m$$

where the L_i's are linear functionals. The estimator, x_λ, of x is the minimizer of the functional

$$\frac{1}{m} \sum_{i=1}^{m} [z_i - L_i x]^2 + \lambda \int [x''(t)]^2 dt \qquad \lambda > 0 \ .$$

It turns out that the residuals, $z_i - \hat{z}_i$, can be written as a linear transform of the data.

$$z - \hat{z} = [I - A(\lambda)]z$$

where $A(\lambda)$ is an $m \times m$ matrix defined in terms of the linear functionals. Formally $A(\lambda)_{ij} = \frac{\partial L_i}{\partial z_j}(x_\lambda)$. The GCV estimate of λ is the minimizer of the function

$$V(\lambda) = \frac{\frac{1}{m} \sum_{i=1}^{n} [z_i - \hat{z}_i]^2}{(\frac{1}{m} \operatorname{tr}[I - A(\lambda)])^2}$$

$$= \frac{\frac{1}{m}\,\mathrm{RSS}(\lambda)}{[1-p(\lambda)]^2}$$

$V(\lambda)$ is an assessment of λ combining goodness of fit and model dimensionality (Wahba refers to $\mathrm{tr}[I - A(\lambda)]$ as the "equivalent degrees of freedom" of $\mathrm{RSS}(\lambda)$) of model. The goodness of fit is measured by the residual sum of squares, $\frac{1}{m}\,\mathrm{RSS}(\lambda)$, while the dimensionality of x_λ is measured by the "effective" number of parameters describing x_λ relative to the sample size, i.e. $p(\lambda)$. As $\lambda \to \infty$, x_λ tends to a straight line regression making $\frac{1}{m}\,\mathrm{RSS}(\lambda)$ take on its maximum and $p(\lambda)$ be a minimum. On the other hand, as $\lambda \to 0$, x_λ tends to a data interpolant minimizing $\frac{1}{m}\,\mathrm{RSS}(\lambda)$ and maximizing $p(\lambda)$. It has been shown that the GCV estimate of λ, for large sample sizes, comes close to minimizing the integrated squared error between x and x_λ:

$$R(\lambda) = \frac{1}{m}\sum_{i=1}^{m}[L_i x - L_i x_\lambda]^2 \ .$$

For further discussion see Craven and Wahba (1979), Speckman (1983) and Wahba (1982).

Extended GCV Estimator of the Smoothing Parameter

In the linear functional model the definition of dimensionality is facilitated by linearity. Even though the radiance functionals are not linear a simple approximation can be used to develop a dimensionality measure similar in spirit to $p(\lambda)$ above. This approximation makes use of the fact that the optimal correction under the smoothness and regional climatology priors minimize quadratic functionals:

$$\frac{1}{m}\sum_{i=1}^{m} w_i[z_i - R_{\nu_i}(T_\lambda) + R'_{\nu_i}(T_\lambda)\delta_\lambda - R'_{\nu_i}(T_\lambda)\delta]^2 + \lambda J(\delta) \qquad (*)$$

so that the residuals are defined, implicitly, by a linear equation.

$$z - R_\nu(T_\lambda) = [I - A_1(\lambda)]z^*$$

where $z_i^* = z_i - R_{\nu_i}(T_\lambda) + R'_{\nu_i}(T_\lambda)\delta_\lambda$ and $A_1(\lambda)$ is the A-matrix arising in the minimization of (*).

We therefore have, that to a first order approximation, the effective number of parameters in δ_λ relative to the sample size is given by:

$$p_1(\lambda) = \mathrm{tr}[A_1(\lambda)]/m$$

which leads us to the GCV-type assessment of λ as

$$V(\lambda) = \frac{\frac{1}{m}\,\mathrm{RSS}(\lambda)}{[1-p_1(\lambda)]^2}\,.$$

Note that the matrix $A_1(\lambda)$ is already available from the final stage of the Gauss-Newton iteration. After some algebra one can show that $V(\lambda)$ for the smoothness prior is given by:

$$V(\lambda) = \frac{\|\hat{c}\|^2_{R^m}}{\frac{1}{m}\,[\mathrm{tr}(Q_2^t K Q_2 + m\lambda I)^{-1}]^2}$$

and for the regional climatology prior $V(\lambda)$ becomes:

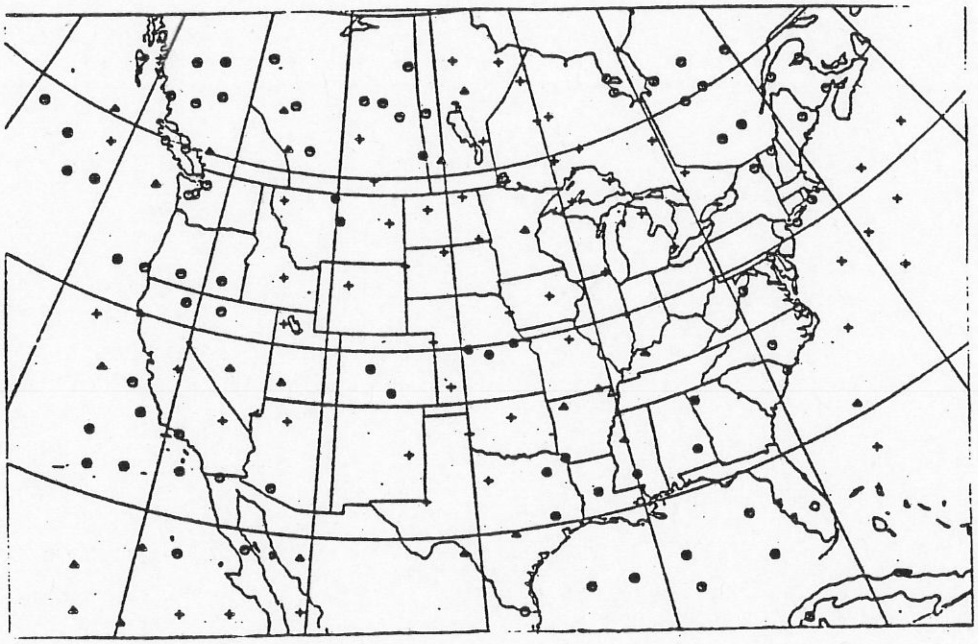
$$V(\lambda) = \frac{\frac{1}{m}\sum_{i=1}^{m} w_i [z_i - R_{\nu_i}(T_\lambda)]^2}{[\frac{1}{m}\,\mathrm{tr}\,\Sigma_\alpha^{-1}[X^t D_\omega X + m\lambda\Sigma_\alpha^{-1}]^{-1}]^2} \quad .$$

5.4 PERFORMANCE OF THE TECHNIQUE

In order to assess the potential of the Bayesian method described in section 2, analyses with real and simulated radiance measurements were performed. For each of 15 locations shown in Figure 5.4.1, an operational profile (obtained via the statistical eigenvector method of Smith and Woolf 1976) and a verification profile (obtained from radiosonde data) were kindly made available to us by Dr. Tom Koehler of the Meteorology Department at UW-Madison. Dr. Koehler also provided us with the HIRS (the seven 15μm and five 4.3μm band channels) and MSU (the three O_2 channels) satellite

Figure 5.4.1

0027 January 1980 TIROS-N Sounding Locations



○ clear, Δ partly cloudy, + cloudy.

The 15 shaded circles were the locations used in the study.

radiance data corresponding to the operational profile. These soundings were given "cloud free" classification by the HIRS radiometer.

The radiance data appeared to have systematic biases by channel. Figure 5.4.2 plots the equivalent brightness temperature differences between the observed radiance measurements and their "true" values, i.e. the radiances predicted by the verification profile. (The brightness temperature of a radiance measurement z at wavenumber ν is defined to be $B_{\nu}^{-1}(z)$ i.e. the equivalent black-body temperature corresponding to z .) Table 5.4.1 gives the median bias and the scale of measurement error computed from Figure 5.4.2. Using the table below, a simple correction was made to the radiance data - a raw radiance measurement was corrected by subtracting off the median bias found in the channel. In addition to these bias corrected radiance data, simulated radiance measurements were generated from the verification profile at each location according to:

$$z_i = R_{\nu_i}(T) + \sqrt{w_i}\,\varepsilon_i$$

where $\varepsilon_i \sim N(0,1)$ and the $\sqrt{w_i}$'s corresponded to the standard channel weights given in column 4 of Table 5.4.1 i.e. column 4 converted into radiance units. T is the verification profile.

Figure 5.4.2

Brightness temperature differences between satellite radiance measurements and "true" radiance i.e. those predicted by the verification profiles.

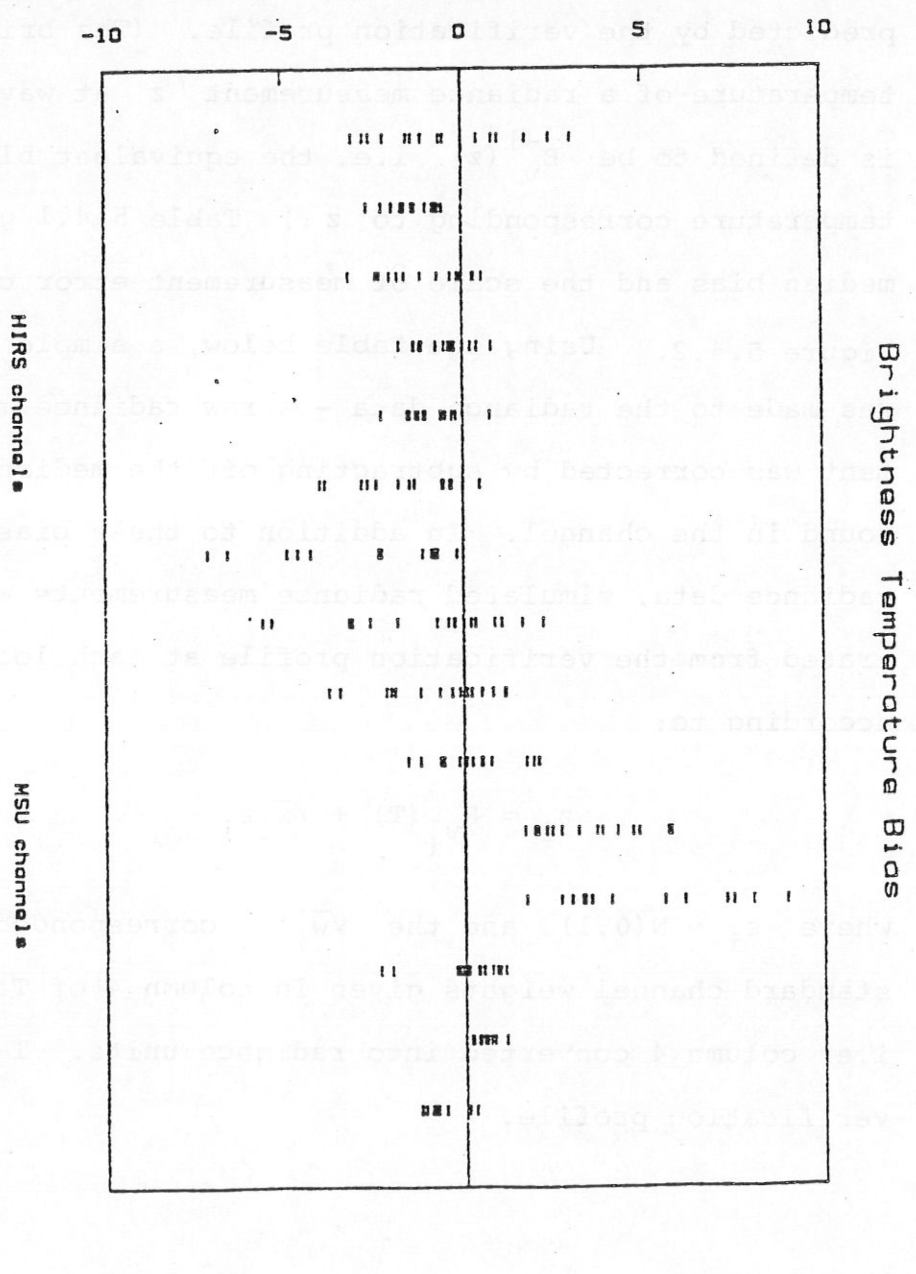


TABLE 5.4.1

SUMMARY STATISTICS FROM FIGURE 5.4.2

Channel Number from Smith et al (1979)	Channel Wave Number	Estimated Bias from Figure 1	Estimated Noise level from Figure 1	Noise Level from Past Experience
HIRS Channels				
1	668	-0.62	1.54	2.50
2	679	-1.38	1.32	1.20
3	691	-0.68	0.68	1.30
4	704	-0.38	0.38	0.75
5	716	-1.02	0.47	1.00
6	732	-1.34	0.66	1.00
7	748	-2.25	1.49	1.50
13	2190	-0.40	1.74	1.00
14	2213	-0.30	1.14	0.70
15	2240	0.08	0.82	1.00
16	2276	3.16	3.46	1.40
17	2361	4.07	4.34	3.00
MSU Channels				
2	1.792	0.11	1.12	0.60
3	1.833	0.52	1.01	0.50
4	1.933	-0.84	0.50	1.00

Using the algorithm in section 2, with the general purpose prior and λ chosen by the modified GCV criterion, two temperature profiles were estimated at each location - one from bias corrected satellite data, and the other from simulated radiances. Throughout these calculations transmittances corresponding to the verification profile were used, and the surface pressure was set at 1000mb. The operational sounding was used as the first guess.

The individual retrievals are given in an appendix at the end of this chapter. Figures 5.4.3 - 5.4.5 summarize the results of the study; three average bias and variability plots are presented. Bias and variability were computed as follows: for the i^{th} map location, let δ_i be the difference between the verification profile and the retrieved profile. The bias, $b(p)$, at pressure p is the 20% trimmed mean (mean taken over the middle 60% of values) of $\{\delta_i(p)\}_{i=1}^{15}$, while the variability is the 20% trimmed mean of the absolute deviations $\{|\delta_i(p) - b(p)|\}_{i=1}^{15}$. Because of the presence of a few "wild" δ_i's it was felt that these were more accurate estimates of bias and variability than the usual mean and standard deviation (see Tukey 1977).

Figure 5.4.3

Performance of the Bayesian Method on Simulated Radiance Data

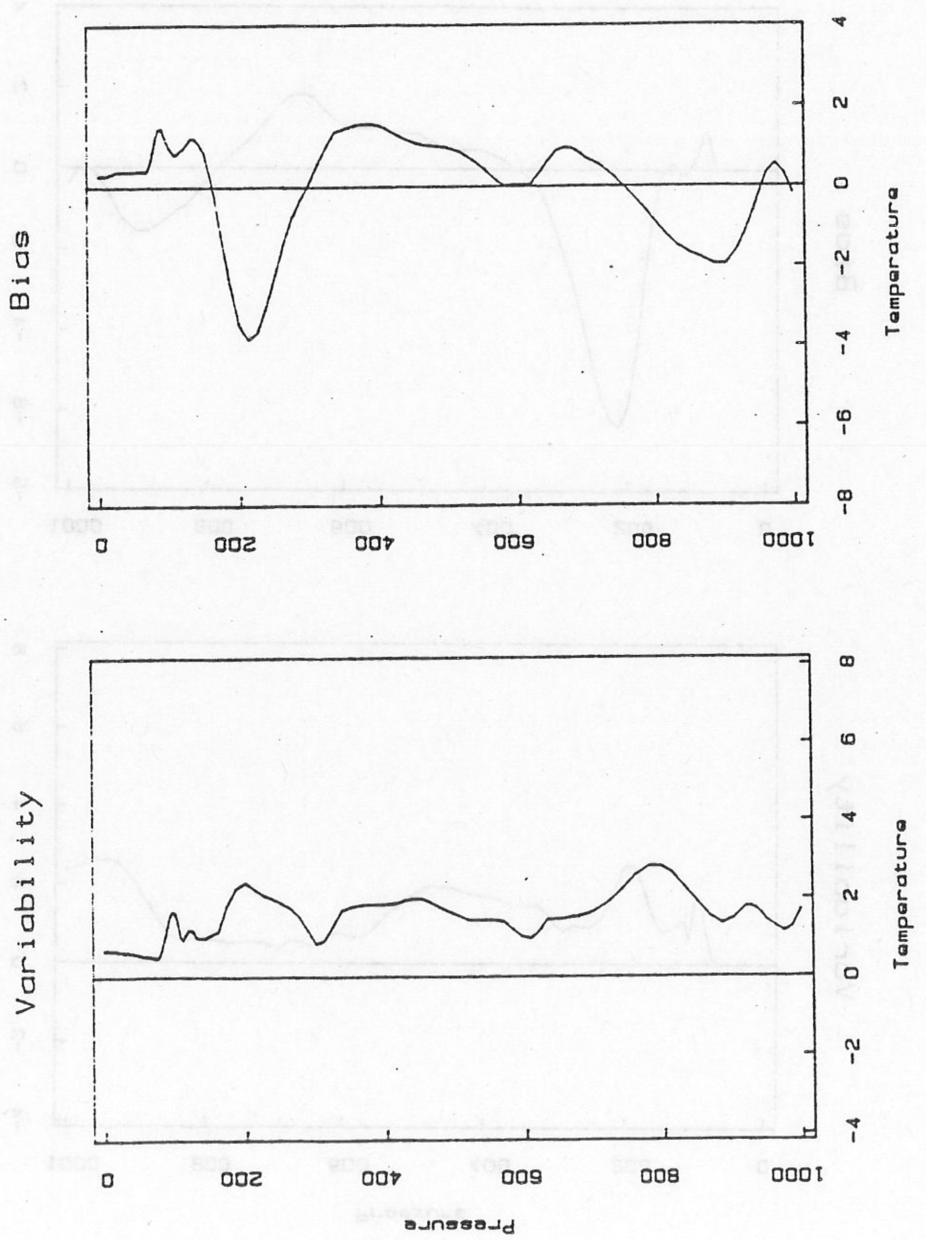


Figure 5.4.4

Performance of the Operational Method on Actual Satellite Radiance Data

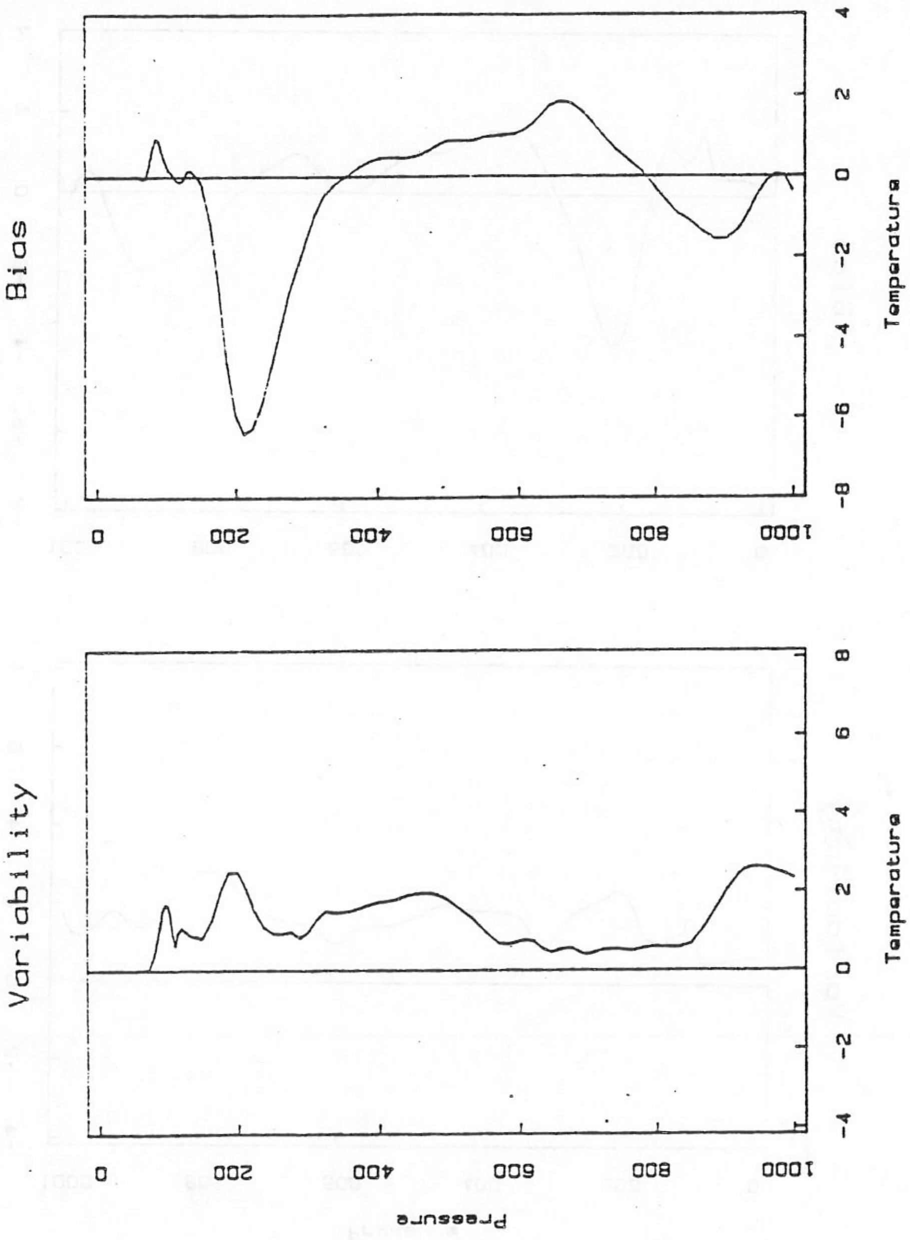
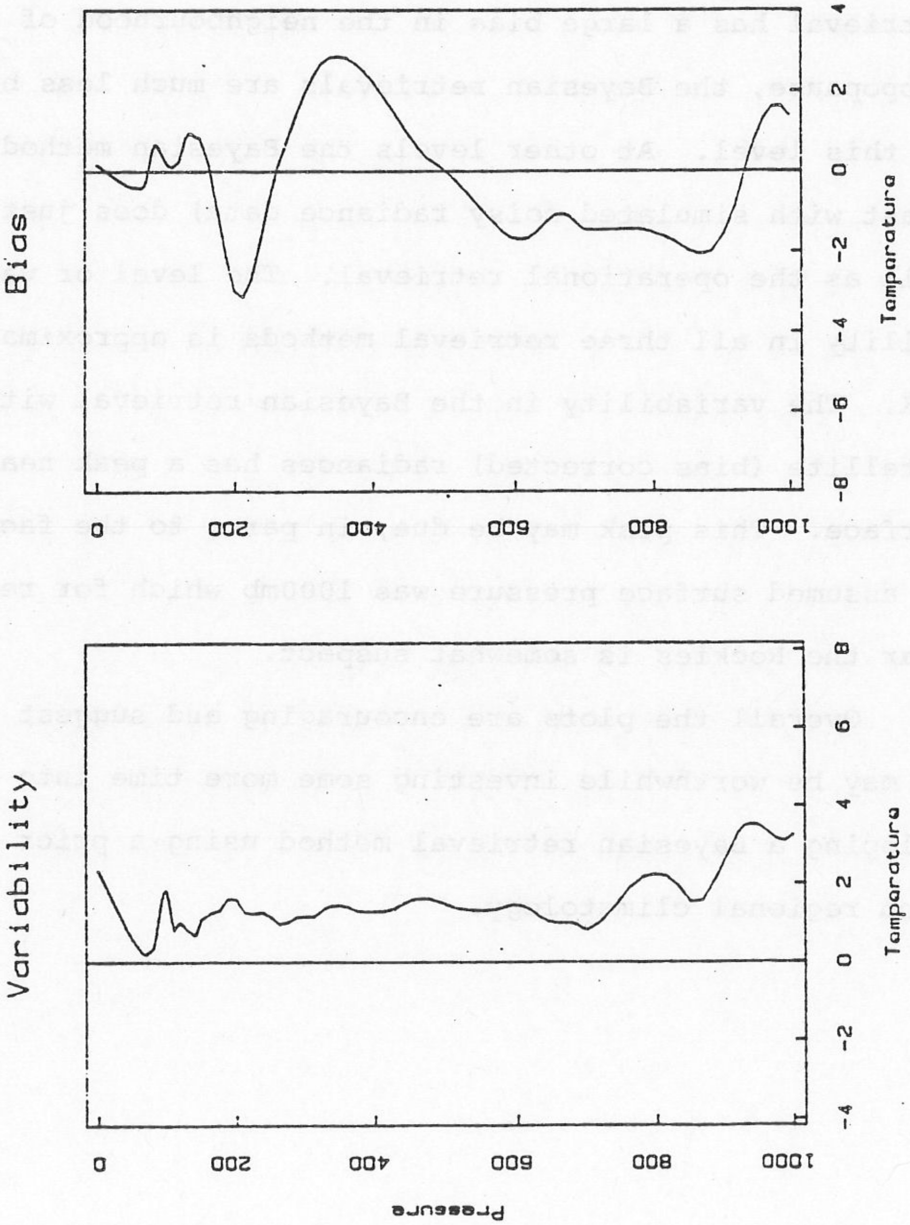


Figure 5.4.5

Performance of the Bayesian Method on Bias-corrected Satellite Radiance Data



The Bayesian retrievals perform quite well, especially since no real regional climatology has been incorporated into the prior specification. While the operational retrieval has a large bias in the neighbourhood of the tropopause, the Bayesian retrievals are much less biased at this level. At other levels the Bayesian method (at least with simulated noisy radiance data) does just as well as the operational retrieval. The level of variability in all three retrieval methods is approximately 2°K. The variability in the Bayesian retrieval with satellite (bias corrected) radiances has a peak near the surface. This peak may be due, in part, to the fact that we assumed surface pressure was 1000mb which for retrievals near the Rockies is somewhat suspect.

Overall the plots are encouraging and suggest that it may be worthwhile investing some more time into developing a Bayesian retrieval method using a prior derived from regional climatology.

APPENDIX B. RESULTS OF INDIVIDUAL RETRIEVALS

Fifteen figures follow. Each figure has two plots corresponding to the retrievals from simulated radiance data and bias-corrected satellite radiance data respectively. The true profile (solid), the operational profile (dot), and the estimated profile (dash) are all given.

The latitude (degrees North) and longitude (degrees East) corresponding to each retrieval are as follows:

Retrieval #	Latitude	Longitude
1	30.48	95.36
2	31.28	92.51
3	31.68	89.95
4	31.99	85.99
5	32.83	95.41
6	33.37	94.08
7	34.41	85.15
8	37.73	105.27
9	39.52	98.05
10	39.95	96.15
11	39.04	106.74
12	39.91	99.51
13	47.39	109.69
14	48.73	110.28
15	51.36	101.99

Figure 5.A.1: Results for Retrieval 1

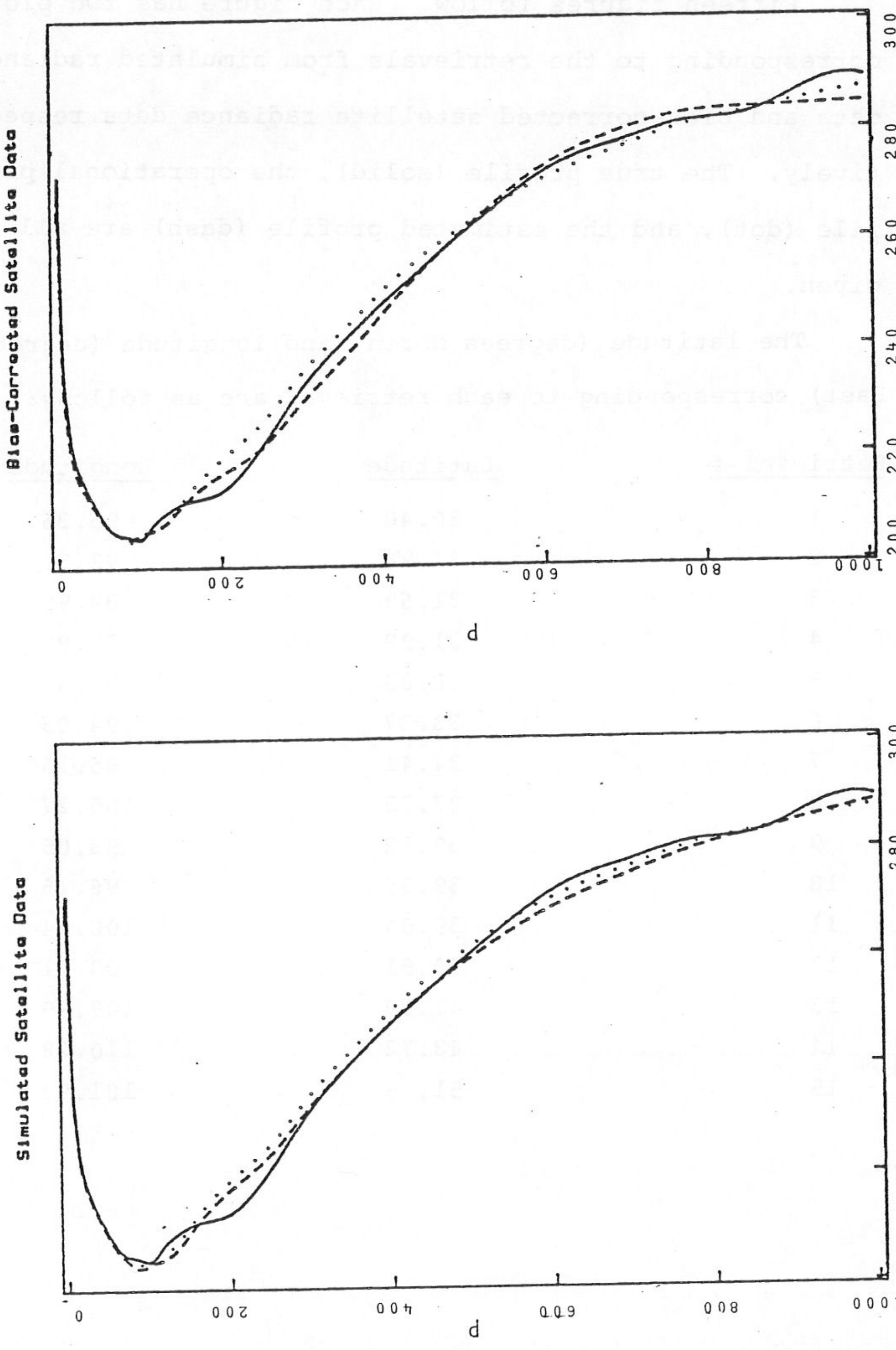
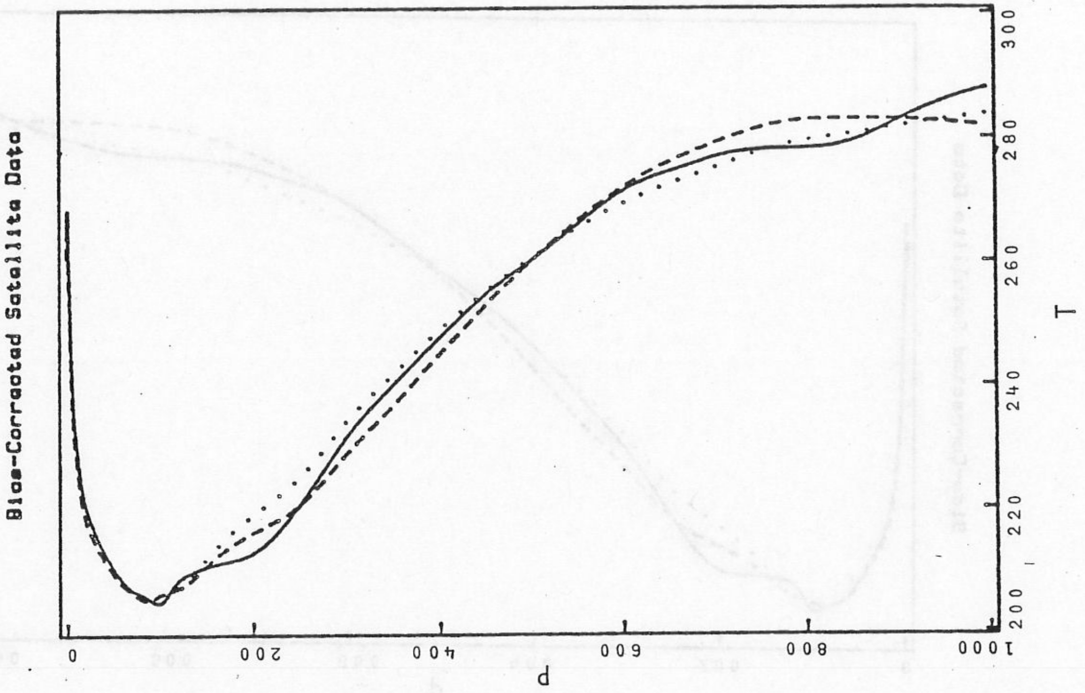


Figure 5.A.2: Results for Retrieval 2



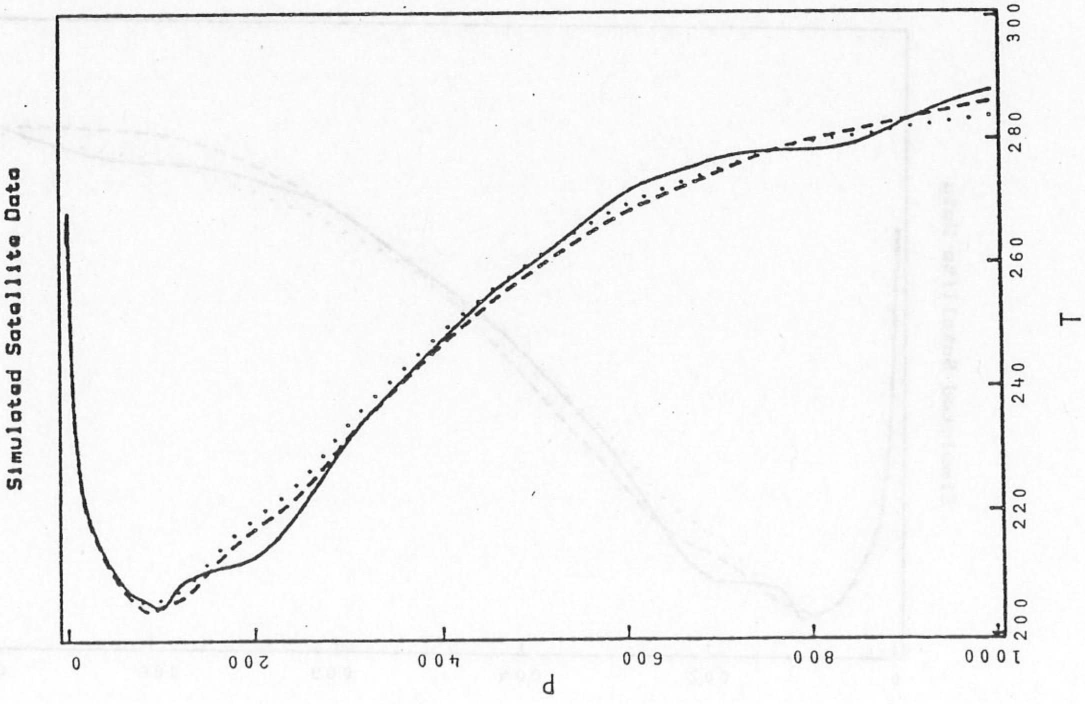


Figure 5.A.3: Results for Retrieval 3

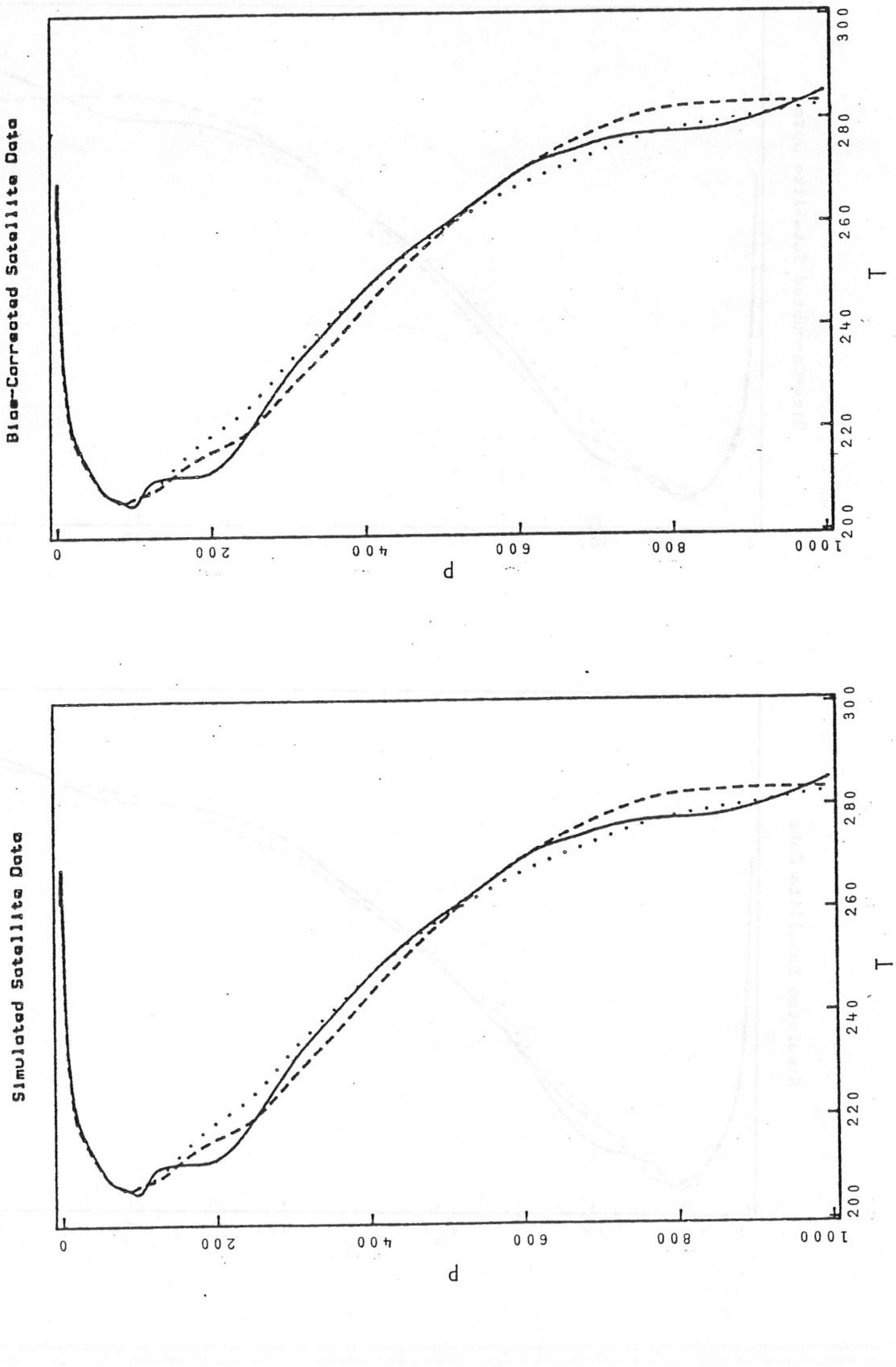
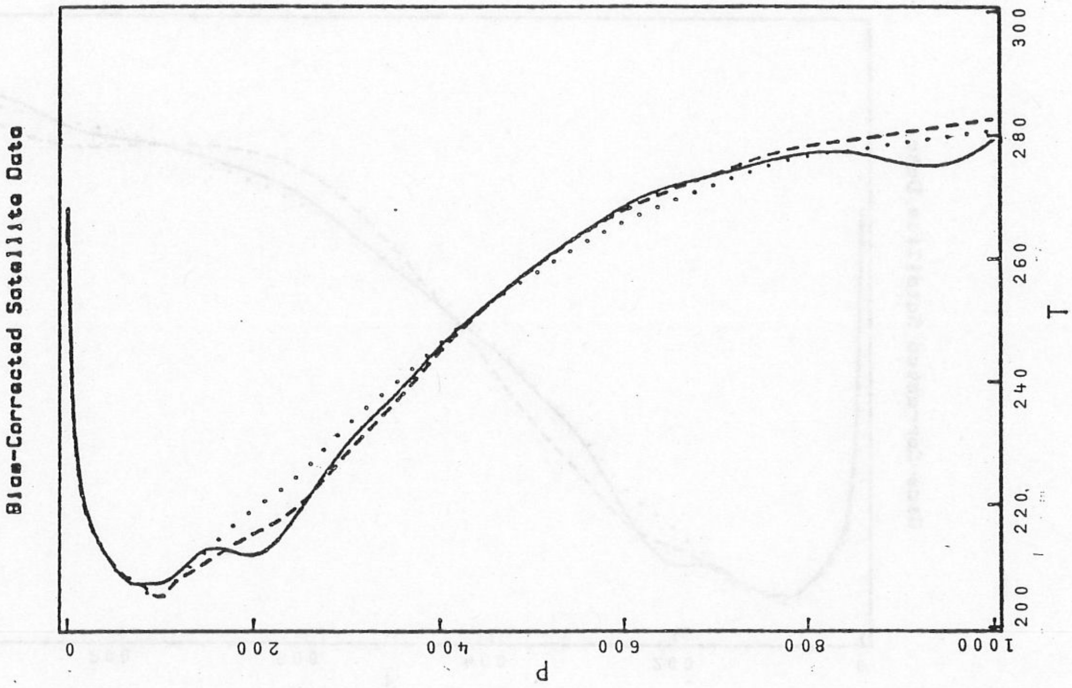


Figure 5.A.4: Results for Retrieval 4



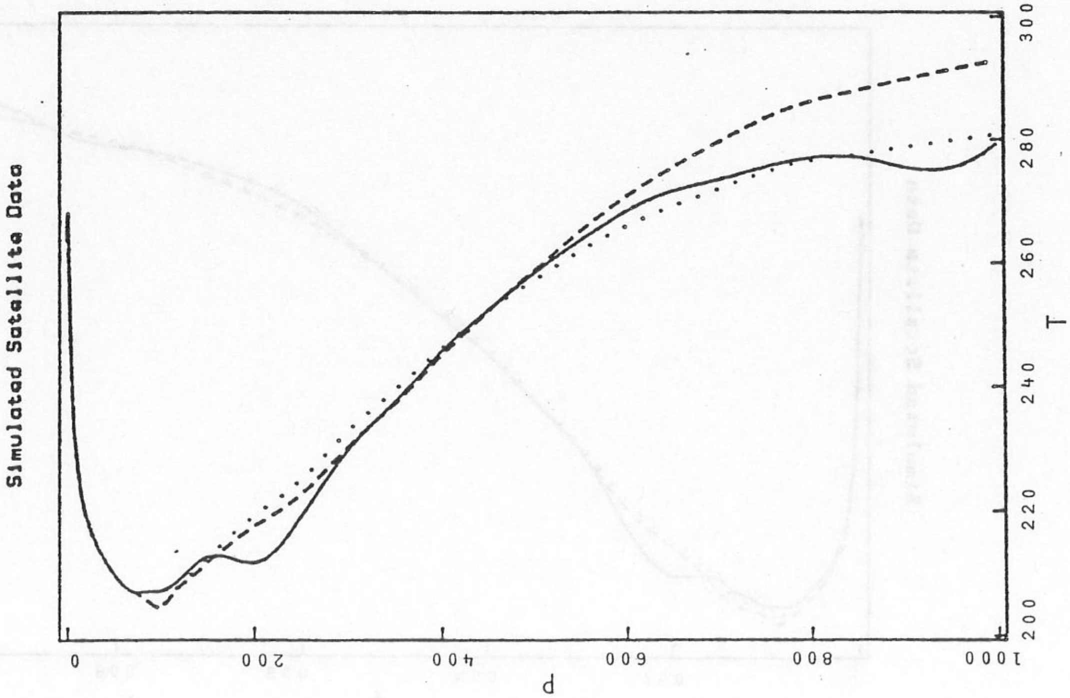


Figure 5.A.5: Results for Retrieval 5

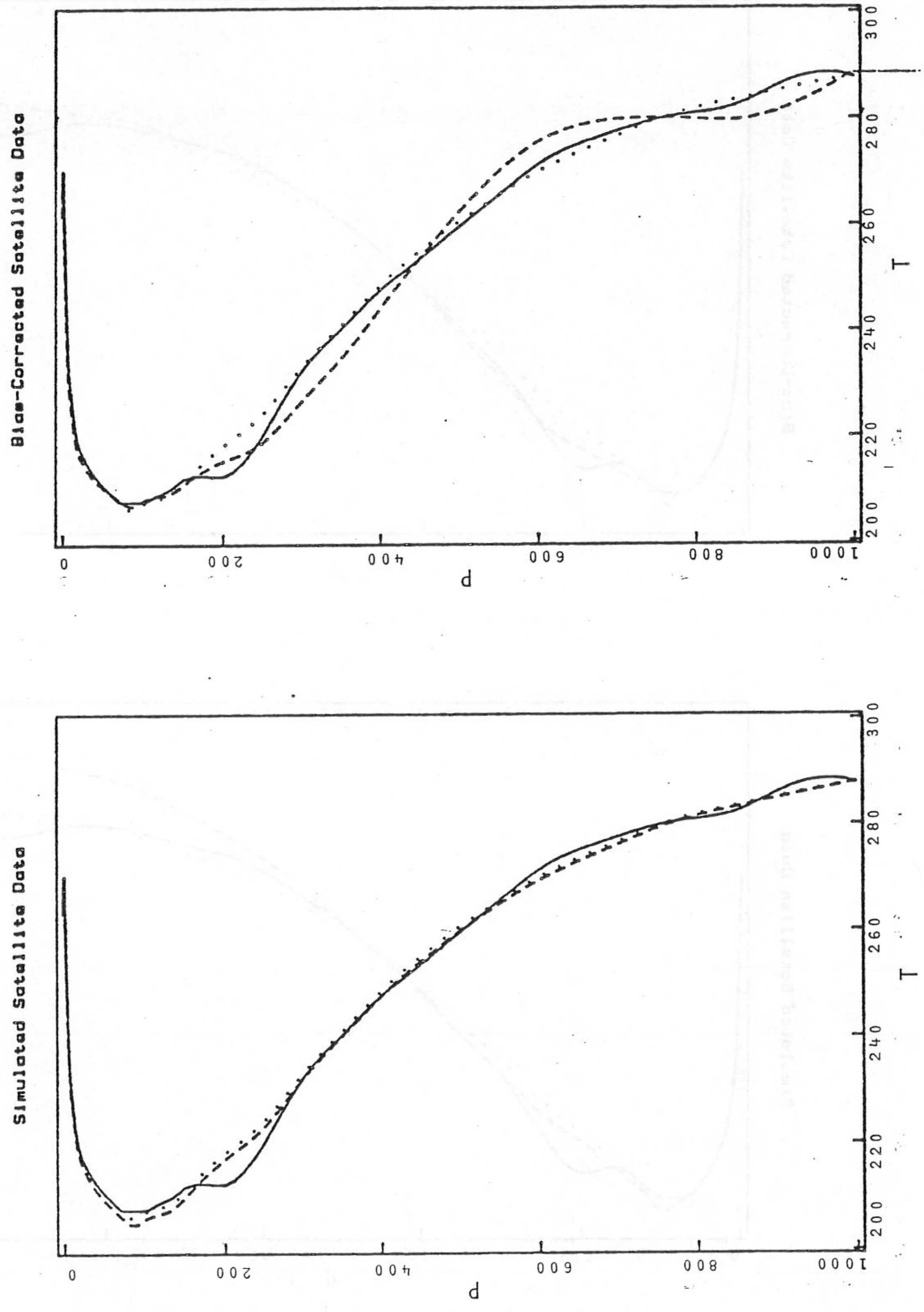


Figure 5.A.6: Results for Retrieval 6

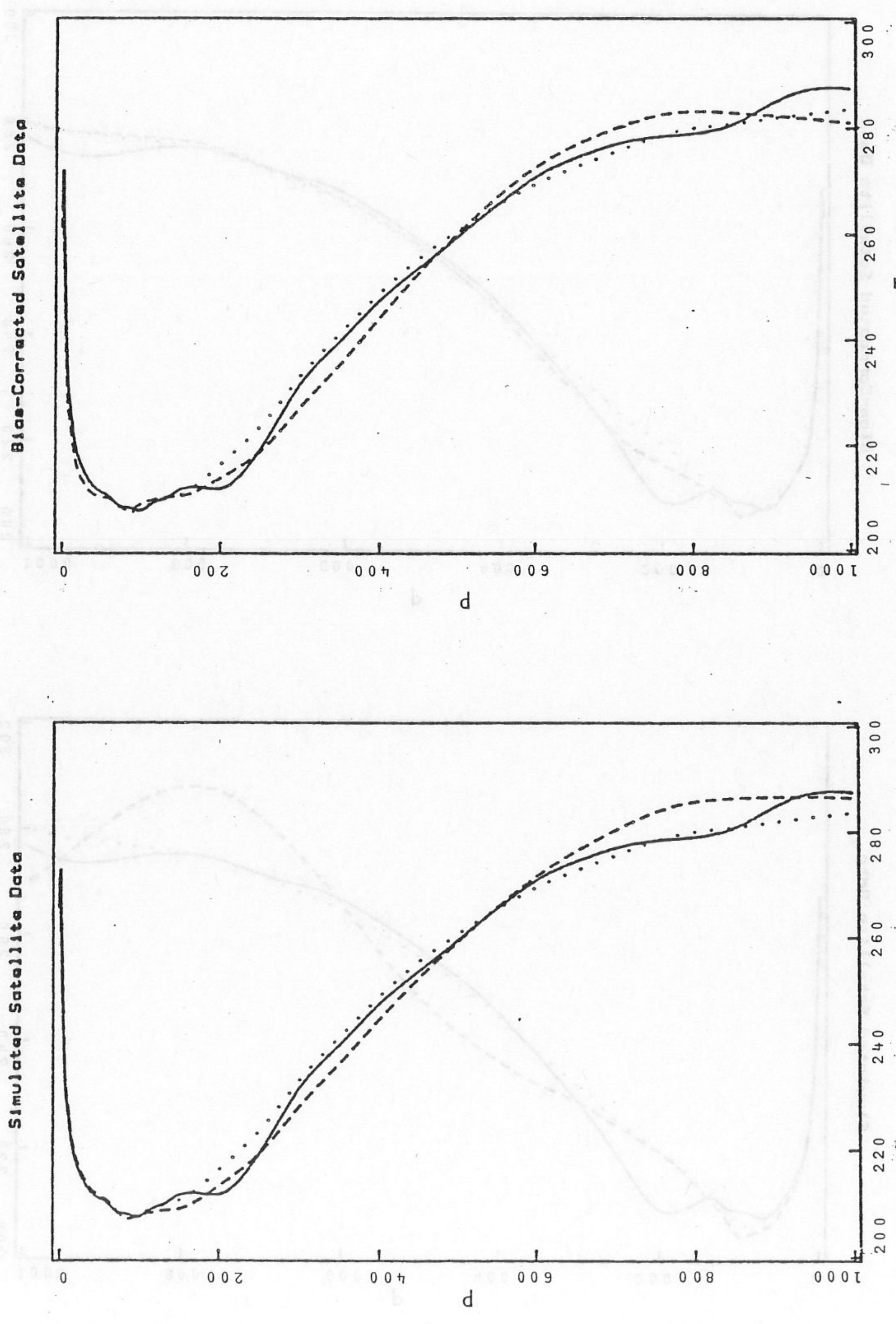
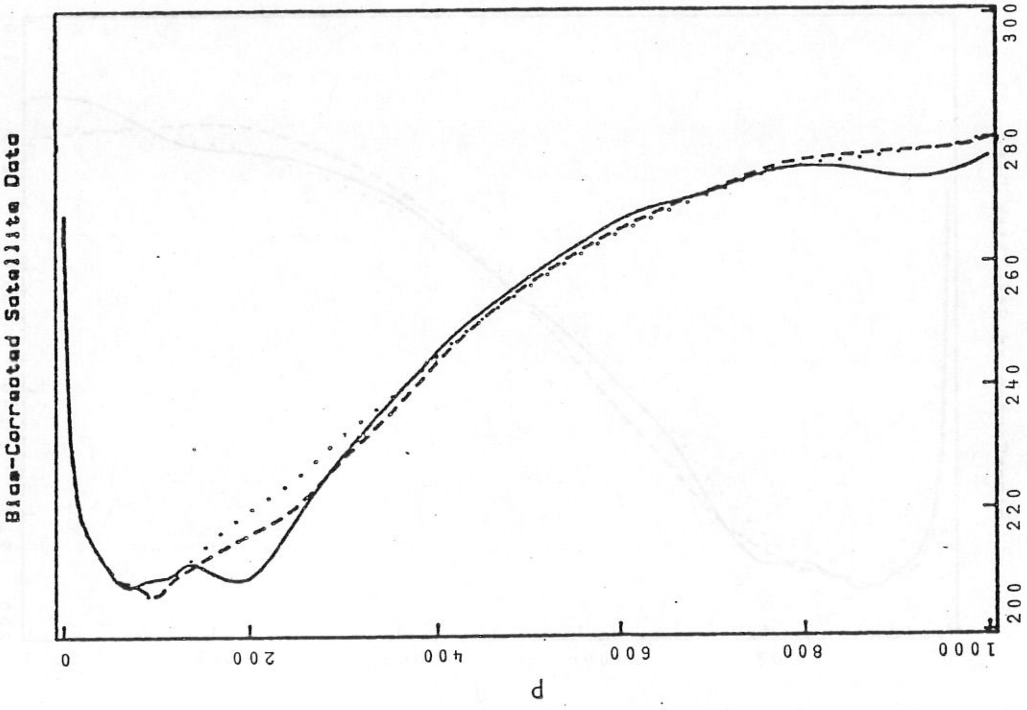


Figure 5.A.7: Results for Retrieval 7



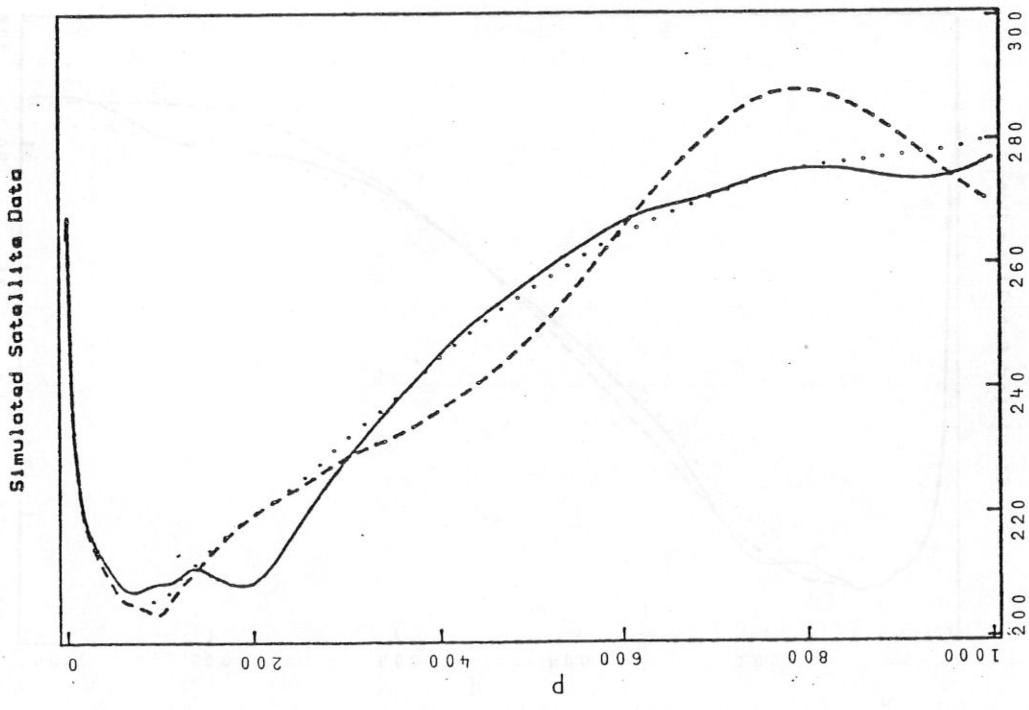
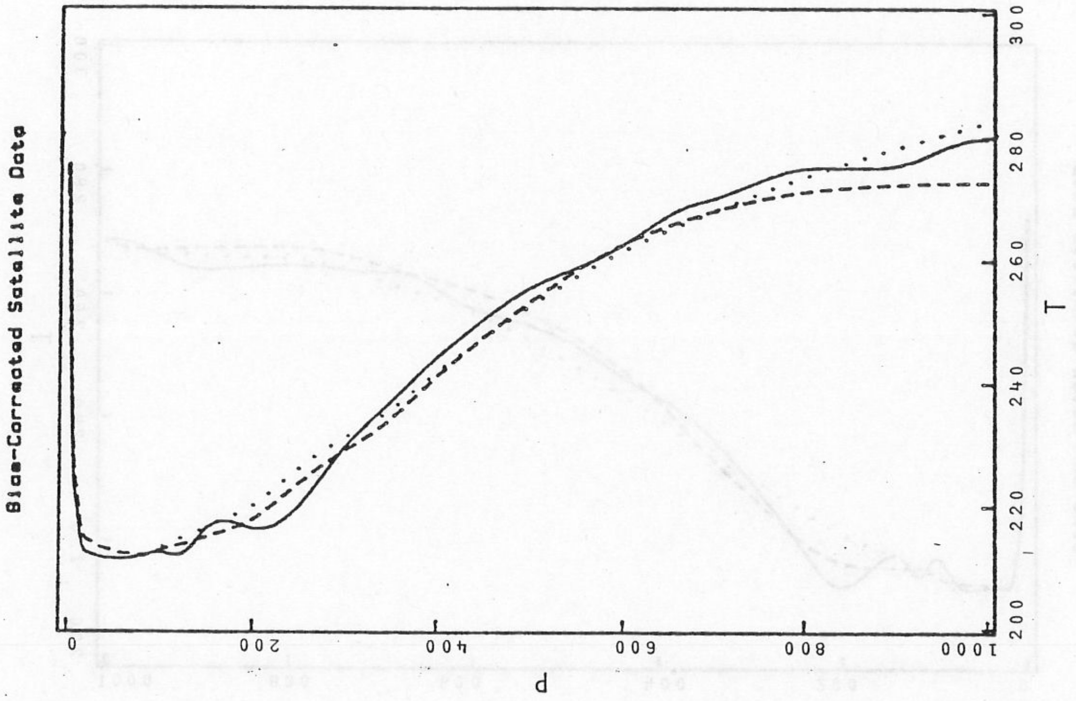
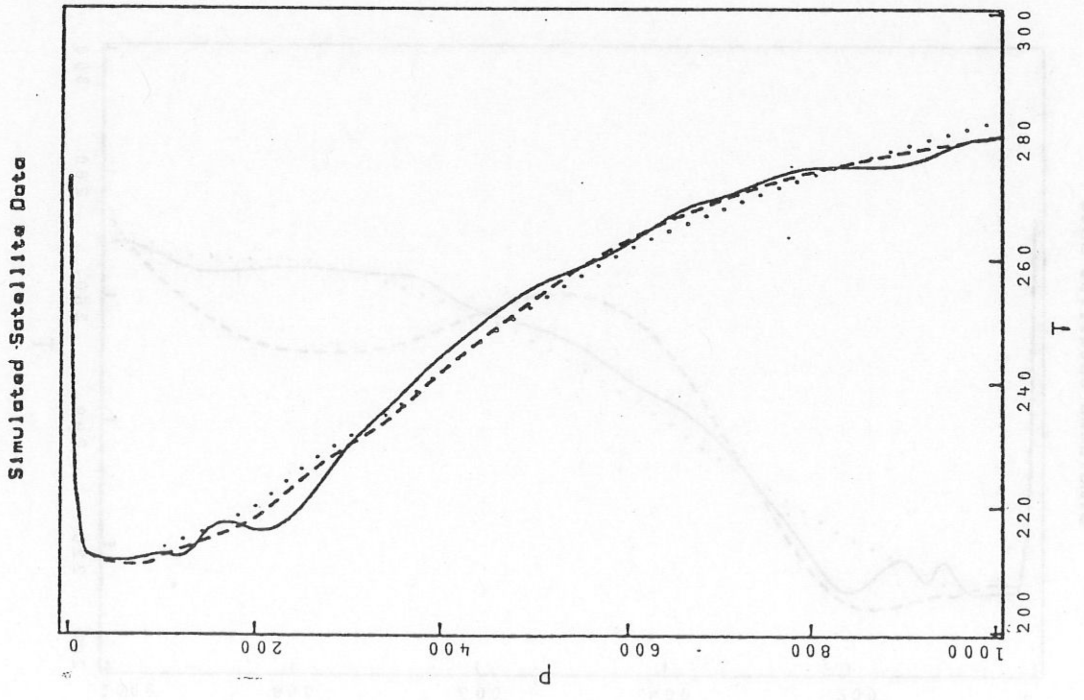


Figure 5.A.8: Results for Retrieval 8



Bias-Corrected Satellite Data



Simulated Satellite Data

Figure 5.A.9: Results for Retrieval 9

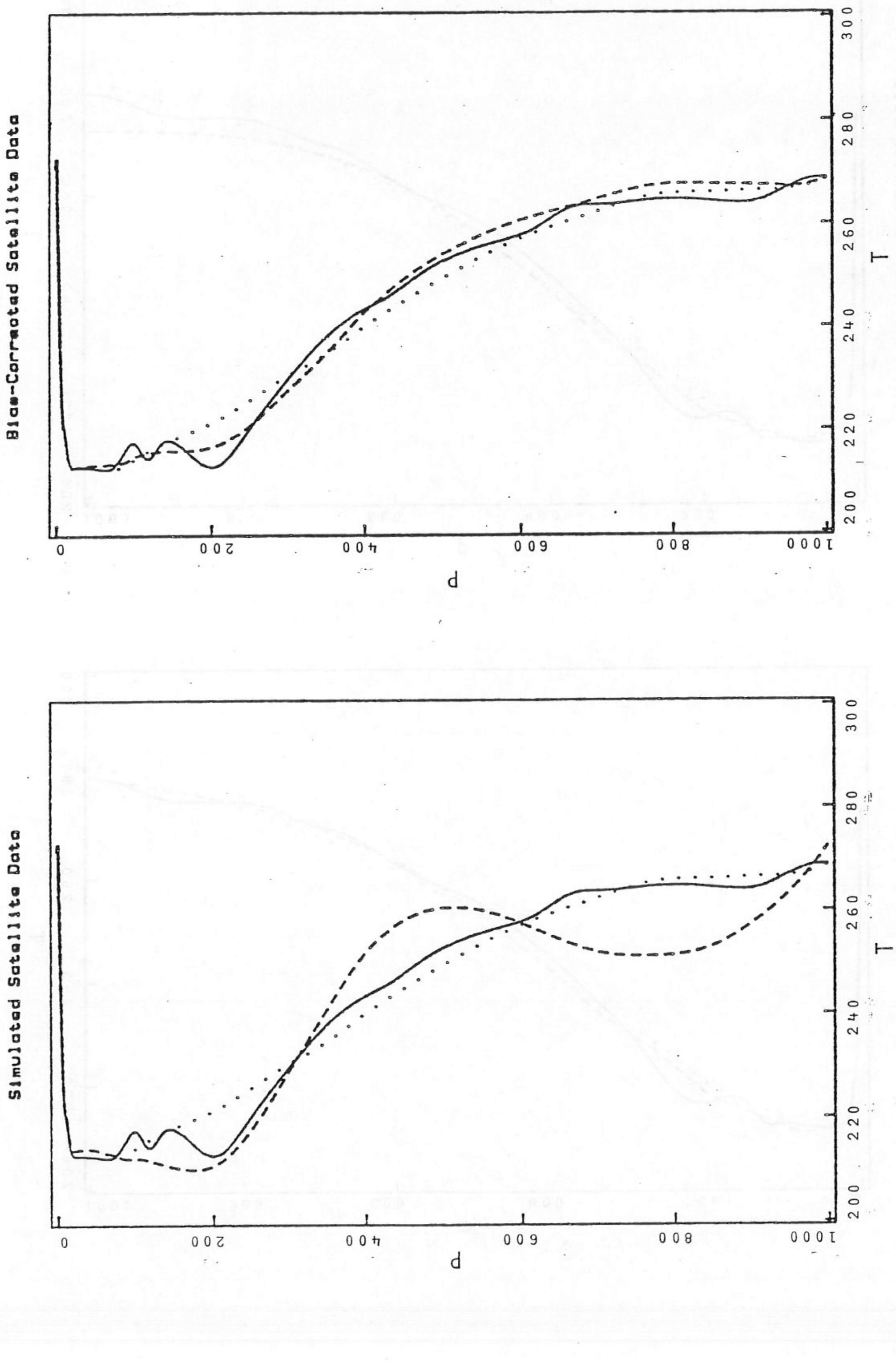


Figure 5.A.10: Results for Retrieval 10

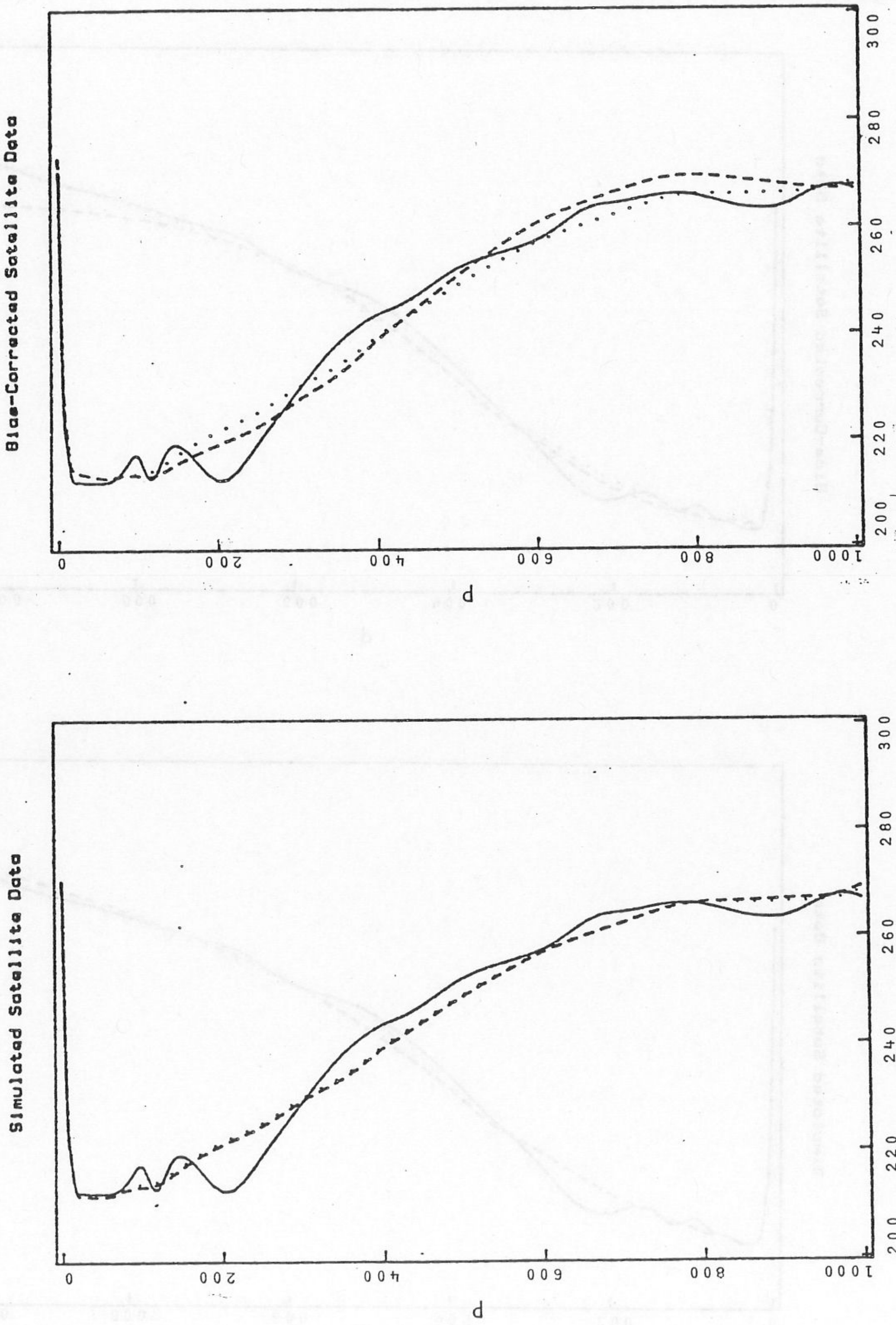
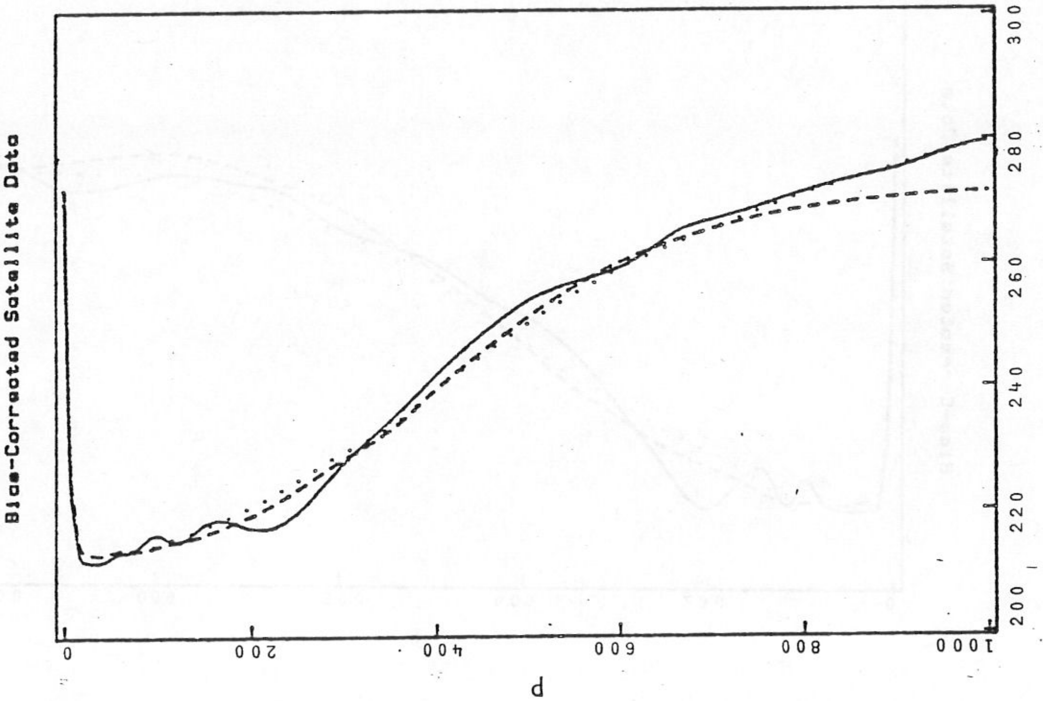


Figure 5.A.11: Results for Retrieval 11



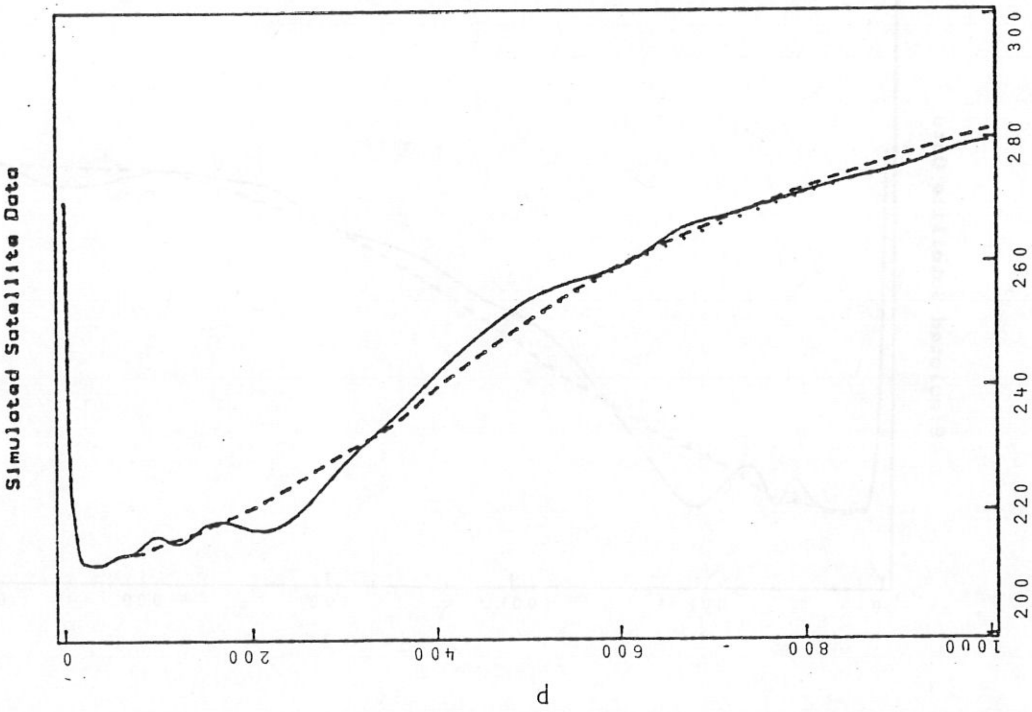


Figure 5.A.12: Results for Retrieval 12

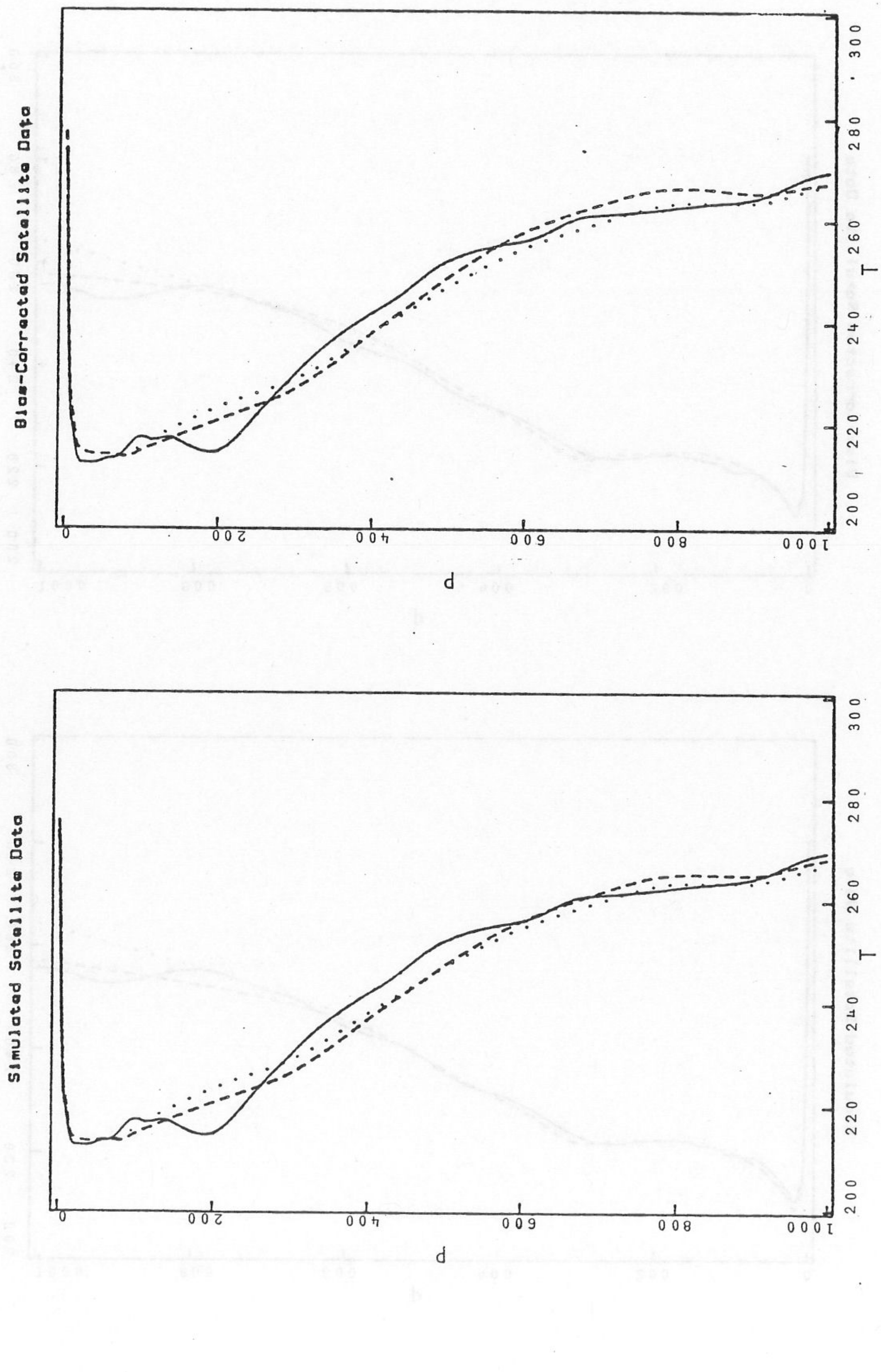
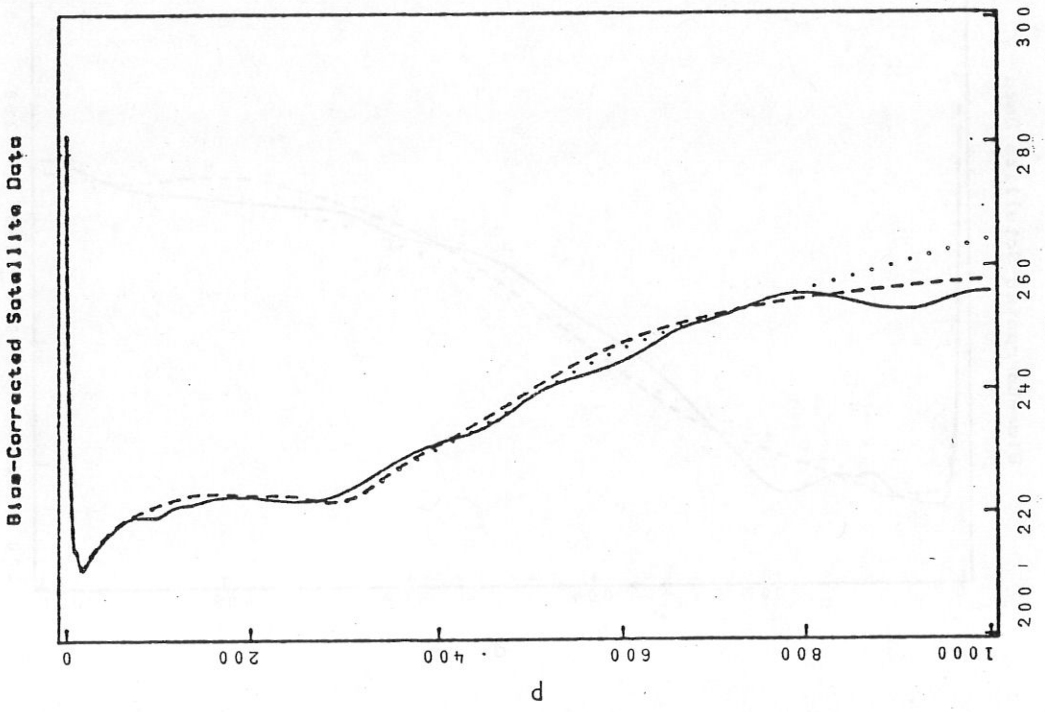


Figure 5.A.13: Results for Retrieval 13



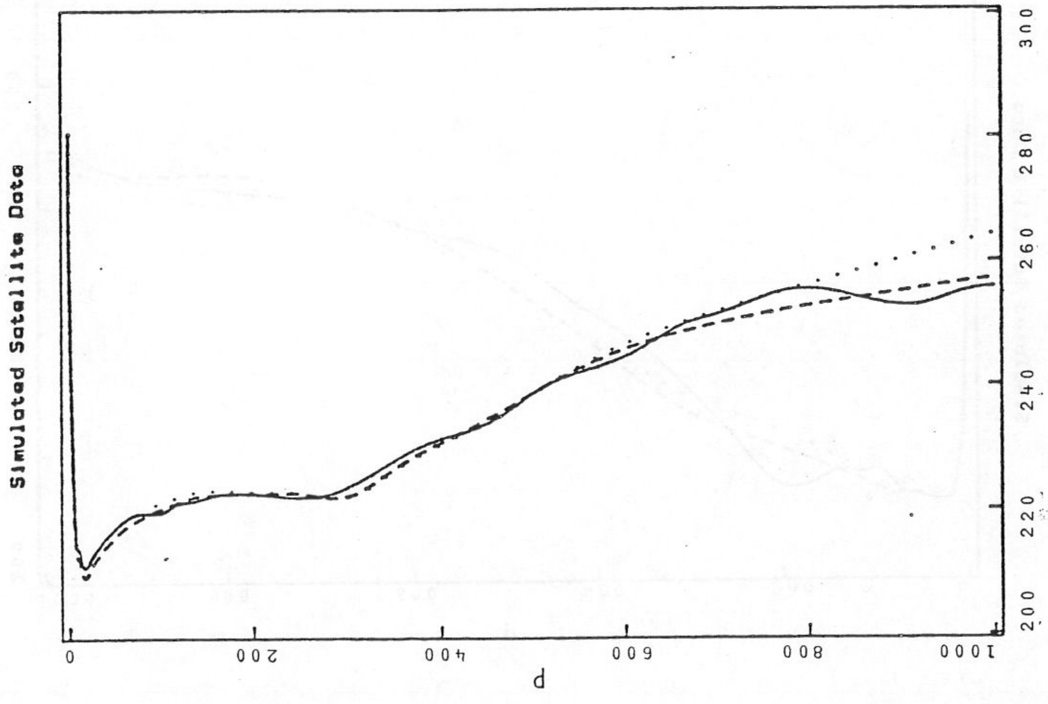


Figure 5.A.14: Results for Retrieval 14

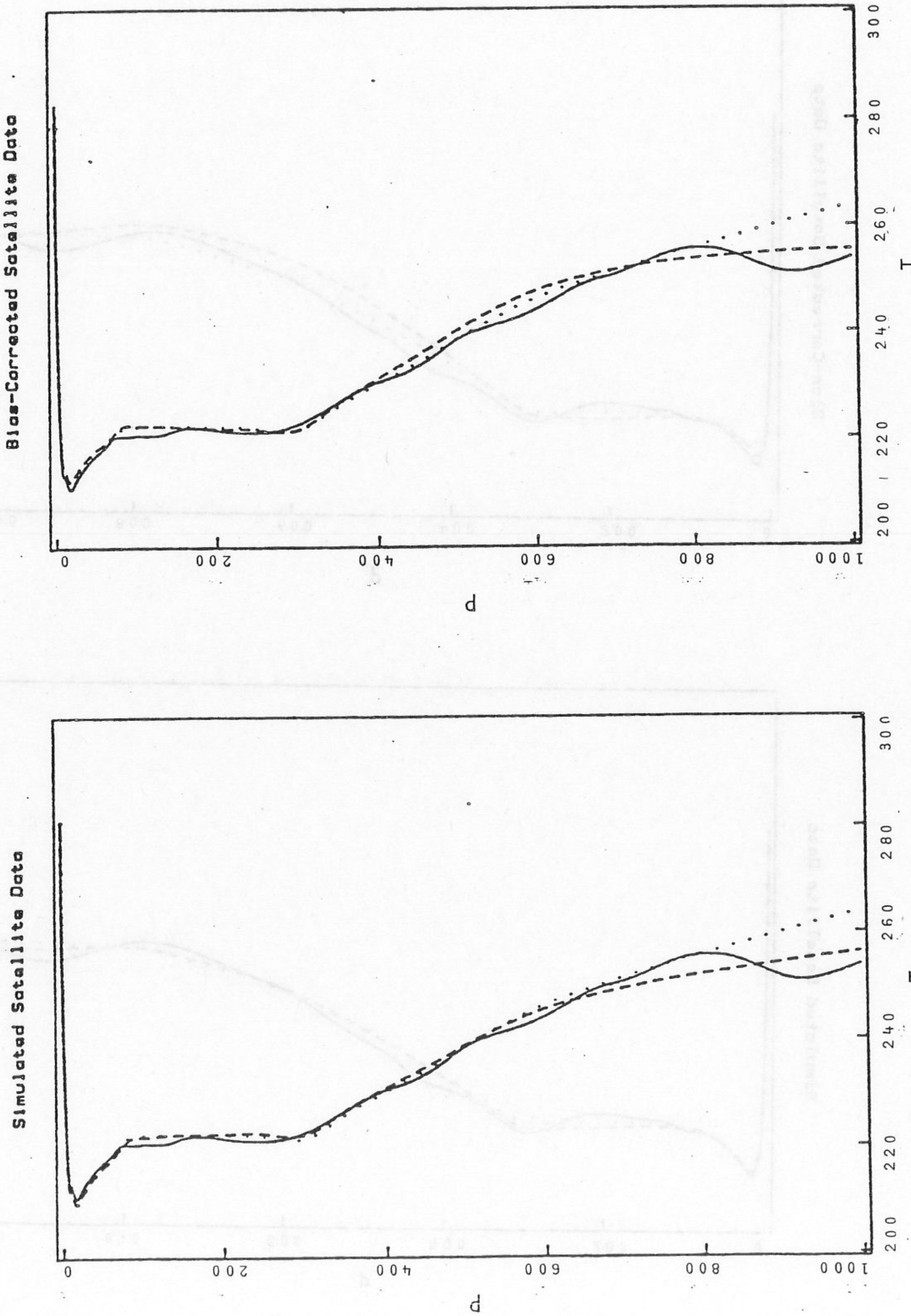
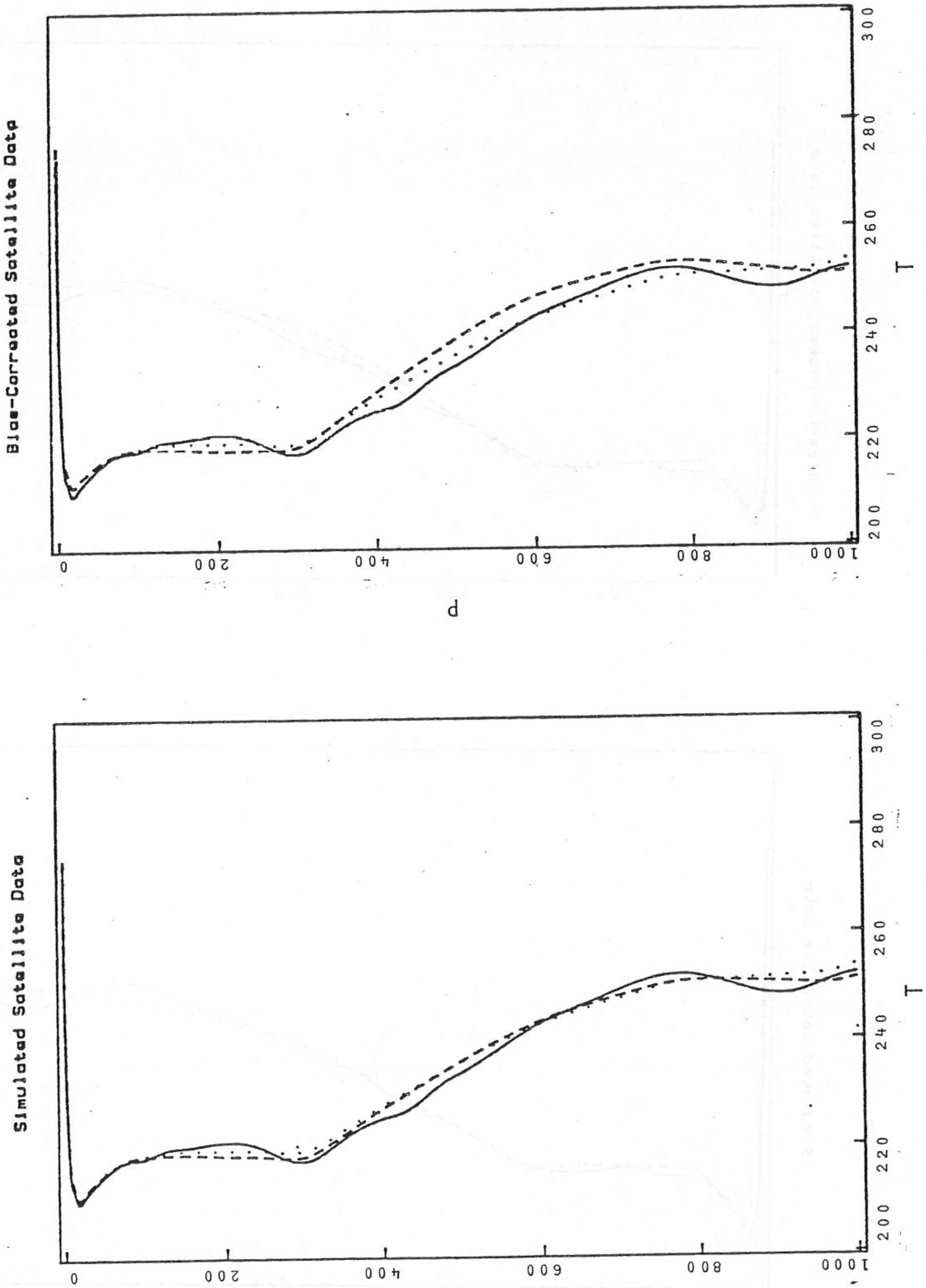


Figure 5.A.15: Results for Retrieval 15



ACKNOWLEDGMENTS

I am grateful to my thesis committee for the time and effort which they spent in evaluating my work. I thank Professor D. D. Cox and Professor G. Wahba for their frequent suggestions and advice over the last few years.

APPENDIX C

EXISTENCE OF SMOOTH SOLUTIONS TO THE RADIATIVE TRANSFER EQUATIONS

Here we show how the non-linear regression theory developed in section 3 of Chapter 2 can be used to establish the existence of smooth (smoothness prior) solutions to the ill posed problem associated with the radiative transfer equations. Recall that the objective functional here has the form

$$I_\lambda(\delta) = \sum_{i=1}^{n} [z_i - R_{\nu_i}(T_0 + \delta)]^2 + \lambda \int_0^{p_0} [\delta^{(m)}(p)]^2 dp \qquad (B.1)$$

and we are looking for minimizer of this functional in C_{T_0} where $C_{T_0} = \{\delta \in W_2^m[0,p_0] \mid T_0 + \delta \geq 0\}$; T_0 is the initial guess profile and $R_\nu(T)$ is given by:

$$R_\nu(T) = B_\nu[T(p_0)]\tau_\nu(p_0) - \int_0^{p_0} B_\nu[T(p)]T_\nu'(p)dp$$

where B_ν is Plank's function and τ_ν is atmospheric transmittance as a function of pressure.

We note from physical considerations that atmospheric transmittance is a continuously differentiable and monotonic decreasing function of pressure. We assume that the initial guess profile is continuous in pressure. These facts will be made use of in the sequel.

For each $m \geq 1$, C_{T_0} is clearly a convex set but since evaluation is a continuous linear functional in W_2^m for $m \geq 1$ it follows that C_{T_0} is also weakly closed. Let us now make the following three claims:

(c.1) R_ν is weakly continuous on C_{T_0} $\forall\ \nu > 0$

(c.2) $R_\nu(T_0 + \delta) \quad 0 \quad \forall\ \delta \in C_{T_0}$ with equality iff $T_0 + \delta \equiv 0$

(c.3) $R_\nu(T_0 + \delta)$ is convex as a functional of δ on C_{T_0}

Our existence result will follow from these claims.

Theorem B.1.

If (c.1)-(c.3) are true then $\exists\ \delta_\lambda \in C_{T_0}$ such that

$$I_\lambda(\delta_\lambda) = \min_{\delta \in C_{T_0}} I_\lambda(\delta) .$$

Proof:

From (c.1) we have by Theorem 2.2 that I_λ is weakly lower semi-continuous on C_{T_0} also using the continuity of evaluation C_{T_0} satisfies property 2 of section 2 with $z \equiv -T_0$. Letting $\phi(x) = |x|$ in Theorem 2.3.5 we have from (c.2) and (c.3) that $\sum_{i=1}^{n} \phi[R_{\nu_i}(T_0 + \delta)]$ is convex and $\sum_{i=1}^{n} \phi[R_{\nu_i}(T_0 + \delta)] = 0 \Longleftrightarrow T_0 + \delta = 0$ (i.e. $\sum_{i=1}^{n} R_{\nu_i}(T_0 + \delta)$

has a proper minimizer, $-T_0$, in $[-T_0 + H_0] \cap C_{T_0}$). Hence I_λ is coercive on C_{T_0} and the result follows by Theorem 2.2.1. □

The remainder will be devoted to proving (c.1)-(c.3). We begin with a few elementary lemmas concerning the Plank's function B_ν.

<u>Lemma B.2.</u>

$\forall\ \nu > 0$ B_ν is a convex function on $[0,\infty)$.

<u>Proof</u>:

$$B_\nu(t) = c_1\nu^3/[\exp(c_2\nu/t) - 1] .$$

Clearly it is enough to show that

$$f(x) = 1/[\exp(1/x) - 1] \quad \text{is convex for} \quad x \geq 0.$$

A direct calculation reveals that

$$f''(x) = \frac{\exp(1/x)}{x^3[\exp(1/x)-1]^2}\, g(1/x)$$

where

$$\begin{aligned} g(y) &= -2 - y + 2y\exp(y)/[\exp(y) - 1] \\ &= [(y - 2)(\exp(y) - 1) + 2y]/[\exp(y) - 1] \end{aligned}$$

but $1/[\exp(y) - 1] > 0$ $\forall\ y > 0$ and

$$\begin{aligned}(y-2)[\exp(y)-1]+2y &= \sum_{j=3}^{\infty} \frac{y^j[j!-2(j-1)!]}{j!(j-1)!} \\ &= \sum_{j=3}^{\infty} \frac{y^j(j-2)}{j!} \\ &\geq 0 \ \forall\, y \geq 0 .\end{aligned}$$

Consequently $g(1/x) \geq 0 \ \forall\, x \geq 0$ and this implies that $f''(x) \geq 0 \ \forall\, x \geq 0$. It follows that f and therefore B_ν are convex on $[0,\infty)$. □

Lemma B.3.

B_ν' is bounded on $[0,\infty)$.

Proof:

$$B_\nu'(t) = \frac{c_1 c_2 \nu^4 \exp(c_2\nu/t)}{t^2[\exp(c_2\nu/t)-1]^2} .$$

This is continuous in t and tends to $c_1\nu^2/c_2$ as $t \to \infty$. Consequently B_ν' is bounded on $[0,\infty)$. □

Lemma B.4.

[Justification of (c.1)] $\forall\, \nu > 0$, $R_\nu(T_0+\delta)$ is weakly continuous on C_{T_0}.

Proof:

$$R_\nu(T) = B_\nu[T(p_0)]\tau_\nu(p_0) - \int_0^{p_0} B_\nu[T(p)]\tau_\nu'(p)\,dp$$

Now evaluation is a continuous linear functional on $W_2^m[0,p_0]$ and B_ν is continuous on $[0,\infty)$ it follows that $B_\nu[T(p_0)]\tau_\nu(p_0)$ is weakly continuous on C_{T_0}. It remains to show that $\int_0^{p_0} B_\nu[T(p)]\tau_\nu'(p)dp$ is weakly continuous. Let $\{\delta_n\}$ be a sequence of functions in C_{T_0} converging weakly to δ (δ must also lie in C_{T_0} by weak closedness) and let $T_n = T_0 + \delta_n$ and $T_0 + \delta$.

$$\left|\int_0^{p_0} B_\nu[T_n(p)]\tau_\nu'(p)dp - \int_0^{p_0} B_\nu[T(p)]\tau_\nu'(p)dp\right|$$

$$\leq \int_0^{p_0}|B_\nu[T_n(p)] - B_\nu[T(p)]||\tau_\nu(p)|dp$$

$$\leq \text{constant} \int_0^{p_0} |T_n(p) - T(p)||\tau_\nu'(p)|dp$$

by Taylor's theorem and the boundedness of B_ν' (Lemma B.3). τ_ν is a transmittance function so τ_ν' is continuous which implies that τ_ν' is bounded on $[0,p_0]$. Consequently

$$\int_0^{p_0}|T_n(p) - T(p)||\tau_\nu'(p)|dp \leq \text{constant} \int_0^{p_0}|T_n(p) - T(p)|dp$$

Hence

$$\int_0^{p_0}B_\nu[T_n(p)]\tau_\nu'(p)dp \to \int_0^{p_0}B_\nu[T(p)]\tau_\nu'(p)dp \quad \text{as} \quad n \to \infty$$

i.e. $\int_0^{p_0} B_\nu[T(p)]\tau_\nu'(p)dp$ is weakly continuous on C_{T_0}. Finally since the sum of two weakly continuous functionals is weakly continuous it follows that R_ν is weakly continuous on C_{T_0}. □

Lemma B.5.

[Justification of (c.2)] $\forall\ \nu > 0,\ R_\nu(T_0 + \delta) \geq 0$ $\forall\ \delta \in C_{T_0}$ with equality iff $T_0 + \delta = 0$.

Proof:

Positivity follows from the positivity of B_ν on $[0,\infty)$ and the fact that τ_ν is a positive strictly monotonic decreasing function of pressure. Clearly $R_\nu(0) = 0$ since $B_\nu(0) = 0$. Now suppose $R_\nu(T_0 + \delta) = 0$ for some $\delta \in C_{T_0}$. Then $B_\nu[T_0(p_0) + \delta(p_0)]\tau_\nu(p_0) = 0$ $(\Longrightarrow T_0(p_0) + \delta(p_0) = 0)$ and $\int_0^{p_0} B_\nu[T_0(p) + \delta(p)]\tau_\nu'(p)dp = 0$. If $\exists\ p^* \in [0,p_0)$ such that $T_0(p^*) + \delta(p^*) > 0$ then by continuity of $T_0 + \delta$ $\exists\ \varepsilon > 0$ such that $\forall\ p \in [p^*, p^* + \varepsilon]$ $T_0(p) + \delta(p) > 0$. But, since B_ν and $-\tau_\nu'$ are both strictly positive functions on their respective ranges, we would then have that

$$-\int_0^{p_0} B_\nu[T_0(p) + \delta(p)]\tau_\nu'(p)dp > 0 \Longrightarrow R_\nu(T_0 + \delta) > 0 .$$

This is a contradiction. Consequently $R_\nu(T_0 + \delta) = 0$ iff $T_0 + \delta = 0$. □

BIBLIOGRAPHY

Anderssen, B. and Bloomfield, P. (1974). "Numerical differentiation procedures for non-exact data," *Numer. Math.* 22.

Atilgan, T. (1983). Ph.D. Dissertation (to appear), Department of Statistics, University of Wisconsin, Madison, Wisconsin

Baker, K. J. and Nelder, J. A. (1978). *The GLIM Manual-Release 3*, Numerical Algorithms Group. Oxford.

Box, G. E. P. and Tiao, G. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley Publishing Company.

Bunch, J. R., Dongarra, J. J., Moler, C. B., and Stewart, G. W. (1979). *Linpack User's Guide*, SIAM, Philadelphia.

Cox, D. D. (1983). "A Penalty Method for Non-parametric estimation of the Logarithmic Derivative of a Density Function." Statistics Department, University of Wisconsin, TR#704.

Cox, D. D. and O'Sullivan, F. (1983 (under consideration)). "Asymptotic Analysis of the Roots of Penalized Likelihood Equations." Statistics Department, University of Wisconsin, Madison.

Craven, P. and Wahba, G. (1979). "Smoothing Noisy Data with Spline Functions," *Numerisch Mathematic 31*, pp. 377-403.

Crowley, J. and O'Sullivan, F., (1983 (under consideration)). Smoothing the Proportional Hazard Regression Function.

Daniel, J. W. (1971). *The Approximate Minimization of Functionals*, Prentice-Hall.

Ekeland, J. and Teman, R. (1973). *Analyse Convexe et Problems Variationelles*, Herman, Paris.

Fleming, H. E. (1982). "Satellite Remote Sensing by the Technique of Computed Tomography," *Journal of Applied Meteorology*, 21, 10, pp. 1538-1549.

Good, I. J. and Gaskins, R. A. (1971). "Non-parametric Roughness Penalties for Probability Densities," Biometrika 58, pp. 255-277.

Kimeldorf, G. S. and Wahba, G. (1970). "A correspondence between Bayesian Estimation in Stochastic Processes and Smoothing by Splines," Annals of Mathematical Statistics 41, pp. 495-562.

Leonard, T. (1978). "Density Estimation, Stochastic Processes and Prior Information," Journal of the Royal Statistical Society B, 40, pp. 113-146.

Leonard, T. (January 1982). "An Empirical Bayesian Approach to the Smooth Estimation of Unknown Functions," MRC Technical Summary Report.

Liou, K. (1979). Introduction to Atmospheric Radiation, Academic Press, London.

Miller, R. G., Jr., (1981). Survival Analysis. John Wiley & Sons.

Nelder, J. A. and Wedderburn, R. W. M. (1972). "Generalized Linear Interactive Models," Journal of the Royal Statistical Society A, 135, p. 370.

Nychka, D. (1983). Ph.D. Dissertation, Department of Statistics, University of Wisconsin, Madison.

Ortega, J. M. and Rheinbold, W. C. (1970). Iterative Solutions of Nonlinear Equations in Several Variables, Academic Press.

Purser, R. (1983). Personal Communication, Department of Meteorology, University of Wisconsin, Madison.

Rall, L. B. (1969). Computational Solution of Nonlinear Operator Equations, John Wiley & Sons.

Raynor, W. J. and Bates, D. M. (1983). "Spline Smoothing of Binary Regressions in Large Data Sets." Statistics Department, University of Wisconsin, Madison ,TR 724.

Silverman, B. (1978). "Choosing the window width when estimating a density," Biometrika 65.

Silverman, B. W. (1982). "On the estimation of a Probability Density Function by the Maximum Penalized Likelihood Method," The Annals of Statistics 10, 3, pp. 795-810.

Smith, W. (1983). "The retrieval of atmospheric profiles from VAS geostationary radiance observations. J. Atmospheric Sciences, 40, 2025-2035.

Smith, W. L. and Woolf, H. M. (July 1976). "The Use of Eigenvectors of Statistical Covarianc- Matrices for Interpreting Satellite Sounding Radiometer Observations," Journal of the Atmospheric Sciences 33, 7, pp. 1127-1140.

Smith, W. L., Woolf, H. M., Hayden, C. M., Wark, D. Q., and McMillin, L. M. (October 1979). "The TIROS-N Operational Vertical Sounder," Bulletin of the American Meteorological Society 50, 10, pp. 1177-1187.

Spcckman, P. (1983). "Cross Validated Smoothing Splines" (to appear), Annals of Statistics.

Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions," Journal of the Royal Statistical Society B, 39.

Villalobos, M. A. (1983). "Estimation of Posterior Probabilities Using Multivariate Smoothing Splines and Generalized Cross Validation," Statistics Department, University of Wisconsin, Madison, TR 725.

Wahba, G. (1982). "Constrained Regularization for Ill-Posed Linear Operator Equations with Applications in Meteorology and Medicine," pp. 383-418. In Statistical Decision Theory and Related Topics III, Ed. J. O. Berger, Academic Press.

Wegman, E. J. (June 15, 1982). "Optimal Nonparametric Function Estimation." In The Proceedings of a Conference held at the University of Kentucky, Ed. Z. Govindarajulu, University of Kentucky.

Wendelberger, J. (1982). "Computational methods for calculating smoothing splines in one or more dimensions." Ph.D. thesis in Department of Statistics, University of Wisconsin, Madison.

Westwater, E. R. (October 1979). "Ill-Posed Problems in Remote Sensing of the Earth's Atmosphere by Microwave Radiometry," International Symposium on Ill-Posed Problems: Theory and Practice.

Whittaker, E. (1923). "On a new method of graduation." Proc. Edinburgh Math. Soc. 41.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
Technical Report No. 726

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
The Analysis of Some Penalized Likelihood Schemes

5. TYPE OF REPORT & PERIOD COVERED
Scientific Interim

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Finbarr O'Sullivan

8. CONTRACT OR GRANT NUMBER(s)
ONR N00014-77-C-0675

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Statistics
University of Wisconsin, 1210 W. Dayton St.
Madison, WI 53706

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
800 N. Quincy
Arlington, VA

12. REPORT DATE
September 1983

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
penalized likelihood; generalized linear models; nonlinear integral equations; numerical methods, generalized cross validation; information; satellite meteorology; temperature profile retrieval

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
see attached

unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT

There are many areas in applied science where the non-parametric estimation of regression functions is important.

In this thesis a general penalized likelihood method for non-parametrically estimating regression functions under a variety of observational models is developed. The existence and numerical approximation of the estimators is studied and a cross-validatory method for estimating the smoothing parameter is presented. Implementation of the method is algorithmically straightforward.

The procedures developed are applied to the estimation of atmospheric temperature profiles from satellite radiance data and are found to compare favorably with the currently used methodology.